Imad Abugessaisa
Takeya Kasukawa *Editors*

Practical Guide to Life Science Databases

MOREMEDIA

Springer

Practical Guide to Life Science Databases

Imad Abugessaisa • Takeya Kasukawa
Editors

Practical Guide to Life Science Databases

Springer

Editors
Imad Abugessaisa
Laboratory for Large-Scale Biomedical Data Technology
RIKEN Center for Integrative Medical Sciences (IMS)
Yokohama, Kanagawa, Japan

Takeya Kasukawa
Laboratory for Large-Scale Biomedical Data Technology
RIKEN Center for Integrative Medical Sciences (IMS)
Yokohama, Kanagawa, Japan

ISBN 978-981-16-5811-2 ISBN 978-981-16-5812-9 (eBook)
https://doi.org/10.1007/978-981-16-5812-9

© The Editor(s) (if applicable) and The Author(s), under exclusive license to Springer Nature Singapore Pte Ltd. 2021
Chapter 2 is licensed under the terms of the Creative Commons Attribution 4.0 International License (http://creativecommons.org/licenses/by/4.0/). For further details see license information in the chapter.
This work is subject to copyright. All rights are solely and exclusively licensed by the Publisher, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, reuse of illustrations, recitation, broadcasting, reproduction on microfilms or in any other physical way, and transmission or information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed.
The use of general descriptive names, registered names, trademarks, service marks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.
The publisher, the authors, and the editors are safe to assume that the advice and information in this book are believed to be true and accurate at the date of publication. Neither the publisher nor the authors or the editors give a warranty, expressed or implied, with respect to the material contained herein or for any errors or omissions that may have been made. The publisher remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

This Springer imprint is published by the registered company Springer Nature Singapore Pte Ltd.
The registered company address is: 152 Beach Road, #21-01/04 Gateway East, Singapore 189721, Singapore

Preface

Revealing the structure of DNA has changed the way life science research is carried out; it was to become a new era of study and understanding of biological systems. Since the last century, life science research has adapted to become more and more interdisciplinary, which encompasses biology, engineering, and data science (Bioinformatics). Together, this collaborative effort enables us to generate enormous amounts of knowledge, and in tandem, enormous and complex datasets. Life science datasets, generated by research groups, bring us several challenges. Genomics data, as an example, doubles every 8 months, which consequently brings several challenges for data acquisition, processing, and retrieval.[1] The research community needs to maximize utilization and reuse of published knowledge and datasets to make discoveries, and advance our understanding of biological systems under health and disease.

To this end, *Practical Guide to Life Science Databases* considers two categories of databases. The first category was the knowledgebase which aims at collection and integration of the knowledge and findings in life science (e.g., The GENCODE Portal, RefEX, CHIP Atlas etc.). The second category was the repository of life science dataset (e.g., The International Human Epigenome Consortium (IHEC) portal), a repository of datasets. By discussing these types of life science databases the book aims at providing life science researchers with the necessary knowledge about the untapped opportunities available in these databases, and how it could be used to advance basic research and applied research findings; all of this in the aim of transforming them to the benefit of human life.

The book in its current edition brings together expertise from renowned researchers in the field of life science databases and brings their experience and tools to the fingertips of the researcher. The book takes a bottom-up approach to explain the structure, content, and usability of life science databases. For each of the databases, the chapter authors provide a detailed explanation of the content,

[1] Broad Institute.

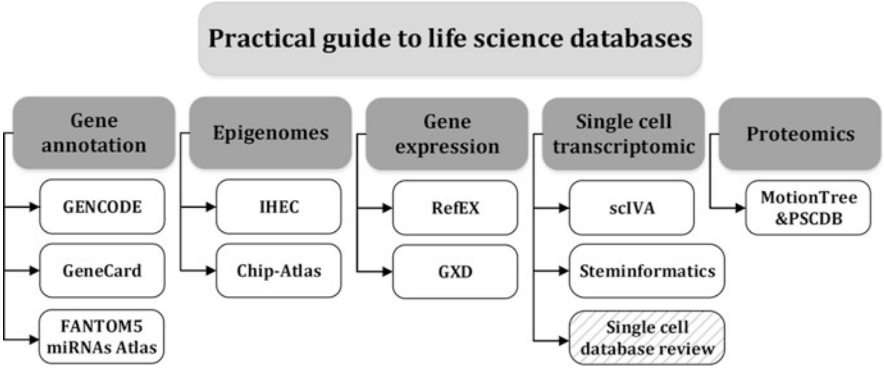


Fig. 1 Current edition of *Practical Guide to Life Science Databases* covers five types of life science databases. The shadow box indicates a review chapter rather than a database

structure, query, and data retrieval methods. The composition of the book chapters was summarized in Fig. 1. As the figure illustrates, the chapters cover Genomics, Proteomics, and Epigenomics database for both human and mouse data. The current edition introduces a new single-cell transcriptomic database scIVA (Single Cell Interactive Visualisation and Analysis Data Mining Platform). In addition, we dedicate a chapter which provides an up-to-date comprehensive review of single-cell RNA-seq databases.

This book will provide an indispensable tool for life science researchers. The collection of chapters will cover the necessary information for a wide audience in this interdisciplinary field, covering all bases. The reader will be made aware of available life science databases, their features, and how they could benefit from their usage in their area of expertise. The book is aimed at providing an understanding of the practical use of life science databases and enables the reader to use the provided tools in practice.

The following are some points summarizing the main features of this edition:

- It provides the latest, comprehensive information of a set of life science databases that are centered in life science research and drive the development of the field.
- It systematically introduces the fundamental principles, rationales, and methodologies of creating and updating life science databases.
- It include a user-oriented illustration of how to interact with each database by giving demonstration examples and use cases.
- It give the reader the necessary knowledge and provides any researcher with a fundamental grip on the subject to be able to publish and share their data with collaborators all over the world.

To meet the ever-evolving and rapid development in life science databases, in the near future, we plan to make the book as Springer Major Reference Works (or MRWs) by expanding the content to cover more life science databases.

We hope that the reader of the book will benefit greatly from the knowledge and will utilize it in their research, and for aspiring life science students, we hope this will provide excellent material in their course work at university.

We would like to thank the contributors to this edition of *Practical Guide to Life Science Databases* for their excellent contributions and effort to help us publish this book.

Yokohama, Kanagawa, Japan

Imad Abugessaisa
Takeya Kasukawa

Contents

Chapter 1
GENCODE Annotation for the Human and Mouse Genome: A User Perspective

Saleh Musleh, Meshari Alazmi, and Tanvir Alam

Abstract The GENCODE project provides comprehensive annotation of the functional elements in human and mouse genomes with high accuracy. The annotations are released for the benefit of biomedical and genomic research domain. In this initiative, we have provided a basic user manual or roadmap to facilitate the exploration of GENCODE annotation. We have provided a brief history of GENCODE and the general working principles that GENCODE adopts for their annotation. Then, we have introduced few workflows to guide users in the extraction and exploration of GENCODE resources for downstream analysis. The structure of this chapter is as follows. We started by introducing the GENCODE from a historical perspective, the needs and objectives that led to its creation, and being one of the most reliable sources for human and mouse genome functional elements. Afterward, we provided an overview of the GENCODE database. Mainly, different types of annotated genes, their description, basic statistics, and how they were created with emphasis on the latest four releases. Following this database overview, we described different annotation methods adopted by the GENCODE consortium for both human and mouse genomes along with validation methods. Besides GENCODE annotation methods, the user can find GENCODE annotation data format fields and definitions as they appear in the GTF and GFF3 files. Then we described three different ways to access GENCODE annotations via the GENCODE portal, Ensembl Genome Browser, and UCSC Genome Browser. We concluded with three use cases showcasing how to explore the GENCODE annotation for answering research questions. Source code, interactive user guide, and other files are made available for users at https://github.com/smusleh/BookChapterGENCODE.

Supplementary Information The online version of this chapter (https://doi.org/10.1007/978-981-16-5812-9_1) contains supplementary material, which is available to authorized users.

S. Musleh · T. Alam (✉)
College of Science and Engineering, Hamad Bin Khalifa University, Doha, Qatar
e-mail: talam@hbku.edu.qa

M. Alazmi
College of Computer Science and Engineering, University of Ha'il, Ha'il, Saudi Arabia

© The Author(s), under exclusive license to Springer Nature Singapore Pte Ltd. 2021

I. Abugessaisa, T. Kasukawa (eds.), *Practical Guide to Life Science Databases*,
https://doi.org/10.1007/978-981-16-5812-9_1

Keywords GENCODE · Genomics · Long noncoding · RNA · DNA elements · Protein · Identification · Database · Genes · Bioinformatics tool

1.1 Introduction

In 2003, the National Human Genome Research Institute (NHGRI) launched a project named The ENCyclopedia Of DNA Elements (ENCODE) to discover structural and functional elements (e.g., genes, transcripts) in the human genome (ENCODE Project Consortium 2004). It was essential to conduct such an effort as the protein-coding sequences, the best-defined functional element of the human genome, was still incomplete mainly due to the gap in the human genome. The initial pilot study covered nearly 1% of the human genome covering 44 different regions totaling 30 Mb sequence data (ENCODE Project Consortium 2004). Then, the GENCODE consortium (https://www.gencodegenes.org/) was built to identify and map protein-coding genes in the ENCODE regions (Harrow et al. 2006). The initial release of GENCODE covered 487 loci (420 were coding) having 2087 transcripts (1087 were coding) (Harrow et al. 2006). After completing the pilot study, the Wellcome Sanger Institute was awarded a research grant to scale up this project, GENCODE, to cover all types of gene features. In 2013, ENCODE consortium was awarded another grant to extend the GENCODE project to cover the mouse genome as well (Mudge and Harrow 2015). Currently, GENCODE is one of the largest and most reliable sources for human and mouse functional elements. Therefore, it is considered as reference annotation by many consortia like ENCODE, the 1000 Genomes Project (Siva 2008), GTEx (GTEx Consortium 2013), Human Cell Atlas (Regev et al. 2017), Exome Aggregation Consortium (Lek et al. 2016), International Human Epigenome Consortium (Stunnenberg and Hirst 2016), etc.

1.2 GENCODE Database Overview

In GENCODE, there are mainly four types of genes that are annotated, they are: (a) protein-coding genes, (b) pseudogenes, (c) long non-coding RNAs (lncRNAs), and (d) small non-coding RNAs (sncRNAs) (Frankish et al. 2019). Recently GENCODE introduced a new type of annotation for immunoglobulin/T-cell receptor genes as well. According to the GENCODE protocol, a locus is considered as protein-coding genes where sufficient evidence of coding sequence (CDS) is present. GENCODE provides three levels of confidence for protein-coding genes as well as other types of gene annotation. Levels 1, 2, and 3 represent "validate," "manual annotation," and "automated annotation," respectively, for the gene annotations (Harrow et al. 2012). GENCODE also annotates pseudogenes which are derived from protein-coding genes. Still, it contains variations in stop codons, frameshift

Table 1.1 Summary of pseudogenes from GENCODE annotation

Pseudogene class	Event	Mutation	Evidence level
Processed pseudogene	Retrotransposition	Turn-off mutation after the event	No transcription
Transcribed processed pseudogene	Retrotransposition	Turn-off mutation after the event	Only transcribed
Translated processed pseudogene	Retrotransposition	Turn-off mutation after the event	Translated
Unprocessed pseudogene	Duplication	Turn-off mutation after the event	No transcription
Transcribed unprocessed Pseudogene	Duplication	Turn-off mutation after the event	Only transcribed
Translated unprocessed pseudogene	Duplication	Turn-off mutation after the event	Translated
Unitary pseudogene	A fixed disabling mutation	Unambiguous functional ortholog	Based on previous functional protein-coding genes
Polymorphic pseudogene	A fixed disabling mutation	The presence of a validated dbSNP (Sherry et al. 2001) entry	Based on the previous functional protein-coding gene

insertion or deletion (INDEL), aberrant insertion, or truncation, which fails to provide any evidence of proper transcription process. Pseudogenes identification is important task for GENCODE (Frankish and Harrow 2014). They represent inactive protein-coding regions due to mutations and they frequently occur in many different organisms and species (Zheng et al. 2007). Under GENCODE annotations, pseudogenes are classified into different categories based on the event, mutation, evidence level, as shown in Table 1.1.

GENCODE does not apply strict length threshold of 200 bp for lncRNA annotation, albeit few annotated lncRNAs fall below this threshold (Frankish et al. 2019; Derrien et al. 2012). SncRNAs are entirely annotated by automated GENCODE pipelines that consider the homology to known sncRNAs and their predicted secondary structure. GENCODE provides the annotation for both human (GRCh37/GRCh38) and mouse genome (mm9/mm10). From the release of GENCODE 20 (on August 2014), GENCODE annotations have been exclusively created on the GRCh38 human assembly and those releases have been mapped from GRCh38 to GRCh37 for users (Frankish et al. 2019). Figure 1.1a, b highlights the summary statistics of the recent four GENCODE releases for human (release 35–38) and mouse (release 24–27), respectively. It is important to emphasize that the latest four releases for human annotation were all published in 2020. The total number of protein-coding genes, pseudogenes, sncRNA genes, and lncRNA genes are almost same in the latest four releases (Fig. 1.1a). But the number of human transcripts has

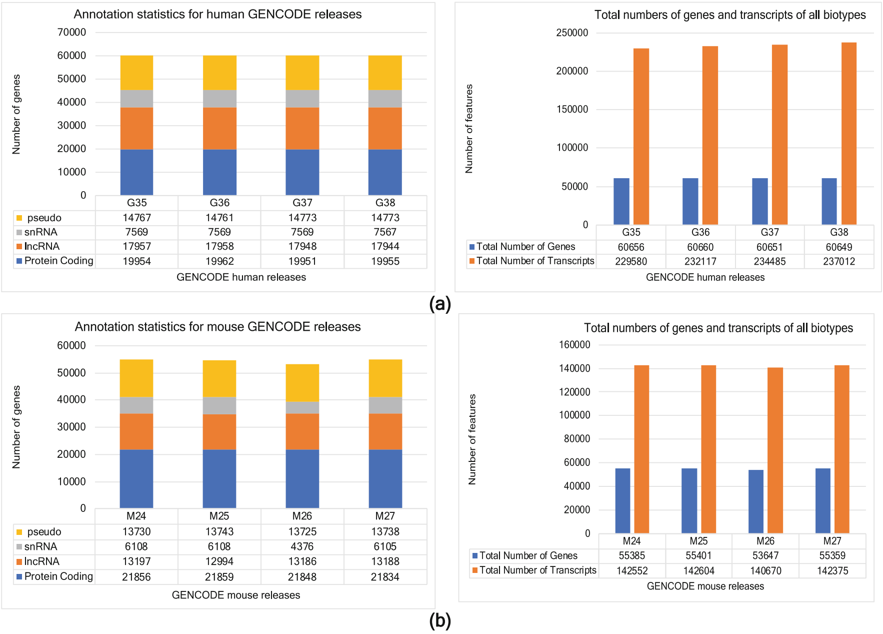


Fig. 1.1 Summary statistics of the available annotation at GENCODE. (**a**) Statistics for the latest releases for human annotation. (**b**) Statistics for the latest four releases for mouse annotation

changed significantly in the latest four releases (Fig. 1.1a). Conversely, the latest four releases for mouse annotation were published in 2019–2020 and the total number of genes and transcripts have not changed much in the latest four mouse releases (Fig. 1.1b).

1.3 Annotation Method Adopted in GENCODE

1.3.1 Overall Annotation Methods Adopted by GENCODE

The whole process of compiling the functional elements in GENCODE is a tedious work that requires the seamless integration of computational analysis, manual curation, and experimental validation from four founding members: Human and Vertebrate Analysis and Annotation (HAVANA) group at the Wellcome Trust Sanger Institute, Yale University, University of California Santa Cruz (UCSC), and Centre for Genomic Regulation (CRG) (Frankish et al. 2019). Later three other groups: Ensembl, Massachusetts Institute of Technology (MIT), and Centro Nacional de Investigaciones Oncológicas (CNIO) joined GENCODE in 2007 (Frankish et al. 2019). HAVANA group having many members and collaborators

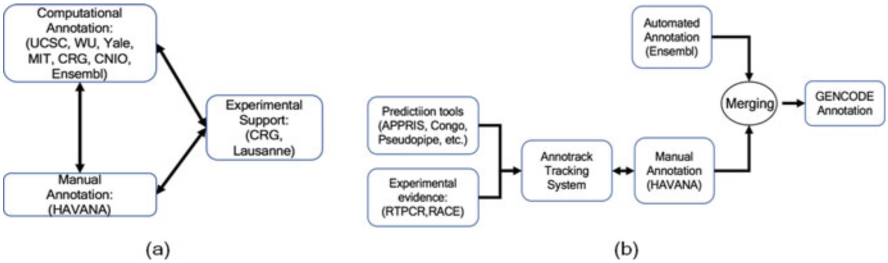

Computational Annotation: (UCSC, WU, Yale, MIT, CRG, CNIO, Ensembl)
Manual Annotation: (HAVANA)
Experimental Support: (CRG, Lausanne)
(a)
Automated Annotation (Ensembl)
Prediction tools (APPRIS, Congo, Pseudopipe, etc.)
Experimental evidence: (RTPCR,RACE)
Annotrack Tracking System
Manual Annotation (HAVANA)
Merging
GENCODE Annotation
(b)

Fig. 1.2 (**a**) Participating groups and their main role in GENCODE. Figure adapted from (Bignell et al. 2009). (**b**) High-level summary of GENCODE pipeline. Figure adapted from (Harrow et al. 2012)

is mainly involved in manual annotation (Bignell et al. 2009). Figure 1.2a highlights the collaborators of GENCODE and their key roles.

Computational predictions are generated by different groups independent of manual annotation. The predictions are used as both a guide for novel annotation and the validation of completed annotation. Experimental validation of novel transcripts is validated by RT-PCR, RACE, and RNA-Seq as well (Bignell et al. 2009). Figure 1.2b highlights the overall pipeline of GENCODE to generate the final annotations.

1.3.2 Automated Annotation Approaches

For automatic annotations, Ensembl gene annotation pipeline (Flicek et al. 2012) was leveraged for protein-coding genes. RefSeq and UniProt (UniProt Consortium 2012) ("protein existence" levels 1 and 2) sequences were considered as input for this pipeline. For novel protein-coding genes, GENCODE leveraged PhyloCSF (Lin et al. 2011) to identify potential genes with evolutionary signature. For long intergenic non-coding RNA (lincRNA) genes cDNA sequences and the regulatory data from the Ensembl project were used for annotation. If an lncRNA overlaps protein-coding genes in any exon at the same strand, it is filtered out (Derrien et al. 2012). The overlapped lncRNA genes ("genic") were further categorized as "exonic," "intronic," or "overlapping." The lncRNAs not intersecting with any protein-coding loci were defined as "intergenic" was further categorized as "sense," "convergent," or "divergent." Although the standard definition of lncRNAs requires a transcript to have more than 200 bases (Alam et al. 2020), the GENCODE lncRNA annotation contains a more than hundred spliced transcripts with length less than 200 bases, and there exists evidence of expression in those positions as well (Harrow et al. 2012). The Rfam (Rangan et al. 2020) and miRbase (Kozomara and Griffiths-Jones 2010) were used as input for Ensembl ncRNA prediction pipelines to annotate short non-coding RNAs (sncRNA). Three different pseudogene annotation pipelines were used to annotate the pseudogenes in the GENCODE consortium

(Frankish and Harrow 2014) and they were RetroFinder (Kent et al. 2003), PseudoPipe (Zhang et al. 2006), and PseudoFinder (Zheng et al. 2007). And manually annotation follows the computational annotation. It is important to emphasize that pseudogenes are computationally annotated as like protein-coding genes and non-coding RNA genes. It is the manual annotation that discriminates pseudogene loci from the functional genes.

1.3.3 Manual Curation Approaches

For manual annotation, GENCODE aligns the transcript to the genome. Then the genomic sequence, rather than the cDNA sequence, is considered as a reference (Harrow et al. 2012). The advantage of considering genomic annotation instead of cDNA annotation is that the alternative spliced variants can be determined with high confidence, as partial ESTs and protein evidence can be considered. But cDNA annotations are constrained to full-length transcripts. Additionally, genomic annotation can support a more robust analysis for pseudogenes. All the manual annotation for protein-coding genes, pseudogenes, and lncRNA genes were performed following the guidelines of the HAVANA (ftp://ftp.sanger.ac.uk/pub/annotation). In brief, HAVANA considers the transcriptomic (mRNA and EST) and proteomic (from UniProt and GenBank) data. Then this data is aligned to bacterial artificial chromosome (BAC) clone using BLAST. Then gene models were manually determined from the alignments by HAVANA annotators considering Otterlace (Searle et al. 2004). Dotter (Sonnhammer and Wootton 2001) was used by HAVANA annotators to resolve any alignment which was unclear or absent from regular Blixem Alignment Viewer (Sonnhammer and Wootton 2001). Short alignments (lower than 15 bases) were detected using Zmap (http://www.sanger.ac.uk/resources/software/zmap/). All canonical splice sites were verified and all non-redundant splicing transcripts from individual locus were used to construct the final transcript model. After constructing the transcript structure, the protein-coding potential of a transcript was determined based on the similarity to known proteins, orthologous and paralogous proteins, based on the presence of Pfam functional domains (Finn et al. 2014). Details of the manual annotation are highlighted in (Harrow et al. 2012). To leverage a detailed investigation of putative pseudogene loci that were identified by three computational tools (RetroFinder, PseudoPipe, and PseudoFinder), GENCODE pseudogenes are all manually validated. For manual annotation, annotators take the support of multiple tools, like Zmap[1] (a genome browser to provide faster access to voluminous data), Dotter[2] (dot-matrix based program to compare two sequences in more detail), Otter[3] (a graphical interactive client for human

[1] https://www.sanger.ac.uk/tool/zmap/

[2] https://sonnhammer.sbc.su.se/Dotter.html

[3] https://www.sanger.ac.uk/tool/otter/

annotators for comparing multiple sources of evidence; previously known as Otterlace), Blixem[4] (an interactive pairwise sequence alignment browser under seqtools), etc. GENCODE manual annotation demonstrated a near perfect specificity for pseudogenes at the cost of sensitivity (Frankish and Harrow 2014).

1.3.4 Merging the Automated and Manual Curation Results

Merging the results from the automated pipeline and manually curated pipeline is a complex task carried out by the GENCODE consortium since last decade. In summary, when transcript models from both the HAVANA and Ensembl agree for all coding exons and all non-coding exons, then the HAVANA model is used for GENCODE annotation. When transcript models from both the HAVANA and Ensembl agree on all exons, but a mismatch in the outer 5′-start or 3′ -end, only then the HAVANA model will be considered for the GENCODE annotation. When transcript models from both the HAVANA and Ensembl agree on all exons, but a structural difference is observed in non-coding regions, both the HAVANA and Ensembl models will be considered for the GENCODE annotation. Moreover, when transcript models from both the HAVANA and Ensembl have unique exon locations and structure, then both the HAVANA and Ensembl models will be considered for the GENCODE annotation.

For more rigorous annotation of the genes, other tools are used under the GENCODE pipeline. For example, APPRIS (Rodriguez et al. 2013) was used to annotate splice isoforms for protein-coding genes. Among all the alternatively spliced variants of protein-coding genes, one of them is labeled as "principal" isoform based on structural, functional, and cross-species conservation of the transcript. The details of this APPRIS "tag"[5] can be found in the 9th column of General Transfer Format/General Feature Format (GTF/GFF3)[6] file shared in GENCODE portal. Moreover AnnoTrack software system was developed by GENCODE consortium for tracking and managing information coming from multiple sources, collaborators and make efficient prioritization and proper resolution of problems during this complex process of annotation (Kokocinski et al. 2010).

1.3.5 Experimental Validation

Multiple experimental approaches RT-PCR, RACESeq, RNA Capture Long Sequence (CLS) (Lagarde et al. 2017) have been leveraged by the GENCODE

[4] https://www.sanger.ac.uk/tool/seqtools/

[5] https://www.gencodegenes.org/pages/tags.html

[6] http://asia.ensembl.org/info/website/upload/gff.html

consortium to annotate as well for the validation of genes. For protein-coding genes and transcripts, multiple mass spectrometry datasets have been used to remove false positives (Frankish et al. 2019). The 5' and 3' end of GENCODE lncRNA annotations were less supported by CAGE (Kodzius et al. 2006) and PET (Hon et al. 2017). To improve the lncRNA annotation, the GENCODE consortium developed the CLS method (Lagarde et al. 2017) which supports to capture long length lncRNA transcripts often representing full length complete 5′-to-3′ RNA molecules which substantially support the manual annotation of the lncRNA transcripts. To validate the annotations, GENCODE leveraged RT-PCR and multiplexed sequence readout (Howald et al. 2012) for the novel and putative transcripts. For human GENCODE version 3 up to 19, eight different human tissues, i.e., liver, lung, brain, kidney, skeletal muscle, spleen, and testis were heart, experimentally tested by RT-PCR and confirmed 78% of all splice junctions tested (Frankish et al. 2019). GENCODE also performed 5′ and 3′ nested RACE experiments in seven different tissues (i.e., liver, lung, brain, kidney, skeletal muscle, spleen, and testis) followed by long-read sequencing revealing 10,380 novel splice junction candidates (Frankish et al. 2019).

1.3.6 Annotation from Mouse Genome

The mouse reference genome annotation was started in 2012 by the GENCODE consortium. Mostly, it has been annotated using the clone-by-clone approach. The entire mouse reference genome was annotated manually completely. There were 133 orthologous loci for human protein-coding genes annotated and identified (Mudge and Harrow 2015). HAVANA group uses bespoke software to align sequences and other information for a genome (Harrow et al. 2012; Harrow et al. 2014). On the contrary, Ensembl produced gene set for mouse genome from computational pipelines (Cunningham et al. 2015). Genome annotation is useful because of the genome scaffolds when aligning of transcriptional evidences. However, errors could occur and propagated in the models due to lack of transcript evidences. Frequently HAVANA uses other species annotations to annotate mouse models (Mudge and Harrow 2015). Usually, the data related to mouse genes are updated every quarter year (Harrow et al. 2012). These releases are results of the combination of the manual and combination models. Even though, the manual is more robust, but the computational model is essential since it fills up the gaps and the undiscovered and unthought-of data (Guigó et al. 2006). HAVANA group annotated pseudogenes and lncRNAs in human more than in mouse (Derrien et al. 2012) (Pei et al. 2012). However, there are more than 2000 protein-coding genes found in mouse compared to human. Consensus Coding Sequence (CCDS) project is collaboration project to annotate protein-coding genes in mouse and the genes that are agreed on by the HAVANA, Ensembl, RefSeq, and the European Conditional Knockout Mouse Consortium (EUCOMM).

1.4 Data Format Available for GENCODE Annotation

Annotations are provided in both GENCODE GTF (General Transfer Formal) format to ENCODE Data Coordination Center (DCC) to integrate with genome browser. Moreover, GFF3 (General Feature Format version 3) format is also provided for all annotation (Harrow et al. 2012). Both GTF and GFF hold nine fields in tab-delimited format (Fig. 1.3a). The fields are (a) chromosome name, (b) source of annotation (HAVANA, Ensembl), (c) feature type (exon, CDS, start codon, stop codon, UTR, selenocysteine, gene, transcript,), (d) start coordinate (1-based index), (e) end coordinate, (f) score (not used for annotation), (g) strand (+, −), (h) CDS (0,1,2 indicates the first, second, or third base of this feature represent the first base on codon, respectively), (i) key-value pairs. Figure 1.3 highlights the overall structure of the GENCODE GFF3 file format to highlight the details of annotation.

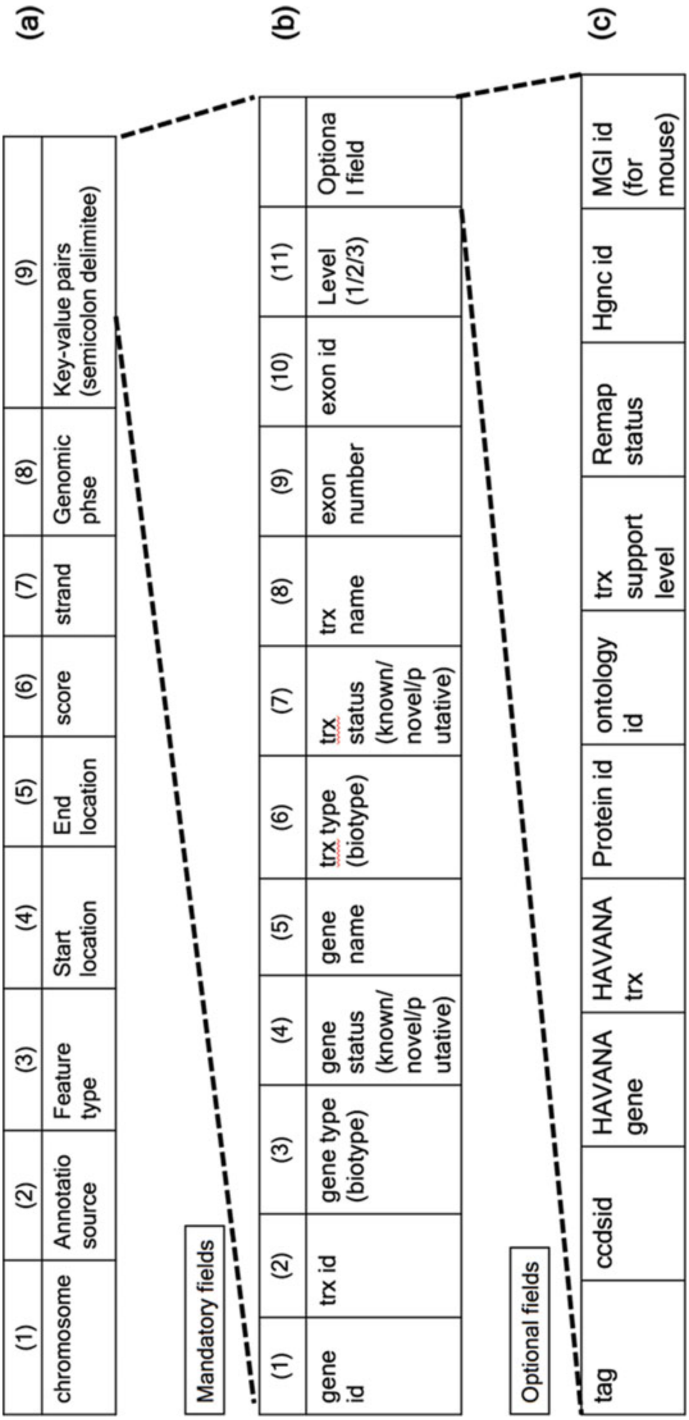
The complete list of key-value pairs for the 9th column of GENCODE GTF files is described in the GENCODE portal.[7] The 9th column (key-value pairs) of the GENCODE GTF file contains eleven mandatory fields and other optional fields (Fig. 1.3b). The mandatory key-value pairs are: (a) gene id, (b) transcript id, (c) gene type, (d) gene status (novel, known, or putative), (e) gene name, (f) transcript type, (g) transcript status (novel, known, or putative), (h) transcript name, (i) exon number (j) exon id, (k) level (1: verified loci, 2: manually annotated loci, 3: automatically annotated loci). The 3rd and the 6th mandatory key-value pairs represent the gene type and transcript type, respectively, and the possible types (also called biotypes) are highlighted in the GENCODE portal.[8] Using these biotypes, user can easily identify lncRNA, lincRNA, pseudogene, protein-coding genes, etc. It is important to emphasize that the gene status and transcript status (the 4th and the 7th mandatory fields are provided until the release 25 and 11 for human and mouse, respectively). There is a provision of optional key-value pairs (also called “tag”) in the 9th column of GENCODE GFF3 file (Fig. 1.3c). The “tags” provides more complete description of the gene and transcript models from the GENCODE annotation. The details of these tags are described in the GENCODE portal.[9]

Ensemble GTF is identical to GFF3. But there is a slight difference between GENCODE GTF and Ensembl GTF format. Genes that are common to the human chromosome X and Y pseudo-autosomal region (PAR) regions are mentioned twice in the GENCODE GTF. In contrast, this type of gene is only mentioned under chromosome X in the Ensembl GTF. Additionally, the 9th column of the GENCODE GTF file contains some attributes that are not present in the Ensembl GTF such as: annotation notes, APPRIS tags, experimental validation by the GENCODE or pseudogenes predicted by HAVANA, Yale or UCSC group.

[7] https://www.gencodegenes.org/pages/data_format.html

[8] https://www.gencodegenes.org/pages/biotypes.html

[9] https://www.gencodegenes.org/pages/tags.html



(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
chromosome	Annotatio source	Feature type	Start location	End location	score	strand	Genomic phse	Key-value pairs (semicolon delimitee)

(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	
gene id	trx id	gene type (biotype)	gene status (known/novel/putative)	gene name	trx type (biotype)	trx status (known/novel/putative)	trx name	exon number	exon id	Level (1/2/3)	Optional field

tag	ccdsid	HAVANA gene	HAVANA trx	Protein id	ontology id	trx support level	Remap status	Hgnc id	MGI id (for mouse)

Fig. 1.3 The GENCODE GFF3 file format to represent annotation in detail

1.5 GENCODE Data Access

The GENCODE consortium team has made huge effort to make the data accessing user-friendly for bioinformatics subject matter experts as well as non-domain expert people. Documentations and examples are available on almost every page of the GENCODE web portal (https://www.gencodegenes.org/). The GENCODE consortium team has catered for many different types of users. They have provided different tools and many different techniques to access the GENCODE databases. A simple user with no programming or database knowledge can interact with the GENCODE main website and accomplish many different tasks like downloading chromosome annotation files (in GFF3 and GTF format) or FASTA sequence files. For more sophisticated users, who have programming experience, the consortium provides tools and packages to help developers programmatically interact with the consortium databases and accomplish different tasks and pipelines supporting their research. They have provided R-Packages to facilitate data discovery and data mining on current and previous datasets within the same species or different ones. The team has provided programming tools and techniques to automate accessing and processing databases. This includes web services, File Transfer Protocol (FTP) services, Application Programming Interface Services (API), and Bioinformatics tools for (R and Python) packages. Bioinformatics teams and user with programming capabilities can use the GENCODE Application Interface (API) to programmatically interact with cloud services and databases to query and download gene data required for the analysis and the research. Other options allow bioinformatics team to download the whole genome database on local server and query and interact with it for the required data. GENCODE dataset can be accessed mainly from three different platforms: (a) GENCODE portal, (b) Ensembl Genome Browser, and (c) UCSC Genome Browser. Figure 1.4 shows the three main ways of accessing the GENCODE dataset and we will briefly describe each of them below. For users' convenience, an interactive user guide highlighting the data access options is shared in Supplementary Data 1.1.

1.5.1 Data Access from GENCODE Web Portal

The current release and the previous release of GENCODE annotations are publicly accessible at www.gencodegenes.org. This included all the annotation files and the corresponding sequence files for both human and mouse genomes. Users can browse this site and easily download the required annotation. All the versions of GENCODE gene set are released every two to four times a year for mouse and human. The current GENCODE human annotation (Release 38) and the current one for the mouse (Release 27), both include annotation files (in GTF and GFF3 formats), FASTA files, and METADATA files associated with the GENCODE annotation on all genomic regions.

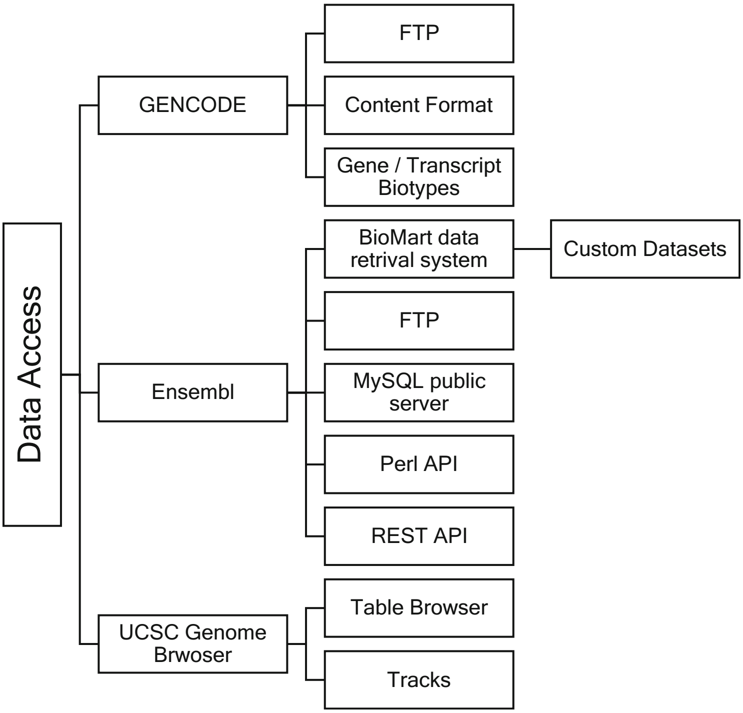


Fig. 1.4 GENCODE Data access platforms

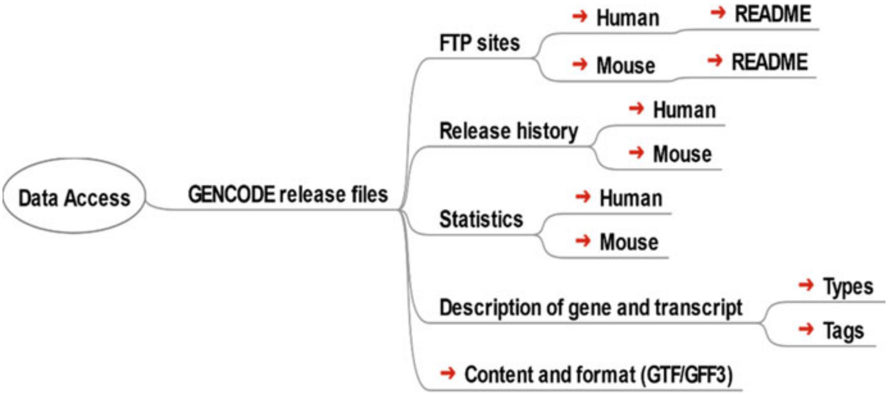


Fig. 1.5 GENCODE Data access from GENCODE web portal and associated ftp site. The figure was generated using open-source software FreeMind (http://freemind.sourceforge.net/wiki/index.php/Download)

As shown in Fig. 1.5, the user can download the annotation files for all previous release of human and mouse as needed. Statistics is always provided for the most recent release and the update of the annotation files for both and mouse human

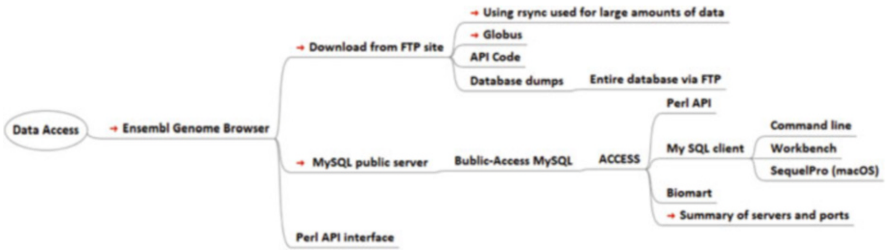


Fig. 1.6 GENCODE data access from Ensembl Genome Browser. The figure was generated using open-source software FreeMind

(Release date 05.2021 with Freeze data 12.2020). In the Documentation tab, the user can find Annotation Data Format, Tags, and Biotypes of the gene set. GENCODE data can be downloaded from the ftp site for human (http://ftp.ebi.ac.uk/pub/databases/gencode/Gencode_human/) and mouse (http://ftp.ebi.ac.uk/pub/databases/gencode/Gencode_mouse).

For parsing the GENCODE annotation (in GTF or GFF3 format) some tools, i.e., gtfparse,[10] gffutils[11] are already available in Python. Moreover, users can use BioJava[12] and BioPerl[13] for parsing and analyzing GENCODE dataset as these packages are considered as standard bioinformatics tools for sequence analysis.

1.5.2 Data Access from Ensembl Genome Browser

As GENCODE is the default gene set for Ensembl, different ways are provided to access the annotation from Ensembl Genome Browser (http://www.ensembl.org/). Figure 1.6 highlights different ways, depending on the amount and the type of data the user is interested in, of accessing the GENCODE dataset from Ensembl Genome Browser.

If a user is interested about the entire database, then it is also available via FTP as MySQL dumps.[14] But for small data quantities like a single gene sequence or single transcript, the Ensembl Genomic Browser (http://asia.ensembl.org/index.html) offers an “Export data”[15] (mainly on the left-hand menu of Ensemble Genome Brower pages) to export either FASTA sequence or GTF/GFF features. The user

[10] https://github.com/openvax/gtfparse

[11] https://github.com/daler/gffutils

[12] https://biojava.org/

[13] https://bioperl.org/

[14] https://asia.ensembl.org/info/docs/webcode/mirror/install/ensembl-data.html

[15] http://asia.ensembl.org/info/data/export.html

can click on "Export data" and have it for further downstream processing. The exports are also available to download from the Ensembl FTP site at: http://ftp.ensembl.org/pub/. Readers are suggested to check the use case 2 of this article to know more about this export option.

The Ensembl REST server is a great resource and service for language-agnostic programmatic access to the GENCODE annotation. The user request data via GET command and information is returned in JSON or XML format. Many examples[16] are provided in Ensembl REST server highlighting the command and the expected outcome returned using Java, R, Perl, Python, Ruby, Curl, and Wget command. The Ensembl REST API Endpoints are available for the users at https://rest.ensembl.org.

Ensembl also provides a user-friendly data mining tool called BioMart (http://asia.ensembl.org/info/data/biomart/index.html), and it can be used for more complex cross database queries based on GENCODE annotation. The BioMart is considered a user-friendly web-based tool that allows the extraction of annotation data without programming language background or understanding of the details of the database structure. The consortium team provides three ways to interact with BioMart: (a) via the biomaRt R package,[17] (b) via RESTful access[18] (using Perl and Wget), and (c) via Perl API.[19]

1.5.3 Data Access from UCSC Genome Browser

The UCSC Genome Browser (https://genome.ucsc.edu/) hosts selected releases from GENCODE. GENCODE annotation data can also be downloaded from the Table Browser[20] option of UCSC Genome Browser. Usually, there is a short time lag between the latest GENCODE release and corresponding track in the UCSC browser due to the effort required by the UCSC team to link the latest GENCODE annotation to massive amount of other relevant tracks available under UCSC Genome Browser.

1.6 Use Cases to Utilize GENCODE

In this section, we would like to highlight a use case on the usage of GENCODE reference annotation.

[16] https://rest.ensembl.org/documentation/info/data

[17] http://asia.ensembl.org/info/data/biomart/biomart_r_package.html

[18] http://asia.ensembl.org/info/data/biomart/biomart_restful.html

[19] http://asia.ensembl.org/info/data/biomart/biomart_perl_api.html

[20] http://genome.cse.ucsc.edu/cgi-bin/hgTables

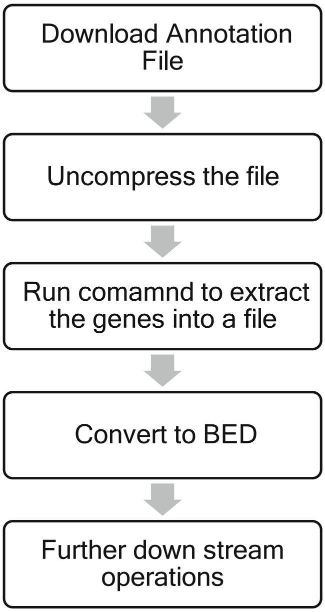

Download Annotation File
Uncompress the file
Run comamnd to extract the genes into a file
Convert to BED
Further down stream operations

Fig. 1.7 Outline to download, parse, and convert into BED format

1.6.1 Use Case 1: Extracting Different Types of Genes from GENCODE for Downstream Analysis

If a user is interested in specific types of genes from the Human GENCODE v38 annotation, we will show how to download, extract the annotation, and convert into browser extensible data (BED) format for downstream analysis. This use case is outlined in Fig. 1.7.

The general overall steps need to accomplish tasks in this use case are summarized in the following workflow:

Step 1:. The user can download the GFF3 annotation file from GENCODE (https://www.gencodegenes.org/human/). The recent release for human annotation is version 38. The user can use the following command to download the annotation file:

```
wget
```

http://ftp.ebi.ac.uk/pub/databases/gencode/Gencode_human/release_38/gencode.v38.annotation.gff3.gz

Step 2:. The user can decompress the archive using the following command:

```
gunzip gencode.v38.annotation.gff3.gz
```

This will generate the uncompressed file gencode.v38.annotation.gff3.

Step 3:. The user can view the first 10 lines of the uncompressed line using the following command:

```
head gencode.v38.annotation.gff3
```

Step 4:. Now the user can select the desired gene type (e.g., protein-coding, lncRNA, snRNA, or the pseudogenes).

For protein-coding genes, the user can use the following command:

```
  awk '{if($3=="gene" && $9~"protein_coding" ){print $0}}'
gencode.v38.annotation.gff3 > pc_genes.gff3
```

For lncRNA genes, the user can use the following command:

```
  awk '{if($3=="gene" && $9~"gene_type=TEC|lncRNA" ){print
$0}}' gencode.v38.annotation.gff3 > lncRNA_genes.gff3
```

For snRNA genes, the user can use the following command:

```
  awk '{if($3=="gene" && $9 ~ "gene_type=(miRNA|misc_RNA|
Mt_rRNA|Mt_tRNA|ribozyme|rRNA|rRNA_pseudogene|scaRNA|
scRNA|snoRNA|snRNA|sRNA|vault_RNA)" ) {print $0}} ' gencode.
v38.annotation.gff3 > snRNA_genes.gff3
```

For pseudogenes, the user can use the following command:

```
  awk '{if($3=="gene" && $9~"gene_type=(pseudogene|
processed_pseudogene|transcribed_processed_pseudogene|
translated_processed_pseudogene|
transcribed_unprocessed_pseudogene|
translated_unprocessed_pseudogene|unprocessed_pseudogene|
unitary_pseudogene|transcribed_unitary_pseudogene|
polymorphic_pseudogene|transcribed_unprocessed_pseudogene|
unprocessed_pseudogene|IG_C_pseudogene|IG_J_pseudogene|
IG_V_pseudogene|TR_J_pseudogene|TR_V_pseudogene)" ){print
$0}}' gencode.v38.annotation.gff3 > pseudo_genes.gff3
```

Step 5:. Now the user can select the desired gene type (e.g., protein-coding, lncRNA, snRNA, or the pseudogenes) and convert the GFF3 file into BED file using the following command (Assuming that gff2bed Linux package[21] is installed):

[21] BEDOPS: the fast, highly scalable and easily-parallelizable genome analysis toolkit — BEDOPS v2.4.39

```
gff2bed --keep-header < pc_genes.gff3 > pc_genes.bed
```

The default usage of the command gff2bed strips the leading header (##gff-version 3) but adding the --keep-header option will preserve this as a BED element that uses _header as a chromosome name. For more details, please refer to gfft2bed usage manual page.[22]

The rest of the gene bed files can be produced the same way. Although we have applied the workflow on the human genome, the same process applies to the mouse genome. The shell commands for this use case are available as Supplementary Data 1.2.

1.6.2 Use Case 2: Exploration of the Annotation of lncRNA MALAT1

In this use case, we will demonstrate how a user can explore GENCODE annotation of lncRNA gene MALAT1 using Ensembl Genome Browser (Fig. 1.8).

Step 1:. The user can download the comprehensive annotation file and find the annotation of MATLA1 by using shell command on the input annotation file gencode.v38.annotation.gff3 in the following way:

```
grep MALAT1 gencode.v38.annotation.gff3 > MALAT1.gff3
```

The above-mentioned shell command will return the annotation of MALAT1 gene, 17 transcripts associated with MALT1 genes and associated exons (Supplementary Data 1.3).

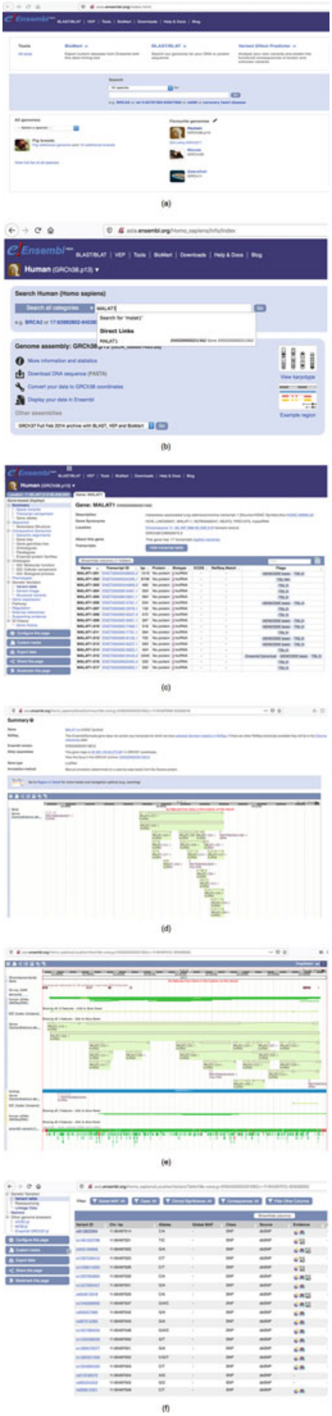
Step 2:. Now, user can use Ensembl Genome Browser (http://asia.ensembl.org/index.html) to explore the GENCODE annotation of MALAT1. Ensembl Genome Browser considers GENCODE annotation as its default annotation. Then the user can select human as the organism (Fig. 1.8a) to load the annotation of human genes.

Step 3:. Then the user can type "MALAT1" in the search box and it will automatically suggest the name MALAT1 (Fig. 1.8b).

Step 4:. Once the user clicks on "GO" button, this will load the summary of MALAT1 in Ensembl Genome Browser (Fig. 1.8c). On the left side panel, there is a panel where users can go into the details of comparative genomics, genetic variation, markers, etc. The is an "Export data" option on the left panel and user can download MALAT1 related annotations as well as the sequence from this "Export data" option.

[22] https://bedops.readthedocs.io/en/latest/content/reference/file-management/conversion/gff2bed.html#downloads

Fig. 1.8 Discovering MALAT1 and associated annotation using Ensemble Genome Browser


(a)
Ensembl
BLAST/BLAT | VEP | Tools | BioMart | Downloads | Help & Docs | Blog
Human (GRCh38.p13)
Search Human (Homo sapiens)
Search all categories
MALAT1
Search for 'malat1'
Direct Links
MALAT1
e.g. BRCA2 or 17:63992802-64038237
Genome assembly: GRCh38.p13
More information and statistics
Download DNA sequence (FASTA)
Convert your data to GRCh38 coordinates
Display your data in Ensembl
Other assemblies
GRCh37 Full Feb 2014 archive with BLAST, VEP and BioMart
View karyotype
Example region
(b)
Gene: MALAT1
(c)
Summary
(d)
(e)
(f)

Step 5:. If user scrolls down, then more tails of the MALAT1 can be found where the user can see the location of all the 17 transcripts, exons under each transcript and the directionality of MALAT1 transcripts (Fig. 1.8d).

Step 6:. Then the user can click “Region in Detail” (Fig. 1.8d) to find the more tracks and navigation options of MALAT1 (Fig. 1.8e). Figure 1.8e shows some of the available tracks (e.g., sequence and assembly, genes and transcripts, mRNA and protein alignments, variations (SNP, INDEL, structural variants), regulation, comparative genomics, etc.) that are available in Ensemble Genome Browser.

Step 7:. If a user is interested to explore the genetic variants of MALAT1, he/she can click the “Genetic Variation” on the left panel (Fig. 1.8c) and it will load the details of all variants related to MALAT1 (Fig. 1.8f).

1.6.3 Use Case 3: Exploration of SUMO1P3 (Small Ubiquitin Like Modifier 1 Pseudogene 3)

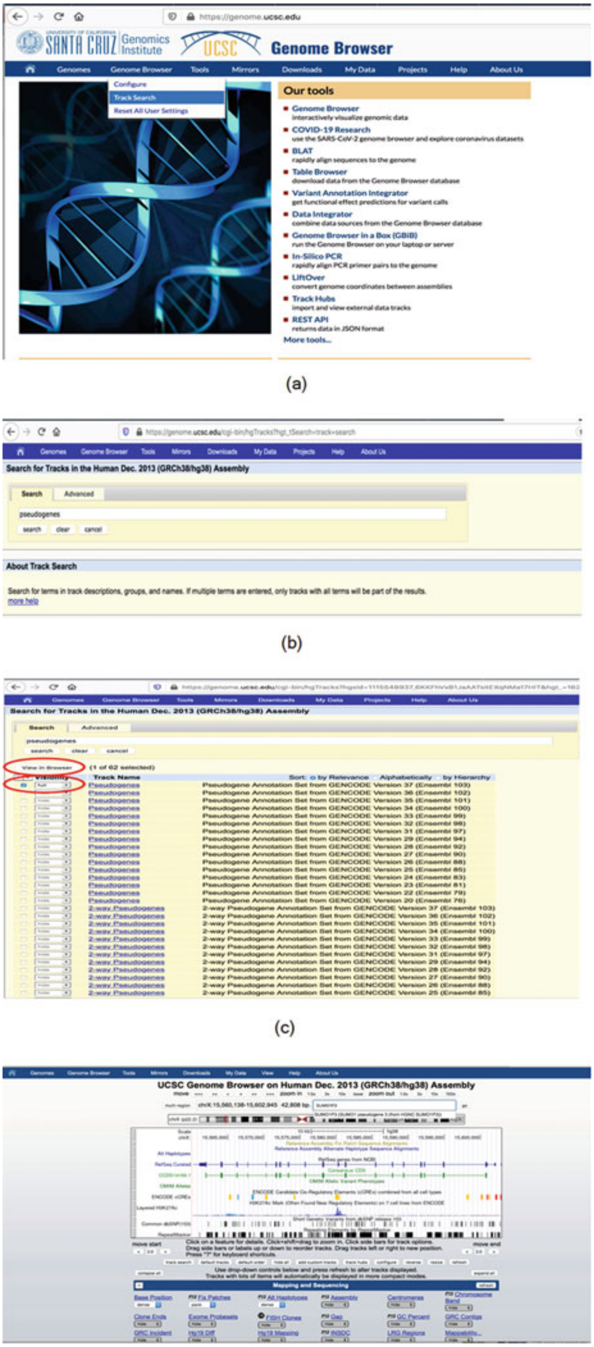
In this use case, we will demonstrate how a user can explore the GENCODE annotation of SUMO1P3 pseudogene. SUMO1P3 pseudogenes are upregulated for patients with gastric cancer and can be used to determine cancer comparing against the benign gastric disease (Emadi-Baygi et al. 2017). Figure 1.9 demonstrates how a user can leverage UCSC Genome Browser to explore the loci and related mutations on SUMO1P3.

Step 1:. The user is advised to go to UCSC Genome Browser via the main webpage at http://genome-euro.ucsc.edu/ and choose “Genome Browser” tab. From the drop-down menu, user needs to select “Track Search” tab (Fig. 1.9a).

Step 2:. Then, the user searches for “pseudogenes” and clicks “search” button (Fig. 1.9b). Then, it will load all the tracks related to pseudogenes. User needs to uncheck all tracks and select on the latest GENCODE pseudogene track at the top of this page (Fig. 1.9c). The user needs to click on “View on Browser” button.

Step 3:. Then, the user will see the default view of UCSC Genome Browser (Fig. 1.9d). Then the user needs to search for “SUMO1P3” and click go button. Then, website shows the default information on the figure at the beginning of the webpage (Fig. 1.9e). The SUMO1P3 pseudogene will be highlighted as a pink colored bar in the UCSC Genome Browser (Fig. 1.9e).

Step 4:. Then user can control the information that needed to be displayed by selecting/unselecting different tracks. For example, by selecting “TCGA Pan-cancer” (under Phenotype and Literature) and “GTEx gene” (under Expression) and hiding other options, user can check the mutations and the expression shown in Fig. 1.9f. To know more details about the gene expression, user can click on the SUMO1P3 to get more details about the



Our tools

- Genome Browser
  interactively visualize genomic data
- COVID-19 Research
  use the SARS-CoV-2 genome browser and explore coronavirus datasets
- BLAT
  rapidly align sequences to the genome
- Table Browser
  download data from the Genome Browser database
- Variant Annotation Integrator
  get functional effect predictions for variant calls
- Data Integrator
  combine data sources from the Genome Browser database
- Genome Browser in a Box (GBiB)
  run the Genome Browser on your laptop or server
- In-Silico PCR
  rapidly align PCR primer pairs to the genome
- LiftOver
  convert genome coordinates between assemblies
- Track Hubs
  import and view external data tracks
- REST API
  returns data in JSON format

More tools...

(a)

Search for Tracks in the Human Dec. 2013 (GRCh38/hg38) Assembly

pseudogenes

About Track Search

Search for terms in track descriptions, groups, and names. If multiple terms are entered, only tracks with all terms will be part of the results.
more help

(b)

https://genome.ucsc.edu/cgi-bin/hgTracks?hgsid=1115548937_6KKFhVvB1JaAATsltEXqNMaf7HfT&hgt_=162
Genomes
Genome Browser
Tools
Mirrors
Downloads
My Data
Projects
Help
About Us

Search for Tracks in the Human Dec. 2013 (GRCh38/hg38) Assembly

pseudogenes

(1 of 62 selected)

Sort: ● by Relevance ○ Alphabetically ○ by Hierarchy

	Visibility	Track Name	
☑	full	Pseudogenes	Pseudogene Annotation Set from GENCODE Version 37 (Ensembl 103)
□	hide	Pseudogenes	Pseudogene Annotation Set from GENCODE Version 36 (Ensembl 102)
□	hide	Pseudogenes	Pseudogene Annotation Set from GENCODE Version 35 (Ensembl 101)
□	hide	Pseudogenes	Pseudogene Annotation Set from GENCODE Version 34 (Ensembl 100)
□	hide	Pseudogenes	Pseudogene Annotation Set from GENCODE Version 33 (Ensembl 99)
□	hide	Pseudogenes	Pseudogene Annotation Set from GENCODE Version 32 (Ensembl 98)
□	hide	Pseudogenes	Pseudogene Annotation Set from GENCODE Version 31 (Ensembl 97)
□	hide	Pseudogenes	Pseudogene Annotation Set from GENCODE Version 29 (Ensembl 94)
□	hide	Pseudogenes	Pseudogene Annotation Set from GENCODE Version 28 (Ensembl 92)
□	hide	Pseudogenes	Pseudogene Annotation Set from GENCODE Version 27 (Ensembl 90)
□	hide	Pseudogenes	Pseudogene Annotation Set from GENCODE Version 26 (Ensembl 88)
□	hide	Pseudogenes	Pseudogene Annotation Set from GENCODE Version 25 (Ensembl 85)
□	hide	Pseudogenes	Pseudogene Annotation Set from GENCODE Version 24 (Ensembl 83)
□	hide	Pseudogenes	Pseudogene Annotation Set from GENCODE Version 23 (Ensembl 81)
□	hide	Pseudogenes	Pseudogene Annotation Set from GENCODE Version 22 (Ensembl 79)
□	hide	Pseudogenes	Pseudogene Annotation Set from GENCODE Version 20 (Ensembl 76)
□	hide	2-way Pseudogenes	2-way Pseudogene Annotation Set from GENCODE Version 37 (Ensembl 103)
□	hide	2-way Pseudogenes	2-way Pseudogene Annotation Set from GENCODE Version 36 (Ensembl 102)
□	hide	2-way Pseudogenes	2-way Pseudogene Annotation Set from GENCODE Version 35 (Ensembl 101)
□	hide	2-way Pseudogenes	2-way Pseudogene Annotation Set from GENCODE Version 34 (Ensembl 100)
□	hide	2-way Pseudogenes	2-way Pseudogene Annotation Set from GENCODE Version 33 (Ensembl 99)
□	hide	2-way Pseudogenes	2-way Pseudogene Annotation Set from GENCODE Version 32 (Ensembl 98)
□	hide	2-way Pseudogenes	2-way Pseudogene Annotation Set from GENCODE Version 31 (Ensembl 97)
□	hide	2-way Pseudogenes	2-way Pseudogene Annotation Set from GENCODE Version 29 (Ensembl 94)
□	hide	2-way Pseudogenes	2-way Pseudogene Annotation Set from GENCODE Version 28 (Ensembl 92)
□	hide	2-way Pseudogenes	2-way Pseudogene Annotation Set from GENCODE Version 27 (Ensembl 90)
□	hide	2-way Pseudogenes	2-way Pseudogene Annotation Set from GENCODE Version 26 (Ensembl 88)
□	hide	2-way Pseudogenes	2-way Pseudogene Annotation Set from GENCODE Version 25 (Ensembl 85)

(c)

UCSC Genome Browser on Human Dec. 2013 (GRCh38/hg38) Assembly

chrX:15,560,138-15,602,945 42,808 bp. SUMO1P3

SUMO1P3 (SUMO1 pseudogene 3 (from HGNC SUMO1P3))

Scale
chrX:
Reference Assembly Fix Patch Sequence Alignments
Reference Assembly Alternate Haplotype Sequence Alignments
Alt Haplotypes
RefSeq genes from NCBI
RefSeq Curated
Consensus CDS
OMIM Allelic Variant Phenotypes
OMIM Alleles
ENCODE Candidate Cis-Regulatory Elements (cCREs) combined from all cell types
ENCODE cCREs
H3K27Ac Mark (Often Found Near Regulatory Elements) on 7 cell lines from ENCODE
Layered H3K27Ac
Short Genetic Variants from dbSNP release 153
Common dbSNP(153)
Repeating Elements by RepeatMasker
RepeatMasker
move start
move end

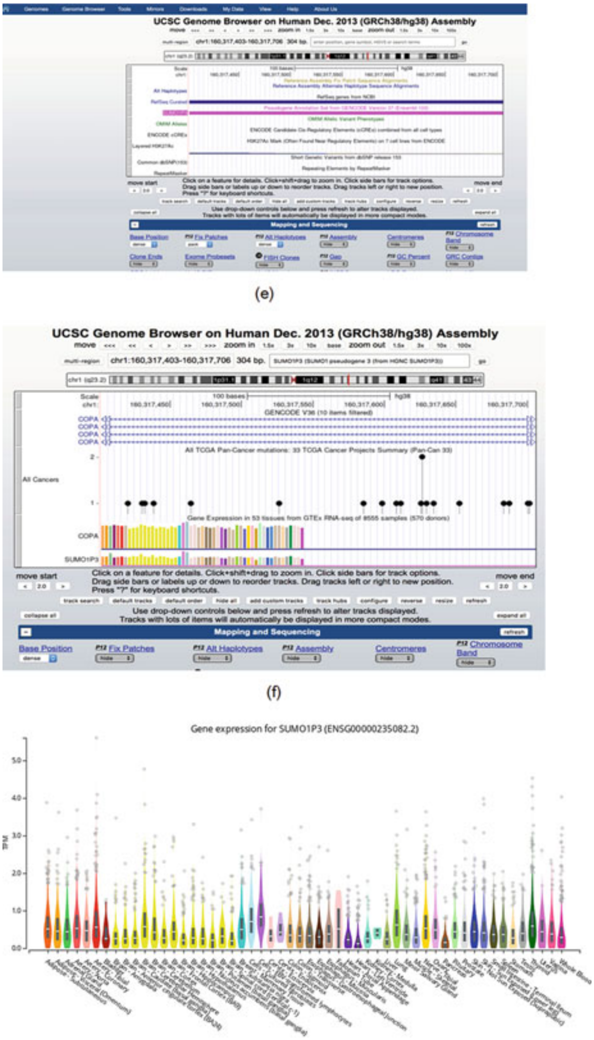
Click on a feature for details. Click+shift+drag to zoom in. Click side bars for track options.
Drag side bars or labels up or down to reorder tracks. Drag tracks left or right to new position.
Press "?" for keyboard shortcuts.

Use drop-down controls below and press refresh to alter tracks displayed.
Tracks with lots of items will automatically be displayed in more compact modes.

Mapping and Sequencing

(d)

Fig. 1.9 Exploration of the pseudogene SUMO1P3 using UCSC Genome Browser. Figure (**a**) shows TCGA Pan-Cancer mutations in SUMO1P3 pseudogene. Figure (**b**) shows the gene expression of SUMO1P3 pseudogene compared with the reverse gene (COPA). Figure (**c**) shows SUMO1P3 pseudogene expression in different tissues as shown in GTEx portal



UCSC Genome Browser on Human Dec. 2013 (GRCh38/hg38) Assembly

(e)

UCSC Genome Browser on Human Dec. 2013 (GRCh38/hg38) Assembly

chr1:160,317,403-160,317,706 304 bp.

Click on a feature for details. Click+shift+drag to zoom in. Click side bars for track options.
Drag side bars or labels up or down to reorder tracks. Drag tracks left or right to new position.
Press "?" for keyboard shortcuts.

Use drop-down controls below and press refresh to alter tracks displayed.
Tracks with lots of items will automatically be displayed in more compact modes.

(f)

Gene expression for SUMO1P3 (ENSG00000235082.2)

(g)

Fig. 1.9 (continued)

expression of SUMO1P3. User can even view the details of SUMO1P3 from GTEx portal (https://www.gtexportal.org/home/gene/ENSG00000235082) as well (Fig. 1.9g).

1.7 Latest Update of the GENCODE Annotation

GENCODE provides the standard and well-accepted annotation for human and mouse genome functional elements for nearly two decades. Still, they are improving their pipeline to deliver more robust and accurate annotation. Recently GENCODE has developed a new pipeline, TAGENE which is capable of supporting long-read transcripts generated by Pacific Biosciences (PacBio) and Oxford Nanopore Technologies (ONT) (Frankish et al. 2021). GENCODE has generated more than 36M ONT reads and 2M PacBio Sequel reads to identify nearly 1600 human and 4500 mouse potential novel loci. The details of TAGENE can be found in (Frankish et al. 2021). GENCODE also initiated The Matched Annotation from NCBI and EMBL-EBI project (MANE) collaboration project between Ensembl, GENCODE, and RefSeq to identify a default protein-coding transcript from each human protein-coding locus that could be considered as a representative considering the underlying biology, overall expression, and conservation (Frankish et al. 2021). The recent release of MANE (version 0.91) provides an overall coverage of 84% of all the protein-coding genes of GENCODE (Frankish et al. 2021). Users can use the "Ensembl_canonical" tag in GENCODE annotation to find the representative protein-coding transcript. Under GIFTS (Genome Integration with FuncTion and Sequence) effort, GENCODE is trying to improve the interoperability between UniProt and GENCODE annotation (Frankish et al. 2021). Under GIFTS initiative, GENCODE investigated 1044 unmapped human and mouse proteins from UniProt and identified specific cases to update respective GENCODE annotation (Frankish et al. 2021). To tackle the COVID-19 pandemic, GENCODE reviewed and released the annotation protein-coding genes that are associated with SARS-CoV-2 infection (Frankish et al. 2021). Considering the human proteins that could be physically associated with SARS-CoV-2 protein (Gordon et al. 2020) or could be considered for drug repurposing (Zhou et al. 2020), the GENCODE released "COVID-19 genes track hub"[23] for the exploration associated protein-coding genes. As of now (up to 22 May 2021), GENCODE enlisted 282 COVID-19 associated genes under their annotation.

GENCODE annotations are current available on the reference assembly of GRCh38 for human and GRCm38 for the mouse. The resulted mapping on GRCh37 (for human) was not manually verified by GENCODE and this may contain some errors in complicated regions (e.g., gapped regions, repeated regions, etc.) of the human genome. GENCODE, therefore, recommends the GRCh38 annotations for the users.

[23] http://ftp.ebi.ac.uk/pub/databases/gencode/covid19_trackhub/data/

1.8 Conclusion

The GENCODE project provides comprehensive gene annotation for the human and mouse genomes, as part of the wider Ensembl project. In this chapter we have attempt to give the user a roadmap to facilitate genome data access huge amounts of information. We have introduced workflows to guide the user in producing datasets for getting genome data and information. The bioinformatics teams can use these datasets for more downstream analysis should they choose to do so. This chapter serves not only as a user guide or user roadmap to GENCODE website, but also to exchange ideas with bioinformatics community on future practical guides to life science databases.

Acknowledgement **Funding:** None

References

Alam T, Al-Absi HRH, Schmeier S (2020) Deep learning in LncRNAome: contribution, challenges, and perspectives. Noncoding RNA 6(4):47. https://doi.org/10.3390/ncrna6040047

Bignell A et al (2009) GENCODE: creating a validated manually annotated geneset for the whole human genome. Nat Preced:1756-0357

Cunningham F et al (2015) Ensembl 2015. Nucleic Acids Res 43(Database issue):D662–D669. https://doi.org/10.1093/nar/gku1010

Derrien T et al (2012) The GENCODE v7 catalog of human long noncoding RNAs: analysis of their gene structure, evolution, and expression. Genome Res 22(9):1775–1789. https://doi.org/10.1101/gr.132159.111

Emadi-Baygi M, Sedighi R, Nourbakhsh N, Nikpour P (2017) Pseudogenes in gastric cancer pathogenesis: a review article. Brief Funct Genomics 16(6):348–360. https://doi.org/10.1093/bfgp/elx004

ENCODE Project Consortium (2004) The ENCODE (ENCyclopedia Of DNA Elements) Project. Science 306(5696):636–640. https://doi.org/10.1126/science.1105136

Finn RD et al (2014) Pfam: the protein families database. Nucleic Acids Res 42(Database issue): D222–D230. https://doi.org/10.1093/nar/gkt1223

Flicek P et al (2012) Ensembl 2012. Nucleic Acids Res 40(Database issue):D84–D90. https://doi.org/10.1093/nar/gkr991

Frankish A, Harrow J (2014) GENCODE pseudogenes. Methods Mol Biol 1167:129–155. https://doi.org/10.1007/978-1-4939-0835-6_10

Frankish A et al (2019) GENCODE reference annotation for the human and mouse genomes. Nucleic Acids Res 47(D1):D766–D773. https://doi.org/10.1093/nar/gky955

Frankish A et al (2021) GENCODE 2021. Nucleic Acids Res 49(D1):D916–D923. https://doi.org/10.1093/nar/gkaa1087

Gordon DE et al (2020) A SARS-CoV-2 protein interaction map reveals targets for drug repurposing. Nature 583(7816):459–468. https://doi.org/10.1038/s41586-020-2286-9

GTEx Consortium (2013) The Genotype-Tissue Expression (GTEx) project. Nat Genet 45(6):580–585. https://doi.org/10.1038/ng.2653

Guigó R et al (2006) EGASP: the human ENCODE Genome Annotation Assessment Project. Genome Biol 7(Suppl 1):S2.1–S231. https://doi.org/10.1186/gb-2006-7-s1-s2

Harrow J et al (2006) GENCODE: producing a reference annotation for ENCODE. Genome Biol 7(1):S4.1–S4.9. https://doi.org/10.1186/gb-2006-7-s1-s4

Harrow J et al (2012) GENCODE: the reference human genome annotation for The ENCODE Project. Genome Res 22(9):1760–1774. https://doi.org/10.1101/gr.135350.111

Harrow JL et al (2014) The vertebrate genome annotation browser 10 years on. Nucleic Acids Res 42(Database issue):D771–D779. https://doi.org/10.1093/nar/gkt1241

Hon CC et al (2017) An atlas of human long non-coding RNAs with accurate 5′ ends. Nature 543(7644):199–204. https://doi.org/10.1038/nature21374

Howald C et al (2012) Combining RT-PCR-seq and RNA-seq to catalog all genic elements encoded in the human genome. Genome Res 22(9):1698–1710. https://doi.org/10.1101/gr.134478.111

Kent WJ, Baertsch R, Hinrichs A, Miller W, Haussler D (2003) Evolution's cauldron: duplication, deletion, and rearrangement in the mouse and human genomes. Proc Natl Acad Sci U S A 100(20):11484–11489. https://doi.org/10.1073/pnas.1932072100

Kodzius R et al (2006) CAGE: cap analysis of gene expression. Nat Methods 3(3):211–222. https://doi.org/10.1038/nmeth0306-211

Kokocinski F, Harrow J, Hubbard T (2010) AnnoTrack—a tracking system for genome annotation. BMC Genomics 11:538. https://doi.org/10.1186/1471-2164-11-538

Kozomara A, Griffiths-Jones S (2010) miRBase: integrating microRNA annotation and deep-sequencing data. Nucleic Acids Res 39(Suppl 1):D152–D157

Lagarde J et al (2017) High-throughput annotation of full-length long noncoding RNAs with capture long-read sequencing. Nat Genet 49(12):1731–1740. https://doi.org/10.1038/ng.3988

Lek M et al (2016) Analysis of protein-coding genetic variation in 60,706 humans. Nature 536(7616):285–291. https://doi.org/10.1038/nature19057

Lin MF, Jungreis I, Kellis M (2011) PhyloCSF: a comparative genomics method to distinguish protein coding and non-coding regions. Bioinformatics 27(13):i275–i282. https://doi.org/10.1093/bioinformatics/btr209

Mudge JM, Harrow J (2015) Creating reference gene annotation for the mouse C57BL6/J genome assembly. Mamm Genome 26(9–10):366–378. https://doi.org/10.1007/s00335-015-9583-x

Pei B et al (2012) The GENCODE pseudogene resource. Genome Biol 13(9):R51. https://doi.org/10.1186/gb-2012-13-9-r51

Rangan R et al (2020) RNA genome conservation and secondary structure in SARS-CoV-2 and SARS-related viruses: a first look. RNA 26(8):937–959. https://doi.org/10.1261/rna.076141.120

Regev A et al (2017) The human cell atlas. Elife 6:e27041. https://doi.org/10.7554/eLife.27041

Rodriguez JM et al (2013) APPRIS: annotation of principal and alternative splice isoforms. Nucleic Acids Res 41(Database issue):D110–D117. https://doi.org/10.1093/nar/gks1058

Searle SM, Gilbert J, Iyer V, Clamp M (2004) The otter annotation system. Genome Res 14(5):963–970. https://doi.org/10.1101/gr.1864804

Sherry ST et al (2001) dbSNP: the NCBI database of genetic variation. Nucleic Acids Res 29(1):308–311. https://doi.org/10.1093/nar/29.1.308

Siva N (2008) 1000 Genomes project. Nat Biotechnol 26(3):256

Sonnhammer EL, Wootton JC (2001) Integrated graphical analysis of protein sequence features predicted from sequence composition. Proteins 45(3):262–273. https://doi.org/10.1002/prot.1146

Stunnenberg HG, Hirst M (2016) The International Human Epigenome Consortium: a blueprint for scientific collaboration and discovery. Cell 167(5):1145–1149. https://doi.org/10.1016/j.cell.2016.11.007

UniProt Consortium (2012) Reorganizing the protein space at the Universal Protein Resource (UniProt). Nucleic Acids Res 40(Database issue):D71–D75. https://doi.org/10.1093/nar/gkr981

Zhang Z, Carriero N, Zheng D, Karro J, Harrison PM, Gerstein M (2006) PseudoPipe: an automated pseudogene identification pipeline. Bioinformatics 22(12):1437–1439. https://doi.org/10.1093/bioinformatics/btl116

Zheng D et al (2007) Pseudogenes in the ENCODE regions: consensus annotation, analysis of transcription, and evolution. Genome Res 17(6):839–851. https://doi.org/10.1101/gr.5586307

Zhou Y, Hou Y, Shen J, Huang Y, Martin W, Cheng F (2020) Network-based drug repurposing for novel coronavirus 2019-nCoV/SARS-CoV-2. Cell Discov 6:14. https://doi.org/10.1038/s41421-020-0153-3

Chapter 2
The GeneCards Suite

Marilyn Safran, Naomi Rosen, Michal Twik, Ruth BarShir, Tsippi Iny Stein, Dvir Dahary, Simon Fishilevich, and Doron Lancet

Abstract The GeneCards® database of human genes was launched in 1997 and has expanded since then to encompass gene-centric, disease-centric, and pathway-centric entities and relationships within the GeneCards Suite, effectively navigating the universe of human biological data—genes, proteins, cells, regulatory elements, biological pathways, and diseases—and the connections among them. The knowledgebase amalgamates information from >150 selected sources related to genes, proteins, ncRNAs, regulatory elements, chemical compounds, drugs, splice variants, SNPs, signaling molecules, differentiation protocols, biological pathways, stem cells, genetic tests, clinical trials, diseases, publications, and more and empowers the suite's Next Generation Sequencing (NGS), gene set, shared descriptors, and batch query analysis tools.

Keywords GeneCards · Bioinformatics · Biological database · Diseases · Gene prioritization · Integrated information retrieval · Next generation

2.1 Introduction

The GeneCards® database of human genes was launched in 1997 (Rebhan et al. 1997) and has expanded since then to encompass gene-centric, disease-centric, and pathway-centric entities and relationships within the GeneCards Suite, effectively navigating the universe of human biological data—genes, proteins, cells, regulatory elements, biological pathways, and diseases—and the connections among them. The suite's integrated biomedical knowledgebase includes GeneCards (Stelzer et al. 2016a), the integrated human gene database, MalaCards (Rappaport et al. 2017a),

M. Safran · N. Rosen · M. Twik · R. BarShir · T. I. Stein · S. Fishilevich · D. Lancet (✉)
Department of Molecular Genetics, The Weizmann Institute of Science, Rehovot, Israel
e-mail: Doron.Lancet@weizmann.ac.il

D. Dahary
LifeMap Sciences Inc., Marshfield, MA, USA

© The Author(s) 2021

I. Abugessaisa, T. Kasukawa (eds.), *Practical Guide to Life Science Databases*,
https://doi.org/10.1007/978-981-16-5812-9_2

the unified human disease database, PathCards (Belinky et al. 2015), the consolidated human pathways database, LifeMap Discovery (Edgar et al. 2013), the embryonic development and stem cell compendium, GeneLoc (Rosen et al. 2003), the human genomic neighborhood location-based database, and GeneHancer (Fishilevich et al. 2017), an innovative and growing regulatory element database with ~250,000 enhancer and promoter entries. The knowledgebase amalgamates information from >150 selected sources related to genes, proteins, ncRNAs, regulatory elements, chemical compounds, drugs, splice variants, SNPs, signaling molecules, differentiation protocols, biological pathways, stem cells, genetic tests, clinical trials, diseases, publications, and more, and empowers Next Generation Sequencing (NGS) analysis by highlighting associations between genes and phenotypes, providing supporting evidence for immediate evaluation via the suite's NGS analysis tools: VarElect (Stelzer et al. 2016b), the phenotype interpreter, receives a list of genes and phenotypes as input and computes prioritized direct (keyword-based) and indirect (inferred from gene-to-gene associations) gene/disease connections; TGex, the VCF-to-report clinical analyzer, incorporates VarElect's algorithms and automatically generates clinical case reports. Rounding out the suite are GeneAnalytics (Ben-Ari Fuchs et al. 2016), for gene set analysis, GenesLikeMe (Stelzer et al. 2009) for finding genes with shared descriptors, and GeneALaCart (Stelzer et al. 2016a) for batch queries.

The suite's websites, data dumps, APIs, publications, and collaborations are enjoyed by >3.5 million users, including research and applied scientists, doctors, geneticists, and lay-people, in >3000 institutions worldwide, encompassing academia, national patent offices, leading biopharma and diagnostic companies, and hospitals.

2.2 Database Overview

2.2.1 Importance and Current Status

Historically, users have characterized GeneCards as being their user-friendly "first port of call" to "orient their understanding" when coming across unfamiliar genes. Its popularity encouraged the expansion of the knowledgebase to provide the same functionality for diseases and pathways. Together with this growth came the realization that the depth and breadth of the data itself, while extremely useful in its own right, could be leveraged to solve problems. Today, there is increasing recognition by the scientific community that NGS is a pivotal technology for diagnosing the genetic cause of many human diseases; several large-scale projects implement NGS as a key instrument for elucidating the genetic components of rare diseases and cancer (Bamshad et al. 2012). Other clinical studies aimed at deciphering monogenic and complex diseases have also demonstrated the effectiveness of NGS approaches including whole genome, whole exome, and gene panel sequencing (van den Veyver and Eng 2015; Yang et al. 2013; Gilissen et al. 2014; Zheng et al. 2015; Stranneheim and Wedell 2016). Primary analysis of disease NGS results includes sequence read mapping and variant calling, with results stored in a Variant Call Format (VCF) file.

The VCF file typically contains ~20,000–50,000 positions that differ from the reference genome exome regions (“variant long list”). Subsequently, analysis pipelines sift these SNPs and indels by populating the VCF file with annotation data, such as segregation in affected families, genetic linkage information (Smith et al. 2011), population frequency (Ramos et al. 2012), and missense protein impact (Adzhubei et al. 2010; Sim et al. 2012; Hecht et al. 2015), all facilitating variant filtration (secondary analysis). This helps generate a “variant medium list” of typically dozens to a few hundred entries, depending on the assumed mode of inheritance and on the employed filtering cutoffs. In these analyses, variants are analyzed without regard to the disease phenotype of the sequenced individual. As a first step in introducing phenotype relationships, many pipelines use variant-disease relationships (e.g. from ClinVar (Landrum et al. 2014) and/or COSMIC (Forbes et al. 2015)) for further filtration of the sequence variants. But a typical gene can have a multitude of variants that have not yet been documented to have a relationship with a disease or a phenotype. In many cases, none of the annotated variant-disease relations appears relevant to the sequenced subject. The GeneCards suite’s rich knowledgebase facilitates gene-based interpretation. The strategy entails finding disease or phenotype relationships for the gene itself, instead of only for the variant contained within it. VarElect (ve.genecards.org), the suite’s web-based phenotype-dependent NGS variant prioritizer, leverages the wealth of information in GeneCards and its affiliated databases. VarElect’s algorithm computes prioritized direct (keyword-based) and indirect (inferred from comprehensive gene-to-gene associations) gene/disease connections. The avalanche of variants residing in genomic non-coding “dark matter,” available via whole genome sequencing (WGS), contributes three classes of functional genomic elements to variant analyses: promoters, enhancers, and ncRNAs, all central to tissue-related gene expression, with many underlying diseases. Together they amount to >20% of such “novel” DNA territories, unexplored in exome sequencing. Judiciously incorporated into the knowledgebase, the suite’s GeneHancer and upgraded ncRNA data is leveraged by its WGS disease interpretation platform and provides a comprehensive route to clinical significance of coding and non-coding single nucleotide and structural genomic variations, often elucidating unsolved clinical cases.

2.2.2 Future Update and Availability of the Database

Major synchronized new versions of the suite sites are currently deployed every four months. This weighty effort involves regenerating the gene and diseases lists, updating data from all of the knowledgebase’s sources, annotating each of the entities, re-computing the relationships, and quality assurance testing to ensure that all sites are in sync, that data integrity was maintained, and that nothing broke during the process due to changes in source formats and/or other pipeline technicalities. Further, new scientific features are provided by incorporating information from new and/or existing sources and developing/tweaking heuristics and algorithms when warranted. Minor revisions, providing incremental updates for a subset of the data and suite sites, are deployed as needed (typically within 1–2 months), for crucial time-dependent annotations like new publications, localized features, and hot bug fixes. We continue to work on increasing the frequency and content of our releases and expect significant speedup in 2019.

2.3 Content and Architecture of the Database

2.3.1 Main Database Features and Types of Data Stored

Figure 2.1 and Table 2.1 provide an overview of the major entities and relationships in GeneCards and MalaCards, in schematic and tabular forms, respectively. Some of the data include straightforward annotations (e.g. summary information about TP53 from NCBI's Entrez Gene database (Brown et al. 2015), the GeneCards Inferred Functionality Score (GIFtS) for APOA1, the KEGG pathway (Kanehisa et al. 2019) associated with Alzheimer's Disease, companies that provide antibody products for EGFR, publications associated with a gene or disease, and so on). Others reflect sophisticated behind-the-scenes data amalgamation: Compound groups, unified from 12 sources, with drug-specific and drug-gene annotations; GeneHancer (Fishilevich et al. 2017) regulatory element clusters, integrated from 7 sources based on location, with scored GeneHancer elements and GeneHancer-gene annotations; SuperPaths (Belinky et al. 2015), consolidated from 12 sources based on gene content, finding a balance between reducing pathway redundancies and optimizing pathway-related informativeness for individual genes; GeneCards genes (Safran et al. 2010), hierarchically choosing a symbol from HGNC (Yates et al. 2017), Entrez Gene (Brown et al. 2015), Ensembl (Zerbino et al. 2018), or GeneLoc (Rosen et al. 2003), and associating all relevant aliases, descriptions, and external identifiers; MalaCards diseases, canonicalizing, transforming, lexically manipulating, and unifying names from 10 primary and 5 secondary ranked sources (Rappaport et al. 2013).

Data collection methods: The GeneCards data collection process is a pipeline that starts with defining the full set of GeneCards genes, obtained from four primary sources as follows: First, the complete current snapshot of HGNC-approved symbols (Yates et al. 2017) is used as the core gene list. Second, human Entrez Gene (Brown et al. 2015) entries that are different from the HGNC genes are added. Next, human Ensembl (Zerbino et al. 2018) records are matched against the emerging gene list via GeneLoc's exon-based unification algorithm (Rosen et al. 2003); those that are not found to be equivalent to others in the set are included as novel Ensembl-based GeneCards gene entries. Finally, our RNA genes identification and unification facility ((Belinky et al. 2013) and work in progress) adds new ncRNAs not available in the other sources. These primary sources provide annotations for aliases, descriptions, previous symbols, gene category, location, summaries, paralogs, and ncRNA details. Once the gene list is in place with these significant annotations, over 150 data sources, including those noted above and others (Bateman et al. 2017; Gene Ontology Consortium 2015; Smith et al. 2018; Chalifa-Caspi et al. 2004) are mined for thousands of additional descriptors.

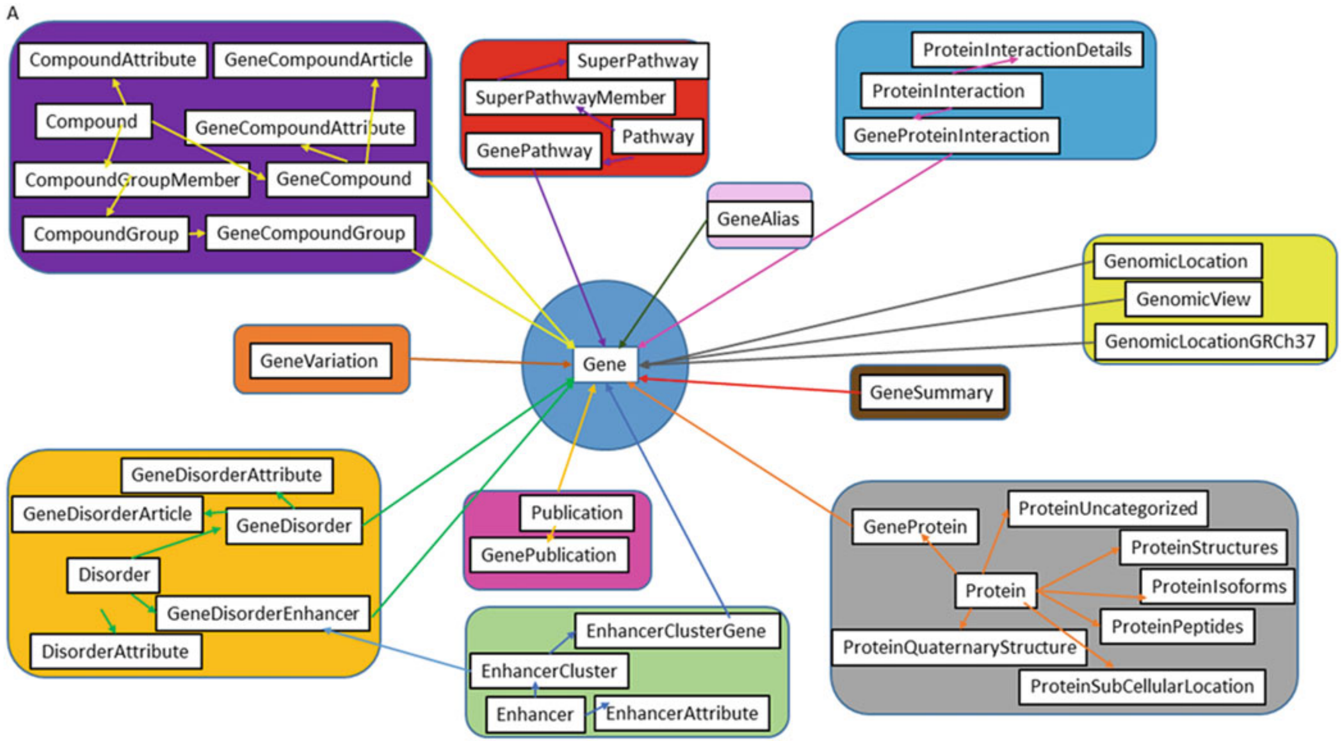


Fig. 2.1 Schematic representation of major GeneCards (**a**) and MalaCards (**b**) entities and relationships. Omitted GeneCards sections include domains, expression, function, localization, orthologs, paralogs, products, sources, and transcripts. Omitted MalaCards sections include summaries, genetic tests, anatomical context, expression, GO terms, and sources

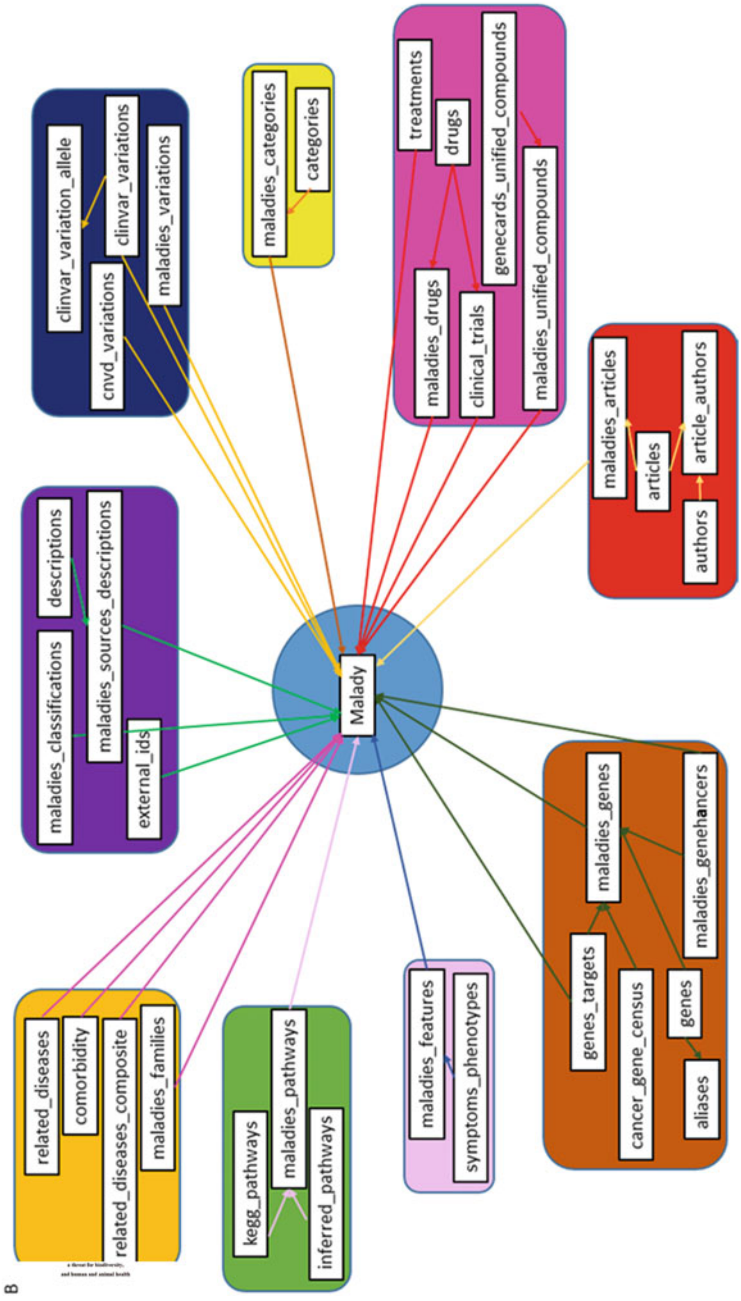


Fig. 2.1 (continued)

Table 2.1 GeneCards and MalaCards entity and relationship tables: (**a**) subset of major entities' tables and their fields; (**b**) types and quantities of tables

(a)

Entity	Table	Fields
Compounds	Compound	id
		sourceId
		sourceAccession
		Name
	CompoundAttribute	id
		compoundId
		type
		Value
	CompoundGroup	id
		Name
	CompoundGroupMember	id
		groupId
		compoundId
Superpathways	SuperPathway	id
		name
		sourceId
		sourceAccession
	SuperPathwayMember	id
		score
		superPathId
		pathwayId
	Pathway	id
		name
		sourceAccession
		sourceID
Enhancers	Enhancer	id
		clusterId
		sourceId
		Chromosome
		Start
		End
		Identifier
		FriendlyName
		HasLink
		Version
		Classification
Maladies	Maladies	id
		symbol
		acronymId
		descId
		slug

(continued)

Table 2.1 (continued)

(a)		
Entity	Table	Fields

(b)	
Category	**Number of tables**
Major entities	57
Relationships	67
Annotation tables	23

MalaCards builds its comprehensive-integrated list of diseases by hierarchically mining heterogeneous, partially overlapping naming sources (15 primary and 29 secondary), unifying disease names and acronyms, initially transforming each name to a canonical form while simultaneously retaining original strings for the alias list. This canonical form is constructed by a series of steps (conversion to lowercase; removal of words like “disease,” “syndrome,” “deficiency,” “failure,” “type,” as well as conjunctions, articles, and prepositions); merging equivalent words (e.g. “juvenile” and “childhood,” “kidney,” and “renal”); handling of different number formats (Roman versus Indian/Arabic), and of plurals and possessives; word stemming, using the porter stemming algorithm (Porter 2006), and others (Rappaport et al. 2013) to enable textual comparison. Diseases with names that are identical except for type specification (e.g. “Alzheimer disease type 3”) are grouped into parent/child families. Once the disease list is in place with these significant annotations, over 70 data sources, including those noted above, and others including GeneCards, MalaCards, and the suite’s gene set analysis capabilities (Ben-Ari Fuchs et al. 2016; Stelzer et al. 2009) are interrogated to yield thousands of additional descriptors and relationships.

2.3.2 *Data Collection and Curation Methods*

The knowledgebase, for the most part, is automatically generated. Our data sources range from those that are manually curated, (e.g. UniProt/SwissProtKB (Bateman et al. 2017)) to those that rely on text mining algorithms (e.g. DISEASES (Pletscher-Frankild et al. 2015)). Our generation software and portals rank the information in the various sections accordingly, giving greater weight to curated over inferred annotations. If the QA process (see below) and/or user feedback uncovers anomalies which cannot immediately be addressed by the relevant sources, we edit the data or use a “cheat list” of corrections to compensate.

2.3.3 *Dataset Indexing/Accession Number/Identification*

Alphabetical human gene and disease database indices appear at the footer of each respective GeneCards and MalaCards page, providing linked lists of symbols/

disease names. Clicking on a letter in the index, say D, brings up a page that lists all genes/diseases that start with "D," each linked to the relevant GeneCard or MalaCard.

GeneCards gene symbols, used in accessing the GeneCards pages for particular genes, are derived from HGNC (Yates et al. 2017), Entrez Gene (Brown et al. 2015), Ensembl (Zerbino et al. 2018), and GeneCards identifiers (GCIDs) (Rosen et al. 2003). GCIDs are unique, informative, and stable, provided by the GeneLoc Algorithm (see http://www.genecards.org/Images/Guide/GeneLocAlgorithm.jpg) as follows.

- The id begins with GC, which is followed by the chromosome number (where "00" indicates unknown chromosome and "MT" indicates the mitochondria), "P" or "M" for orientation (Plus or Minus strand), and approximate kilobase start coordinate.

  For example: OXA1L, with GC id **GC14P022766** is on chromosome **14** on the **plus** strand, starting at **22766** kilobases.
- Genes that are currently placed on a specific chromosome, but whose exact location on the chromosome is not yet known, receive a modified GC id, consisting of the chromosome and strand information, followed by a number, which indicates uncertain location, followed by a letter representing the specific contig containing the gene, and the gene's kilobase position on that contig.

  For example: ENSG00000278198, with GC id **GC07P9O0173** is on chromosome **7** on the **plus** strand of **contig GL000195.1**, starting at **173** kilobases.
- Genes located on the alternative reference sequences (haplotypes—see NCBI (https://www.ncbi.nlm.nih.gov/mapview/static/humansearch.html#assembly) for a full explanation) have a special GC id made up of the chromosome and strand information, followed by a letter, and the gene's approximate kilobase start coordinate.

  For example: KIR2DS5, with GC id **GC19MA00037** is on chromosome **19** on the **minus** strand of **ALT_REF_LOCI_18**, starting at **37** kilobases.
- Genes whose positional information includes only the chromosome need a further modified GC id, which includes the chromosome number, followed by "U9," indicating lack of strand and positional information, followed by five digits, assigned sequentially.

  For example: GUK2, with GC id **GC01U990078** is on chromosome **1**. Its **strand** and **position** are currently unknown.

  If an id needs to change in future versions because the previously reported position is refined, the superseded id remains associated with the gene, along with the new one, so it cannot be assigned to any other gene, and so that users can still find the gene by that id.

MalaCards identifiers, used in its URLs, are its *main disease names* supplied by primary sources (Rappaport et al. 2013) (e.g. Pick Disease) converted to lowercase, with spaces replaced by underscores (pick_disease for this example). To be as consistent as possible across versions, all such URLs are preserved, even if the disease name has changed or the disease was merged with another. In situations like these, old URLs are redirected to new ones. If a disease was removed completely from MalaCards, the old link is redirected to the search results page generated by querying the old disease name. In addition, a unique internal MCID is generated for each malady, composed of the first letter of its name, followed by the next two consonants, followed by a sequence number. For example, the MCID for "rett syndrome" is RTT001.

PathCards SuperPath identifiers, used in its URLs, are the names of the SuperPaths (e.g. glucose metabolism) converted to lowercase with spaces replaced by underscores (glucose_metabolism for this example).

2.3.4 *Quality Control Methods*

Before releasing a version of the knowledgebase, the system undergoes a semi-automated QA process. An in-house tool verifies the integrity of the GeneCards database by comparing it with that of the previous version, and it highlights inconsistencies and extreme results. The anomalies are then manually reviewed. Web cards and their links for a sample set of genes and diseases are manually checked by our QA professionals and a medical doctor consultant. As our heuristics are still evolving, problematic disease names (e.g. "Interferon" or "memory") are entered into a "cheat list" and removed from the system. VarElect and GeneHancer have their own set of automated QA scenarios, wherein deviations from expected results are reported and followed up by manual scrutiny. Test scenarios, bugs, and suggestions for improvements are all ticketed in our JIRA tracking system (https://www.atlassian.com/software/jira) and mapped to target releases.

2.3.5 *Database Update and Maintenance Strategy*

The knowledgebase is regenerated from scratch for each major version. For incremental updates, source-specific generation modules are rerun using the latest data. In both situations, the search index is regenerated for the benefit of the database portals themselves, as well as for usage by VarElect and TGex.

2.4 Database Access and Mining Methods

2.4.1 *Tools and Techniques to Access, Discover, and Mine the Content of the Database*

Gene-centric, disease-centric, location-centric, and pathway-centric information are, respectively, available and searchable from the GeneCards, MalaCards, GeneLoc, and PathCards portals, each with their own entity-specific web "card" and powerful search engine. GeneHancer data is incorporated in the knowledgebase, and in GeneCards, MalaCards, GeneLoc, VarElect, and TGex. The extensive knowledgebase (Ben-Ari Fuchs et al. 2016) is exploited to provide NGS interpretation and gene set analysis solutions as follows:

2.4.1.1 VarElect: The NGS Phenotyper of the GeneCards Suite

A key challenge in the interpretation of NGS in genetic disease studies is to effectively associate the identified variant-containing genes with a patient's disease phenotypes. This is addressed by VarElect (Stelzer et al. 2016b), the GeneCards Suite powered NGS interpretation tool, leveraging the broad knowledgebase for gene prioritization. VarElect is a comprehensive search tool that helps to effectively and rapidly identify and prioritize direct and indirect associations between genes and user-supplied disease terms, joined with providing extensive evidence for such associations.

Typical NGS analyses of a patient discover tens of thousands non-reference coding single nucleotide variants (SNVs), but only one or very few are expected to be significant for the relevant disease. In a filtering stage, various approaches, such as family segregation, frequency in the population, predicted protein impact, and evolutionary conservation are combined to shorten the variant list. A major challenge is the interpretation of the remaining (typically) few hundred genes, aiming to further focus on the most viable disease-causing candidate genes.

To cope with genes that have no direct association to the phenotype terms on their own, VarElect infers indirect (or "guilt by association") relationships between genes and phenotype keywords exploiting the GeneCards Suite diverse gene-to-gene relationships. Gene-to-gene relationships are generated using the GeneCards search engine, by searching gene symbols in selected GeneCards sections. The integrated pathway information from PathCards is a major contribution to the gene-to-gene relationships.

2.4.1.2 TGex: The Knowledge-Driven Clinical Genetics Analysis Platform of the GeneCards Suite

Clinical genetics analysis of thousands of variants requires a user interface that will enable browsing, viewing, filtering, and interpretation interactively. To this aim, TGex, the GeneCards Suite Knowledge-Driven Clinical Genetics Analysis platform, combines VarElect strength with comprehensive variant annotation and filtering capabilities in a consolidated view, which enables the genetic analyst to quickly pinpoint the strongest candidates. The comprehensive reporting system of TGex leverages the capabilities of VarElect and the vast amount of structured data available in the GeneCards Suite to automatically generate a full clinical report. TGex supports comprehensive data scrutiny, from raw patient genetic data (a VCF file), through intermediate annotations and interpretations, to detailed final reports.

2.4.1.3 Analysis of Genomic Structural Variants (SVs) Enabled by GeneHancer

A major source of pathogenic genomic alterations are structural variants (SVs), comprising both balanced modifications (inversions and translocations) and unbalanced variations—copy number variants (CNVs), including deletions, duplications, and insertions (Hurles et al. 2008; Weischenfeldt et al. 2013). Evaluation of the impact of SVs with respect to phenotype or disease relies on the genomic functional units associated with the SVs. Disease-related functional consequences of SVs involve changes in gene expression, which might occur when the SV encompasses the gene territory, either completely or partially. In this vein, the GeneCards Suite tools are useful for SVs interpretation, by helping to identify and prioritize SVs using the potential disease-causing genes damaged in each SV.

Often SVs do not overlap the coding regions of the disease-associated gene. SVs might influence genes over large distances by altering non-coding functional components such as regulatory elements and non-coding RNA genes. Tackling variations in non-coding regulatory elements to decipher the genetic underpinnings of human diseases is a great challenge in the analysis of both SNVs and SVs. Addressing this challenge necessitates the ability to map variants to regulatory elements such as promoters and enhancers. The mapping program requires access to a comprehensive database of regulatory elements. Since the biomedical knowledge directly linking regulatory elements to a disease/phenotype is obscure, the variant mapping step needs to be complemented by annotative information regarding a relationship between such an element and its target gene, for which a phenotype relationship is already known.

These capabilities are the core of GeneHancer, the GeneCards Suite database of regulatory elements and their gene targets. GeneHancer's comprehensive-integrated and scored set of regulatory elements and their gene-associations enables translating the finding of a WGS variant in a non-coding region into a variant-to-gene annotation, along with a confidence indication. Thus, integrating GeneHancer into the WGS annotation and filtering functions of VarElect and TGex assists in the mapping of non-coding variants to regulatory elements and via the gene targets forms a basis for variant-phenotype interpretation of whole genome sequences in health and disease.

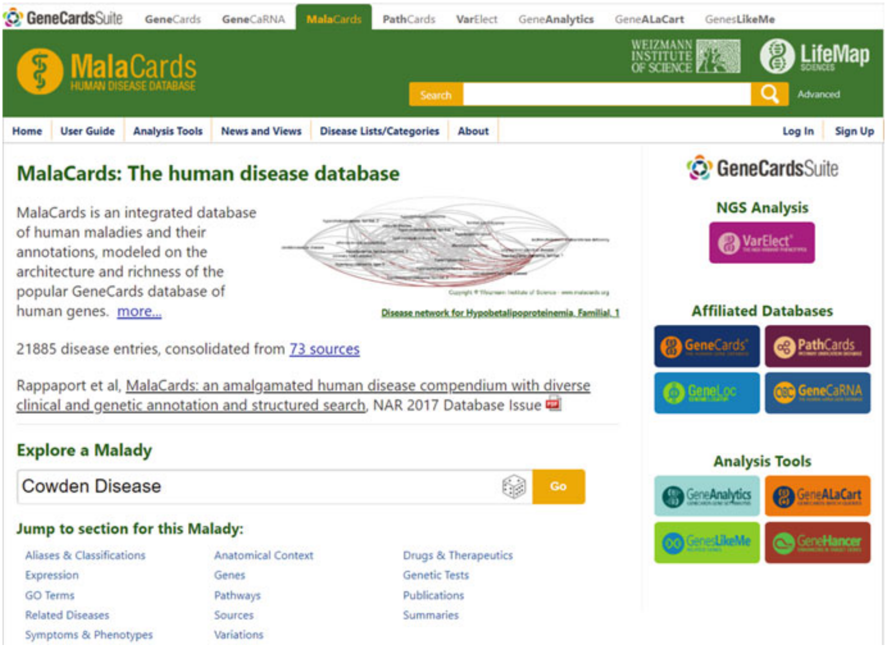
2.4.1.4 Gene Set Enrichment Analysis

GeneAnalytics (Ben-Ari Fuchs et al. 2016) is an analysis tool for finding commonalities within gene sets resulting from NGS, RNAseq, and microarray experiments. Using in-depth evidence-based scoring algorithms and taking advantage of the

GeneCards Suite knowledgebase, GeneAnalytics identifies cell types, diseases, pathways, and functions related to the gene set and provides supporting evidence links for matched biological terms in the GeneCards Suite.

2.4.2 *How to Explore and Browse the Database*

We illustrate exploring and browsing of the various suite sites by describing the MalaCards (Rappaport et al. 2014) compendium of human diseases portal (www.malacards.org), which features ~22,000 human diseases, with annotations integrated from 73 sources and shown in 14 sections. The homepage (Fig. 2.2a) is a common entry point to the Web site, showcasing most of the features and tools including exploring a particular (sample, random, or specified) malady, jumping to a particular section within it, quick searches, a disease index, statistics, a menu bar with links to documentation and disease list/category pages, and links to the other GeneCards



MalaCards: The human disease database

MalaCards is an integrated database of human maladies and their annotations, modeled on the architecture and richness of the popular GeneCards database of human genes. more...

21885 disease entries, consolidated from 73 sources

Rappaport et al, MalaCards: an amalgamated human disease compendium with diverse clinical and genetic annotation and structured search, NAR 2017 Database Issue

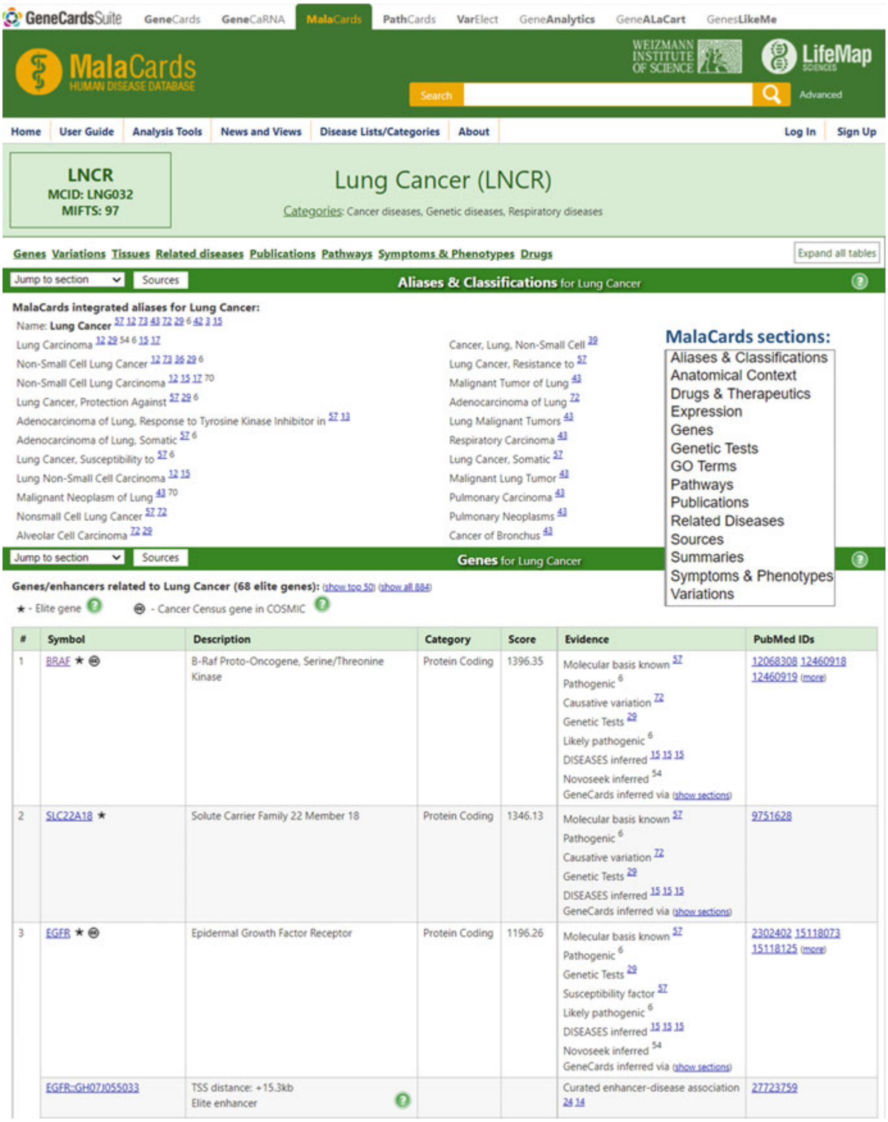
Explore a Malady

Cowden Disease

Jump to section for this Malady:

Aliases & Classifications
Expression
GO Terms
Related Diseases
Symptoms & Phenotypes
Anatomical Context
Genes
Pathways
Sources
Variations
Drugs & Therapeutics
Genetic Tests
Publications
Summaries

Fig. 2.2 (**a**) The homepage of MalaCards, the human disease database. (**b**) The MalaCard for Lung Cancer includes the Genes section, which provides the list of the affiliated genes and enhancers found to be associated with the disease. MalaCards "elite" genes (marked with *) are those likely to be associated with causing the disease, since their gene-disease associations are supported by manually curated and trustworthy sources. The cancer COSMIC Gene Census list is an ongoing effort to catalog those genes for which mutations have been causally implicated in cancer. Cancer census gene list genes are marked with a CC icon



LNCR
MCID: LNG032
MIFTS: 97

Lung Cancer (LNCR)

Categories: Cancer diseases, Genetic diseases, Respiratory diseases

Aliases & Classifications for Lung Cancer

MalaCards integrated aliases for Lung Cancer:

Name: **Lung Cancer** [57, 12, 73, 43, 72, 29, 6, 42, 3, 15]

Lung Carcinoma [12, 29, 54, 6, 15, 17]	Cancer, Lung, Non-Small Cell [39]
Non-Small Cell Lung Cancer [12, 73, 36, 29, 6]	Lung Cancer, Resistance to [57]
Non-Small Cell Lung Carcinoma [12, 15, 17, 70]	Malignant Tumor of Lung [43]
Lung Cancer, Protection Against [57, 29, 6]	Adenocarcinoma of Lung [72]
Adenocarcinoma of Lung, Response to Tyrosine Kinase Inhibitor in [57, 13]	Lung Malignant Tumors [43]
Adenocarcinoma of Lung, Somatic [57, 6]	Respiratory Carcinoma [43]
Lung Cancer, Susceptibility to [57, 6]	Lung Cancer, Somatic [57]
Lung Non-Small Cell Carcinoma [12, 15]	Malignant Lung Tumor [43]
Malignant Neoplasm of Lung [43, 70]	Pulmonary Carcinoma [43]
Nonsmall Cell Lung Cancer [57, 72]	Pulmonary Neoplasms [43]
Alveolar Cell Carcinoma [72, 29]	Cancer of Bronchus [43]

MalaCards sections:

Aliases & Classifications
Anatomical Context
Drugs & Therapeutics
Expression
Genes
Genetic Tests
GO Terms
Pathways
Publications
Related Diseases
Sources
Summaries
Symptoms & Phenotypes
Variations

Genes for Lung Cancer

Genes/enhancers related to Lung Cancer (68 elite genes): (show top 50) (show all 884)

★ - Elite gene   ⊛ - Cancer Census gene in COSMIC

#	Symbol	Description	Category	Score	Evidence	PubMed IDs
1	BRAF ★ ⊛	B-Raf Proto-Oncogene, Serine/Threonine Kinase	Protein Coding	1396.35	Molecular basis known [57] Pathogenic [6] Causative variation [72] Genetic Tests [29] Likely pathogenic [6] DISEASES inferred [15, 15, 15] Novoseek inferred [54] GeneCards inferred via (show sections)	12068308 12460918 12460919 (more)
2	SLC22A18 ★	Solute Carrier Family 22 Member 18	Protein Coding	1346.13	Molecular basis known [57] Pathogenic [6] Causative variation [72] Genetic Tests [29] DISEASES inferred [15, 15, 15] GeneCards inferred via (show sections)	9751628
3	EGFR ★ ⊛	Epidermal Growth Factor Receptor	Protein Coding	1196.26	Molecular basis known [57] Pathogenic [6] Genetic Tests [29] Susceptibility factor [57] Likely pathogenic [6] DISEASES inferred [15, 15, 15] Novoseek inferred [54] GeneCards inferred via (show sections)	2302402 15118073 15118125 (more)
	EGFR::GH07J055033	TSS distance: +15.3kb Elite enhancer			Curated enhancer-disease association [24, 14]	27723759

Fig. 2.2 (continued)

Suite members. MalaCards can be navigated in a variety of ways. The search box is typically the initial starting point, where one can submit free text as a query string, including Boolean expressions. It is centrally located on the homepage, as well as at the top right corner of every page comprising the Web site.

A MalaCards disease page (Web "card" or simply MalaCard) is where one can find all available information pertaining to a disease of interest. The information within a MalaCard is divided into 14 sections: Aliases and Classifications, Summaries, Related Diseases, Symptoms and Phenotypes, Drugs and Therapeutics, Genetic Tests, Anatomical Context, Publications, Genes, Variations, Expression, Pathways, GO Terms, and Sources. Documentation is accessible via hyperlinks, often context-specific, from within many parts of the MalaCard, to the right of the section, by clicking on the question mark icon. Each section displays disease-specific information and contains deep links to supporting sources, often with superscripts when multiple sources contain details about the datum. Different sections contain ranking and scoring of the elements, including genes in the Genes section, diseases in the "Related diseases" section, and pathways in the Pathways section. Figure 2.2b shows portions of the MalaCard for Lung Cancer, including the Genes section, which provides the list of the affiliated genes and enhancers found to be associated with the disease. MalaCards "elite" genes (marked with *) are those likely to be associated with causing the disease, since their gene-disease associations are supported by manually curated and trustworthy sources. The cancer Gene Census list from COSMIC is an ongoing effort to catalogue those genes for which mutations have been causally implicated in cancer. Genes listed in the cancer census gene list are marked with a CC icon. When relevant, shown GeneHancers are genomic regulatory elements-gene-disease associations provided by GeneHancer. Initially, at least 10 affiliated genes are shown (all of the elite genes are always shown), with an option to see the complete list.

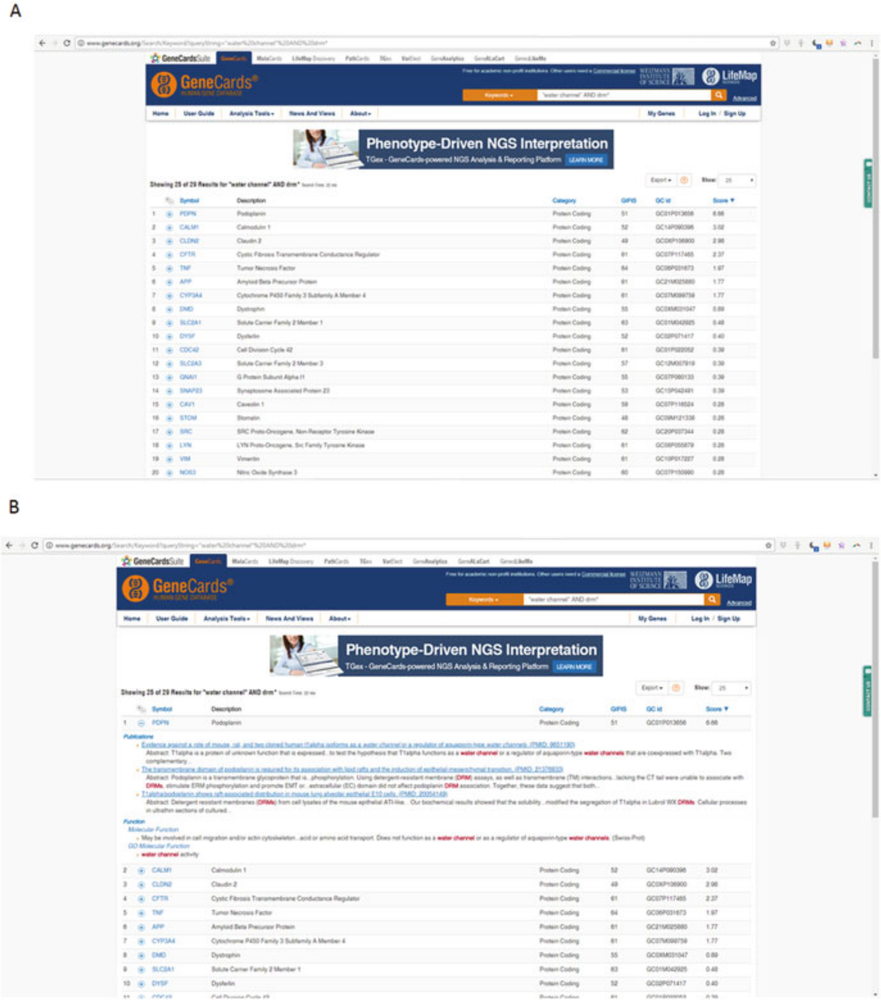
The ranked genes list is composed by taking into account: (1) genetic testing resources supplying specific genetic tests for the disease: (2) genetic variations resources supplying specific causative variations in genes for the disease; (3) resources that manually curate the association of the disease with genes; (4) searches within GeneCards, providing inferred associations.

The section's genes table shows gene symbols, descriptions, category, relevance scores, the context according to which the gene is related to the disease, and Pubmed ids. The relevance score is computed by factoring in the importance of the different resources associating the gene with the disease.

Long lists within the card sections are partially hidden by default (initially showing only the most relevant information for efficiency), with a "show all" option to display the complete list. Pressing "Expand all tables" activates "see all" in all of the sections and enables convenient searches within the card.

2.4.3 How to Query the Database

We illustrate the search capabilities of the various suite sites by describing GeneCards searches. In the top right corner of the GeneCards banner on each of its pages, enter your search terms into the search box and click the magnifying glass icon to submit the query. The query term may be a disease name, gene name, or any other keyword. Boolean operators (AND/OR) can be used to query GeneCards, as can wildcards (*) when placed at the end of a word. Note that Boolean operators



Showing 25 of 29 Results for "water channel" AND drm*

	Symbol	Description	Category	GIFtS	GC id	Score ▼
1	PDPN	Podoplanin	Protein Coding	51	GC01P013656	6.66
2	CALM1	Calmodulin 1	Protein Coding	52	GC14P090396	3.02
3	CLDN2	Claudin 2	Protein Coding	49	GC0XP106900	2.98
4	CFTR	Cystic Fibrosis Transmembrane Conductance Regulator	Protein Coding	61	GC07P117465	2.37
5	TNF	Tumor Necrosis Factor	Protein Coding	64	GC06P031673	1.97
6	APP	Amyloid Beta Precursor Protein	Protein Coding	61	GC21M025880	1.77
7	CYP3A4	Cytochrome P450 Family 3 Subfamily A Member 4	Protein Coding	61	GC07M099758	1.77
8	DMD	Dystrophin	Protein Coding	55	GC0XM031047	0.69
9	SLC2A1	Solute Carrier Family 2 Member 1	Protein Coding	63	GC01M042925	0.48
10	DYSF	Dysferlin	Protein Coding	52	GC02P071417	0.40
11	CDC42	Cell Division Cycle 42	Protein Coding	61	GC01P022052	0.39
12	SLC2A3	Solute Carrier Family 2 Member 3	Protein Coding	57	GC12M007919	0.39
13	GNAI1	G Protein Subunit Alpha I1	Protein Coding	55	GC07P080133	0.39
14	SNAP23	Synaptosome Associated Protein 23	Protein Coding	53	GC15P042491	0.39
15	CAV1	Caveolin 1	Protein Coding	58	GC07P116524	0.28
16	STOM	Stomatin	Protein Coding	48	GC09M121338	0.28
17	SRC	SRC Proto-Oncogene, Non-Receptor Tyrosine Kinase	Protein Coding	62	GC20P037344	0.28
18	LYN	LYN Proto-Oncogene, Src Family Tyrosine Kinase	Protein Coding	61	GC08P055879	0.28
19	VIM	Vimentin	Protein Coding	61	GC10P017227	0.28
20	NOS3	Nitric Oxide Synthase 3	Protein Coding	60	GC07P150990	0.28

Showing 25 of 29 Results for "water channel" AND drm*

	Symbol	Description	Category	GIFtS	GC id	Score ▼
1	PDPN	Podoplanin	Protein Coding	51	GC01P013656	6.66

Publications

- Evidence against a role of mouse, rat, and two cloned human T1alpha isoforms as a water channel or a regulator of aquaporin-type water channels. (PMID: 9651190)
  Abstract: T1alpha is a protein of unknown function that is expressed... to test the hypothesis that T1alpha functions as a water channel or a regulator of aquaporin-type water channels that are coexpressed with T1alpha. Two complementary...
- The transmembrane domain of podoplanin is required for its association with lipid rafts and the induction of epithelial-mesenchymal transition. (PMID: 21376833)
  Abstract: Podoplanin is a transmembrane glycoprotein that is... phosphorylation. Using detergent-resistant membrane (DRM) assays, as well as transmembrane (TM) interactions... lacking the CT tail were unable to associate with DRMs, stimulate ERM phosphorylation and promote EMT or... extracellular (EC) domain did not affect podoplanin DRM association. Together, these data suggest that both...
- T1alpha/podoplanin shows raft-associated distribution in mouse lung alveolar epithelial E10 cells. (PMID: 20054149)
  Abstract: Detergent resistant membranes (DRMs) from cell lysates of the mouse epithelial ATI-like... Our biochemical results showed that the solubility ...modified the segregation of T1alpha in Lubrol WX DRMs. Cellular processes in ultrathin sections of cultured...

Function

Molecular Function

- May be involved in cell migration and/or actin cytoskeleton...acid or amino acid transport. Does not function as a water channel or as a regulator of aquaporin-type water channels. (Swiss-Prot)

GO Molecular Function

- water channel activity

	Symbol	Description	Category	GIFtS	GC id	Score ▼
2	CALM1	Calmodulin 1	Protein Coding	52	GC14P090396	3.02
3	CLDN2	Claudin 2	Protein Coding	49	GC0XP106900	2.98
4	CFTR	Cystic Fibrosis Transmembrane Conductance Regulator	Protein Coding	61	GC07P117465	2.37
5	TNF	Tumor Necrosis Factor	Protein Coding	64	GC06P031673	1.97
6	APP	Amyloid Beta Precursor Protein	Protein Coding	61	GC21M025880	1.77
7	CYP3A4	Cytochrome P450 Family 3 Subfamily A Member 4	Protein Coding	61	GC07M099758	1.77
8	DMD	Dystrophin	Protein Coding	55	GC0XM031047	0.69
9	SLC2A1	Solute Carrier Family 2 Member 1	Protein Coding	63	GC01M042925	0.48
10	DYSF	Dysferlin	Protein Coding	52	GC02P071417	0.40

Fig. 2.3 MalaCards search results: (**a**) sorted, scored gene hits. (**b**) with Minicards including hit context

must be capitalized to yield expected results: For example, specifying "*water channel" AND drm** yields 29 results.

Searches result in a list of genes, each with its description, category, GeneCards Inferred Functionality Score (GIFtS) (Harel et al. 2009), and GeneCards identifier (Rosen et al. 2003), sorted by Elastic search relevance score (Fig. 2.3a). Clicking the plus to the left of the symbol opens a "MiniCard," which shows the hit context of the search terms (Fig. 2.3b). Clicking on the symbol opens the gene's card.

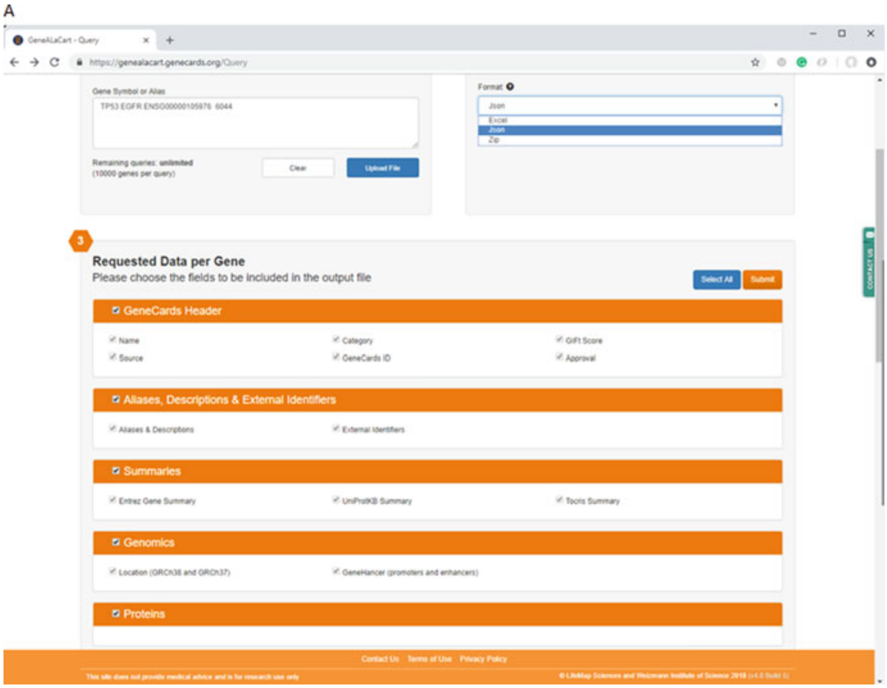
GeneCards can also be searched for a specific symbol, using the search dropdown (choose "Symbols"). When searching for a symbol that might not be the gene's official symbol (from a paper, for example), and when using a gene identifier from another database, the other dropdown options should be used ("Symbols/Aliases" and "Symbols/Aliases/Identifiers," respectively).

To use the GeneCards advanced search, click on the "Advanced" link to the right of the search box. The advanced search allows complex queries in which each keyword can be restricted to a specific section of the GeneCard.

MalaCards and PathCards have similar querying facilities.

2.4.4 *How to Upload/Download Data*

Registered users have a variety of download facilities. GeneALaCart (https://genealacart.genecards.org/), the GeneCards batch query portal generates a file of GeneCards annotations associated with input gene lists. For each query, one supplies the "batch" of gene symbols or identifiers and selects the annotations of interest (Fig. 2.4a); GeneALaCart then extracts the information from the knowledgebase and produces a customized results file in Excel [Fig. 2.4b] or JSON format [Fig. 2.4c].


A
GeneALaCart - Query
https://genealacart.genecards.org/Query
Clear
Upload File
Excel
Json
Zip
3
Select All
Submit
CONTACT US
Contact Us Terms of Use Privacy Policy

Gene Symbol or Alias: TP53 EGFR ENSG00000105976 6044

Remaining queries: unlimited
(10000 genes per query)

Format: Json

Requested Data per Gene
Please choose the fields to be included in the output file

☑ GeneCards Header
- ☑ Name
- ☑ Category
- ☑ GIFt Score
- ☑ Source
- ☑ GeneCards ID
- ☑ Approval

☑ Aliases, Descriptions & External Identifiers
- ☑ Aliases & Descriptions
- ☑ External Identifiers

☑ Summaries
- ☑ Entrez Gene Summary
- ☑ UniProtKB Summary
- ☑ Tocris Summary

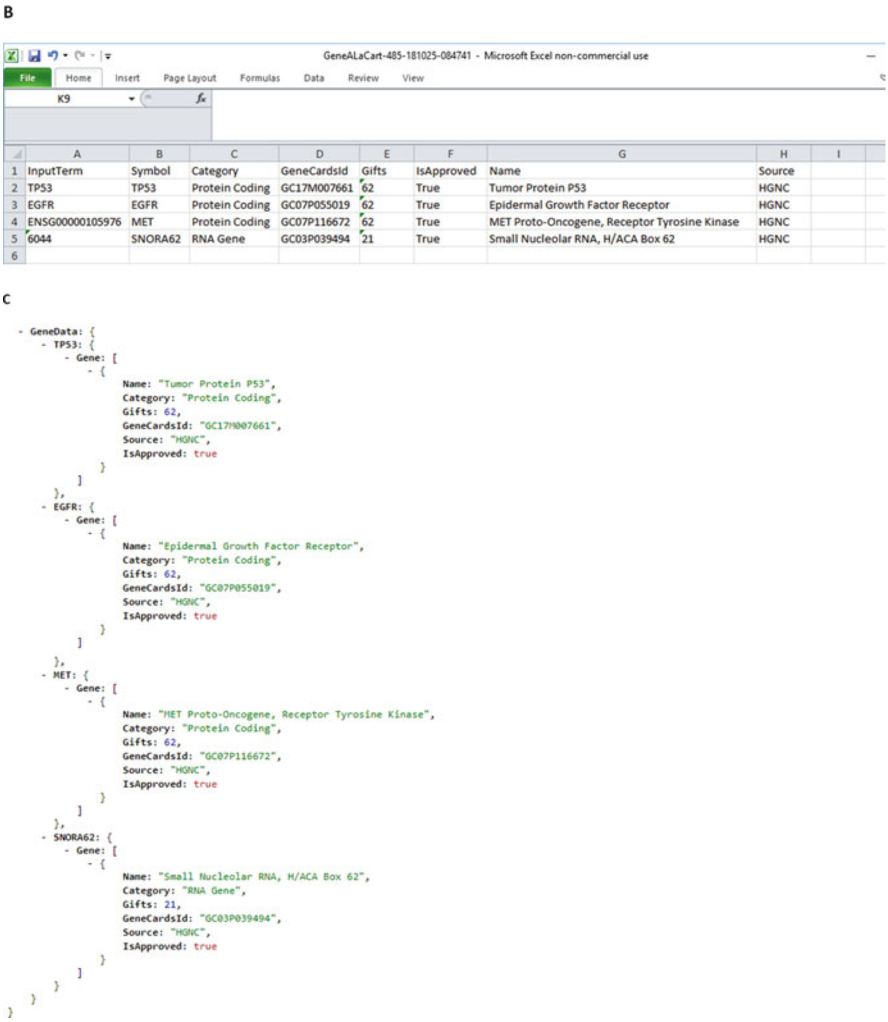
☑ Genomics
- ☑ Location (GRCh38 and GRCh37)
- ☑ GeneHancer (promoters and enhancers)

☑ Proteins

This site does not provide medical advice and is for research use only

© LifeMap Sciences and Weizmann Institute of Science 2018

Fig. 2.4 GeneAlaCart input and output: (**a**) user inputs genes/identifiers of choice, selected annotations, and output file format; (**b**) sample Excel sheet output; (**c**) sample JSON output


B
GeneALaCart-485-181025-084741 - Microsoft Excel non-commercial use
File
Home
Insert
Page Layout
Formulas
Data
Review
View
K9
fx
C

	A	B	C	D	E	F	G	H	I
1	InputTerm	Symbol	Category	GeneCardsId	Gifts	IsApproved	Name	Source	
2	TP53	TP53	Protein Coding	GC17M007661	62	True	Tumor Protein P53	HGNC	
3	EGFR	EGFR	Protein Coding	GC07P055019	62	True	Epidermal Growth Factor Receptor	HGNC	
4	ENSG00000105976	MET	Protein Coding	GC07P116672	62	True	MET Proto-Oncogene, Receptor Tyrosine Kinase	HGNC	
5	6044	SNORA62	RNA Gene	GC03P039494	21	True	Small Nucleolar RNA, H/ACA Box 62	HGNC	
6									

```
 - GeneData: {
     - TP53: {
         - Gene: [
             - {
                   Name: "Tumor Protein P53",
                   Category: "Protein Coding",
                   Gifts: 62,
                   GeneCardsId: "GC17M007661",
                   Source: "HGNC",
                   IsApproved: true
               }
           ]
       },
     - EGFR: {
         - Gene: [
             - {
                   Name: "Epidermal Growth Factor Receptor",
                   Category: "Protein Coding",
                   Gifts: 62,
                   GeneCardsId: "GC07P055019",
                   Source: "HGNC",
                   IsApproved: true
               }
           ]
       },
     - MET: {
         - Gene: [
             - {
                   Name: "MET Proto-Oncogene, Receptor Tyrosine Kinase",
                   Category: "Protein Coding",
                   Gifts: 62,
                   GeneCardsId: "GC07P116672",
                   Source: "HGNC",
                   IsApproved: true
               }
           ]
       },
     - SNORA62: {
         - Gene: [
             - {
                   Name: "Small Nucleolar RNA, H/ACA Box 62",
                   Category: "RNA Gene",
                   Gifts: 21,
                   GeneCardsId: "GC03P039494",
                   Source: "HGNC",
                   IsApproved: true
               }
           ]
       }
   }
}
```

Fig. 2.4 (continued)

Other download capabilities within the suite sites include exporting GeneCards search results, details about MalaCards diseases, GeneLikeMe functional partners with evidence, GeneHancer details, VarElect prioritized results, GeneAnalytics enriched gene sets, and TGex annotated reports. Facilities for database acquisition for the purposes of further analyses and integration include a variety of knowledgebase dumps and APIs. For more details, please contact the authors.

2.5 Use Cases

As noted above, discovery within the GeneCards Suite is exemplified by how VarElect and TGex leverage the extensive knowledgebase to provide NGS interpretation. The following use cases illustrate this.

2.5.1 Interpretation of Single Nucleotide Variants (SNVs)

VarElect is useful for variant interpretation in genetic disease studies by helping to identify and prioritize associations between variant-containing genes and phenotype keywords. VarElect helped us solve clinical cases in our own laboratory (Alkelai et al. 2016, 2017; Oz-Levi et al. 2015; Heimer et al. 2016, 2018) and was further used in numerous studies worldwide (Yang et al. 2017; Einhorn et al. 2017; Ekhilevitch et al. 2016; Jia et al. 2017; Bafunno et al. 2018; Zhang et al. 2016; Azim et al. 2019; Carneiro et al. 2018; Feliubadalo et al. 2017; Syama et al. 2018). VarElect exploits the GeneCards Suite diverse gene-to-gene relationships to pinpoint the relevance of genes that have no direct association to the phenotype keywords on their own (using the indirect, or "guilt by association" mode). The indirect approach proved crucial to solving a case of systemic capillary leak syndrome (Stelzer et al. 2016b). Figure 2.5a depicts an example of another VarElect case solved in our group (Rappaport et al. 2017b). In this example, the genome of a 6 year old boy, who suffered from atypical epilepsy combined with retinitis pigmentosa, was sequenced. Eighty-one rare homozygous variants, which were heterozygous in both parents, were identified in the patient. The list of 63 variant-containing genes was submitted to VarElect, along with the phenotype search terms; "epilepsy OR macular OR retinitis." VarElect's top scoring gene was *CLN6*. The patient had a homozygous missense variation (V148D) in this gene with zero population frequency and a high predicted protein damage impact. Following this discovery, the patient was clinically diagnosed with accuracy, enabling appropriate genetic counseling and preimplantation diagnosis for the family in the event of future pregnancies.

VarElect can be used stand-alone as described above, or within TGex, the GeneCards Suite Knowledge-Driven Clinical Genetics Analysis platform. TGex requires two inputs: (Rebhan et al. 1997) A VCF file; (Stelzer et al. 2016a) disease/phenotype/symptom terms for VarElect gene-phenotype interpretation. With TGex (Fig. 2.5b), thousands of variants within the uploaded patient VCF file are analyzed in an interactive web-based interface, allowing the user to browse, view, and filter input variants. Those capacities are combined with VarElect's gene-phenotype interpretation strength, allowing one to effectively identify disease-causing candidates. Top candidate variants, along with disease association evidence, are automatically pulled into the detailed clinical report.

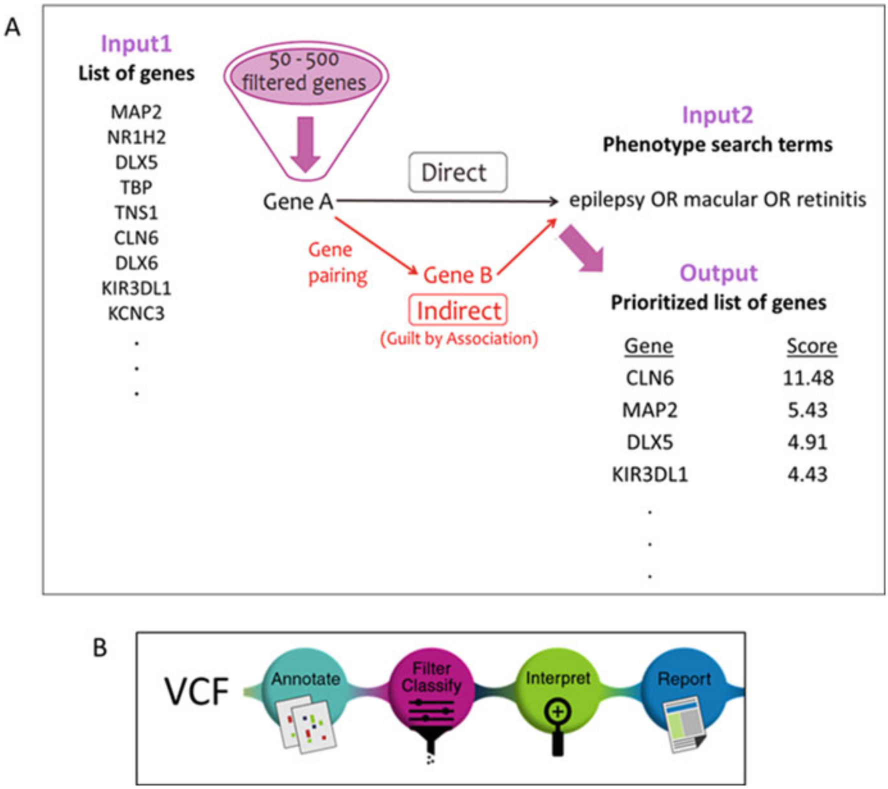
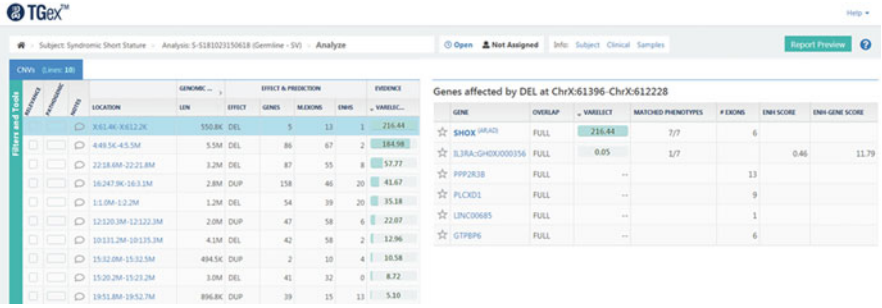


Fig. 2.5 The GeneCards Suite NGS analysis tools VarElect and TGex. (**a**) Example of a VarElect case solved in our group (Rappaport et al. 2017b); (**b**) NGS data analysis with TGex. TGex allows data scrutiny and analysis, starting from raw patient genetic data (a VCF file) to a detailed report. Variants are annotated using information from the GeneCards knowledgebase, allowing interactive filtering. These variant annotation and filtering steps are strengthened by gene-phenotype interpretation using VarElect. Hence, TGex allows the examination of variants using both variant-based annotations and variant-containing-genes-based interpretation, presenting this information for optimal candidate variant selection for the clinical report

2.5.2 Interpretation of Genomic Structural Variants (SVs)

VarElect is useful for structural variants interpretation by the identification and prioritization of SVs via the potential disease-causing genes damaged by each SV. In this workflow, the gene list submitted to VarElect includes genes residing (completely or partially) within the detected SVs. This mode of analysis using VarElect helped solve a number of cases (Homma et al. 2018; Fidalgo et al. 2016). One study aimed to diagnose recurrent CNVs associated with syndromic short stature of unknown cause (Homma et al. 2018). Two hundred and twenty-nine patients were genotyped by chromosomal microarray analysis, leading to identification of candidate CNVs. The gene content of those CNVs was submitted to VarElect


TGex™
Help ▾
Subject: Syndromic Short Stature
Analysis: S-S181023150618 (Germline - SV)
Analyze
Open
Not Assigned
Info: Subject Clinical Samples
Report Preview
?
CNVs (Lines: 10)
Filters and Tools

				GENOMIC ...	EFFECT & PREDICTION				EVIDENCE
RELEVANCE	PATHOGENIC	NOTES	LOCATION	LEN	EFFECT	GENES	M.EXONS	ENHS	VARELEC...
			X:61.4K-X:612.2K	550.8K	DEL	5	13	1	216.44
□	□		4:49.5K-4:5.5M	5.5M	DEL	86	67	2	184.98
□	□		22:18.6M-22:21.8M	3.2M	DEL	87	55	8	57.77
□	□		16:247.9K-16:3.1M	2.8M	DUP	158	46	20	41.67
□	□		1:1.0M-1:2.2M	1.2M	DEL	54	39	20	35.18
□	□		12:120.3M-12:122.3M	2.0M	DUP	47	58	6	22.07
□	□		10:131.2M-10:135.3M	4.1M	DEL	42	58	2	12.96
□	□		15:32.0M-15:32.5M	494.5K	DUP	2	10	4	10.58
□	□		15:20.2M-15:23.2M	3.0M	DEL	41	32	0	8.72
□	□		19:51.8M-19:52.7M	896.8K	DUP	39	15	13	5.10

Genes affected by DEL at ChrX:61396-ChrX:612228

	GENE	OVERLAP	VARELECT	MATCHED PHENOTYPES	# EXONS	ENH SCORE	ENH-GENE SCORE
☆	SHOX (AR,AD)	FULL	216.44	7/7	6		
☆	IL3RA:GH0XJ000356	FULL	0.05	1/7		0.46	11.79
☆	PPP2R3B	FULL	--		13		
☆	PLCXD1	FULL	--		9		
☆	LINC00685	FULL	--		1		
☆	GTPBP6	FULL	--		6		

Fig. 2.6 SV analysis with TGex

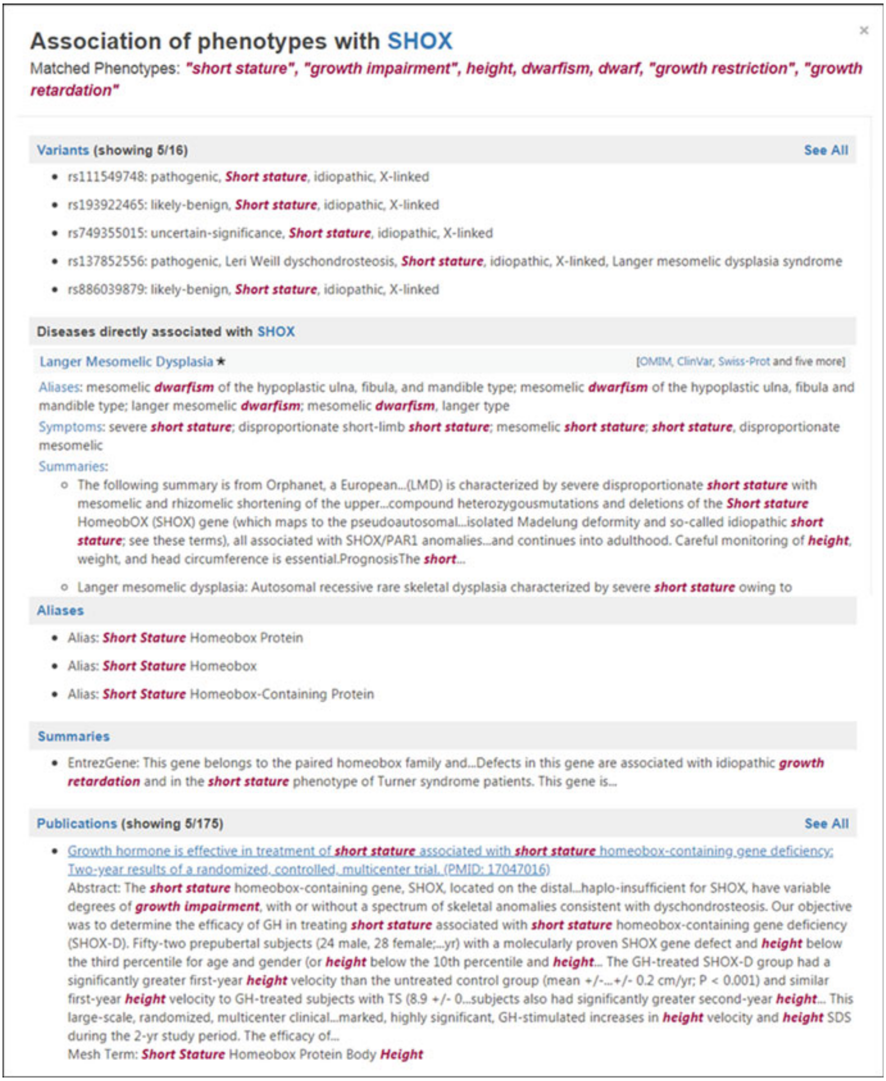
to find and prioritize phenotype related genes, leading to identification of pathogenic CNVs. We demonstrate this workflow using the TGex SVs module (Fig. 2.6).

The user inputs to TGex are: (Rebhan et al. 1997) a list of SVs; (Stelzer et al. 2016a) disease/phenotype/symptom terms. The analysis screen allows the user to browse and interpret the SVs. The list of entered SVs (Fig. 2.6, left pane) is presented along with annotations, such as the genomic location and length, SV type, number of genes in the region, and more. Those annotations are amplified with the VarElect score, which is also used as the default sort column for the SVs list. The value in this column is the highest VarElect phenotype score of the gene pool in each SV gene list. In this analysis the highest scoring SV is a 550kb deletion on chromosome X, overlapping 5 genes and one enhancer element.

The user can click on any of the SVs in the list (left pane) for the detailed view of each SV. In this view (Fig. 2.6, right pane) functional genomic elements in overlap with the SV region are shown (including not only protein coding genes, but also ncRNA genes, enhancers, and promoters), with annotations such as the overlap type (full/partial), the number of exons in overlap (for genes), and GeneHancer confidence scores for regulatory elements (see below). For the selected SV, the gene *SHOX* (Short Stature Homeobox) is the VarElect top scoring gene for the submitted keyword list (“short stature” OR “growth impairment” OR height OR dwarfism OR dwarf OR “growth restriction” OR “growth retardation”). Clicking on the VarElect score opens the “MiniCard,” which shows the hit context of the search terms within different sections of the *SHOX* gene in GeneCards, and diseases related to *SHOX* in MalaCards (Fig. 2.7).

2.5.3 *GeneHancer-Powered Interpretation of SVs*

GeneHancer, the GeneCards Suite database of regulatory elements and their gene targets, has been used by the community as an annotation standard for enhancers and promoters in the human genome, as well as for the associations of those elements with their gene targets (Quigley et al. 2018; Zhang et al. 2018; Holzinger et al. 2017;



Association of phenotypes with SHOX

Matched Phenotypes: ***"short stature", "growth impairment", height, dwarfism, dwarf, "growth restriction", "growth retardation"***

Variants **(showing 5/16)** See All

- rs111549748: pathogenic, ***Short stature***, idiopathic, X-linked
- rs193922465: likely-benign, ***Short stature***, idiopathic, X-linked
- rs749355015: uncertain-significance, ***Short stature***, idiopathic, X-linked
- rs137852556: pathogenic, Leri Weill dyschondrosteosis, ***Short stature***, idiopathic, X-linked, Langer mesomelic dysplasia syndrome
- rs886039879: likely-benign, ***Short stature***, idiopathic, X-linked

Diseases directly associated with SHOX

Langer Mesomelic Dysplasia ★ [OMIM, ClinVar, Swiss-Prot and five more]

Aliases: mesomelic ***dwarfism*** of the hypoplastic ulna, fibula, and mandible type; mesomelic ***dwarfism*** of the hypoplastic ulna, fibula and mandible type; langer mesomelic ***dwarfism***; mesomelic ***dwarfism***, langer type

Symptoms: severe ***short stature***; disproportionate short-limb ***short stature***; mesomelic ***short stature***; ***short stature***, disproportionate mesomelic

Summaries:

- The following summary is from Orphanet, a European...(LMD) is characterized by severe disproportionate ***short stature*** with mesomelic and rhizomelic shortening of the upper...compound heterozygousmutations and deletions of the ***Short stature*** HomeobOX (SHOX) gene (which maps to the pseudoautosomal...isolated Madelung deformity and so-called idiopathic ***short stature***; see these terms), all associated with SHOX/PAR1 anomalies...and continues into adulthood. Careful monitoring of ***height***, weight, and head circumference is essential.PrognosisThe ***short***...
- Langer mesomelic dysplasia: Autosomal recessive rare skeletal dysplasia characterized by severe ***short stature*** owing to

Aliases

- Alias: ***Short Stature*** Homeobox Protein
- Alias: ***Short Stature*** Homeobox
- Alias: ***Short Stature*** Homeobox-Containing Protein

Summaries

- EntrezGene: This gene belongs to the paired homeobox family and...Defects in this gene are associated with idiopathic ***growth retardation*** and in the ***short stature*** phenotype of Turner syndrome patients. This gene is...

Publications **(showing 5/175)** See All

- Growth hormone is effective in treatment of ***short stature*** associated with ***short stature*** homeobox-containing gene deficiency: Two-year results of a randomized, controlled, multicenter trial. (PMID: 17047016)
  Abstract: The ***short stature*** homeobox-containing gene, SHOX, located on the distal...haplo-insufficient for SHOX, have variable degrees of ***growth impairment***, with or without a spectrum of skeletal anomalies consistent with dyschondrosteosis. Our objective was to determine the efficacy of GH in treating ***short stature*** associated with ***short stature*** homeobox-containing gene deficiency (SHOX-D). Fifty-two prepubertal subjects (24 male, 28 female;...yr) with a molecularly proven SHOX gene defect and ***height*** below the third percentile for age and gender (or ***height*** below the 10th percentile and ***height***... The GH-treated SHOX-D group had a significantly greater first-year ***height*** velocity than the untreated control group (mean +/-...+/- 0.2 cm/yr; P < 0.001) and similar first-year ***height*** velocity to GH-treated subjects with TS (8.9 +/- 0...subjects also had significantly greater second-year ***height***... This large-scale, randomized, multicenter clinical...marked, highly significant, GH-stimulated increases in ***height*** velocity and ***height*** SDS during the 2-yr study period. The efficacy of...
  Mesh Term: ***Short Stature*** Homeobox Protein Body ***Height***

Fig. 2.7 MiniCards—evidence for gene-phenotype associations. This figure shows selected parts of the MiniCard for the gene SHOX and the phenotypes used in the short stature study. A list of matched phenotypes is shown in red in the top part. This is followed by several gene-centric evidence for queried phenotype association, e.g. from the GeneCards Variants, Aliases, Summaries, and Publication sections. This evidence is combined by MalaCards-based evidence, showing queried phenotype associations in diseases associated with the gene SHOX, from various MalaCards sections, e.g. Aliases, Symptoms, and Summaries. For all sections, only partial evidence list is shown here

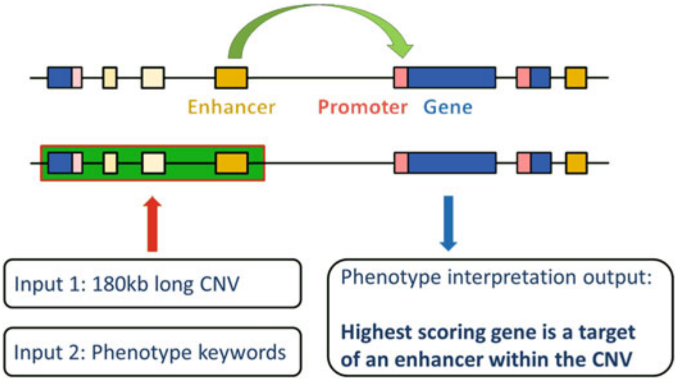

Enhancer
Promoter
Gene
Input 1: 180kb long CNV
Input 2: Phenotype keywords
Phenotype interpretation output:
Highest scoring gene is a target of an enhancer within the CNV

Fig. 2.8 Solving a genetic disease with GeneCards Suite NGS tools SV analysis capacities. GeneHancer enriches the GeneCards Suite NGS tools VarElect and TGex, providing the ability to map SVs to non-coding functional regulatory elements such as enhancers and promoters. This mapping, combined with GeneHancer's information on the association of those elements with target genes, enables pinpointing variant-phenotype relationships that otherwise might be undiscovered, increasing the potential solve rate of genetic disease studies

Singh et al. 2018; Huang et al. 2018; Yang et al. 2018; Bermejo et al. 2019; Erlangsen et al. 2020; Nikulin et al. 2018; Slater et al. 2018). With the growing understanding of the importance of non-coding variants for NGS interpretation, GeneHancer-enriched VarElect and TGex offer novel modes of analysis for tackling this challenge.

First, we augmented VarElect to be able to process GeneHancer element identifiers. For a given element, VarElect performs gene-phenotype prioritization for its GeneHancer gene targets. The phenotype prioritization in this workflow is performed by combining the VarElect gene-phenotype score with the GeneHancer element and gene-association confidence scores. This mode of analysis allows users to perform phenotype interpretation of mixed lists of genes and regulatory elements after both SNV and SV primary data analysis steps.

Second, we enhanced TGex to include regulatory elements in SV interpretation. User-submitted SVs are mapped to both genes and regulatory elements, followed by VarElect interpretation of the mixed list of genes and enhancers/promoters. This mode of analysis helped our lab solve a genetic disease study (Fig. 2.8). In this case, a family with a rare congenital autosomal dominant genetic skin disease was genotyped, leading to the identification of a CNV shared by all affected individuals. Phenotype interpretation of this CNV discovered that it overlaps an enhancer, whose gene target, albeit not residing within the CNV, is extremely relevant for the studied phenotype.

2.5.4 Other VarElect Use Cases

While interpretation of genetic disease NGS analyses was the focus of our described use cases, VarElect is also a potent tool for supporting the interpretation of other experimental results. In such scenarios, VarElect is utilized to analyze gene lists retrieved from various methodologies, helping to focus on more affordable candidate gene lists based on gene-phenotype information. Scenarios benefitting from the gene-phenotype prioritization capacities of VarElect include gene expression (RNAseq/Microarrays), protein expression (mass spectrometry), and other multi-OMICS downstream analyses (Hulst et al. 2017; Yang et al. 2016; Biro et al. 2017; Voisey et al. 2017; Amorim et al. 2017; Fonseca et al. 2018); genome-wide association studies (Luzon-Toro et al. 2015); Quantitative Trait Locus (QTL) gene targets downstream analysis (Martinez-Montes et al. 2018); and others (Chen et al. 2016; Alvarez-Castelao et al. 2017; Butler et al. 2016; Makler and Narayanan 2017; Hashemi et al. 2017).

2.6 Summary and Future Development of the Database

The tools and databases in the GeneCards Suite synergistically work in concert to provide information, elucidate relationships, and facilitate solving clinical cases. Each suite member provides deep insights about particular facets of biological research. Specifically, GeneCards is gene-centered, the one-stop-shop for comprehensive details related to genes of interest. MalaCards focuses on diseases and disorders, presenting a detailed view of each malady, with annotations and links including symptoms, drugs, articles, genes, clinical trials, related diseases/disorders, and more. LifeMap Discovery concentrates on gene expression, providing data on the developmental ontology of organ/tissues, anatomical compartments, and cells. It also presents manually curated gene expression at all developmental stages, as well as data extracted from high-throughput experiments and large-scale in situ databases. Users who want to explore human pathways data will find it in PathCards, an integrated database of human biological pathways and their annotations, wherein each record presents a SuperPath that represents one or more human pathways, their gene content, and relationships within member pathways. GeneLoc consolidates genes from major worldwide sources, merging them by location and assigning each GeneCards gene a unique GeneCards Identifier. The GeneLoc site provides a tabular view of a gene's genomic context, including neighboring genes, EST cluster, and markers. GeneHancer, an innovative and growing regulatory element database, focuses on enhancers and promoters, central to tissue-related gene expression, with many known strong connections to diseases. GenesLikeMe measures how genes are related to a target gene, based on shared characteristics, including expression, ontologies, or disorders. Using a gene set from the results of a GeneCards search, or any set of genes of interest, one can extract GeneCards annotations for all genes in the set using GeneALaCart, the suite's batch query facility. The set can be further analyzed using GeneAnalytics, which can identify cell types, diseases, pathways, and functions enriched in the gene set, and provides tools for further in-depth analysis of all of the genes in the set. VarElect identifies and prioritizes genes and variants according to their relevance to diseases and phenotypes of interest and allows one to explore relationships between genes and gene variants and selected diseases, phenotypes, or any pertinent biological term via relevant pathways, interaction networks, and publications. TGex, the suite's end-to-end NGS solution, is a VCF-to-report clinical analyzer which incorporates VarElect's algorithms.

Future plans include continuing to build on the efforts of the last twenty years, ensuring that information from current sources is kept up-to-date, relevant, and provided in a user-friendly manner, in parallel with continuing to innovate in the "dark matter" arena of regulatory elements and RNA genes. The GeneCards Suite's extensive KnowledgeBase and disease interpretation platform fortifies its capacities to relate diseases to non-coding variants identified by WGS, towards providing a comprehensive route to clinical significance of coding and non-coding single nucleotide and structural genomic variations, in order to elucidate unsolved clinical cases and enable accurate clinical diagnosis and comprehensive genetic counseling.

Acknowledgements **Funding**: This work was supported by a research and licensing grant from LifeMap Sciences, Inc. (California, USA) and by the Crown Human Genome Center at the Weizmann Institute of Science (Rehovot, Israel).

References

Adzhubei IA, Schmidt S, Peshkin L, Ramensky VE, Gerasimova A, Bork P, Kondrashov AS, Sunyaev SR (2010) A method and server for predicting damaging missense mutations. Nat Methods 7(4):248–249

Alkelai A, Olender T, Haffner-Krausz R, Tsoory MM, Boyko V, Tatarskyy P, Gross-Isseroff R, Milgrom R, Shushan S, Blau I, Cohn E, Beeri R, Levy-Lahad E, Pras E, Lancet D (2016) A role for TENM1 mutations in congenital general anosmia. Clin Genet 90(3):211–219

Alkelai A, Olender T, Dode C, Shushan S, Tatarskyy P, Furman-Haran E, Boyko V, Gross-Isseroff R, Halvorsen M, Greenbaum L, Milgrom R, Yamada K, Haneishi A, Blau I, Lancet D (2017) Next-generation sequencing of patients with congenital anosmia. Eur J Hum Genet 25(12):1377–1387

Alvarez-Castelao B, Schanzenbacher CT, Hanus C, Glock C, Tom Dieck S, Dorrbaum AR, Bartnik I, Nassim-Assir B, Ciirdaeva E, Mueller A, Dieterich DC, Tirrell DA, Langer JD, Schuman EM (2017) Cell-type-specific metabolic labeling of nascent proteomes in vivo. Nat Biotechnol 35(12):1196–1201

Amorim IS, Graham LC, Carter RN, Morton NM, Hammachi F, Kunath T, Pennetta G, Carpanini SM, Manson JC, Lamont DJ, Wishart TM, Gillingwater TH (2017) Sideroflexin 3 is an alpha-synuclein-dependent mitochondrial protein that regulates synaptic morphology. J Cell Sci 130(2):325–331

Azim MK, Mehnaz A, Ahmed JZ, Mujtaba G (2019) Exome sequencing identifies a novel frameshift variant causing hypomagnesemia with secondary hypocalcemia. CEN Case Rep 8(1):42–47

Bafunno V, Firinu D, D'Apolito M, Cordisco G, Loffredo S, Leccese A, Bova M, Barca MP, Santacroce R, Cicardi M, Del Giacco S, Margaglione M (2018) Mutation of the angiopoietin-1 gene (ANGPT1) associates with a new type of hereditary angioedema. J Allergy Clin Immunol 141(3):1009–1017

Bamshad MJ, Shendure JA, Valle D, Hamosh A, Lupski JR, Gibbs RA, Boerwinkle E, Lifton RP, Gerstein M, Gunel M, Mane S, Nickerson DA, Centers for Mendelian Genomics (2012) The Centers for Mendelian Genomics: a new large-scale initiative to identify the genes underlying rare Mendelian conditions. Am J Med Genet A 158A(7):1523–1525

Bateman A, Martin MJ, O'Donovan C, Magrane M, Alpi E, Antunes R, Bely B, Bingley M, Bonilla C, Britto R, Bursteinas B, Bye-A-Jee H, Cowley A, Da Silva A, De Giorgi M, Dogan T, Fazzini F, Castro LG, Figueira L, Garmiri P et al (2017) UniProt: the universal protein knowledgebase. Nucleic Acids Res 45(D1):D158–D169

Belinky F, Bahir I, Stelzer G, Zimmerman S, Rosen N, Nativ N, Dalah I, Iny Stein T, Rappaport N, Mituyama T, Safran M, Lancet D (2013) Non-redundant compendium of human ncRNA genes in GeneCards. Bioinformatics 29(2):255–261

Belinky F, Nativ N, Stelzer G, Zimmerman S, Iny Stein T, Safran M, Lancet D (2015) PathCards: multi-source consolidation of human biological pathways. Database (Oxford) 2015:bav006

Ben-Ari Fuchs S, Lieder I, Stelzer G, Mazor Y, Buzhor E, Kaplan S, Bogoch Y, Plaschkes I, Shitrit A, Rappaport N, Kohn A, Edgar R, Shenhav L, Safran M, Lancet D, Guan-Golan Y, Warshawsky D, Shtrichman R (2016) GeneAnalytics: an integrative gene set analysis tool for next generation sequencing, RNAseq and microarray data. OMICS 20(3):139–151

Bermejo JL, Huang G, Manoochehri M, Mesa KG, Schick M, Silos RG, Ko Y-D, Bruning T, Brauch H, Lo W-Y, Hoheisel JD, Hamann U (2019) Long intergenic noncoding RNA 299 methylation in peripheral blood is a biomarker for triple-negative breast cancer. Epigenomics 11(1):81–93

Biro O, Nagy B, Rigo J Jr (2017) Identifying miRNA regulatory mechanisms in preeclampsia by systems biology approaches. Hypertens Pregnancy 36(1):90–99

Brown GR, Hem V, Katz KS, Ovetsky M, Wallin C, Ermolaeva O, Tolstoy I, Tatusova T, Pruitt KD, Maglott DR, Murphy TD (2015) Gene: a gene-centered information resource at NCBI. Nucleic Acids Res 43(D1):D36–D42

Butler MG, McGuire AB, Masoud H, Manzardo AM (2016) Currently recognized genes for schizophrenia: high-resolution chromosome ideogram representation. Am J Med Genet B Neuropsychiatr Genet 171B(2):181–202

Carneiro TN, Krepischi AC, Costa SS, da Silva IT, Vianna-Morgante AM, Valieris R, Ezquina SA, Bertola DR, Otto PA, Rosenberg C (2018) Utility of trio-based exome sequencing in the elucidation of the genetic basis of isolated syndromic intellectual disability: illustrative cases. Appl Clin Genet 11:93–98

Chalifa-Caspi V, Yanai I, Ophir R, Rosen N, Shmoish M, Benjamin-Rodrig H, Shklar M, Stein TI, Shmueli O, Safran M, Lancet D (2004) GeneAnnot: comprehensive two-way linking between oligonucleotide array probesets and GeneCards genes. Bioinformatics 20(9):1457–1458

Chen P, Mancini M, Sonis ST, Fernandez-Martinez J, Liu J, Cohen EE, Toback FG (2016) A novel peptide for simultaneously enhanced treatment of head and neck cancer and mitigation of oral mucositis. PLoS One 11(4):e0152995

Edgar R, Mazor Y, Rinon A, Blumenthal J, Golan Y, Buzhor E, Livnat I, Ben-Ari S, Lieder I, Shitrit A, Gilboa Y, Ben-Yehudah A, Edri O, Shraga N, Bogoch Y, Leshansky L, Aharoni S, West MD, Warshawsky D, Shtrichman R (2013) LifeMap discovery: the embryonic development, stem cells, and regenerative medicine research portal. PLoS One 8(7):e66629

Einhorn Y, Weissglas-Volkov D, Carmi S, Ostrer H, Friedman E, Shomron N (2017) Differential analysis of mutations in the Jewish population and their implications for diseases. Genet Res 99: e3

Ekhilevitch N, Kurolap A, Oz-Levi D, Mory A, Hershkovitz T, Ast G, Mandel H, Baris HN (2016) Expanding the MYBPC1 phenotypic spectrum: a novel homozygous mutation causes arthrogryposis multiplex congenita. Clin Genet 90(1):84–89

Erlangsen A, Appadurai V, Wang Y, Turecki G, Mors O, Werge T, Mortensen PB, Starnawska A, Borglum AD, Schork A, Nudel R, Baekvad-Hansen M, Bybjerg-Grauholm J, Hougaard DM, Thompson WK, Nordentoft M, Agerbo E (2020) Genetics of suicide attempts in individuals with and without mental disorders: a population-based genome-wide association study. Mol Psychiatry 25(10):2410–2421

Feliubadalo L, Tonda R, Gausachs M, Trotta JR, Castellanos E, Lopez-Doriga A, Teule A, Tornero E, del Valle J, Gel B, Gut M, Pineda M, Gonzalez S, Menendez M, Navarro M, Capella G, Gut I, Serra E, Brunet J, Beltran S et al (2017) Benchmarking of whole exome sequencing and Ad Hoc designed panels for genetic testing of hereditary cancer. Sci Rep 7: 37984

Fidalgo F, Rodrigues TC, Silva AG, Facure L, de Sa BC, Duprat JP, Achatz MI, Rosenberg C, Carraro DM, Krepischi AC (2016) Role of rare germline copy number variation in melanoma-prone patients. Future Oncol 12(11):1345–1357

Fishilevich S, Nudel R, Rappaport N, Hadar R, Plaschkes I, Iny Stein T, Rosen N, Kohn A, Twik M, Safran M, Lancet D, Cohen D (2017) GeneHancer: genome-wide integration of enhancers and target genes in GeneCards. Database (Oxford) 2017:bax028

Fonseca PAS, Id-Lahoucine S, Reverter A, Medrano JF, Fortes MS, Casellas J, Miglior F, Brito L, Carvalho MRS, Schenkel FS, Nguyen LT, Porto-Neto LR, Thomas MG, Canovas A (2018) Combining multi-OMICs information to identify key-regulator genes for pleiotropic effect on fertility and production traits in beef cattle. PLoS One 13(10):e0205295

Forbes SA, Beare D, Gunasekaran P, Leung K, Bindal N, Boutselakis H, Ding MJ, Bamford S, Cole C, Ward S, Kok CY, Jia MM, De TS, Teague JW, Stratton MR, McDermott U, Campbell PJ (2015) COSMIC: exploring the world's knowledge of somatic mutations in human cancer. Nucleic Acids Res 43(D1):D805–D811

Gene Ontology Consortium (2015) Gene Ontology Consortium: going forward. Nucleic Acids Res 43(Database issue):D1049–D1056

Gilissen C, Hehir-Kwa JY, Thung DT, van de Vorst M, van Bon BWM, Willemsen MH, Kwint M, Janssen IM, Hoischen A, Schenck A, Leach R, Klein R, Tearle R, Bo T, Pfundt R, Yntema HG, de Vries BBA, Kleefstra T, Brunner HG, Vissers LELM et al (2014) Genome sequencing identifies major causes of severe intellectual disability. Nature 511(7509):344–347

Harel A, Inger A, Stelzer G, Strichman-Almashanu L, Dalah I, Safran M, Lancet D (2009) GIFtS: annotation landscape analysis with GeneCards. BMC Bioinformatics 10:348

Hashemi S, Fernandez Martinez JL, Saligan L, Sonis S (2017) Exploring genetic attributions underlying radiotherapy-induced fatigue in prostate cancer patients. J Pain Symptom Manage 54(3):326–339

Hecht M, Bromberg Y, Rost B (2015) Better prediction of functional effects for sequence variants. BMC Genomics 16:S1

Heimer G, Oz-Levi D, Eyal E, Edvardson S, Nissenkorn A, Ruzzo EK, Szeinberg A, Maayan C, Mai-Zahav M, Efrati O, Pras E, Reznik-Wolf H, Lancet D, Goldstein DB, Anikster Y, Shalev SA, Elpeleg O, Ben Zeev B (2016) TECPR2 mutations cause a new subtype of familial dysautonomia like hereditary sensory autonomic neuropathy with intellectual disability. Eur J Paediatr Neurol 20(1):69–79

Heimer G, Eyal E, Zhu X, Ruzzo EK, Marek-Yagel D, Sagiv D, Anikster Y, Reznik-Wolf H, Pras E, Oz Levi D, Lancet D, Ben-Zeev B, Nissenkorn A (2018) Mutations in AIFM1 cause an X-linked childhood cerebellar ataxia partially responsive to riboflavin. Eur J Paediatr Neurol 22(1):93–101

Holzinger ER, Li Q, Parker MM, Hetmanski JB, Marazita ML, Mangold E, Ludwig KU, Taub MA, Begum F, Murray JC, Albacha-Hejazi H, Alqosayer K, Al-Souki G, Albasha Hejazi A, Scott AF, Beaty TH, Bailey-Wilson JE (2017) Analysis of sequence data to identify potential risk variants for oral clefts in multiplex families. Mol Genet Genomic Med 5(5):570–579

Homma TK, Krepischi ACV, Furuya TK, Honjo RS, Malaquias AC, Bertola DR, Costa SS, Canton AP, Roela RA, Freire BL, Kim CA, Rosenberg C, Jorge AAL (2018) Recurrent Copy Number Variants Associated with Syndromic Short Stature of Unknown Cause. Horm Res Paediatr 89(1):13–21

Huang H, Zhang C, Wang B, Wang F, Pei B, Cheng C, Yang W, Zhao Z (2018) Transduction with lentiviral vectors altered the expression profile of host microRNAs. J Virol 92(18):e00503-18

Hulst M, Jansman A, Wijers I, Hoekman A, Vastenhouw S, van Krimpen M, Smits M, Schokker D (2017) Enrichment of in vivo transcription data from dietary intervention studies with in vitro data provides improved insight into gene regulation mechanisms in the intestinal mucosa. Genes Nutr 12:11

Hurles ME, Dermitzakis ET, Tyler-Smith C (2008) The functional impact of structural variation in humans. Trends Genet 24(5):238–245

Jia Z, Mao FB, Wang L, Li MZ, Shi YY, Zhang BR, Gao GL (2017) Whole-exome sequencing identifies a de novo mutation in TRPM4 involved in pleiotropic ventricular septal defect. Int J Clin Exp Pathol 10(5):5092–5104

Kanehisa M, Sato Y, Furumichi M, Morishima K, Tanabe M (2019) New approach for understanding genome variations in KEGG. Nucleic Acids Res 47(D1):D590–D595

Landrum MJ, Lee JM, Riley GR, Jang W, Rubinstein WS, Church DM, Maglott DR (2014) ClinVar: public archive of relationships among sequence variation and human phenotype. Nucleic Acids Res 42(Database issue):D980–D985

Luzon-Toro B, Bleda M, Navarro E, Garcia-Alonso L, Ruiz-Ferrer M, Medina I, Martin-Sanchez M, Gonzalez CY, Fernandez RM, Torroglosa A, Antinolo G, Dopazo J, Borrego S (2015) Identification of epistatic interactions through genome-wide association studies in sporadic medullary and juvenile papillary thyroid carcinomas. BMC Med Genomics 8:83

Makler A, Narayanan R (2017) Mining exosomal genes for pancreatic cancer targets. Cancer Genomics Proteomics 14(3):161–172

Martinez-Montes AM, Fernandez A, Munoz M, Noguera JL, Folch JM, Fernandez AI (2018) Using genome wide association studies to identify common QTL regions in three different genetic backgrounds based on Iberian pig breed. Plos One 13(3):e0190184

Nikulin SV, Knyazev EN, Poloznikov AA, Shilin SA, Gazizov IN, Zakharova GS, Gerasimenko TN (2018) Expression of SLC30A10 and SLC23A3 transporter mRNAs in Caco-2 cells correlates with an increase in the area of the apical membrane. Mol Biol 52(4):577–582

Oz-Levi D, Weiss B, Lahad A, Greenberger S, Pode-Shakked B, Somech R, Olender T, Tatarsky P, Marek-Yagel D, Pras E, Anikster Y, Lancet D (2015) Exome sequencing as a differential diagnosis tool: resolving mild trichohepatoenteric syndrome. Clin Genet 87(6):602–603

Pletscher-Frankild S, Palleja A, Tsafou K, Binder JX, Jensen LJ (2015) DISEASES: text mining and data integration of disease-gene associations. Methods 74:83–89

Porter MF (2006) An algorithm for suffix stripping. Program-Electronic Library and Information Systems 40(3):211–218

Quigley DA, Dang HX, Zhao SG, Lloyd P, Aggarwal R, Alumkal JJ, Foye A, Kothari V, Perry MD, Bailey AM, Playdle D, Barnard TJ, Zhang L, Zhang J, Youngren JF, Cieslik MP, Parolia A, Beer TM, Thomas G, Chi KN et al (2018) Genomic hallmarks and structural variation in metastatic prostate cancer. Cell 174(3):758–769. e9

Ramos E, Levinson BT, Chasnoff S, Hughes A, Young AL, Thornton K, Li AL, Vallania FLM, Province M, Druley TE (2012) Population-based rare variant detection via pooled exome or custom hybridization capture with or without individual indexing. BMC Genomics 13:683

Rappaport N, Nativ N, Stelzer G, Twik M, Guan-Golan Y, Stein TI, Bahir I, Belinky F, Morrey CP, Safran M, Lancet D (2013) MalaCards: an integrated compendium for diseases and their annotation. Database (Oxford) 2013:bat018

Rappaport N, Twik M, Nativ N, Stelzer G, Bahir I, Stein TI, Safran M, Lancet D (2014) MalaCards: a comprehensive automatically-mined database of human diseases. Curr Protoc Bioinformatics 47:1.24.1–1.24.19

Rappaport N, Twik M, Plaschkes I, Nudel R, Iny Stein T, Levitt J, Gershoni M, Morrey CP, Safran M, Lancet D (2017a) MalaCards: an amalgamated human disease compendium with diverse clinical and genetic annotation and structured search. Nucleic Acids Res 45(D1):D877–D887

Rappaport N, Fishilevich S, Nudel R, Twik M, Belinky F, Plaschkes I, Stein TI, Cohen D, Oz-Levi D, Safran M, Lancet D (2017b) Rational confederation of genes and diseases: NGS interpretation via GeneCards. MalaCards and VarElect Biomed Eng Online 16(Suppl 1):72

Rebhan M, Chalifa-Caspi V, Prilusky J, Lancet D (1997) GeneCards: integrating information about genes, proteins and diseases. Trends Genet 13(4):163

Rosen N, Chalifa-Caspi V, Shmueli O, Adato A, Lapidot M, Stampnitzky J, Safran M, Lancet D (2003) GeneLoc: exon-based integration of human genome maps. Bioinformatics 19(Suppl 1): i222–i224

Safran M, Dalah I, Alexander J, Rosen N, Iny Stein T, Shmoish M, Nativ N, Bahir I, Doniger T, Krug H, Sirota-Madi A, Olender T, Golan Y, Stelzer G, Harel A, Lancet D (2010) GeneCards Version 3: the human gene integrator. Database (Oxford) 2010:baq020

Sim NL, Kumar P, Hu J, Henikoff S, Schneider G, Ng PC (2012) SIFT web server: predicting effects of amino acid substitutions on proteins. Nucleic Acids Res 40(Web Server issue):W452–W457

Singh G, Bhat B, Jayadev MSK, Madhusudhan C, Singh A (2018) mutTCPdb: a comprehensive database for genomic variants of a tropical country neglected disease-tropical calcific pancreatitis. Database (Oxford) 2018:bay043

Slater SC, Jover E, Martello A, Mitic T, Rodriguez-Arabaolaza I, Vono R, Alvino VV, Satchell SC, Spinetti G, Caporali A, Madeddu P (2018) MicroRNA-532-5p regulates pericyte function by targeting the transcription regulator BACH1 and angiopoietin-1. Mol Ther 26(12):2823–2837

Smith KR, Bromhead CJ, Hildebrand MS, Shearer AE, Lockhart PJ, Najmabadi H, Leventer RJ, McGillivray G, Amor DJ, Smith RJ, Bahlo M (2011) Reducing the exome search space for Mendelian diseases using genetic linkage analysis of exome genotypes. Genome Biol 12(9):R85

Smith CL, Blake JA, Kadin JA, Richardson JE, Bult CJ, Mouse Genome G (2018) Database, Mouse Genome Database (MGD)-2018: knowledgebase for the laboratory mouse. Nucleic Acids Res 46(D1):D836–D842

Stelzer G, Inger A, Olender T, Iny-Stein T, Dalah I, Harel A, Safran M, Lancet D (2009) GeneDecks: paralog hunting and gene-set distillation with GeneCards annotation. OMICS 13(6):477–487

Stelzer G, Rosen N, Plaschkes I, Zimmerman S, Twik M, Fishilevich S, Stein TI, Nudel R, Lieder I, Mazor Y, Kaplan S, Dahary D, Warshawsky D, Guan-Golan Y, Kohn A, Rappaport N, Safran M, Lancet D (2016a) The GeneCards Suite: from gene data mining to disease genome sequence analyses. Curr Protoc Bioinformatics 54:1.30.1–1.30.33

Stelzer G, Plaschkes I, Oz-Levi D, Alkelai A, Olender T, Zimmerman S, Twik M, Belinky F, Fishilevich S, Nudel R, Guan-Golan Y, Warshawsky D, Dahary D, Kohn A, Mazor Y, Kaplan S, Iny Stein T, Baris HN, Rappaport N, Safran M et al (2016b) VarElect: the phenotype-based variation prioritizer of the GeneCards Suite. BMC Genomics 17(Suppl 2):444

Stranneheim H, Wedell A (2016) Exome and genome sequencing: a revolution for the discovery and diagnosis of monogenic disorders. J Intern Med 279(1):3–15

Syama A, Sen S, Kota LN, Viswanath B, Purushottam M, Varghese M, Jain S, Panicker MM, Mukherjee O (2018) Mutation burden profile in familial Alzheimer’s disease cases from India. Neurobiol Aging 64:158 e7–158 e13

van den Veyver IB, Eng CM (2015) Genome-wide sequencing for prenatal detection of fetal single-gene disorders. Cold Spring Harb Perspect Med 5(10):a023077

Voisey J, Mehta D, McLeay R, Morris CP, Wockner LF, Noble EP, Lawford BR, Young RM (2017) Clinically proven drug targets differentially expressed in the prefrontal cortex of schizophrenia patients. Brain Behav Immun 61:259–265

Weischenfeldt J, Symmons O, Spitz F, Korbel JO (2013) Phenotypic impact of genomic structural variation: insights from and for human disease. Nat Rev Genet 14(2):125–138

Yang YP, Muzny DM, Reid JG, Bainbridge MN, Willis A, Ward PA, Braxton A, Beuten J, Xia F, Niu ZY, Hardison M, Person R, Bekheirnia MR, Leduc MS, Kirby A, Pham P, Scull J, Wang M, Ding Y, Plon SE et al (2013) Clinical whole-exome sequencing for the diagnosis of mendelian disorders. N Engl J Med 369(16):1502–1511

Yang WE, Suchindran S, Nicholson BP, McClain MT, Burke T, Ginsburg GS, Harro CD, Chakraborty S, Sack DA, Woods CW, Tsalik EL (2016) Transcriptomic analysis of the host response and innate resilience to enterotoxigenic Escherichia coli infection in humans. J Infect Dis 213(9):1495–1504

Yang C, Xu Y, Yu M, Lee D, Alharti S, Hellen N, Ahmad Shaik N, Banaganapalli B, Sheikh Ali Mohamoud H, Elango R, Przyborski S, Tenin G, Williams S, O'Sullivan J, Al-Radi OO, Atta J, Harding SE, Keavney B, Lako M, Armstrong L (2017) Induced pluripotent stem cell modelling of HLHS underlines the contribution of dysfunctional NOTCH signalling to impaired cardiogenesis. Hum Mol Genet 26(16):3031–3045

Yang C, Lim W, Bazer FW, Song G (2018) Avobenzone suppresses proliferative activity of human trophoblast cells and induces apoptosis mediated by mitochondrial disruption. Reprod Toxicol 81:50–57

Yates B, Braschi B, Gray KA, Seal RL, Tweedie S, Bruford EA (2017) Genenames.org: the HGNC and VGNC resources in. Nucleic Acids Res 45(D1):D619–D625

Zerbino DR, Achuthan P, Akanni W, Amode MR, Barrell D, Bhai J, Billis K, Cummins C, Gall A, Giron CG, Gil L, Gordon L, Haggerty L, Haskell E, Hourlier T, Izuogu OG, Janacek SH, Juettemann T, J.K. To, Laird MR et al (2018) Ensembl 2018. Nucleic Acids Res 46(D1):D754–D761

Zhang L, Jia Z, Mao F, Shi Y, Bu RF, Zhang B (2016) Whole-exome sequencing identifies a somatic missense mutation of NBN in clear cell sarcoma of the salivary gland. Oncol Rep 35(6):3349–3356

Zhang W, Bojorquez-Gomez A, Velez DO, Xu G, Sanchez KS, Shen JP, Chen K, Licon K, Melton C, Olson KM, Yu MK, Huang JK, Carter H, Farley EK, Snyder M, Fraley SI, Kreisberg JF, Ideker T (2018) A global transcriptional network connecting noncoding mutations to changes in tumor gene expression. Nat Genet 50(4):613–620

Zheng HF, Forgetta V, Hsu YH, Estrada K, Rosello-Diez A, Leo PJ, Dahia CL, Park-Min KH, Tobias JH, Kooperberg C, Kleinman A, Styrkarsdottir U, Liu CT, Uggla C, Evans DS, Nielson CM, Walter K, Pettersson-Kymmer U, McCarthy S, Eriksson J et al (2015) Whole-genome sequencing identifies EN1 as a determinant of bone density and fracture. Nature 526(7571):112–117

Open Access This chapter is licensed under the terms of the Creative Commons Attribution 4.0 International License (http://creativecommons.org/licenses/by/4.0/), which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.

BY

Chapter 3
Atlas of miRNAs and Their Promoters in Human and Mouse

Michiel Jan Laurens de Hoon

Keywords Gene expression · MicroRNA biogenesis · Regulatory sites · Noncoding RNAs · Cap Analysis Gene Expression (CAGE) · Transcription start sites · Transcription · Functional annotation · Vertebrate

3.1 Preamble

MicroRNAs (miRNAs) are a class of short (typically 21–24 nucleotides) noncoding RNAs that inhibit the expression of specific genes by binding to complementary target sequences on RNA transcripts, usually located in the 3′ untranslated region (UTR), and repressing translation or inducing RNA degradation. The miRNA biogenesis pathway consists of excision by Drosha of the precursor miRNA (pre-miRNA) from a primary miRNA (pri-miRNA) transcript, followed by processing of the pre-miRNA by Dicer to release the mature miRNA duplex consisting of a guide RNA and a passenger strand RNA. Mature miRNAs have a phosphate group on their 5′ end and a hydroxyl group on their 3′ end; short RNA (sRNA) sequencing libraries for miRNA expression profiling by next-generation sequencing instruments can be produced by ligating adapters to the 5′ and the 3′ end of the mature miRNA, followed by reverse transcription, PCR amplification, and gel purification (Illumina, Inc. 2011).

As pri-miRNAs are long transcripts with a 5′ cap, their expression can be profiled using Cap Analysis Gene Expression (CAGE; Takahashi et al. 2012). In CAGE, transcripts are reverse-transcribed using random priming to include both polyadenylated and non-polyadenylated RNAs, followed by cap-trapping to capture capped transcripts such as mRNAs, long noncoding RNAs, and pri-miRNAs while avoiding ribosomal RNAs. In addition to measuring expression levels, CAGE

M. J. L. de Hoon (✉)
RIKEN Center for Integrative Medical Sciences, Laboratory for Applied Computational Genomics, Yokohama, Kanagawa, Japan
e-mail: michiel.dehoon@riken.jp

© The Author(s), under exclusive license to Springer Nature Singapore Pte Ltd. 2021

I. Abugessaisa, T. Kasukawa (eds.), *Practical Guide to Life Science Databases*,
https://doi.org/10.1007/978-981-16-5812-9_3

identifies the exact 5′ end of the profiled transcript and therefore its transcription start site and promoter region.

In the fifth edition (FANTOM5) of the FANTOM (Functional Annotation of the Mammalian Genome) project (http://fantom.gsc.riken.jp/), RNA samples from human and mouse, mostly from primary cells, were subjected to CAGE profiling to create an expression atlas of transcription initiation at single-nucleotide resolution (Forrest et al. 2014). A subset of 422 human and 78 mouse RNA samples in FANTOM5 were selected for sRNA library production and sequencing to produce a complementary atlas of miRNA expression in human and mouse (De Rie et al. 2017). By making use of the CAGE data in FANTOM5, for each miRNA the associated pri-miRNA and its promoter was identified. Importantly, each short RNA library in the FANTOM5 collection had a matching CAGE library produced from the same RNA sample. As the expression levels of mature miRNAs observed by sRNA sequencing were correlated to the expression levels of the corresponding pri-miRNAs in the matching CAGE library, pri-miRNA CAGE expression levels could be used as a proxy for the expression level of mature miRNAs, allowing the miRNA expression atlas to be extended to the 1829 human and 1029 mouse CAGE libraries included in FANTOM5 (De Rie et al. 2017).

The FANTOM5 expression atlas of miRNAs and their promoters (http://fantom.gsc.riken.jp/5/suppl/De_Rie_et_al_2017/) thus created is a comprehensive resource of miRNAs in human and mouse, their expression levels in primary cells, tissues, and cell lines, as well as their promoters and associated CAGE expression levels. Using this atlas, the expression pattern across samples of miRNAs can be evaluated as an indication of the cell types in which the miRNA is biologically most relevant. Additionally, sequence analysis of the promoter region around the transcription start site of the identified pri-miRNA enabled an analysis of how the cell type specific expression patterns of each miRNA are encoded in the regulatory control region of the corresponding pri-miRNA (De Rie et al. 2017).

Target users of this atlas are scientists focusing on specific miRNAs or specific cell types, as well as system biologists interested in a global analysis of the cellular regulatory network and the role of miRNAs therein.

3.2 Database Content

Table 3.1 shows an overview of the sRNA data in FANTOM5. Most samples are from human, and most human samples were derived from primary cells. As described previously (De Rie et al. 2017), all sRNA sequencing data were generated using the same protocol to prepare barcoded Illumina TruSeq Small RNA libraries (Illumina, Inc. 2011) and the same sequencing protocol on the Illumina HiSeq2000 sequencer, allowing direct comparison of the expression level of each miRNA between different samples. Similarly, CAGE sequencing data were generated using the same library preparation and sequencing protocol (Kanamori-Katayama et al. 2011) as described in the corresponding publications (Forrest et al. 2014; Arner

Table 3.1 Overview of sRNA libraries in FANTOM5

Organism	Origin	RNA source	Size	# of samples
Human	Primary cells	Total RNA	Short	286
	Primary cells, ES, iPS	Nuclear RNA	Short	16
	Primary cells, ES, iPS	Cytoplasmic RNA	Short	9
	Primary cells, ES, iPS	Nuclear RNA	Long	9
	Primary cells, ES, iPS	Cytoplasmic RNA	Long	9
	Tissue	Total RNA	Short	6
	Time course	Total RNA	Short	87
	Total: 422			
Mouse	Primary cells	Total RNA	Short	1
	Whole body	Total RNA	Short	14
	Time course	Total RNA	Short	27
	Primary cells, ES, iPS	Nuclear RNA	Short	9
	Primary cells, ES, iPS	Cytoplasmic RNA	Short	9
	Primary cells, ES, iPS	Nuclear RNA	Long	9
	Primary cells, ES, iPS	Cytoplasmic RNA	Long	9
	Total: 78			
Rat	Primary cells	Total RNA	Short	3
	Whole body	Total RNA	Short	3
	Total: 6			
Dog	Primary cells	Total RNA	Short	3
	Whole body	Total RNA	Short	3
	Total: 6			
Chicken	Primary cells	Total RNA	Short	3
	Whole body	Total RNA	Short	2
	Total: 5			

RNAs with sizes between 15 and 40 or 50 nucleotides (“short”) or between 80 and 280 bp (“long”) were selected for sequencing (Fort et al. 2014; De Rie et al. 2017). The sRNA libraries for rat, dog, and chicken samples are unpublished and have not yet been integrated in the database
ES embryonic stem cells, *iPS* induced pluripotent stem cells

et al. 2015). Detailed information on each sample is provided in the FANTOM5 Semantic catalog of Samples, Transcription initiation And Regulators (SSTAR; http://fantom.gsc.riken.jp/5/sstar; Abugessaisa et al. 2016).

To generate a miRNA expression table, sequence reads were mapped using bwa (Li and Durbin 2009) to genome assembly hg19 for human and mm9 for mouse and assigned to miRNAs previously annotated in miRBase release 21 (Kozomara and Griffiths-Jones 2014) or to candidate novel miRNAs identified using miRDeep2 (Friedländer et al. 2012) based on genomic overlap. Expression values were converted to c.p.m. (counts-per-million) by normalizing against the total miRNA expression in each sample (De Rie et al. 2017). The FANTOM5 CAGE data was used to identify the pri-miRNA transcript associated with each miRNA by applying a computational pipeline followed by manual curation (De Rie et al. 2017) and to create an expression table for all identified pri-miRNAs. Cell ontology enrichment analysis of the human sRNA and CAGE data in FANTOM5 was performed by evaluating the statistical significance of expression enrichment or depletion of each miRNA or pri-miRNA in cell ontology clusters of primary cell types, retaining the three most enriched and depleted cell ontology terms as a systematic annotation of cell type specific expression.

3.3 Database Architecture

CAGE expression data are referred to by the RNA sample ID (a 4- or 5-digit number) of the RNA sample from which the CAGE library was produced. Each RNA sample number is associated with a FANTOM5 sample ontology ID (FF ontology ID) for human and mouse. Short RNA expression data are provided per RNA sample identified by concatenating the sRNA library ID (of the form SRhi*nnnnn*, where *nnnnn* is a 5-digit number), a 6-nucleotide barcode, and the RNA sample number. In a few cases, RNA samples from the same cellular origin were pooled before sRNA library construction; in such cases, the RNA sample numbers and FF ontology IDs are concatenated by a + sign.

Figure 3.1 shows a schematic view of the data stored in the FANTOM5 miRNA atlas. Each table in this schema corresponds to one flat file available at the FANTOM5 miRNA atlas website. At the core is the miRNA promoter annotation table, which associates each pre-miRNA with the corresponding mature miRNA as well as the promoter of the predicted pri-miRNA. Each pre-miRNA is identified by its pre-miRNA ID (i.e. the pre-miRNA name in miRBase) and its miRBase accession number, as well as by its genomic coordinates on genome assembly hg19 (for human) or mm9 (for mouse). Likewise, the mature miRNA is identified by its miRBase miRNA ID and miRBase accession number. Candidate novel miRNAs are indicated by their number in the accompanying publication (De Rie et al. 2017). The promoter of the predicted pri-miRNA is specified by its CAGE peak ID in FANTOM5 (Forrest et al. 2014), as well as by the genomic coordinate of the transcription start site.

The sRNA expression table provides the expression level, normalized to counts-per-million, of each mature miRNA (identified by the miRNA ID) in the FANTOM5 sRNA samples. A table of sRNA library descriptions shows the RNA sample (specified by the FANTOM5 sample ontology ID and corresponding sample

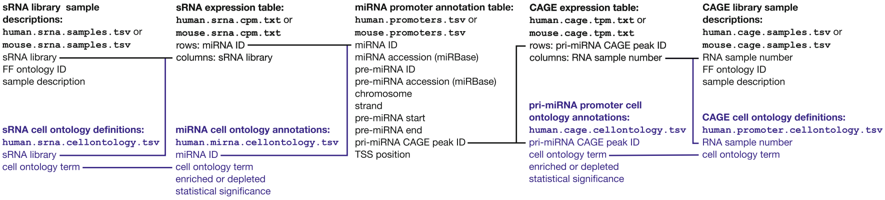


Fig. 3.1 Database architecture of the FANTOM5 miRNA atlas

Table 3.2 Accession numbers for raw sequencing data at DDBJ

Accession number	Data description
DRA000991	CAGE data, human and mouse (Forrest et al. 2014)
DRA001101	sRNA data, human (Andersson et al. 2014)
DRA002711	CAGE data, mouse (Arner et al. 2015)
DRA002747	CAGE data, human (Arner et al. 2015)
DRA002748	CAGE data, mouse (Arner et al. 2015)
DRA003804	sRNA data, human (De Rie et al. 2017)
DRA003807	sRNA data, mouse (De Rie et al. 2017)
DRA000914	sRNA data, human and mouse; CAGE data, human and mouse (Fort et al. 2014)

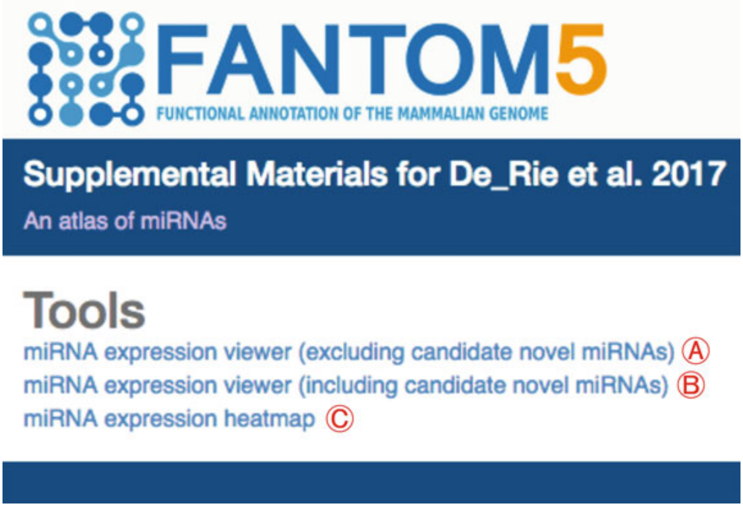
description) from which the sRNA library was produced. The sRNA cell ontology definitions table lists the sRNA libraries associated with each of the cell ontology terms; the miRNA cell ontology annotation table shows the three most enriched and depleted cell ontology terms for each miRNA, together with the statistical significance found.

The CAGE expression table provides the expression level, normalized to tags-per-million (t.p.m.), of all CAGE peaks associated with predicted pri-miRNAs for each CAGE library in FANTOM5, specified by their RNA sample number. The table of CAGE library descriptions shows the FANTOM5 sample ontology number and sample name for each CAGE library. The CAGE cell ontology definitions table shows the CAGE libraries (referenced by their RNA sample number) associated with each cell ontology term, while the pri-miRNA promoter cell ontology annotations show the three most enriched and depleted cell ontology terms for each pri-miRNA (identified by the CAGE peak ID of its associated promoter), together with the statistical significance found.

Raw sequencing data are available from the DNA Data Bank of Japan (DDBJ; https://www.ddbj.nig.ac.jp) as shown in Table 3.2. Sequence alignments to the human and mouse genome are available as part of the FANTOM5 data files (http://fantom.gsc.riken.jp/5/datafiles/). Expression tables as well as promoter, cell ontology, and sample annotations can be downloaded directly from the FANTOM5 miRNA atlas website, as described below.

3.4 Using the FANTOM5 miRNA Atlas Interactively

The FANTOM5 miRNA atlas is available at http://fantom.gsc.riken.jp/5/suppl/De_Rie_et_al_2017/. The landing page (Fig. 3.2) provides links to the miRNA expression viewer including novel miRNAs (Fig. 3.2Ⓑ) or excluding them (Fig. 3.2Ⓐ) for faster loading. The landing page also provides a link to an interactive miRNA



FANTOM5

FUNCTIONAL ANNOTATION OF THE MAMMALIAN GENOME

Supplemental Materials for De_Rie et al. 2017

An atlas of miRNAs

Tools

miRNA expression viewer (excluding candidate novel miRNAs) Ⓐ
miRNA expression viewer (including candidate novel miRNAs) Ⓑ
miRNA expression heatmap Ⓒ

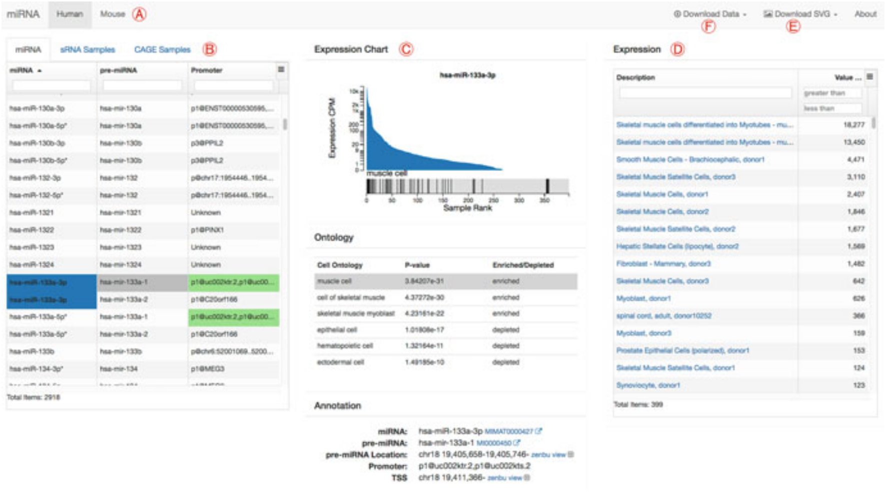
Fig. 3.2 Landing page of the FANTOM5 miRNA atlas at http://fantom.gsc.riken.jp/5/suppl/De_Rie_et_al_2017/

expression heatmap (Fig. 3.2Ⓒ), showing the expression profile of mature miRNAs in the robust set in the human primary cells.

3.4.1 Using the miRNA Expression Viewer

The miRNA expression viewer (Fig. 3.3) visualizes the miRNA expression data and annotation files shown in Fig. 3.1. At the top of the miRNA expression viewer, the user can select to access either the human or the mouse data (Fig. 3.3Ⓐ). The three panels below show, from left to right, the miRNA or sample list (Fig. 3.3Ⓑ), the expression chart, cell ontology analysis, and annotation data (Fig. 3.3Ⓒ), and the expression table (Fig. 3.3Ⓓ).

In the left panel, the user can select to list miRNAs, the sRNA samples, or the CAGE samples (Fig. 3.4Ⓐ). Selecting miRNAs will show a list of all mature miRNAs (both guide RNA and passenger strand RNA), the pre-miRNA from which they originate, and the associated pri-miRNA promoter (Fig. 3.4). Paralogous miRNAs will be listed once for each instance on the genome. The miRNAs can be sorted alphabetically by mature miRNA name (Fig. 3.4Ⓑ), pre-miRNA name (Fig. 3.4Ⓒ), or promoter name (Fig. 3.4Ⓓ) by clicking on the corresponding label. To search for miRNAs, the name of the mature miRNA, of the pre-miRNA, or of the promoter of the pri-miRNA can be entered in the boxes below the label (Fig. 3.4Ⓑ–Ⓓ). Selecting a miRNA from the list will highlight all instances of the mature miRNA (Fig. 3.4Ⓔ), the pre-miRNA associated with the selected miRNA (Fig. 3.4Ⓕ), and the promoter of the associated pri-miRNA both for the guide RNA



miRNA ▲	pre-miRNA	Promoter
hsa-miR-130a-3p	hsa-mir-130a	p1@ENST00000530595,...
hsa-miR-130a-5p*	hsa-mir-130a	p1@ENST00000530595,...
hsa-miR-130b-3p	hsa-mir-130b	p3@PPIL2
hsa-miR-130b-5p*	hsa-mir-130b	p3@PPIL2
hsa-miR-132-3p	hsa-mir-132	p@chr17:1954446..1954...
hsa-miR-132-5p*	hsa-mir-132	p@chr17:1954446..1954...
hsa-miR-1321	hsa-mir-1321	Unknown
hsa-miR-1322	hsa-mir-1322	p1@PINX1
hsa-miR-1323	hsa-mir-1323	Unknown
hsa-miR-1324	hsa-mir-1324	Unknown
hsa-miR-133a-3p	hsa-mir-133a-1	p1@uc002ktr.2,p1@uc00...
hsa-miR-133a-3p	hsa-mir-133a-2	p1@C20orf166
hsa-miR-133a-5p*	hsa-mir-133a-1	p1@uc002ktr.2,p1@uc00...
hsa-miR-133a-5p*	hsa-mir-133a-2	p1@C20orf166
hsa-miR-133b	hsa-mir-133b	p@chr6:52001069..5200...
hsa-miR-134-3p*	hsa-mir-134	p1@MEG3

Total Items: 2918

Cell Ontology	P-value	Enriched/Depleted
muscle cell	3.84207e-31	enriched
cell of skeletal muscle	4.37272e-30	enriched
skeletal muscle myoblast	4.23161e-22	enriched
epithelial cell	1.01808e-17	depleted
hematopoietic cell	1.32164e-11	depleted
ectodermal cell	1.49185e-10	depleted

miRNA: hsa-miR-133a-3p MIMAT0000427
pre-miRNA: hsa-mir-133a-1 MI0000450
pre-miRNA Location: chr18 19,405,658-19,405,746- zenbu view
Promoter: p1@uc002ktr.2,p1@uc002kts.2
TSS chr18 19,411,366- zenbu view

Description	Value ...
Skeletal muscle cells differentiated into Myotubes - mu...	18,277
Skeletal muscle cells differentiated into Myotubes - mu...	13,450
Smooth Muscle Cells - Brachiocephalic, donor1	4,471
Skeletal Muscle Satellite Cells, donor3	3,110
Skeletal Muscle Cells, donor1	2,407
Skeletal Muscle Cells, donor2	1,846
Skeletal Muscle Satellite Cells, donor2	1,677
Hepatic Stellate Cells (lipocyte), donor2	1,569
Fibroblast - Mammary, donor3	1,482
Skeletal Muscle Cells, donor3	642
Myoblast, donor1	626
spinal cord, adult, donor10252	366
Myoblast, donor3	159
Prostate Epithelial Cells (polarized), donor1	153
Skeletal Muscle Satellite Cells, donor1	124
Synoviocyte, donor1	123

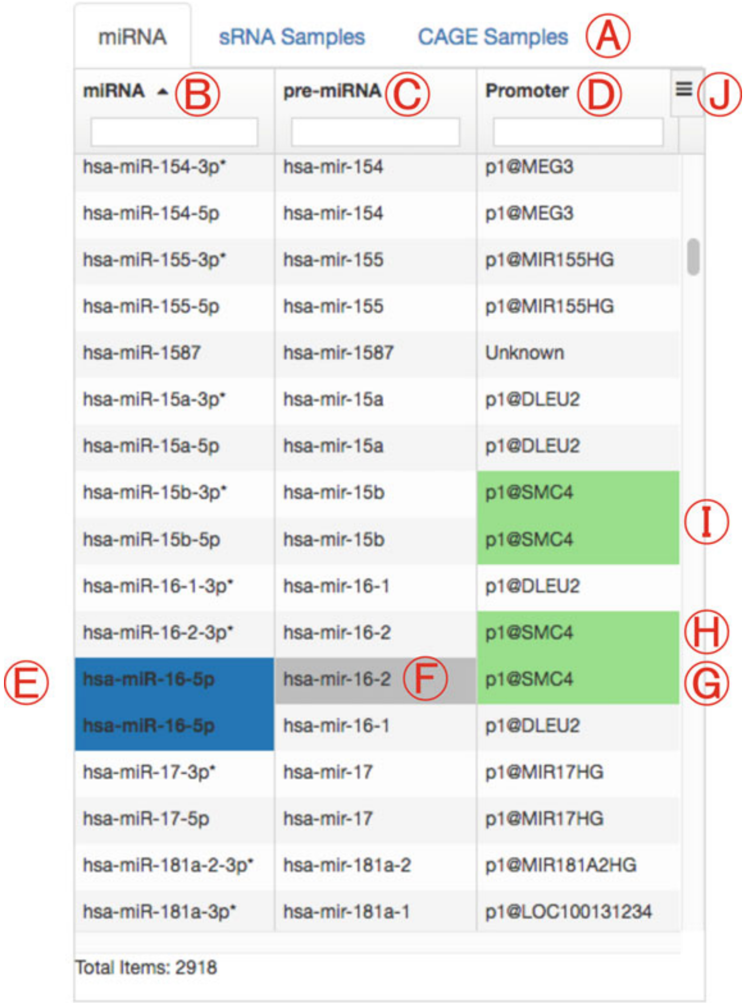
Total Items: 399

Fig. 3.3 FANTOM5 miRNA expression viewer showing human pre-miRNA hsa-mir-133a-1 with its associated mature miRNA hsa-miR-133a-5p (guide) and pri-miRNA promoter p1@uc002ktr.2, p1@uc002kts.2

(Fig. 3.4Ⓕ) and for the passenger strand RNA (Fig. 3.4Ⓕ). The promoter is also highlighted for any other miRNAs originating from the same pri-miRNA (Fig. 3.4Ⓘ). In addition to the columns shown, the pre-miRNA ID (Fig. 3.5Ⓐ) and the miRNA ID as defined by miRBase (Fig. 3.5Ⓑ) can be included in this table by clicking on the options button (Fig. 3.4Ⓙ) to open the options menu (Fig. 3.5).

The center panel (Fig. 3.6) shows the expression chart (Fig. 3.6Ⓐ) for the miRNA selected in the left panel, with the expression in counts-per-million (c.p.m.) on the vertical axis on a logarithmic scale, and the samples sorted by the expression of the miRNA on the horizontal axis. The expression chart can be downloaded as an editable vector image file in the Scalable Vector Graphics (SVG) format by clicking on “Download SVG” (Fig. 3.3Ⓔ).

For guide strand mature miRNAs, below the expression chart the cell ontology panel (Fig. 3.6Ⓑ) shows a table with the cell ontology clusters in which expression of the miRNA is most enriched or depleted, with the statistical significance shown as the P-value. Selecting a cell ontology term from this table will indicate the expression rank of the associated samples on the sample rank bar of the expression chart (Fig. 3.6Ⓒ) as a visual representation of the expression enrichment or depletion of the miRNA in the selected cell ontology cluster samples. Figures 3.6 and 3.7 show the examples of hsa-miR-16-5p and hsa-miR-100-5p with enriched and depleted, respectively, expression in leukocytes. Further below, the annotation panel (Fig. 3.6Ⓓ) provides links to the mature miRNA and the pre-miRNA in the miRBase (Kozomara and Griffiths-Jones 2014) database (Fig. 3.6Ⓔ), the genomic coordinates of the pre-miRNA, the FANTOM5 name of the promoter associated with the pri-miRNA, the coordinates of the transcription start site (TSS), and links to



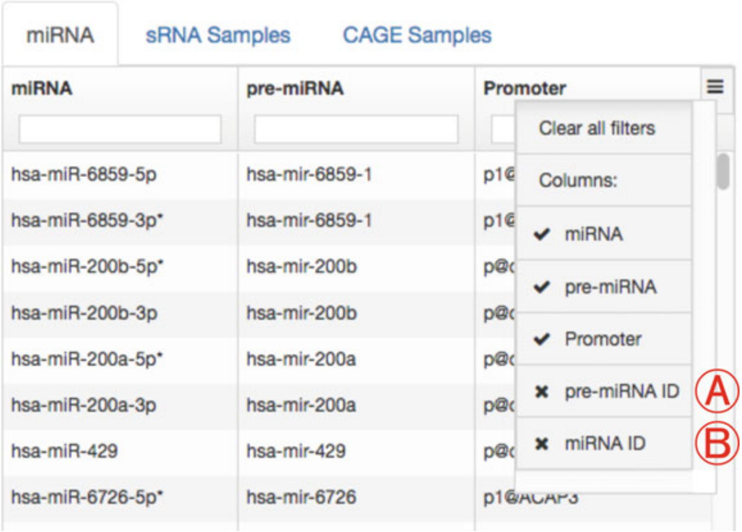
miRNA ▲	pre-miRNA	Promoter
hsa-miR-154-3p*	hsa-mir-154	p1@MEG3
hsa-miR-154-5p	hsa-mir-154	p1@MEG3
hsa-miR-155-3p*	hsa-mir-155	p1@MIR155HG
hsa-miR-155-5p	hsa-mir-155	p1@MIR155HG
hsa-miR-1587	hsa-mir-1587	Unknown
hsa-miR-15a-3p*	hsa-mir-15a	p1@DLEU2
hsa-miR-15a-5p	hsa-mir-15a	p1@DLEU2
hsa-miR-15b-3p*	hsa-mir-15b	p1@SMC4
hsa-miR-15b-5p	hsa-mir-15b	p1@SMC4
hsa-miR-16-1-3p*	hsa-mir-16-1	p1@DLEU2
hsa-miR-16-2-3p*	hsa-mir-16-2	p1@SMC4
hsa-miR-16-5p	hsa-mir-16-2	p1@SMC4
hsa-miR-16-5p	hsa-mir-16-1	p1@DLEU2
hsa-miR-17-3p*	hsa-mir-17	p1@MIR17HG
hsa-miR-17-5p	hsa-mir-17	p1@MIR17HG
hsa-miR-181a-2-3p*	hsa-mir-181a-2	p1@MIR181A2HG
hsa-miR-181a-3p*	hsa-mir-181a-1	p1@LOC100131234

Total Items: 2918

Fig. 3.4 Left panel of the miRNA expression viewer, showing the list of miRNAs

the pre-miRNA and TSS region in the ZENBU (Severin et al. 2014) genome browser (Fig. 3.6Ⓕ).

The right panel (Fig. 3.8) shows an expression table with the expression level in c.p.m. of the selected miRNA in each of the FANTOM5 samples. By default, samples are sorted by expression level of the miRNA. The samples can be sorted alphabetically by clicking on the description label (Fig. 3.8Ⓐ) and sorted by increasing or decreasing expression level by clicking on the value label (Fig. 3.8Ⓑ). Samples with miRNA expression levels greater than or less than a user-specified value can be selected by entering the desired maximum and minimum values in the boxes below the value label (Fig. 3.8Ⓒ). Clicking on the option button

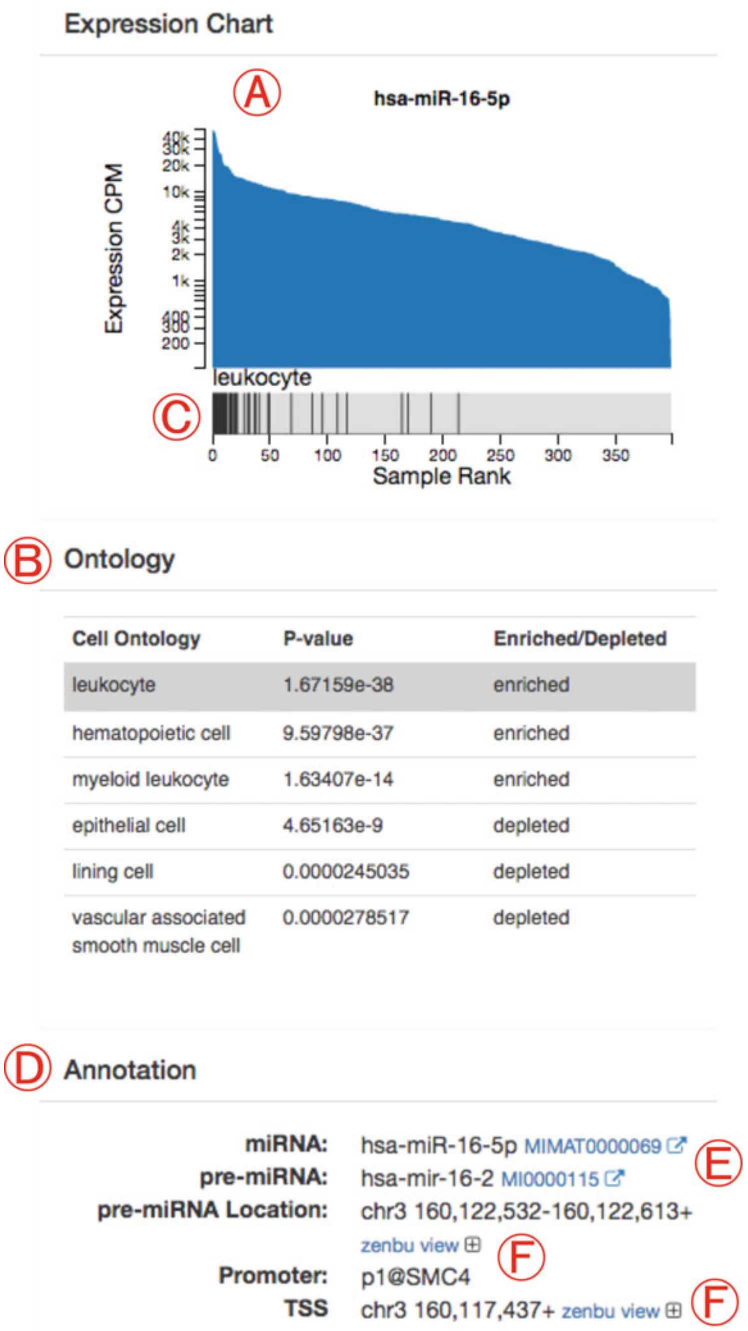


miRNA	pre-miRNA	Promoter
hsa-miR-6859-5p	hsa-mir-6859-1	p1@
hsa-miR-6859-3p*	hsa-mir-6859-1	p1@
hsa-miR-200b-5p*	hsa-mir-200b	p@c
hsa-miR-200b-3p	hsa-mir-200b	p@c
hsa-miR-200a-5p*	hsa-mir-200a	p@c
hsa-miR-200a-3p	hsa-mir-200a	p@c
hsa-miR-429	hsa-mir-429	p@c
hsa-miR-6726-5p*	hsa-mir-6726	p1@ACAP3

Fig. 3.5 Options menu for the left panel of the miRNA expression viewer, in which the columns to be shown in the list of miRNAs can be selected

(Fig. 3.8Ⓓ) to the right of the value label will display open the options menu displaying a list of columns to be shown in this panel (Fig. 3.9), which allows including the rank (Fig. 3.9Ⓐ) and name (Fig. 3.9Ⓑ) of each sample as additional columns in the expression table. Here, the expression ranks are numbered starting from 0, and the name consists of the sRNA library, barcode, and RNA sample number concatenated by periods. This panel also allows exporting the expression table as a downloadable file of comma-separated values (csv) (Fig. 3.9Ⓒ).

Selecting sRNA samples or CAGE samples in the left panel (Fig. 3.4Ⓐ). The sample ID (Fig. 3.11Ⓐ) and SSTAR ID (Fig. 3.11Ⓑ) can be shown as additional columns by selecting them in the options menu (Fig. 3.11) accessible by clicking the options button (Fig. 3.10Ⓐ). Choosing one of the samples will show the expression levels of miRNAs in the expression chart in the central panel, together with the sample annotation information (Fig. 3.12); the expression table in the right panel will show the expression levels numerically for each mature miRNA in c.p.m. (Fig. 3.13). Clicking the options button (Fig. 3.13Ⓐ) will display the options menu from which the miRNA expression rank (Fig. 3.14Ⓐ) and miRBase name (Fig. 3.14Ⓑ) can be shown as additional columns. The options menu also allows exporting the data visible in the expression table as a downloadable file with comma-separated values (csv) (Fig. 3.14Ⓒ).



Ontology

Cell Ontology	P-value	Enriched/Depleted
leukocyte	1.67159e-38	enriched
hematopoietic cell	9.59798e-37	enriched
myeloid leukocyte	1.63407e-14	enriched
epithelial cell	4.65163e-9	depleted
lining cell	0.0000245035	depleted
vascular associated smooth muscle cell	0.0000278517	depleted

Annotation

miRNA: hsa-miR-16-5p MIMAT0000069

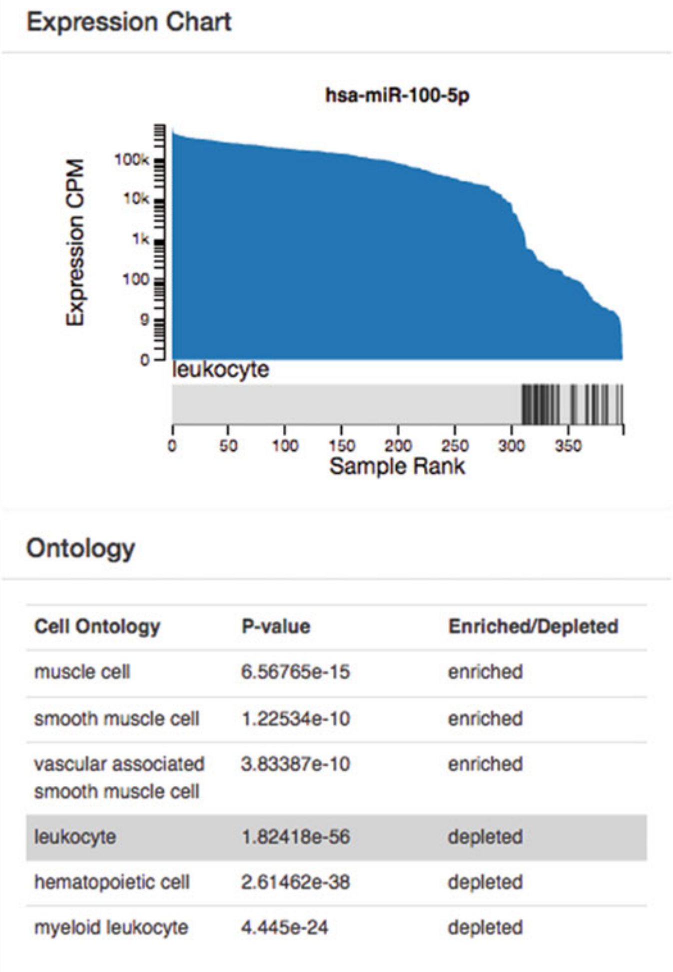
pre-miRNA: hsa-mir-16-2 MI0000115

pre-miRNA Location: chr3 160,122,532-160,122,613+ zenbu view ⊞

Promoter: p1@SMC4

TSS chr3 160,117,437+ zenbu view ⊞

Fig. 3.6 Central panel of the miRNA expression viewer, showing the expression chart and sample rank, the cell ontology results, and the annotation data of the miRNA, pre-miRNA, and promoter. For the selected miRNA, hsa-miR-16-5p, expression is enriched in leukocytes

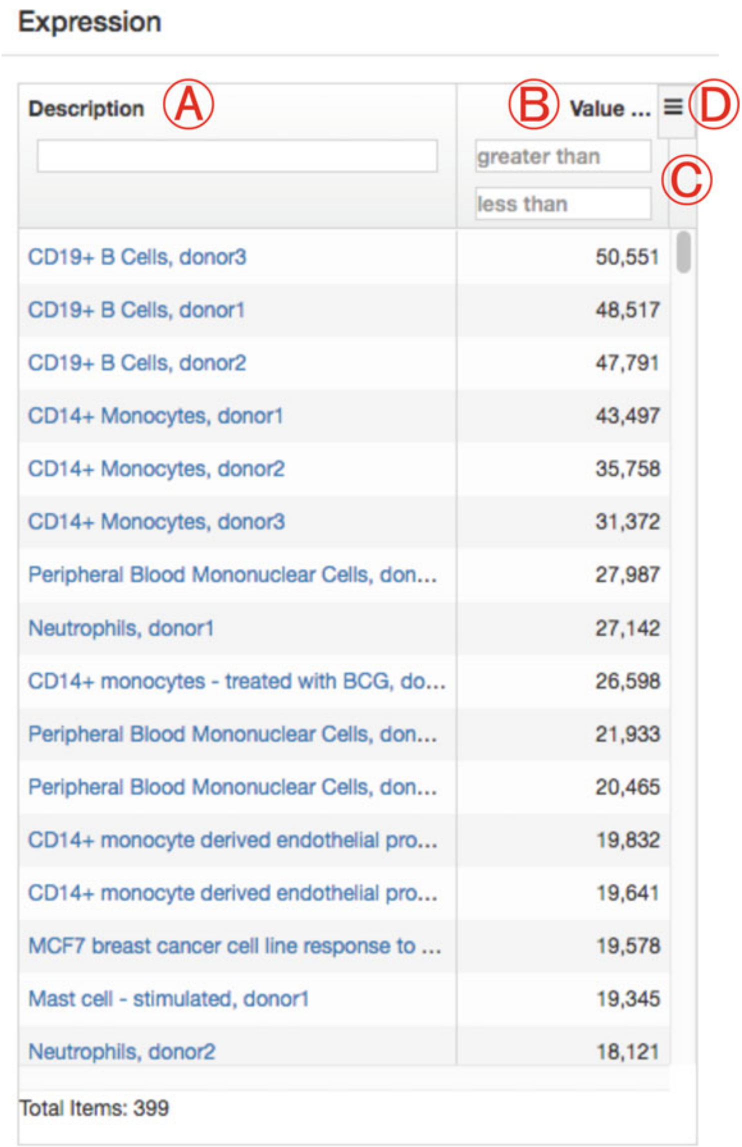


Cell Ontology	P-value	Enriched/Depleted
muscle cell	6.56765e-15	enriched
smooth muscle cell	1.22534e-10	enriched
vascular associated smooth muscle cell	3.83387e-10	enriched
leukocyte	1.82418e-56	depleted
hematopoietic cell	2.61462e-38	depleted
myeloid leukocyte	4.445e-24	depleted

Fig. 3.7 Expression chart, sample rank, and cell ontology results for miRNA hsa-miR-100-5p, for which expression is depleted in leukocytes

3.4.2 Using the Interactive miRNA Expression Heatmap

The interactive miRNA expression heatmap can be accessed by clicking on the link on the landing page of the FANTOM5 miRNA atlas (Fig. 3.2©). The heatmap (Fig. 3.15) shows the expression level of the 735 annotated mature miRNAs (guide strand only) in the robust set (De Rie et al. 2017) as rows, in 118 primary cell types in human, after averaging over donors, as columns. One cell type (“Fibroblast—Pulmonary Artery”) was dropped from the full set of samples (Table 3.1) as the corresponding sRNA library had fewer than 100,000 reads. Cell types were grouped



Description	Value ...
CD19+ B Cells, donor3	50,551
CD19+ B Cells, donor1	48,517
CD19+ B Cells, donor2	47,791
CD14+ Monocytes, donor1	43,497
CD14+ Monocytes, donor2	35,758
CD14+ Monocytes, donor3	31,372
Peripheral Blood Mononuclear Cells, don...	27,987
Neutrophils, donor1	27,142
CD14+ monocytes - treated with BCG, do...	26,598
Peripheral Blood Mononuclear Cells, don...	21,933
Peripheral Blood Mononuclear Cells, don...	20,465
CD14+ monocyte derived endothelial pro...	19,832
CD14+ monocyte derived endothelial pro...	19,641
MCF7 breast cancer cell line response to ...	19,578
Mast cell - stimulated, donor1	19,345
Neutrophils, donor2	18,121

Total Items: 399

Fig. 3.8 Expression table for miRNA hsa-miR-16-5p, which is highly expressed in CD19+ B cells, CD14+ monocytes, and other leukocytes

by category (Fig. 3.15Ⓐ) as indicated by the color bar below the cell type name (Fig. 3.15Ⓑ).

To sort the heatmap, both miRNAs and cell types were clustered using pairwise centroid-linkage hierarchical clustering with the Pearson correlation as the similarity measure (De Hoon et al. 2004) after normalizing the expression of each miRNA to

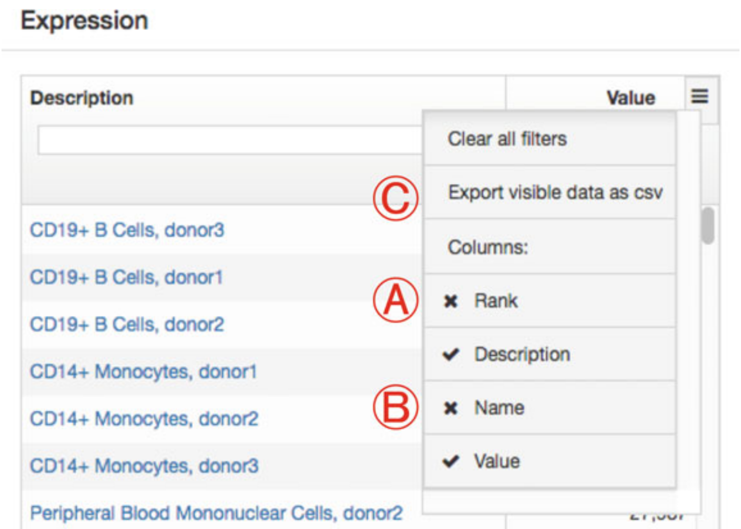


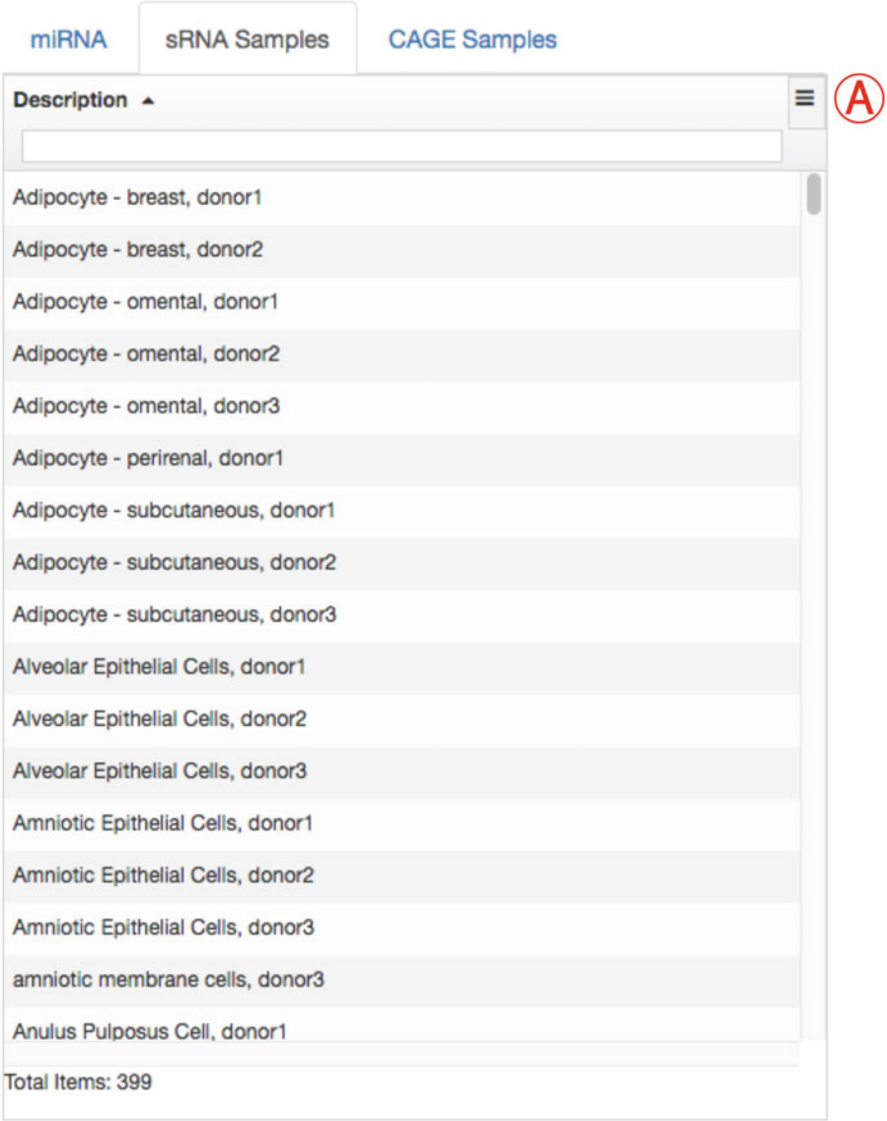
Fig. 3.9 Options menu for the right panel of the miRNA expression viewer, in which the columns to be shown in the expression table can be selected, and data can be exported as a downloadable file in the csv (comma-separated values) format

Z-scores by subtracting the mean and dividing by the standard deviation across samples. The hierarchical clustering tree itself is not displayed. Each cell in the heatmap is colored based on the calculated Z-score of the expression of the miRNA in the cell type (Fig. 3.15Ⓒ). Hovering with the mouse over a cell in the heatmap will show the mature miRNA name, the cell type, the category to which the cell type belongs, and the Z-score of the miRNA expression level in the cell type (Fig. 3.16). Clicking on a cell or on a miRNA name will redirect the browser to the miRNA expression viewer for the corresponding miRNA. The expression data shown in the heatmap can be downloaded as a tab-delimited file by clicking on “Download Data” (Fig. 3.15Ⓓ).

3.5 Database Access and Mining Methods

A flat file for each table shown in Fig. 3.1 can be downloaded from the FANTOM5 miRNA atlas website by clicking on the download button (Fig. 3.3Ⓕ).

The miRNA promoter annotation table is provided as the tab-delimited files `human.promoters.tsv` and `mouse.promoters.tsv` for human and mouse, respectively. This file contains two lines for each pre-miRNA, corresponding to the guide strand and the passenger strand of the mature miRNA. Each line shows the name and miRBase ID of the pre-miRNA, its chromosome, strand, and genomic coordinates, the name and miRBase ID of the guide RNA or passenger strand RNA of the associated mature miRNA, the short description of the FANTOM5 CAGE



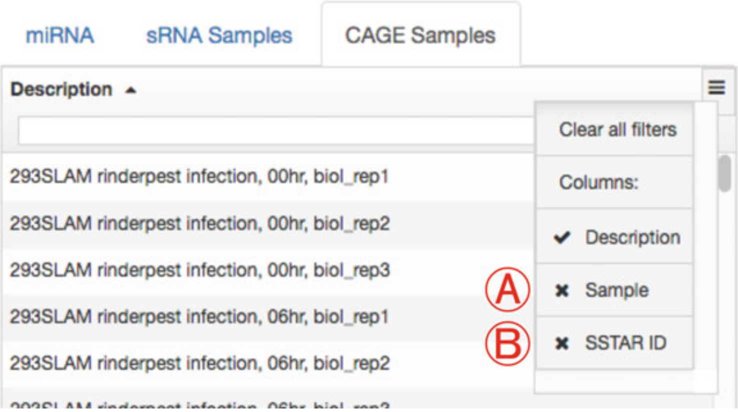
Adipocyte - breast, donor1
Adipocyte - breast, donor2
Adipocyte - omental, donor1
Adipocyte - omental, donor2
Adipocyte - omental, donor3
Adipocyte - perirenal, donor1
Adipocyte - subcutaneous, donor1
Adipocyte - subcutaneous, donor2
Adipocyte - subcutaneous, donor3
Alveolar Epithelial Cells, donor1
Alveolar Epithelial Cells, donor2
Alveolar Epithelial Cells, donor3
Amniotic Epithelial Cells, donor1
Amniotic Epithelial Cells, donor2
Amniotic Epithelial Cells, donor3
amniotic membrane cells, donor3
Anulus Pulposus Cell, donor1

Total Items: 399

Fig. 3.10 Left panel of the miRNA expression viewer, showing the list of sRNA samples

peak of the promoter associated with the pri-miRNA, and the transcription start site of the pri-miRNA.

The sRNA expression tables, `human.srna.cpm.txt` for human and `mouse.srna.cpm.txt` for mouse, are tab-delimited files with the expression of both the guide and passenger strand of each miRNA, identified by its miRBase ID, in each sRNA library. Expression values are normalized to counts-per-million (c.p.m.) in



293SLAM rinderpest infection, 00hr, biol_rep1

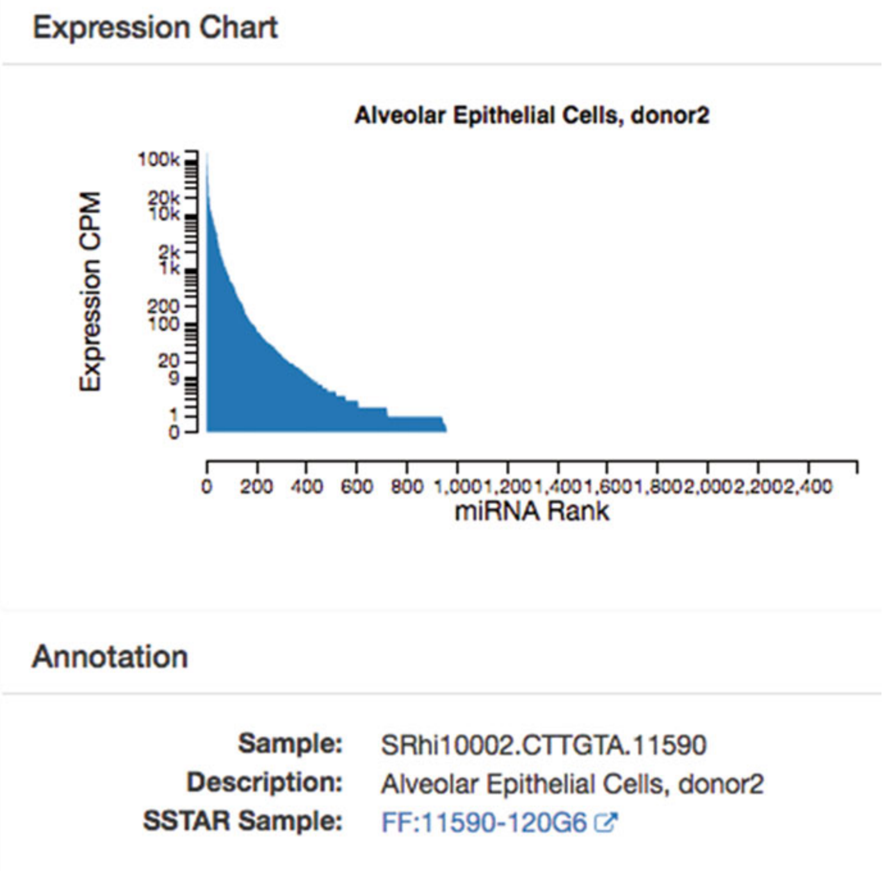
293SLAM rinderpest infection, 00hr, biol_rep2

293SLAM rinderpest infection, 00hr, biol_rep3

293SLAM rinderpest infection, 06hr, biol_rep1

293SLAM rinderpest infection, 06hr, biol_rep2

Fig. 3.11 Left panel of the miRNA expression viewer, showing the list of CAGE samples, together with the options menu in which columns to be included can be selected

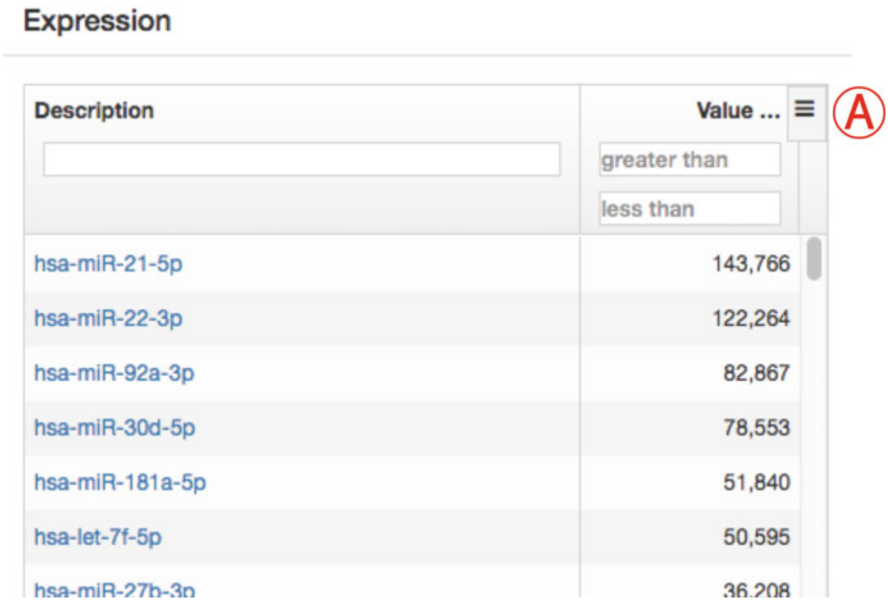


Sample: SRhi10002.CTTGTA.11590

Description: Alveolar Epithelial Cells, donor2

SSTAR Sample: FF:11590-120G6

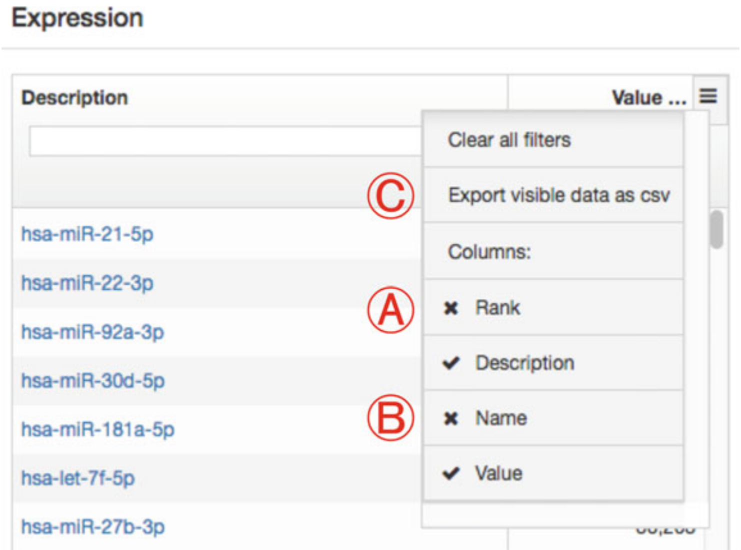
Fig. 3.12 Expression chart for miRNA expression in "alveolar epithelial cells, donor 2" as measured by sRNA sequencing, together with the annotation data of this sample



Expression

Description	Value ... ≡
	greater than less than
hsa-miR-21-5p	143,766
hsa-miR-22-3p	122,264
hsa-miR-92a-3p	82,867
hsa-miR-30d-5p	78,553
hsa-miR-181a-5p	51,840
hsa-let-7f-5p	50,595
hsa-miR-27b-3p	36,208

Fig. 3.13 Right panel of the miRNA expression viewer, showing the expression table for miRNA expression in sample "alveolar epithelial cells, donor 2" as measured by sRNA sequencing



Expression

Description	Value ... ≡
hsa-miR-21-5p	
hsa-miR-22-3p	
hsa-miR-92a-3p	
hsa-miR-30d-5p	
hsa-miR-181a-5p	
hsa-let-7f-5p	
hsa-miR-27b-3p	

Clear all filters

Export visible data as csv

Columns:

✗ Rank

✓ Description

✗ Name

✓ Value

Fig. 3.14 Options menu for the right panel of the miRNA expression viewer, in which the columns to be shown in the expression table can be selected, and data can be exported as a downloadable file in the csv (comma-separated values) format

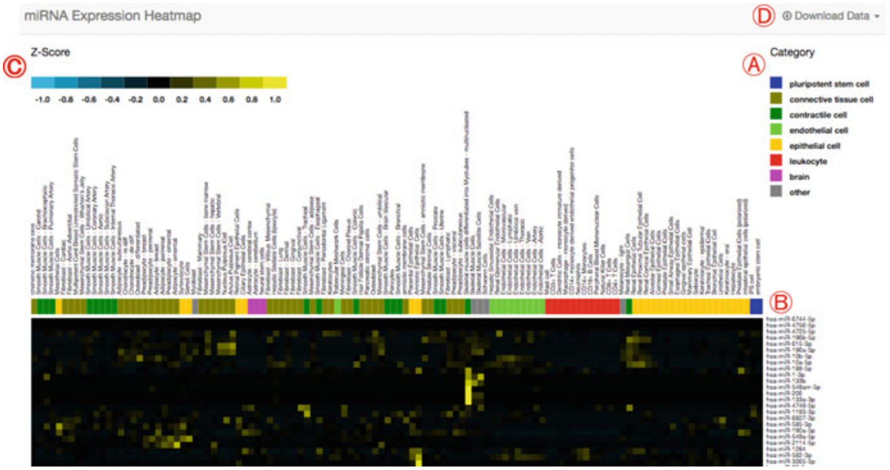
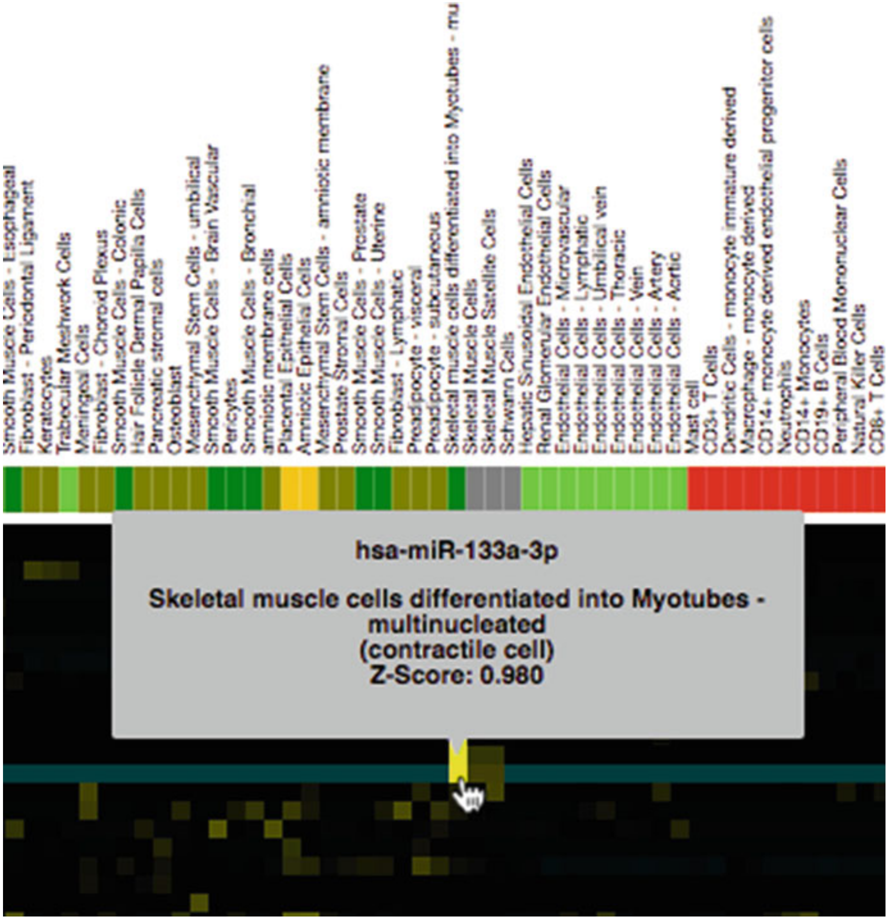


Fig. 3.15 Interactive heatmap showing the expression of mature miRNAs (guide RNA) in the robust set (De Rie et al. 2017), normalized to Z-scores, in human primary cell types after averaging over donors

each library separately. The sRNA library IDs are associated with an FF ontology ID and sample description in the files `human.srna.samples.tsv` and `mouse.srna.samples.tsv`.

The CAGE expression tables, `human.cage.tpm.txt` and `mouse.cage.tpm.txt` for human and mouse, respectively, are tab-delimited files with the CAGE expression level of the FANTOM5 CAGE peaks associated with pri-miRNAs as shown in the miRNA promoter annotation table. Each column corresponds to one CAGE library, as identified by its associated RNA sample number, and is normalized to tags-per-million (t.p.m.). The RNA sample numbers are associated with an FF ontology ID and sample description in the files `human.cage.samples.tsv` and `mouse.cage.samples.tsv`.

The miRNA cell ontology annotations based on the sRNA expression patterns across cell types are available in the file `human.mirna.cellontology.tsv`, listing for each miRNA (identified by miRBase ID) the three cell ontology clusters in which the expression of the miRNA is most enriched, and the three cell ontology clusters in which the expression is most depleted. The sRNA cell ontology definitions are provided in the file `human.srna.cellontology.tsv`, listing the sRNA samples associated with each cell ontology cluster. Similarly, the file `human.promoter.cellontology.tsv` lists for each FANTOM5 CAGE peak associated with a pri-miRNA the three cell ontology clusters in which the CAGE expression levels of the pri-miRNA are most enriched, and the three cell ontology clusters in which the expression is most depleted. The CAGE cell ontology definitions are provided in the file `human.cage.cellontology.tsv`, listing the RNA sample numbers associated with each cell ontology cluster.


Smooth Muscle Cells - Esophageal
Fibroblast - Periodontal Ligament
Keratocytes
Trabecular Meshwork Cells
Meningeal Cells
Fibroblast - Choroid Plexus
Smooth Muscle Cells - Colonic
Hair Follicle Dermal Papilla Cells
Pancreatic stromal cells
Osteoblast
Mesenchymal Stem Cells - umbilical
Smooth Muscle Cells - Brain Vascular
Pericytes
Smooth Muscle Cells - Bronchial
amniotic membrane cells
Placental Epithelial Cells
Amniotic Epithelial Cells
Mesenchymal Stem Cells - amniotic membrane
Prostate Stromal Cells
Smooth Muscle Cells - Prostate
Smooth Muscle Cells - Uterine
Fibroblast - Lymphatic
Preadipocyte - visceral
Preadipocyte - subcutaneous
Skeletal muscle cells differentiated into Myotubes - mu
Skeletal Muscle Cells
Skeletal Muscle Satellite Cells
Schwann Cells
Hepatic Sinusoidal Endothelial Cells
Renal Glomerular Endothelial Cells
Endothelial Cells - Microvascular
Endothelial Cells - Lymphatic
Endothelial Cells - Umbilical vein
Endothelial Cells - Thoracic
Endothelial Cells - Vein
Endothelial Cells - Artery
Endothelial Cells - Aortic
Mast cell
CD3+ T Cells
Dendritic Cells - monocyte immature derived
Macrophage - monocyte derived
CD14+ monocyte derived endothelial progenitor cells
Neutrophils
CD14+ Monocytes
CD19+ B Cells
Peripheral Blood Mononuclear Cells
Natural Killer Cells
CD8+ T Cells

hsa-miR-133a-3p

Skeletal muscle cells differentiated into Myotubes - multinucleated
(contractile cell)
Z-Score: 0.980

Fig. 3.16 Popup shown when hovering the mouse over a particular cell in the interactive heatmap, with the mature miRNA name, sample description, cell type category, and Z-score

3.6 Summary and Future Development of the Database

The FANTOM5 expression atlas of miRNAs and their promoters provides a basis for a detailed analysis of the transcriptional regulation of miRNAs and their role in defining cell types. The atlas will be extended in the near future sRNA sequencing data for rat, dog, and chicken (Table 3.1), together with promoter annotations for miRNAs in these three species, opening the door to cross-species comparisons of miRNA expression and regulation.

Acknowledgements **Funding**: FANTOM5 was made possible by the following grants: Research Grant for RIKEN Omics Science Center from MEXT to Yoshihide Hayashizaki; Grant of the Innovative Cell Biology by Innovative Technology (Cell Innovation Program) from the MEXT to

Yoshihide Hayashizaki; Research Grant from MEXT to the RIKEN Center for Life Science Technologies; Research Grant to RIKEN Preventive Medicine and Diagnosis Innovation Program from MEXT to Yoshihide Hayashizaki.

References

Abugessaisa I, Shimoji H, Sahin S et al (2016) FANTOM5 transcriptome catalog of cellular states based on Semantic MediaWiki. Database (Oxford) 2016:baw105

Andersson R, Gebhard C, Miguel-Escalada I (2014) An atlas of active enhancers across human cell types and tissues. Nature 507(7493):455–461

Arner E, Daub CO, Vitting-Seerup K et al (2015) Transcribed enhancers lead waves of coordinated transcription in transitioning mammalian cells. Science 347(6225):1010–1014

De Hoon MJL, Imoto S, Nolan J, Miyano S (2004) Open source clustering software. Bioinformatics 20(9):1453–1454

De Rie D, Abugessaisa I, Alam T et al (2017) An integrated expression atlas of miRNAs and their promoters in human and mouse. Nat Biotechnol 35(9):872–878

Forrest ARR, Kawaji H, Rehli M et al (2014) A promoter-level mammalian expression atlas. Nature 507(7493):462–470

Fort A, Hashimoto K, Yamada D et al (2014) Deep transcriptome profiling of mammalian stem cells supports a regulatory role for retrotransposons in pluripotency maintenance. Nat Genet 46(6):558–566

Friedländer MR, Mackowiak SD, Li N et al (2012) miRDeep2 accurately identifies known and hundreds of novel microRNA genes in seven animal clades. Nucleic Acids Res 40(1):37–52

Illumina, Inc. (2011) TruSeq® Small RNA Sample Preparation Guide, Catalog # RS-930-1012, Part # 15004197 Rev. C, March 2011. San Diego, California

Kanamori-Katayama M, Itoh M, Kawaji H et al (2011) Unamplified cap analysis of gene expression on a single-molecule sequencer. Genome Res 21(7):1150–1159

Kozomara A, Griffiths-Jones S (2014) miRBase: annotating high confidence microRNAs using deep sequencing data. Nucleic Acids Res 42(Database issue):D68–D73

Li H, Durbin R (2009) Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics 25(14):1754–1760

Severin J, Lizio M, Harshbarger J et al (2014) Interactive visualization and analysis of large-scale sequencing datasets using ZENBU. Nat Biotechnol 32(3):217–219

Takahashi H, Lassmann T, Murata M, Carninci P (2012) 5′ end-centered expression profiling using cap-analysis gene expression and next-generation sequencing. Nat Protoc 7(3):542–561

Chapter 4
IHEC Data Portal

David Bujold, Romain Grégoire, David Brownlee, Ksenia Zaytseva, and Guillaume Bourque

Abstract The International Human Epigenome Consortium (IHEC) coordinates the creation of high-resolution reference epigenome maps for healthy and diseased human tissues. The IHEC Data Portal (http://epigenomesportal.ca/ihec) provides open access to discover, visualize, compare, and download annotation tracks generated for these reference epigenomes, along with their associated metadata. More than 16,000 transcriptome, methylome, and histone tracks, sourced from over 900 tissues and compared against the hg19 and hg38 reference genomes are provided. The standardized IHEC Data Hub submission protocol ensures that complete and compatible data is integrated from the consortium members. The data portal offers a highly visual, interactive, and intuitive data grid interface to facilitate dataset selection across multiple criteria. All data may be correlated, visualized in popular genome browsers, exported to a Galaxy instance, downloaded for further analysis, or saved as a session for future retrieval. Through providing access to IHEC's database of reference epigenomic datasets, the IHEC Data Portal reinforces the consortium's commitment to Open Science.

Keywords API · IHEC · Accession · Analysis · Archival · Dataset · Epigenomics · Metadata · Online resource · Portal · Visualization

4.1 Preamble

4.1.1 Summary

The International Human Epigenome Consortium (Stunnenberg et al. 2016) is composed of several member groups, with the common goal of producing high quality reference epigenomes over a variety of assays for healthy and diseased

D. Bujold · R. Grégoire · D. Brownlee · K. Zaytseva · G. Bourque (✉)
McGill University, Montreal, QC, Canada
e-mail: guil.bourque@mcgill.ca

© The Author(s), under exclusive license to Springer Nature Singapore Pte Ltd. 2021

I. Abugessaisa, T. Kasukawa (eds.), *Practical Guide to Life Science Databases*,
https://doi.org/10.1007/978-981-16-5812-9_4

tissues. These human reference maps typically include information on the transcriptome, methylome, and histone modifications over the whole genome. The IHEC Data Portal (http://epigenomesportal.ca/ihec) contains metadata on the original donor, sample and experimental conditions, and processed experimental results in the form of anonymized annotation tracks. The portal makes these annotations available for convenient viewing in popular visualization and data analysis software, such as the UCSC Genome Browser, Ensembl, and the Galaxy web platform.

4.1.2 Purpose

The IHEC Data Portal (Bujold et al. 2016) is an online database that offers tools to discover, navigate, visualize, and analyze epigenomic datasets produced by IHEC. It is the official resource to access annotations on reference epigenomes produced within the consortium. The visual, navigation, and filtering tools enable users to find datasets of interest based on all the metadata properties provided by data submitters. Central to the portal is a data grid showing counts of available datasets for given sets of cell types and assays.

4.1.3 Source and Type of Dataset

IHEC members contributing data to the IHEC Data Portal include ENCODE, NIH Roadmap, CEEHRC, Blueprint, DEEP, AMED-CREST, and KNIH. The majority of data is from human sources, compared against the hg19 and hg38 reference genomes. A small amount of mouse data (compared against the mm10 reference genome) is also available.

The database stores processed annotation tracks such as bigBed and bigWig files (Kent et al. 2010), and non-personally identifiable metadata on those datasets. Personally identifiable data, including the raw sequencing data, is deposited at controlled access repositories such as EGA (Leinonen et al. 2010) and DDBJ (Tateno et al. 2002). Pointers to these controlled access repositories are provided, should a user want to request access to raw data (FASTQ, BAM files) and sensitive clinical/phenotypic metadata.

4.1.4 Target User Group

The portal targets the broad epigenomic research community seeking reference epigenomic datasets produced by IHEC member consortia and/or other community members. A session creation feature generates a URL to share dataset selections among collaborators.

4.2 Database Overview

4.2.1 Importance of the Dataset

One of the primary goals of IHEC is to produce reference epigenomic maps on a wide array of tissues and cell types, both diseased and healthy. This is done by assessing multiple angles of the epigenomic landscape of a sample, primarily at the levels of histone modification, DNA methylation, and transcriptome. Other assays are also implemented, depending on the consortium who produces the data. The resulting reference maps can be used in multiple ways, such as a point of comparison for similar datasets produced in the community, as a training dataset for algorithms or otherwise. The IHEC Data Portal stores the processed data and annotation tracks produced by the downstream analysis of these reference epigenomes' raw data.

4.2.2 Current Status of Achievements

In operation since 2014, the IHEC Data Portal provides open access to over 16,000 datasets, sourced from over 900 tissues. It is the official platform to navigate the reference epigenomic datasets produced by the IHEC consortium. The platform seamlessly unifies data discovery, visualization, analysis, and sharing.

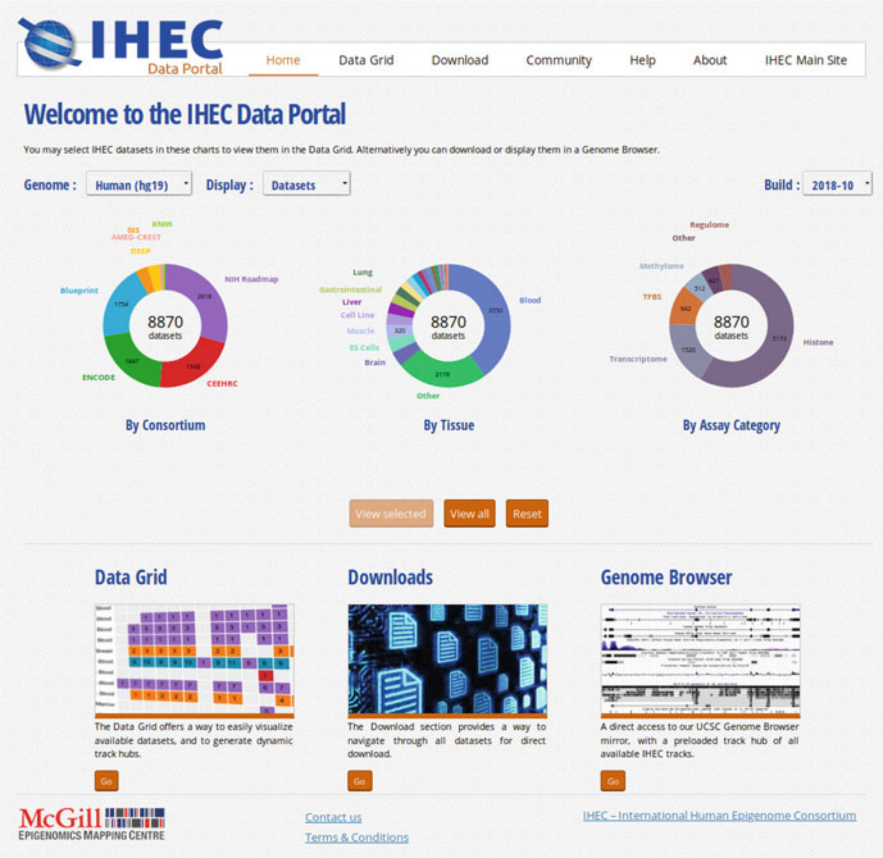
4.2.3 Main Features of the Database

A dynamic web interface responsive to user actions is central to the design of the IHEC Data Portal. When first accessing the IHEC Data Portal, a user is presented with a visual summary showing the volume of data available for a given reference genome (Fig. 4.1). This overview is depicted on data wheels showing the distribution of data over three axes: data producing consortium, tissue type, and assay category. The data wheels are interactive and act as filter shortcuts when selecting datasets to be used during the user session.

Table 4.1 outlines the principal object types stored in the database. All fields within these entities are searchable.

4.2.4 Future Updates and Availability

A new data build is released annually, integrating all new data submitted by member consortia, using IHEC Data Hub documents. Annotations from all versions of datasets are archived and permanently available for retrieval.



Welcome to the IHEC Data Portal

You may select IHEC datasets in these charts to view them in the Data Grid. Alternatively you can download or display them in a Genome Browser.

Data Grid

The Data Grid offers a way to easily visualize available datasets, and to generate dynamic track hubs.

Downloads

The Download section provides a way to navigate through all datasets for direct download.

Genome Browser

A direct access to our UCSC Genome Browser mirror, with a preloaded track hub of all available IHEC tracks.

Fig. 4.1 Database overview, as presented on the home page

4.3 Content and Architecture of the Database

4.3.1 Type of Data Stored

The IHEC Data Portal stores annotation tracks and pertinent metadata for all epigenomic datasets. The annotation tracks are produced by each respective data provider using their own bioinformatics analysis pipelines and are provided to the IHEC Data Portal. At the moment, the platform does not provide a uniformly processed dataset that includes all IHEC members data, although this uniformly processed dataset is currently being generated through an IHEC reanalysis initiative.

Over 200 data fields are collected within the database, though specific fields per dataset vary according to the nature of the sample, assay, and analysis. The minimal required data fields for a dataset description are listed in Table 4.2. A complete data dictionary explaining the stored data properties is found in the IHEC Ecosystem

Table 4.1 Summary of entities stored in the IHEC Data Portal

Entity	Description
Assays	Characterizing an experiment, allows users to understand the nature of the annotations (e.g. dataset is a transcriptome assessed using an mRNA-Seq experiment)
Dataset	A dataset represents one experiment that has been done on one sample (e.g. sample MS000101 methylome, assessed by a whole-genome bisulfite sequencing experiment, is one dataset) Each dataset is characterized by a group of annotation tracks, used in the portal to visualize results
Dataset Track	One specific annotation track for a dataset. For stranded RNA-Seq experiments, one track is the whole-genome signal on the forward strand. There is another track for the reverse strand
Cell type	Information on the nature of the dataset source biological material
Donor	One donor can give multiple samples that will each be mapped as an epigenome. For example, donor McGill0002 gives blood samples at 2 different time points, from which monocytes, T cells, and B cells are characterized, resulting in a total of 6 samples

Metadata Specification: (https://github.com/IHEC/ihec-ecosystems/blob/master/docs/metadata/1.0/Ihec_metadata_specification.md).

4.3.2 Data Collection Methods

The process of including data and metadata in the portal is done through a standardized exchange format called IHEC Data Hubs. These are structured as JSON documents, defined by a JSON schema specification. For IHEC member consortia, a data ingestion mechanism is in place to validate provided metadata, download annotation tracks, and store them on the portal.

For its annual data release, the portal collects IHEC Data Hub documents created by IHEC members with new data. These JSON-formatted documents include all required IHEC metadata properties, alongside URLs of the bigBed and bigWig data files at their remote storage locations. The portal uses these documents to download a copy of all tracks to its servers. This provides some level of protection against datasets that have been modified, moved, or have disappeared from their original source.

4.3.3 Curation Approaches

The geographically distributed and linguistically diverse nature of the IHEC member groups raises challenges for data standardization. As a result, publicly available and established ontologies are relied upon heavily, wherever applicable (Table 4.3).

Table 4.2 Data types detailed description

Top level entity: Dataset	**Annotation track properties**
Sub-element	**Description**
Browser	Type of annotation track
Signal Annotation Tracks	The basic experiment signal, providing information on where the raw readset aligns over the genome. Offered in bigWig file format
ChIP-Seq Peaks	The location of the peaks called by a ChIP-Seq experiment is provided in bigBed file format
Methylation profile	The relative ratio of reads at CpG sites that were methylated. Provided in bigWig file format
big_data_url	The URL from which this dataset track can be obtained online
md5sum	The checksum for this track
Top level entity: Experiment	**Describes metadata captured on experimental conditions, providing more information on how the experiment was conducted**
Sub-element	**Description**
experiment_type	DNA Methylation, mRNA-Seq, ChIP-Seq Input, etc.
experiment_ontology_uri	Link to experiment ontology information
Top level entity: Analysis	**Information on the bioinformatics analysis pipeline that was used to generate annotation tracks**
Sub-element	**Description**
analysis_group	The group that ran the bioinformatics analysis to produce these dataset tracks
alignment_software	The name of the software used for mapping
alignment_software_version	The version of the software used for mapping
analysis_software	The name of the software used for determining signal (read density)
analysis_software_version	The version of the software used for determining signal (read density)
Top level entity: Donor	**Properties captured on the original sample donor**
Sub-element	**Description**
donor_age	The age of the donor that provided the primary cell
donor_life_stage	(Controlled Vocabulary) "fetal," "newborn," "child," "adult," "unknown," "embryonic," "postnatal"
donor_health_status	The health status of the donor that provided the primary cell
donor_sex	(Controlled Vocabulary) "Male," "Female," "Unknown," or "Mixed" for pooled samples
donor_ethnicity	The ethnicity of the donor that provided the primary cell
Top level entity: Sample	**Depending on the biological material type, different properties of the reference epigenome sample are captured**
Sub-element	**Description**
sample_ontology_uri	Link to sample ontology information. The ontology used depends on the sample biological material type
molecule	The type of molecule that was extracted from the biological material, such as total RNA, polyA RNA, cytoplasmic RNA, nuclear RNA, genomic DNA, protein, or other

(continued)

Table 4.2 (continued)

disease_ontology_uri	Link to disease ontology information
disease	More specific disease information
biomaterial_type	(Controlled Vocabulary) “Cell Line,” “Primary Cell,” “Primary Cell Culture,” “Primary Tissue”
line	The name of the cell line
lineage	The developmental lineage to which the cell line belongs
differentiation_stage	The stage in cell differentiation
medium	The medium in which the cell line has been grown
sex	“Male,” “Female,” “Unknown,” or “Mixed” for pooled samples
cell_type	The type of cell
culture_conditions	The conditions under which the primary cell was cultured
tissue_type	The type of tissue
tissue_depot	Details about the anatomical location from which the primary tissue was collected

Table 4.3 Ontologies applied for data standardization

Entity	Ontology applied
Experiment	Experiment type is defined using the Ontology for Biomedical Investigations (The Ontology for Biomedical Investigations 2016)
Cell type	Depending on the type of cell material, different ontologies are used For a tissue: Uberon (Mungall et al. 2016) For primary cells: Cell Ontology (Cell Ontology 2019) For cell lines: Experimental Factor Ontology (Malone et al. 2010)
Sample	Sample condition is defined using the NCI Metathesaurus (NCI Metathesaurus 2019) Type of molecule extracted is defined using Sequence Ontology (Eilbeck et al. 2005)
Donor	Donor health condition is defined using the NCI Metathesaurus (NCI Metathesaurus 2019)

Further, textual data fields are restricted to enumerated controlled vocabulary, when possible.

Non-anonymized unprocessed raw data files (FASTQ files) require storage at controlled access repositories and are not stored in the IHEC Data Portal. A separate database, EpiRR (Epigenome Reference Registry 2019) was created by IHEC as a means for data producers to specify the location of their raw data that is available only upon request. When raw data is deposited at such a controlled access repository, the data provider will create in EpiRR a “reference registry” record that uniquely identifies this reference epigenome and collects metadata and the location of raw data, for each experiment that was run. In this case, they are assigned EpiRR IDs. When importing reference epigenome experiments, the IHEC Data Portal makes the connection between epigenome metadata, processed data such as annotation tracks, and the raw data stored elsewhere.

For track types other than bigWig and bigBed, the IHEC Data Portal should be seen as a means to browse and identify datasets of interest, rather than as the definitive tool to download IHEC's complete data library.

4.3.4 Processing Strategy

Originally, the IHEC Data Portal used UCSC Genome Browser Track Hubs to import its metadata. A data curator was required to review existing UCSC Track Hubs and write a data provider-specific integration plugin that would convert metadata to the expected standards before entering it in the database. This workflow has changed since the adoption of the IHEC Data Hub JSON specification. Currently, the list of expected data values is normalized, ontological references are added, and validation tools are provided to ensure that quality policies are respected.

After successful validation, a local copy of all bigWig and bigBed track files referenced in the IHEC Data Hubs is downloaded. These local copies are processed using specialized software to feed the tools and visualization services of the IHEC Data Portal.

Metadata from the IHEC Data Hub is parsed and stored inside a relational database enabling rapid retrieval. Pre-processing scripts aggregate data by tissue, provider, and assay category.

4.3.5 Dataset Indexing/Accession Number/Identification

Users of the IHEC Data Portal can group their selected datasets for recall, reference, and future refinement with a permanent reference bookmark. After choosing the appropriate datasets from the data grid, an IHEC Data Portal ID is generated by clicking on the "Save Session" button at the bottom of the grid. A URL containing a session identifier is produced that references the specific dataset collection. This permalink may be retained, shared, or used as a reference in publications.

4.3.6 Quality Control Method

The annotation and processed data provided to the IHEC Data Portal by consortium members are integrated into its database directly. While some basic validation is run over tracks submitted to the portal, data providers remain responsible for following the standards established by the various IHEC working groups. Prior to submission, the metadata JSON documents can be validated via an online tool (https://epigenomesportal.ca/metadator/) to ensure proper syntax formatting and that all required properties are present. Further semantic tests check the internal referential integrity of samples and donors. Tracks are restricted to known track types of interest as agreed upon by the IHEC Metadata Standards Workgroup and duplicate tracks are resolved. If the dataset is also registered in the EpiRR reference registry, metadata property values are compared to ensure correspondence.

A series of quality control tools are applied to the tracks to identify potential problems such as incomplete coverage of the whole genome, high background noise, or poor correlation with other tracks of the same category. This quality control pipeline (EpiQC) is currently being enhanced to make further sanity checks on newly submitted data.

4.3.7 Database Update and Maintenance Strategy

Updates to the database are made through IHEC Data Hubs. New data hubs are batch processed for new IHEC Data Portal content releases. Data hubs provided by consortium members are parsed and their relevant data is inserted into the database in a structured manner. No database maintenance is performed outside of these planned uploads since the database is not mutated on any other occasion.

4.4 Database Access and Mining Methods

4.4.1 Tools and Techniques to Access Database Content

The IHEC Data Portal features multiple means of exploring and accessing its data and metadata in an intuitive manner.

The Data Grid page is the primary mechanism of selection. This page provides an interactive interface that can query, filter, and display the quantity of information contained in the database. It uses Javascript libraries such as D3 to produce visually interesting graphics, such as correlation charts. It also offers advanced panels to filter on every kind of available metadata such as the sample source, the cell type, or the assay type.

To achieve session persistence, users can generate a unique accession ID (an IHECDP number) associated with their dataset selection and filtering criteria. Such an accession ID can be shared with collaborators and referenced in publications to illustrate which datasets from the portal are being used.

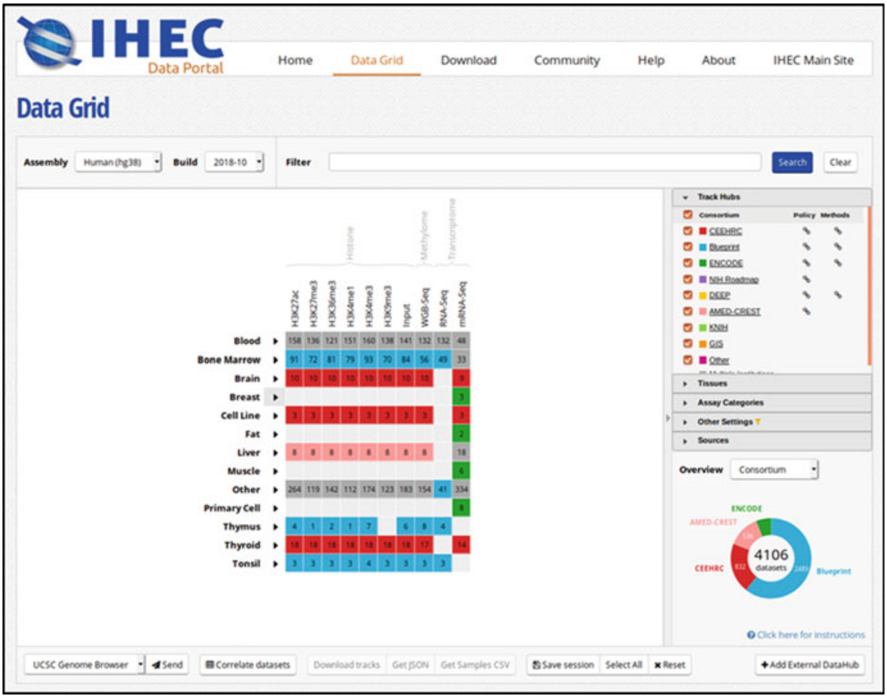
4.4.2 Software and Tools for Discovering and Mining the Database

The IHEC Data Portal server points to the main instance of the UCSC Genome Browser. Selected dataset tracks can be instantly inspected since copies of the associated bigWig and bigBed files are saved on the portal server.

It is possible to assess annotation track similarities over the whole genome by using the epiGeEC (Laperle et al. 2019) Correlation feature. This displays a superimposed 2-dimensional grid, showcasing the Pearson correlation score of each dataset against each other in the grid selection. It enables, for instance, to identify outliers in a group of datasets. A dendrogram is also presented.

4.4.3 How to Explore and Browse the Database

The data grid is the central tool of the portal, allowing the visualization and selection of available datasets (Fig. 4.2). It plots tissue versus assay category and indicates counts of available datasets. Tissue category can be expanded to specific tissue



	H3K27ac	H3K27me3	H3K36me3	H3K4me1	H3K4me3	H3K9me3	Input	WGB-Seq	RNA-Seq	mRNA-Seq
Blood	158	136	121	151	160	138	141	132	132	48
Bone Marrow	91	72	81	79	93	70	84	56	49	33
Brain	10	10	10	10	10	10	10	10		9
Breast										3
Cell Line	3	3	3	3	3	3	3	3		3
Fat										2
Liver	8	8	8	8	8	8	8	8		18
Muscle										6
Other	264	119	142	112	174	123	183	154	41	334
Primary Cell										8
Thymus	4	1	2	1	7		6	8	4	
Thyroid	18	18	18	18	18	18	18	17		14
Tonsil	3	3	3	3	4	3	3	3	3	

Fig. 4.2 The IHEC Data Portal Data Grid page, with filtering panel on the right

types. Data is further filtered through the right-hand panel. Once datasets of interest are chosen in this grid, users can export their selection to multiple tools for visualization, analysis, or data and metadata download. When accessing the data grid for the first time, a tour of the grid features will be offered. It is also possible to access this tour by clicking on the "Click here for instructions" link in the lower right corner of the screen.

4.4.4 How to Query the Database

Much of the data selection may be performed through clickable elements such as the data wheels, drop down menus, and filtering panel. The data grid itself is fundamentally an interactive graphical query tool.

More granular refinement is accomplished through the Filter text input. This input text box is searchable by keyword and dynamically offers autocomplete suggestions such that typing "`Brain`" would propose results like "`cell_type = "Fetal_Brain"`" and "`cell_type_category = "Brain"`." It accepts a sequential Boolean-like syntax such that typing "`donor_sex = "Male" and assay = "mRNA-Seq" and cell_type="liver"`" produces datasets sourced from the livers of male donors that went through mRNA-seq analysis.

4.4.5 How to Upload/Download Data to the Database

It is possible to download bigWig and bigBed annotation tracks directly from the portal, to be used in local analyses. Users should keep in mind that the IHEC Data Portal only stores the processed data annotation tracks and not the raw data itself. The latter is stored at controlled access repositories such as and, EGA, and a data access request must be filled by groups who would want to request access.

The Data Grid page can also produce IHEC Data Hub documents reflecting user-selected subsets of the database content. These data hubs contain all the required information to fetch tracks by providing their metadata and their URL. This is the intended way for users to download content from the database. A Download page is provided to download individual tracks, presented in a directory-like structure directly within the browser.

The portal also offers a mechanism for external groups to add their own datasets to user sessions, through the concept of Community Hubs. Building a community hub involves producing an IHEC Data Hub and hosting it on an external server. The data hub can be included on the Data Grid page for visualization along with the full IHEC database content, without requiring to upload its referenced data.

As previously described, updates and uploads are accomplished through the processing and integration of IHEC Data Hubs generated by consortium members.

4.4.6 Programming and Automated Techniques for Database Access

4.4.6.1 Web Services

The annotation tracks for the portal datasets can be automatically imported into a Galaxy session (Giardine et al. 2005). The portal server's Galaxy instance provides all regular Galaxy tools for comparisons and analyses.

4.4.6.2 FTP

With the abundance of alternative data download mechanisms enabled, no FTP access is offered. Lists of dataset URLs can be downloaded using any common download tool.

4.4.6.3 API

The web API used by the Data Grid page is available for anyone to use. It consists of a few simple endpoints that send their response in JSON format. It enables users to retrieve meta-information on datasets of interest as a machine-readable document. Users can also export parts of the metadata in CSV-formatted documents.

4.4.6.4 Bioinformatics Tools (R/Python) Packages

The API can be used either from another website or by any scripting language such as R or Python. An implementation of the GA4GH rnaget API, used to obtain only slices of transcriptomic experiments, is also offered. For more information on the available rnaget endpoints, please consult (https://github.com/ga4gh-rnaseq/schema).

4.4.7 Database Integration Strategy

Through Community Hubs, the portal is offering ways for the wider community to submit their own datasets for everyone to retrieve. By submitting an appropriate JSON-formatted document, a community data producer is able to upload annotation tracks and their associated metadata directly into a portal grid session. This functions in a similar way to the UCSC Genome Browser that can import external annotation tracks through UCSC Genome Browser track hubs.

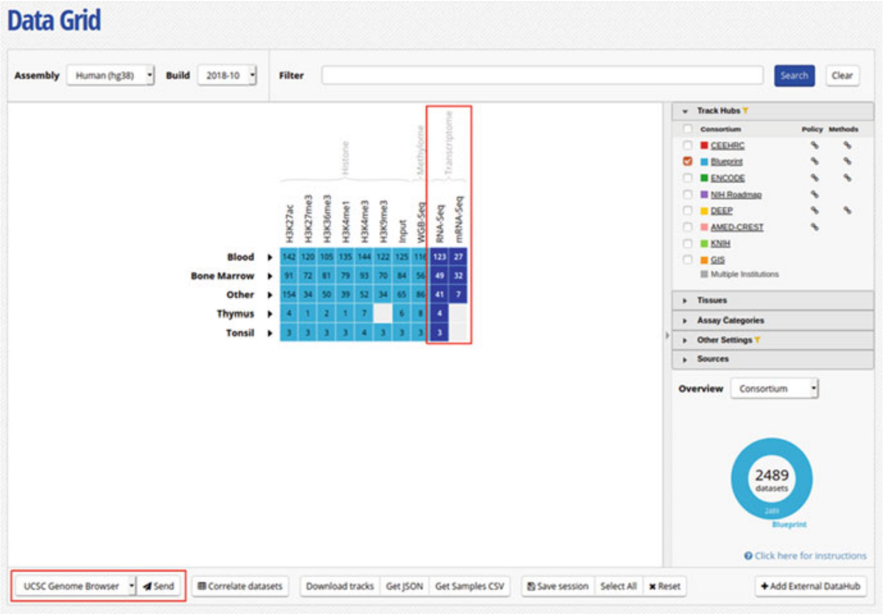
4.5 Use Cases and Demo to Utilize the Database

4.5.1 Use Case 1: Navigating Blueprint hg38 Transcriptomic Data in the UCSC Genome Browser

As a straightforward first use case, all samples' transcriptome data produced by the Blueprint consortium is selected for visualization in the UCSC Genome Browser. To accomplish this, Blueprint is selected from the By Consortium data wheel on the home page and the "View selected" button is pushed. The Data Grid page opens and the "Transcriptome" column header is clicked to select all samples in the two columns below (Fig. 4.3). Subsequently, the "UCSC Genome Browser" option is selected from the dropdown in the bottom-left corner and the Send button is clicked.

4.5.2 Use Case 2: Discovering Available IHEC Datasets Matching Metadata Requirements

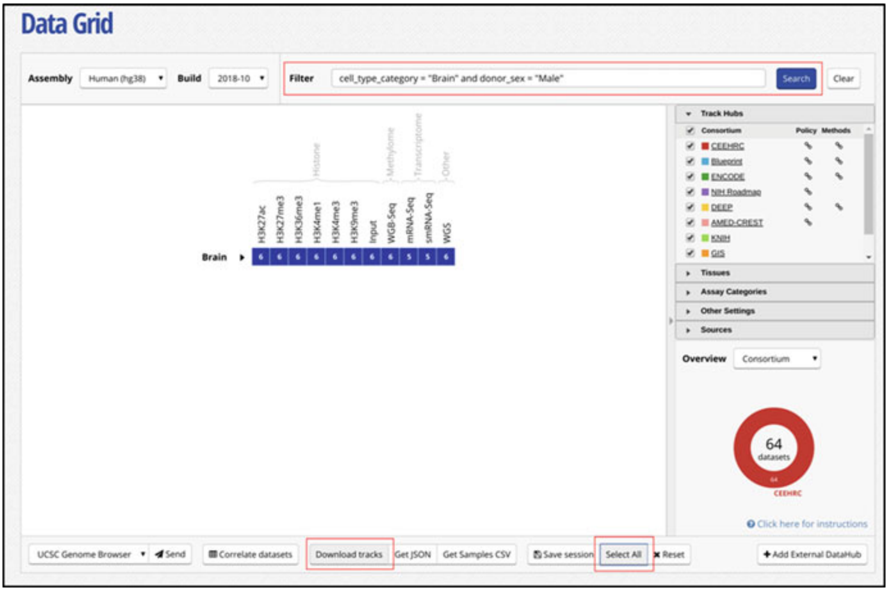
An intermediate use case would be to filter the samples based on metadata properties and to download the resulting filtered data. From the home page, after clicking on the "View all" button, the Data Grid page opens. Further refinement is accomplished



	Histone							Methylome	Transcriptome	
	H3K27ac	H3K27me3	H3K36me3	H3K4me1	H3K4me3	H3K9me3	Input	WGB-Seq	RNA-Seq	mRNA-Seq
Blood ▸	142	120	105	135	144	122	125	116	123	27
Bone Marrow ▸	91	72	81	79	93	70	84	56	49	32
Other ▸	154	34	50	39	52	34	65	86	41	7
Thymus ▸	4	1	2	1	7		6	8	4	
Tonsil ▸	3	3	3	3	4	3	3	3	3	

□	Consortium	Policy	Methods
□	🟥 CEEHRC	✎	✎
☑	🟦 Blueprint	✎	✎
□	🟩 ENCODE	✎	✎
□	🟪 NIH Roadmap	✎	
□	🟨 DEEP	✎	✎
□	🟥 AMED-CREST	✎	
□	🟩 KNIH		
□	🟧 GIS		
	▨ Multiple Institutions		

Fig. 4.3 Selecting all transcriptomic datasets provided by the Blueprint consortium for viewing in the UCSC Genome Browser



Data Grid

Assembly: Human (hg38) Build: 2018-10

Filter: cell_type_category = "Brain" and donor_sex = "Male"

	Histone							Methylome	Transcriptome		Other
	H3K27ac	H3K27me3	H3K36me3	H3K4me1	H3K4me3	H3K9me3	Input	WGB-Seq	mRNA-Seq	smRNA-Seq	WGS
Brain	6	6	6	6	6	6	6	6	5	5	6

Track Hubs

☑ Consortium	Policy	Methods
☑ CEEHRC	✎	✎
☑ Blueprint	✎	✎
☑ ENCODE	✎	✎
☑ NIH Roadmap	✎	
☑ DEEP	✎	✎
☑ AMED-CREST	✎	
☑ KNIH		
☑ GIS		

Overview: Consortium

Fig. 4.4 Filtering based on text search

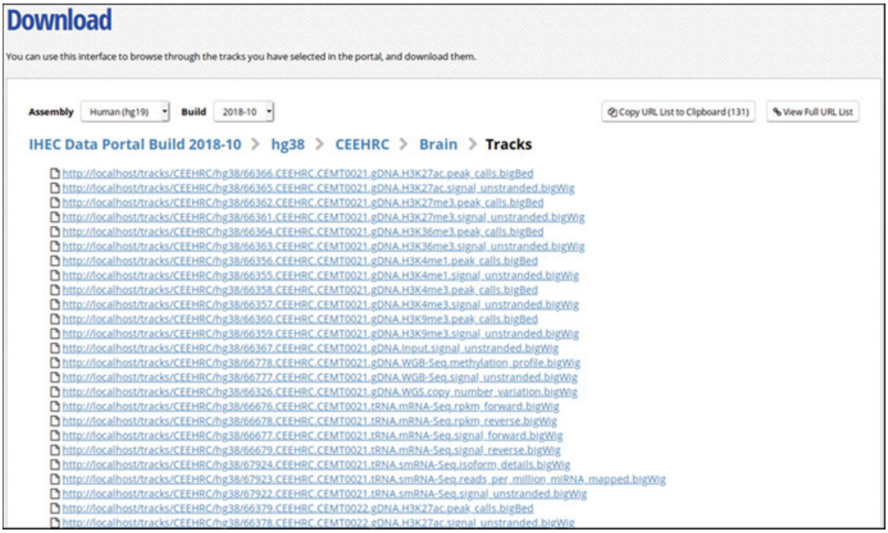
through the Filter text input. Entering, for example, “`cell_type_category = "Brain" and donor_sex = "Male"`” refers to datasets of brain cell samples from male donors. Once the desired search string is composed, pressing the Search button causes the grid to update with the resulting samples (Fig. 4.4).

To download the data, the “Select All,” then “Download tracks” buttons are pushed and the Download page opens (Fig. 4.5). The full list of tracks can be viewed by clicking the “View Full URL List” button. This textual list can be saved, manually entered in the browser URL bar, copied to the clipboard, passed as input to wget or otherwise processed.

4.5.3 *Use Case 3: Assessing Dataset Comparability in the Portal*

A more advanced use case would be to integrate data from two consortia for the same assay/cell type. Their comparability can be assessed using the correlation tool and the results imported into a Galaxy session.

The right panel is used to filter consortia and when the specific cell type/assay datasets are selected from the data grid, their background changes to blue (Fig. 4.6). The correlation matrix is presented (Fig. 4.7) when the Correlate Datasets button is pushed. Finally, the data is sent to a Galaxy session by selecting this option from the bottom-left dropdown and clicking Send.



Download

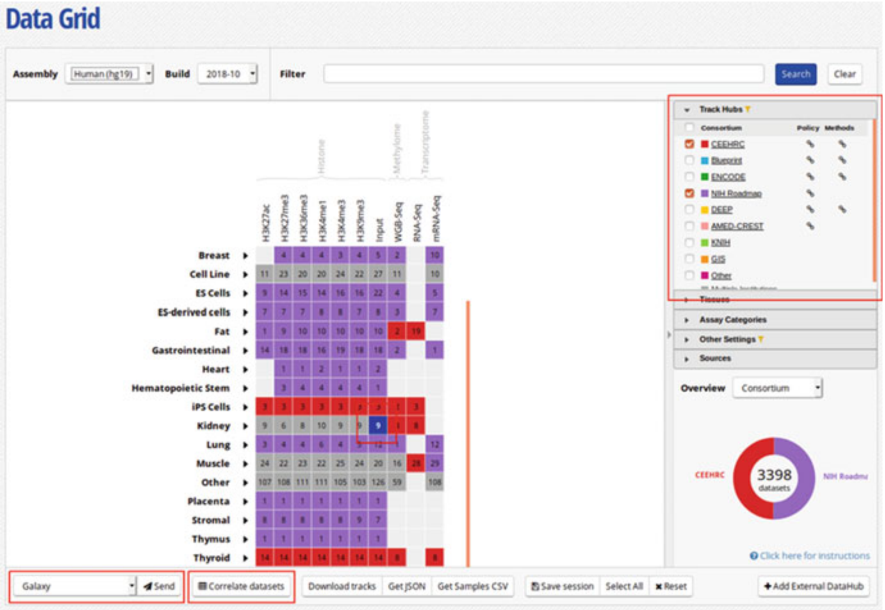
You can use this interface to browse through the tracks you have selected in the portal, and download them.

Assembly: Human (hg19) Build: 2018-10

IHEC Data Portal Build 2018-10 > hg38 > CEEHRC > Brain > Tracks

- http://localhost/tracks/CEEHRC/hg38/66366.CEEHRC.CEMT0021.gDNA.H3K27ac.peak_calls.bigBed
- http://localhost/tracks/CEEHRC/hg38/66365.CEEHRC.CEMT0021.gDNA.H3K27ac.signal_unstranded.bigWig
- http://localhost/tracks/CEEHRC/hg38/66362.CEEHRC.CEMT0021.gDNA.H3K27me3.peak_calls.bigBed
- http://localhost/tracks/CEEHRC/hg38/66361.CEEHRC.CEMT0021.gDNA.H3K27me3.signal_unstranded.bigWig
- http://localhost/tracks/CEEHRC/hg38/66364.CEEHRC.CEMT0021.gDNA.H3K36me3.peak_calls.bigBed
- http://localhost/tracks/CEEHRC/hg38/66363.CEEHRC.CEMT0021.gDNA.H3K36me3.signal_unstranded.bigWig
- http://localhost/tracks/CEEHRC/hg38/66356.CEEHRC.CEMT0021.gDNA.H3K4me1.peak_calls.bigBed
- http://localhost/tracks/CEEHRC/hg38/66355.CEEHRC.CEMT0021.gDNA.H3K4me1.signal_unstranded.bigWig
- http://localhost/tracks/CEEHRC/hg38/66358.CEEHRC.CEMT0021.gDNA.H3K4me3.peak_calls.bigBed
- http://localhost/tracks/CEEHRC/hg38/66357.CEEHRC.CEMT0021.gDNA.H3K4me3.signal_unstranded.bigWig
- http://localhost/tracks/CEEHRC/hg38/66360.CEEHRC.CEMT0021.gDNA.H3K9me3.peak_calls.bigBed
- http://localhost/tracks/CEEHRC/hg38/66359.CEEHRC.CEMT0021.gDNA.H3K9me3.signal_unstranded.bigWig
- http://localhost/tracks/CEEHRC/hg38/66367.CEEHRC.CEMT0021.gDNA.Input.signal_unstranded.bigWig
- http://localhost/tracks/CEEHRC/hg38/66778.CEEHRC.CEMT0021.gDNA.WGB-Seq.methylation_profile.bigWig
- http://localhost/tracks/CEEHRC/hg38/66777.CEEHRC.CEMT0021.gDNA.WGB-Seq.signal_unstranded.bigWig
- http://localhost/tracks/CEEHRC/hg38/66326.CEEHRC.CEMT0021.gDNA.WGS.copy_number_variation.bigWig
- http://localhost/tracks/CEEHRC/hg38/66676.CEEHRC.CEMT0021.tRNA.mRNA-Seq.rpkm_forward.bigWig
- http://localhost/tracks/CEEHRC/hg38/66678.CEEHRC.CEMT0021.tRNA.mRNA-Seq.rpkm_reverse.bigWig
- http://localhost/tracks/CEEHRC/hg38/66677.CEEHRC.CEMT0021.tRNA.mRNA-Seq.signal_forward.bigWig
- http://localhost/tracks/CEEHRC/hg38/66679.CEEHRC.CEMT0021.tRNA.mRNA-Seq.signal_reverse.bigWig
- http://localhost/tracks/CEEHRC/hg38/67924.CEEHRC.CEMT0021.tRNA.smRNA-Seq.isoform_details.bigWig
- http://localhost/tracks/CEEHRC/hg38/67923.CEEHRC.CEMT0021.tRNA.smRNA-Seq.reads_per_million_miRNA_mapped.bigWig
- http://localhost/tracks/CEEHRC/hg38/67922.CEEHRC.CEMT0021.tRNA.smRNA-Seq.signal_unstranded.bigWig
- http://localhost/tracks/CEEHRC/hg38/66379.CEEHRC.CEMT0022.gDNA.H3K27ac.peak_calls.bigBed
- http://localhost/tracks/CEEHRC/hg38/66378.CEEHRC.CEMT0022.gDNA.H3K27ac.signal_unstranded.bigWig

Fig. 4.5 Download URLs may be clicked directly or transferred to secondary software



Data Grid

Assembly: Human (hg19) Build: 2018-10 Filter:

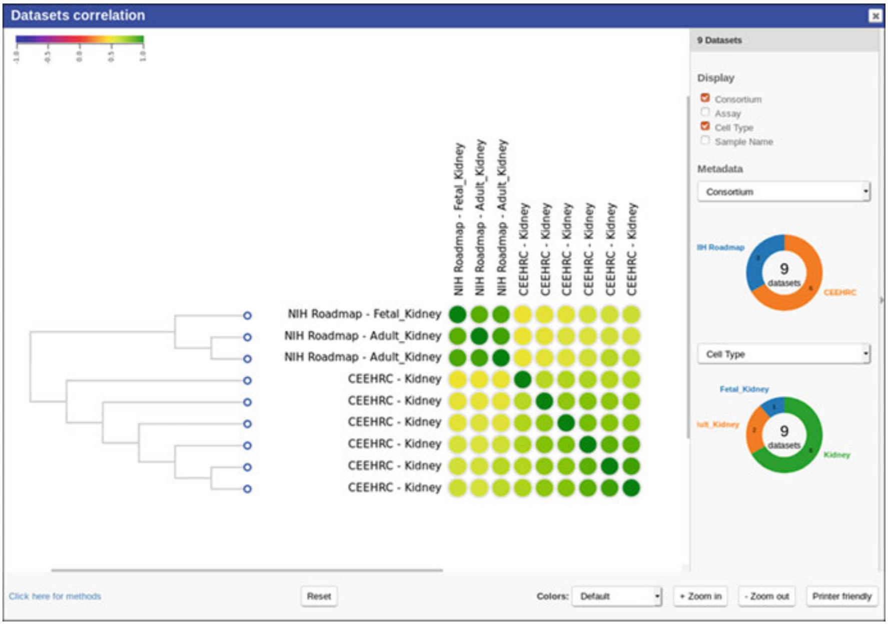
	Histone							Methylome	Transcriptome	
	H3K27ac	H3K27me3	H3K36me3	H3K4me1	H3K4me3	H3K9me3	Input	WGB-Seq	RNA-Seq	mRNA-Seq
Breast ▸		4	4	4	3	4	5	2		10
Cell Line ▸	11	23	20	20	24	22	27	11		10
ES Cells ▸	9	14	15	14	16	16	22	4		5
ES-derived cells ▸	7	7	7	8	8	7	8	3		7
Fat ▸	1	9	10	10	10	10	10	2	19	
Gastrointestinal ▸	14	18	18	16	19	18	18	2		1
Heart ▸		1	1	2	1	1	2			
Hematopoietic Stem ▸		3	4	4	4	4	1			
iPS Cells ▸	3	3	3	3	3	3	3	1	3	
Kidney ▸	9	6	8	10	9	9	9	1	8	
Lung ▸	3	4	4	6	4	3	12	1		12
Muscle ▸	24	22	23	22	25	24	20	16	28	29
Other ▸	107	108	111	111	105	103	126	59		108
Placenta ▸	1	1	1	1	1	1	1			
Stromal ▸	8	8	8	8	8	9	7			
Thymus ▸	1	1	1	1	1	1	1			
Thyroid ▸	14	14	14	14	14	14	14	8		8

▾ Track Hubs

□ Consortium	Policy	Methods
☑ CEEHRC	✎	✎
□ Blueprint	✎	✎
□ ENCODE	✎	✎
☑ NIH Roadmap	✎	
□ DEEP	✎	✎
□ AMED-CREST	✎	
□ KNIH		
□ GIS		
□ Other		

Overview: Consortium

Fig. 4.6 Selection of kidney datasets sourced from two different consortia



Display: ☑ Consortium ☐ Assay ☑ Cell Type ☐ Sample Name

Fig. 4.7 Pearson correlation matrix, with dendrogram

4.6 Summary and Future Development of the Database

The IHEC Data Portal (http://epigenomesportal.ca/ihec/) was initially developed to make the datasets produced by IHEC members more easily discoverable. The tools it offers to query, visualize, download, and analyze epigenomic datasets make the process of identifying and accessing these datasets simpler. With the newly added community hubs, the portal now aims to become a resource to uncover and disseminate other epigenomic datasets that exist beyond IHEC within the rest of the epigenomic research community.

One of the caveats of integrating processed datasets originating from multiple providers is that these providers produce annotation data using different analysis pipelines, tools, and parameters, all of which introduce their own sets of biases. To reduce differences induced by such multi-sites analysis, the IHEC consortium has launched the Epigenome Meta-Analysis Project, an initiative that aims to reprocess the whole IHEC dataset, deposited under controlled access, using a set of consortium-wide accepted bioinformatics analysis pipelines. Through this process, we aim to produce a gold-standard set of quality-controlled, consistently processed, reference epigenomic maps sourced from the different IHEC members. The resulting epigenome will be made available in the IHEC Data Portal, as a special release that will be displayed by default.

Finally, in an effort to increase the shareability of both IHEC and outside research community epigenomic datasets, the EpiShare project has been initiated. The goal of

EpiShare is two-fold. As a driver project for the Global Alliance for Genomics and Health (GA4GH) (Global Alliance for Genomics and Health 2019), the aim is to contribute to the development of widely accepted APIs, standards, and toolkits for storing, querying, and distributing epigenomic data, taking into account data security and ethical considerations. Secondly, the aim is to develop a platform that will enable federated epigenomic data sharing, using the resources developed within GA4GH.

In summary, future versions of the IHEC Data Portal aim to be more accessible, easing the process of storing and distributing epigenomic data sourced from non-IHEC members. The intention is to develop and apply standards facilitating wide-scale, multi-project data analysis.

Funding and Acknowledgements The IHEC Data Portal is hosted by GenAP (http://www.genap.ca) and funded under the CEEHRC, by the CIHR and Genome Quebec, with additional support from Genome Canada. The EpiGeEC correlation matrix computational approach was developed at the Université de Sherbrooke and funded by NSERC. The computing and networking infrastructure, and part of the software development, are provided by Compute Canada and CANARIE.

Many people beyond this book chapter's authors have contributed to the development of the IHEC Data Portal: Michel Barrette, David Anderson de Lima Morais, David Anderson, Catherine Côté, Carol Gauthier, Maxime Caron, Tony Kwan, Kuang Chung Chen, Jonathan Laperle, Alexei Nordell Markovits, Tomi Pastinen, Bryan Caron, Alain Veilleux, Pierre-Étienne Jacques.

References

Bujold D, de Lima Morais DA, Gauthier C, Côté C, Caron M, Kwan T, Chen KC, Laperle J, Markovits AN, Pastinen T, Caron B (2016) The international human epigenome consortium data portal. Cell Syst 3(5):496–499

Cell Ontology (2019). http://cellontology.org

Eilbeck K, Lewis SE, Mungall CJ, Yandell M, Stein L, Durbin R, Ashburner M (2005) The Sequence Ontology: a tool for the unification of genome annotations. Genome Biol 6:R44

Epigenome Reference Registry (2019). https://www.ebi.ac.uk/vg/epirr/

Giardine B, Riemer C, Hardison RC, Burhans R, Elnitski L, Shah P, Zhang Y, Blankenberg D, Albert I, Taylor J, Miller W (2005) Galaxy: a platform for interactive large-scale genome analysis. Genome Res 15(10):1451–1455

Global Alliance for Genomics and Health (2019). https://www.ga4gh.org/

Kent WJ, Zweig AS, Barber G, Hinrichs AS, Karolchik D (2010) BigWig and BigBed: enabling browsing of large distributed datasets. Bioinformatics 26(17):2204–2207

Laperle J, Hébert-Deschamps S, Raby J et al (2019) The epiGenomic Efficient Correlator (epiGeEC) tool allows fast comparison of user datasets with thousands of public epigenomic datasets. Bioinformatics 35(4):674–676. https://doi.org/10.1093/bioinformatics/bty655

Leinonen R, Akhtar R, Birney E, Bower L, Cerdeno-Tárraga A, Cheng Y, Cleland I, Faruque N, Goodgame N, Gibson R, Hoad G (2010) The European nucleotide archive. Nucleic Acids Res 39(Suppl 1):D28–D31

Malone J, Holloway E, Adamusiak T, Kapushesky M, Zheng J, Kolesnikov N, Zhukova A, Brazma A, Parkinson H (2010) Modeling sample variables with an experimental factor ontology. Bioinformatics 26(8):1112–1118

Mungall C, Haendel M, Dahdul W, Ibrahim N, Segerdell E, Blackburn D, Comte A, Niknejad A, Decechi A (2016) Uberon ontology. http://purl.obolibrary.org/obo/uberon/releases/2016-01-26/uberon.owl

NCI Metathesaurus (2019). https://ncim.nci.nih.gov/ncimbrowser/

Stunnenberg HG, Abrignani S, Adams D, de Almeida M, Altucci L, Amin V, Amit I, Antonarakis SE, Aparicio S, Arima T, Arrigoni L (2016) The International Human Epigenome Consortium: a blueprint for scientific collaboration and discovery. Cell 167(5):1145–1149

Tateno Y, Imanishi T, Miyazaki S, Fukami-Kobayashi K, Saitou N, Sugawara H, Gojobori T (2002) DNA Data Bank of Japan (DDBJ) for genome scale research in life science. Nucleic Acids Res 30(1):27–30

The Ontology for Biomedical Investigations (2016) PLoS One 11(4):e0154556. https://doi.org/10.1371/journal.pone.0154556. eCollection 2016

Chapter 5
ChIP-Atlas

Shinya Oki and Tazro Ohta

Abstract In the past decade, large-scale data has been generated with chromatin immunoprecipitation with sequencing (ChIP-seq) technology and is available to every researcher via public domains. Taking full advantage of the large amount of data is challenging, however, due to the need for specialized bioinformatics skills in addition to large computational resources. By assembling and analyzing public ChIP-seq data, ChIP-Atlas (http://chip-atlas.org) was developed as an easy-to-use web service for researchers to visualize genome-wide binding data of transcription factors (TFs) and modified histones. Furthermore, ChIP-Atlas is a unique data-mining suite that provides integrative analysis data for TF–target gene relationships and TF–TF colocalization. It also allows users to perform TF enrichment analysis for given genes and genomic loci. Based on fully integrated public ChIP-seq data, ChIP-Atlas is a useful tool to find novel insights into gene regulatory networks and epigenomics.

Keywords ChIP-seq · DNase-seq · Data mining · Enhancer · Transcription factor

5.1 Preamble

Chromatin immunoprecipitation with sequencing (ChIP-seq) is a powerful method for investigating the genome-wide distribution of transcription factors (TFs) and modified histones (Park 2009). In the past decade, a growing number of researchers have published studies with ChIP-seq technology, which has been established as a gold-standard method in the fields of epigenomics and gene regulatory networks.

S. Oki (✉)
Department of Developmental Biology, Graduate School of Medical Sciences, Kyushu University, Fukuoka, Japan
e-mail: oki.shinya.3w@kyoto-u.ac.jp

T. Ohta
Database Center for Life Science, Joint Support Center for Data Science Research, Research Organization of Information and Systems, Mishima, Shizuoka, Japan

© The Author(s), under exclusive license to Springer Nature Singapore Pte Ltd. 2021
I. Abugessaisa, T. Kasukawa (eds.), *Practical Guide to Life Science Databases*,
https://doi.org/10.1007/978-981-16-5812-9_5

Prior to submission or publication, most academic journals require authors of studies using ChIP-seq to deposit their raw sequence data as sequence read archives (SRAs) in public domains, such as NCBI, DDBJ, and EBI. While these data, in principle, are freely available, actually exploiting the data requires sophisticated skills with command-line processing, such as aligning the sequences to a reference genome and peak-calling in order to define the regions with a statistically significant number of alignments. Furthermore, extensive computational resources are required for data-mining across thousands of public ChIP-seq data to extract novel biological relationships. In order to overcome these challenges, the authors launched the ChIP-Atlas project in 2014 with the following missions:

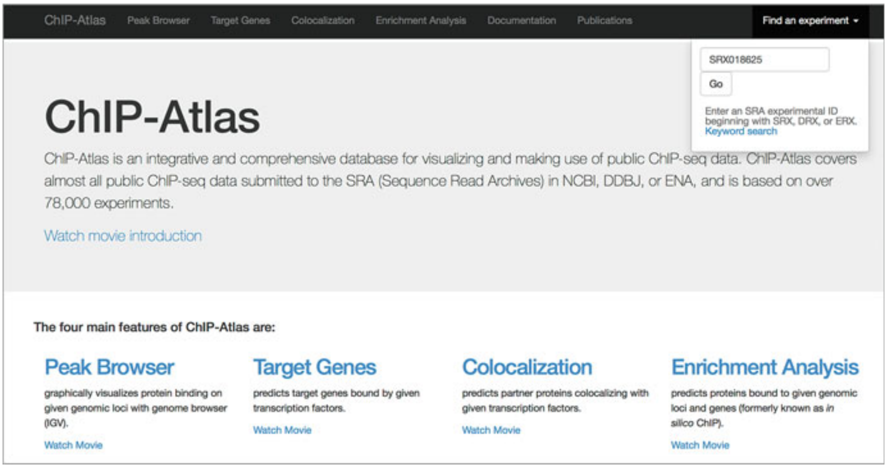
- To process every ChIP-seq SRA for visualizing alignment and peak-call data.
- To perform integrative analyses across entire ChIP-seq datasets for data mining.
- To provide analyzed data through an easy-to-use web service.
- To continuously update the project with the latest ChIP-seq data.

The ChIP-Atlas web service was publicly released in December 2015, and these missions remain. This chapter is a practical guide for readers to make full use of public ChIP-seq data with ChIP-Atlas. Readers interested in the Web tools may directly begin the step-by-step guide in Sect. 5.4, which takes 1–2 h to complete, including a few minutes of hands-on time. For those interested in command-line processing of the data, it is recommended that readers proceed to Sect. 5.5 after learning the data architecture of ChIP-Atlas in Sect. 5.3.

5.2 Database Overview

ChIP-Atlas (Fig. 5.1) is a web server that collects public ChIP-seq and DNase-seq data archived in NCBI SRA (>70,000 experiments as of March 2018). Complete raw sequence data are downloaded from NCBI, aligned to a reference genome, and peak called, which can then be visualized in Integrative Genomics Viewer (IGV, Fig. 5.2; Robinson et al. 2011). ChIP-Atlas also is a web-based data-mining suite powered by integrative analyses of peak-call data. Users are thus able to browse integrative peak-call data derived from thousands of experiments (Fig. 5.3), to view TF–gene and TF–TF interactions, and to perform TF enrichment analysis for given genes and genomic regions. The home page of ChIP-Atlas (Fig. 5.1) is the gateway for these analyses and highlights its four main features, which are described below with examples from tutorial Sect. 5.4:

- **Peak Browser** graphically visualizes protein binding on given genomic loci with genome browser (Sect. 5.4.2).
  - Input: human *KRT19* gene locus
  - Output: enhancer-associated histone modifications (H3K27ac) are apparent near the *KRT19* gene, where multiple TFs are colocalized (Fig. 5.3).



Enter an SRA experimental ID beginning with SRX, DRX, or ERX.

ChIP-Atlas

ChIP-Atlas is an integrative and comprehensive database for visualizing and making use of public ChIP-seq data. ChIP-Atlas covers almost all public ChIP-seq data submitted to the SRA (Sequence Read Archives) in NCBI, DDBJ, or ENA, and is based on over 78,000 experiments.

The four main features of ChIP-Atlas are:

Peak Browser
graphically visualizes protein binding on given genomic loci with genome browser (IGV).

Target Genes
predicts target genes bound by given transcription factors.

Colocalization
predicts partner proteins colocalizing with given transcription factors.

Enrichment Analysis
predicts proteins bound to given genomic loci and genes (formerly known as *in silico ChIP*).

Fig. 5.1 Home page of ChIP-Atlas. The four data-mining tools and their tutorials (movies) can be accessed from the home page of ChIP-Atlas (http://chip-atlas.org). Users can also search for specific data matched with given experimental IDs and keywords

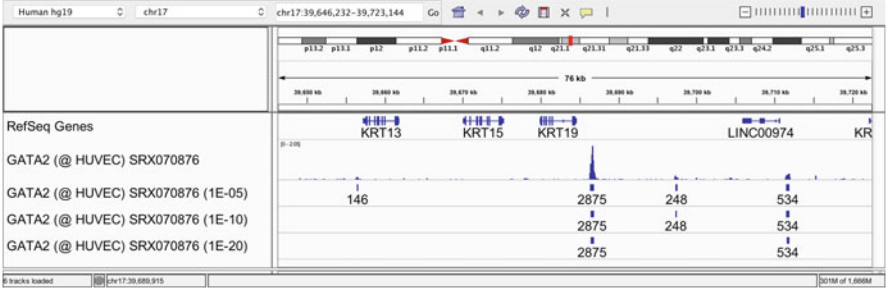

Human hg19
chr17
chr17:39,646,232-39,723,144
Go
p13.2 p13.1 p12 p11.2 p11.1 q11.2 q12 q21.1 q21.31 q21.33 q22 q23.1 q23.3 q24.2 q25.1 q25.3
76 kb
39,650 kb
39,660 kb
39,670 kb
39,680 kb
39,690 kb
39,700 kb
39,710 kb
39,720 kb
RefSeq Genes
KRT13
KRT15
KRT19
LINC00974
KR
GATA2 (@ HUVEC) SRX070876
[0-2.00]
GATA2 (@ HUVEC) SRX070876 (1E-05)
146
2875
248
534
GATA2 (@ HUVEC) SRX070876 (1E-10)
2875
248
534
GATA2 (@ HUVEC) SRX070876 (1E-20)
2875
534
6 tracks loaded
chr17:39,689,915
301M of 1,666M

Fig. 5.2 Browsing data from each experiment. Alignment and peak-call data of GATA2 ChIP-seq data from HUVECs (SRX070876) are shown around the human *KRT19* gene locus on IGV

- **Target Genes** predicts target genes bound by given TFs (Sect. 5.4.3).
  - Input: human GATA2 protein
  - Output: potential target genes of GATA2 such as *ZNF792*, *KRT19*, and *USP54* are shown in Fig. 5.8.

- **Colocalization** predicts partner proteins colocalizing with given TFs (Sect. 5.4.4).
  - Input: human GATA2 protein
  - Output: TFs colocalizing genome-wide with GATA2 such as TAL1, JUNB, and EP300 are listed in Fig. 5.9.

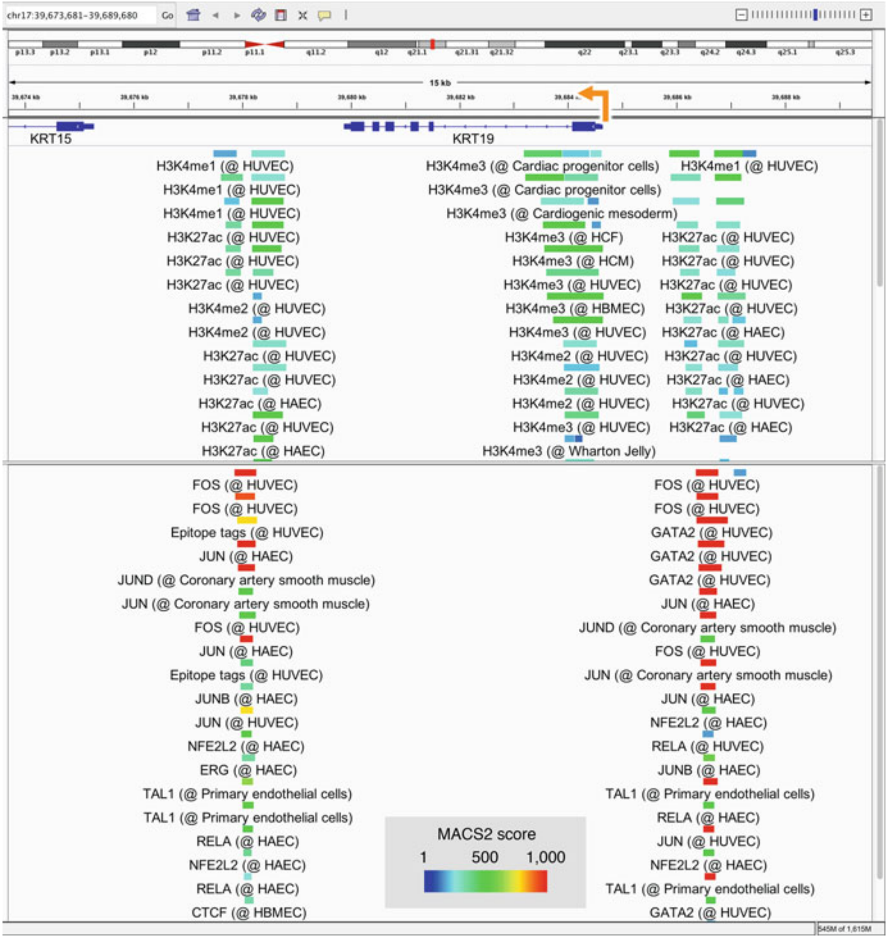
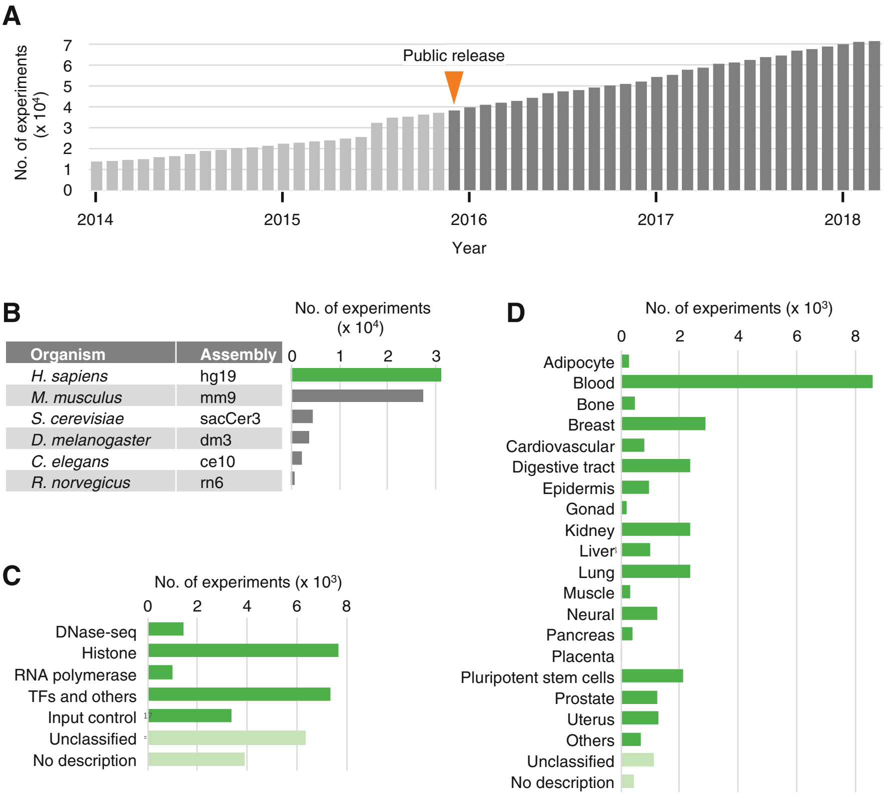


Fig. 5.3 Example view of Peak Browser. Assembled peak-calls (colored bars) for histones (top) and TFs (bottom) for ChIP-seq data from cardiovascular cells are shown around the human *KRT19* gene locus on IGV, with the colors of the bars indicating statistical significance from peak-caller MACS2

- **Enrichment Analysis** predicts proteins with enriched binding to given genomic loci and genes (Sects. 5.4.5 and 5.4.6).
  - Input: hundreds of gene symbols specifically expressed in the liver
  - Output: TFs significantly bound around input gene loci, such as HNF4A/G and FOXA1/2, are shown in Fig. 5.11.

These analyses can be completed without knowledge of command-line processing or bioinformatics analysis. ChIP-Atlas covers ChIP-seq and DNase-seq



Organism	Assembly
H. sapiens	hg19
M. musculus	mm9
S. cerevisiae	sacCer3
D. melanogaster	dm3
C. elegans	ce10
R. norvegicus	rn6

Fig. 5.4 Data content in ChIP-Atlas. (**a**) Cumulative number of SRX-based experiments recorded in ChIP-Atlas are shown, with the light and dark grays indicating the data number before and after the public release of ChIP-Atlas in December 2015, respectively. (**b**–**d**) Numbers of ChIP-seq and DNase-seq experiments recorded in ChIP-Atlas are shown by organism (**b**), by antigen class (**c**), and by cell type class (**d**) of human data (green) as of March 2018

data of six organisms and has been updated monthly since the public release in December 2015 (Fig. 5.4a). Therefore, users can make full use of the latest chromatin-profiling data via an easy-to-use web interface as described in the tutorials in Sect. 5.4. ChIP-Atlas has been used by researchers worldwide in order to make predictions or to design strategies either before or after performing biological experiments (see citation list in http://chip-atlas.org/publications). Furthermore, data generated by ChIP-Atlas are assigned a unique URL and are publicly available and thus can be used for bioinformatics analyses and connections with other biodatabases as shown in Sect. 5.5.

5.3 Content and Architecture of the Database

ChIP-Atlas has been updated monthly via a computational pipeline implemented in the NIG supercomputer system (https://sc2.ddbj.nig.ac.jp/index.php/en/) concurrent with the monthly update of NCBI SRA. Sample metadata and biosample data of all experiments are downloaded from NCBI FTP sites (ftp://ftp.ncbi.nlm.nih.gov/sra/reports/Metadata and ftp://ftp.ncbi.nlm.nih.gov/biosample) in order to extract the experimental ID of updated ChIP-seq and DNA-seq data for the six organisms as shown in Fig. 5.4b. Raw sequence data corresponding to the IDs are automatically downloaded and aligned to a reference genome with bowtie2 (Langmead and Salzberg 2012), before generating browsable alignment data in BigWig format and peak-call data in BED format with the peak-caller MACS2 (Zhang et al. 2008). The NCBI biosample data also includes sample metadata for each experiment, such as antigen and cell type names, written by original data submitters. This text, however, is not standardized; thus, the ChIP-Atlas team manually revises the sample metadata into a standard language according to defined rules, which are shown in the ChIP-Atlas documentation page (https://github.com/inutano/chip-atlas/wiki#data_annotation_doc). Briefly, the names of TFs are mapped to official gene symbols, and the names of cell lines are revised according to a publication that proposed a unified nomenclature (Yu et al. 2015). After completing the curation, the dataset for each experiment includes unique peak-call data and curated sample metadata, which proceeds through a computational pipeline to prepare data files for the following four data-mining tools of ChIP-Atlas.

- **Peak Browser**: All peak-call data in BED format are concatenated into a single file and modified with sample metadata and color values for visualization in the genome browser IGV.
- **Target Genes**: All peak-call data in BED format is evaluated for overlap around the transcription start site (TSS) of every RefSeq coding gene.
- **Colocalization**: Pair-wise comparisons of all peak-call data are performed with the CoLo algorithm (https://github.com/RyoNakaki/CoLo), and similarity scores are computed for all combinations.
- **Enrichment Analysis**: All peak-call data in BED format are concatenated and converted to library files in a format suitable for quickly assessing the overlaps with users' submitted data.

Data collection methods and processing strategies are described on the ChIP-Atlas documentation page (https://github.com/inutano/chip-atlas/wiki).

5.4 Use Cases and Demonstration of the Database

ChIP-Atlas was designed to be used on macOS, Windows, and Linux machines with web browsers such as Safari, Edge, FireFox, and Chrome. The only requisite third-party tool is the genome browser IGV (Robinson et al. 2011) to see the alignment and peak-call data recorded in ChIP-Atlas. Instructions on how to install IGV are described in the download page (https://software.broadinstitute.org/software/igv/download). ChIP-Atlas can be used in the following two modes:

- **Each data mode** to thoroughly understand a single or a small number of experiments (Sect. 5.4.1)
- **data-mining mode** to access integrative analysis data across thousands of experiments (Sects. 5.4.2–5.4.6).

5.4.1 Each Data Mode

In NCBI SRA, data with ChIP-seq or DNA-seq technology for each experiment is assigned an ID with a prefix of SRX, DRX, or ERX (hereafter collectively referred to as SRXs), which are also adopted in ChIP-Atlas for unified management of the records. This tutorial illustrates how to obtain an SRX ID and how to visualize alignment and peak-call data.

1. Click on “Find an experiment” on the top right of the ChIP-Atlas home page (Fig. 5.1).
2. If you know an SRX ID for a ChIP-seq experiment of interest, enter the ID (e.g. SRX070876). Otherwise, SRX IDs can be found via the accession number of the project described in papers. For instance, if you are interested in GATA2 ChIP-seq data in human umbilical vein endothelial cells (HUVECs; Linnemann et al. 2011), find a Gene Expression Omnibus (GEO) accession number “GSE29531” described in the paper; enter the number on the GEO search page (https://www.ncbi.nlm.nih.gov/geo/); find and click on “GSM730701” for the “GATA2_ChIP-seq_A” sample, and then you will find “SRX070876” as the experimental ID assigned in NCBI SRA (Fig. 5.5a). ChIP-Atlas does allow a keyword search for specific SRXs under “Find an Experiment” on the ChIP-Atlas home page: click on “Keyword search” (Fig. 5.1, top right); in the new window, you can enter arbitrary keywords in the search box such as “GATA2 HUVEC”, and find and click on “SRX070876” (Fig. 5.5b).
3. In the new window, you can see detailed information for SRX070876 (Fig. 5.6), including curated antigen and cell type names; descriptions by the original data submitter; and read processing logs such as read number, mapability, and read quality. You will find four blue buttons on the top of the page, which allows you to visualize or download the alignment and peak-call data (“View on IGV” or

"Download," respectively), to see the analyzed data ("View Analysis"), and to open the external pages showing details for the experiment ("Link Out").

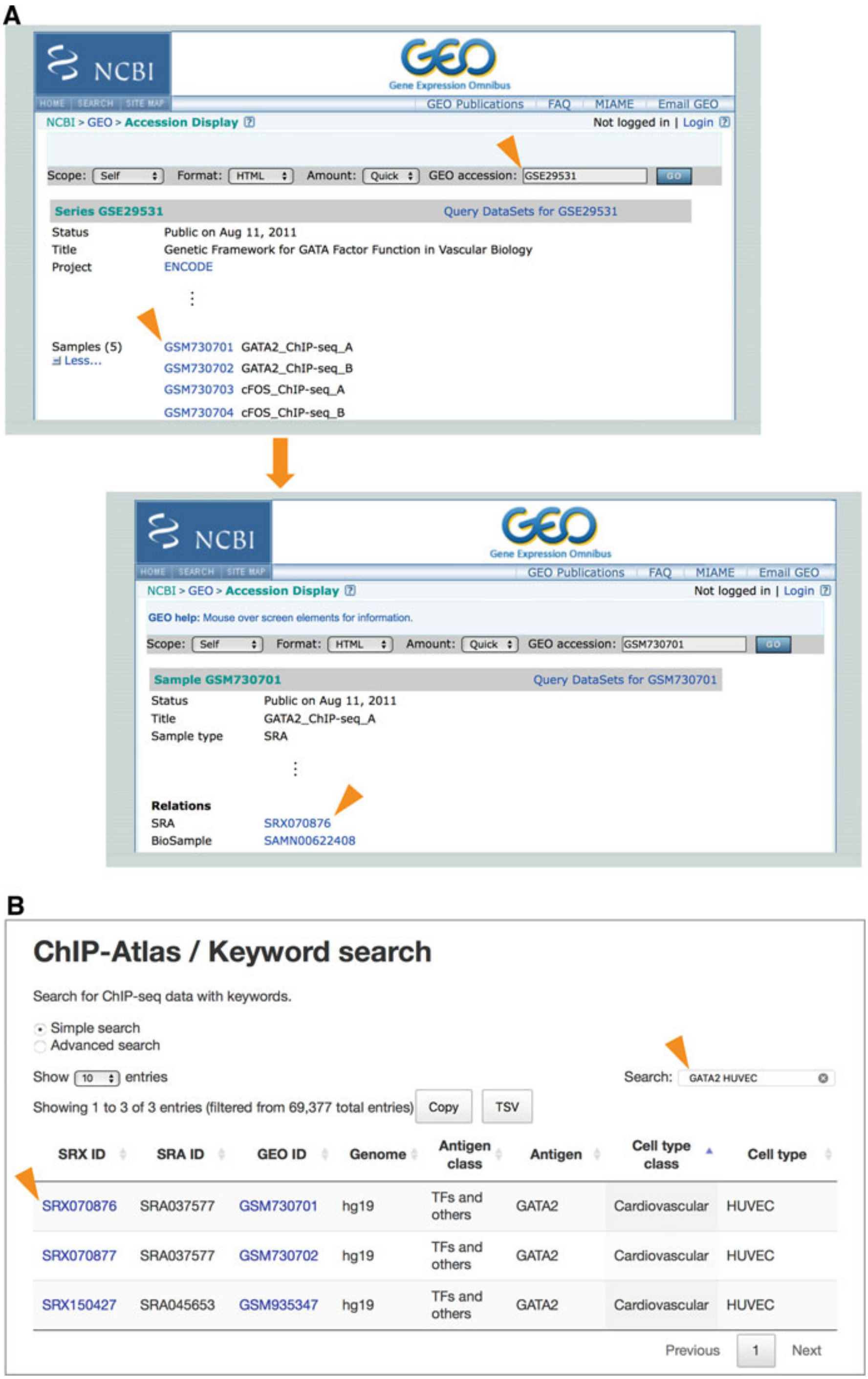
4. After IGV has been launched on your computer and an IGV window has appeared, click on "BigWig" under the "View on IGV" drop-down menu. The alignment data of SRX070876 will appear on IGV. Similarly, clicking on "Peak-call (q < 1E–10)" shows peak-call data on IGV with a MACS2 Q-value less than 1E–10.
5. Enter the gene name of interest in IGV. For example, entering "*KRT19*" and zooming-out will result in the view shown in Fig. 5.2.

Tips: The y-axis of the alignment data indicates the unit RPM (reads per million mapped reads); in other words, the number of mapped reads at a given position was normalized against the total mapped reads. This is useful to compare multiple alignment tracks using a common scale. The values below the peak-call bars indicate statistical significance calculated by the peak-caller MACS2 (–10Log$_{10}$ [Q-value]; hereafter referred to as MACS2 scores). Thus, clicking on "Peak-call (q < 1E−10)" on the web page shows peaks with MACS2 scores greater than 100.

5.4.2 *Peak Browser*

Peak-call data recorded in ChIP-Atlas are fully integrated and can be visualized on IGV with the Peak Browser function. IGV allows the user to graphically determine the binding profile of multiple TFs and histone modifications around a gene or locus of interest. This tutorial illustrates how to visualize protein bindings around the *KRT19* gene in human cardiovascular cells.

1. Click on "Peak Browser" on the ChIP-Atlas home page (Figs. 5.1 and 5.7).
2. Click on the organism tab "H. sapiens" as shown on the top left.
3. Choose "TFs and others" in the "Antigen Class" list.
4. Choose "Cardiovascular" in the "Cell type Class" list.
5. Choose MACS2 score as "100" in the "Threshold for Significance" list.
6. Click on the "View on IGV" button after confirming that IGV has been launched and an IGV window has appeared on your computer. IGV will load the ChIP-seq peak-call data for TFs and other transcriptional regulators in cardiovascular cells with a MACS2 score greater than 100.
7. Go back to the Peak Browser web page and choose "Histone" in the "Antigen Class" list.
8. Click on the "View on IGV" button. IGV will load the ChIP-seq peak-call data for histones in cardiovascular cells with a MACS2 score greater than 100.
9. Enter "KRT19" in the textbox on the top of the IGV window. Right-click on the track names and choose "Expanded" to expand the view.
10. By zooming-out and rearranging the tracks, you will see the view as shown in Fig. 5.3, where the transcription start site (TSS) of *KRT19* is occupied by histone marks for an active promoter (H3K4me3) in HUVECs and other cardiovascular



Scope: Self Format: HTML Amount: Quick GEO accession: GSE29531

Series GSE29531 Query DataSets for GSE29531

Status	Public on Aug 11, 2011
Title	Genetic Framework for GATA Factor Function in Vascular Biology
Project	ENCODE
⋮	
Samples (5) Less...	GSM730701 GATA2_ChIP-seq_A GSM730702 GATA2_ChIP-seq_B GSM730703 cFOS_ChIP-seq_A GSM730704 cFOS_ChIP-seq_B

GEO help: Mouse over screen elements for information.

Scope: Self Format: HTML Amount: Quick GEO accession: GSM730701

Sample GSM730701 Query DataSets for GSM730701

Status	Public on Aug 11, 2011
Title	GATA2_ChIP-seq_A
Sample type	SRA
⋮	
Relations	
SRA	SRX070876
BioSample	SAMN00622408

ChIP-Atlas / Keyword search

Search for ChIP-seq data with keywords.

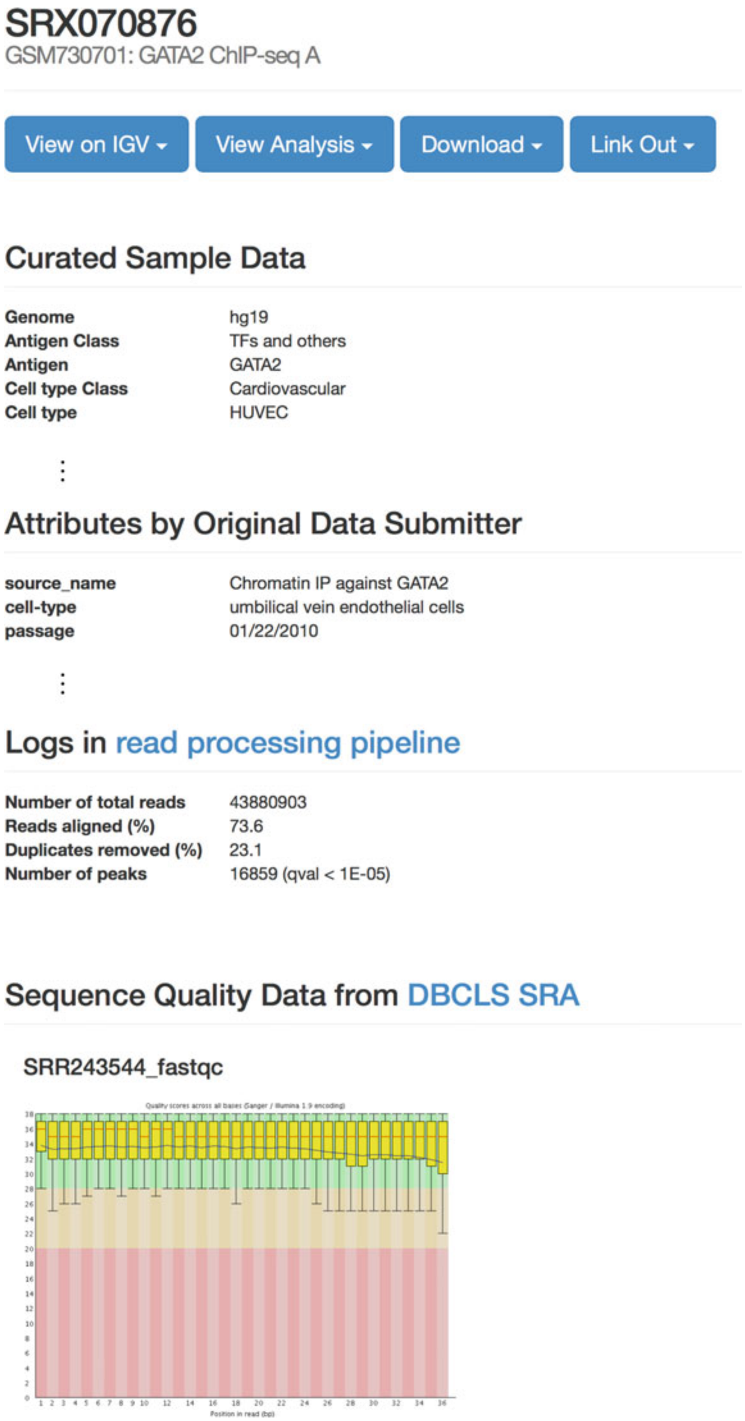
● Simple search
○ Advanced search

Show 10 entries Search: GATA2 HUVEC

Showing 1 to 3 of 3 entries (filtered from 69,377 total entries)

SRX ID	SRA ID	GEO ID	Genome	Antigen class	Antigen	Cell type class	Cell type
SRX070876	SRA037577	GSM730701	hg19	TFs and others	GATA2	Cardiovascular	HUVEC
SRX070877	SRA037577	GSM730702	hg19	TFs and others	GATA2	Cardiovascular	HUVEC
SRX150427	SRA045653	GSM935347	hg19	TFs and others	GATA2	Cardiovascular	HUVEC

Fig. 5.5 Obtaining SRX ID. A specific SRX ID is available from GEO (**a**) or the keyword search page of ChIP-Atlas (**b**), with the arrowheads indicating key values for searching



SRX070876

GSM730701: GATA2 ChIP-seq A

Curated Sample Data

Genome	hg19
Antigen Class	TFs and others
Antigen	GATA2
Cell type Class	Cardiovascular
Cell type	HUVEC

⋮

Attributes by Original Data Submitter

source_name	Chromatin IP against GATA2
cell-type	umbilical vein endothelial cells
passage	01/22/2010

⋮

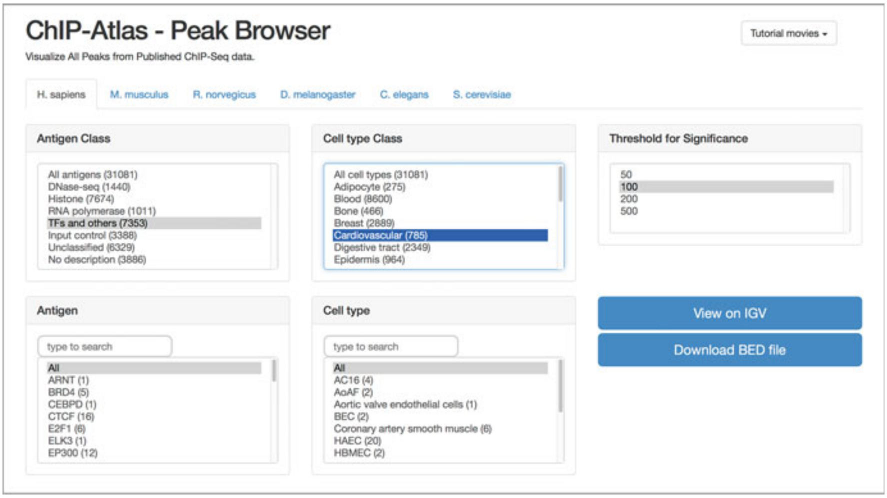
Logs in read processing pipeline

Number of total reads	43880903
Reads aligned (%)	73.6
Duplicates removed (%)	23.1
Number of peaks	16859 (qval < 1E-05)

Sequence Quality Data from DBCLS SRA

SRR243544_fastqc

Fig. 5.6 Example of a metadata page from an experiment. A web page for GATA2 ChIP-seq data from HUVECs (SRX070876) is shown with curated or original sample metadata, processing logs, and buttons for browsing or downloading the analyzed data



ChIP-Atlas - Peak Browser

Visualize All Peaks from Published ChIP-Seq data.

Antigen Class

- All antigens (31081)
- DNase-seq (1440)
- Histone (7674)
- RNA polymerase (1011)
- TFs and others (7353)
- Input control (3388)
- Unclassified (6329)
- No description (3886)

Cell type Class

- All cell types (31081)
- Adipocyte (275)
- Blood (8600)
- Bone (466)
- Breast (2889)
- Cardiovascular (785)
- Digestive tract (2349)
- Epidermis (964)

Threshold for Significance

- 50
- 100
- 200
- 500

Antigen

type to search

- All
- ARNT (1)
- BRD4 (5)
- CEBPD (1)
- CTCF (16)
- E2F1 (6)
- ELK3 (1)
- EP300 (12)

Cell type

type to search

- All
- AC16 (4)
- AoAF (2)
- Aortic valve endothelial cells (1)
- BEC (2)
- Coronary artery smooth muscle (6)
- HAEC (20)
- HBMEC (2)

Fig. 5.7 Settings for Peak Browser function. A Peak Browser web page is shown with the settings to load TF peaks in cardiovascular cells in IGV

cells. In addition, both sides of the *KRT19* gene body are bound by multiple TFs such as FOS, GATA2, and JUN, which are flanked by enhancer-associated histone marks (H3K27ac). This suggests that these two regions have enhancer activity for *KRT19* and/or the neighboring *KRT15* gene in HUVECs.

Tips: The names of antigens and cell types used are shown under the bars representing peak regions, with the color indicating MACS2 scores ranging from 1 (blue), 500 (green), to 1000 (red). These bars shown in IGV are interactive with the following mouse actions: placing the cursor over each bar opens a yellow window showing detailed sample information; left-clicking on the bars opens a new web browser window as shown in Fig. 5.6; and right-clicking on the bars shows a pop-up menu for copying the sample attributes and sequences of the intervals. A useful feature of IGV is the "File > Save Session" menu, which can save the current session as an .xml file. Opening the file via the "File > Open Session" menu will reload the tracks and settings even if IGV is relaunched on a different day or on another machine. To learn more about the Peak Browser function, see the tutorial movie for Peak Browser on the ChIP-Atlas home page (Fig. 5.1).

5.4.3 Target Genes

ChIP-Atlas algorithms examine TF peaks of each SRX for whether they are located around every TSS of RefSeq coding gene (within TSS ± 1, 5, or 10 kb). In the former Sect. 5.4.2, the *KRT19* gene was predicted to be a direct target of GATA2 in HUVECs. To learn more about the target genes of GATA2 (or other TFs of interest), this tutorial demonstrates the “Target Genes” function of ChIP-Atlas, which returns potential target genes of a query TF.

1. Click on “Target Genes” on the ChIP-Atlas home page.
2. Click on the organism tab “H. sapiens” as shown on the top left.
3. Choose “GATA2” as the query TF.
4. Specify the distance from TSS as “±5k.”
5. Clicking on “View Potential Target Genes” will navigate to a result page showing the genes with the TSS ± 5 kb regions bound by GATA2. The potential target genes were sorted by MACS2 score averaged over all the GATA2 ChIP-seq data ($n = 35$). You will find that *WDR74*, *NAA38*, and *TMEM88* are ranked in the top three.
6. To sort GATA2 ChIP-seq data in HUVECs, click on the closed triangle under “SRX070876: HUVEC.” The page will be refreshed as shown in Fig. 5.8, where *ZNF792*, *KRT19*, and *USP54* are ranked as the top three target genes of GATA2 in HUVECs. Click on “SRX070876” and try to check GATA2 binding around the TSS of these three genes in IGV as instructed in the previous Sect. 5.4.1.

Tips: The color of the cells in the matrix indicates the MACS2 score, which ranges from 1 (blue), 500 (green) to 1000 (red), of the peaks located around the TSS within the specified width. The raw values are available by downloading a tab-delimited text file from the “Download: TSV (text)” link on the top of the web page (not shown in Fig. 5.8). To learn more about the Target Gene function, see the tutorial movie for Target Genes on the home page of ChIP-Atlas (Fig. 5.1).

5.4.4 Colocalization

In Sect. 5.4.2, we saw that the upstream region of *KRT19* binds GATA2 and other TFs in HUVECs, suggesting that these TFs may form a complex to colocalize in other loci. Protein–protein interactions (PPI) are often detected by labor-intensive methods such as co-immunoprecipitation experiments and mass spectrometry; however, the “colocalization” function of ChIP-Atlas is able to search for colocalizing proteins with a query TF based on the similarity of genome-wide binding profiles. In ChIP-Atlas, pair-wise comparisons of all TF ChIP-seq data are performed with the CoLo algorithm (https://github.com/RyoNakaki/CoLo), and the similarity scores are pre-computed for all combinations. This tutorial demonstrates how to explore TFs colocalizing with GATA2 in the cells relevant to the cardiovascular system.

1. Click on “Colocalization” on the ChIP-Atlas home page (Fig. 5.1).
2. Specify the organism as “H. sapiens” shown on the top left.
3. Select the “Antigens → Cell Type” radio button to choose a query TF before cell type.
4. Choose “GATA2” from the “Antigen” list.
5. Choose “Cardiovascular” from the “Cell Type Class” list.

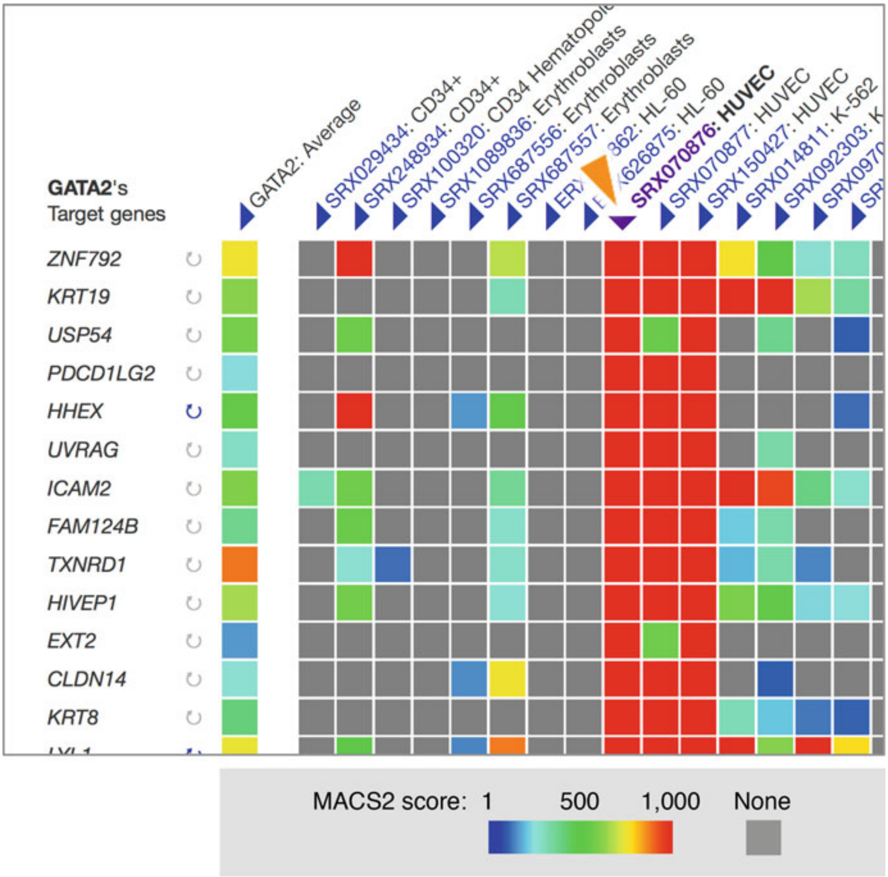
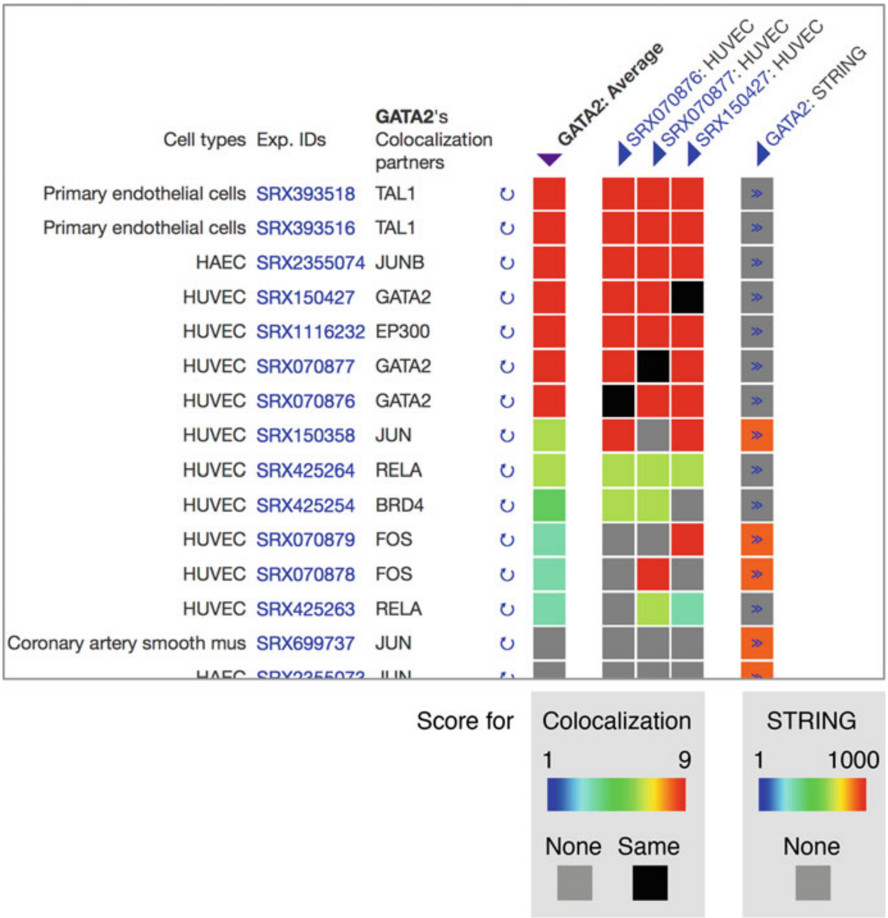


Fig. 5.8 Example of the Target Genes function. Potential target genes of GATA2 in HUVECs are shown on the left, with the cell colors indicating the MACS2 scores of the peaks from GATA2 ChIP-seq data (columns) located within TSS ± 5 kb of RefSeq coding genes (rows). Arrowhead indicates a triangle to sort GATA2 ChIP-seq data in HUVECs (SRX070876)

6. Click on "View Colocalization Data" to navigate to a result page (Fig. 5.9), which shows the ChIP-seq data (rows) with highly similar binding patterns to GATA2 in HUVECs (columns show experiments with query TF). These data suggest that GATA2 exhibits significant genome-wide colocalization with TAL1, JUNB, EP300, RELA, and BRD4 in cardiovascular cell types as well as with JUN and FOS, the known PPI partner with GATA2 according to a PPI database, STRING (shown in the right-most column).

Tips: The color of the cell indicates the scores for colocalization, ranging from low (blue), middle (green), to high (red), calculated using the CoLo algorithm (https://github.com/RyoNakaki/CoLo). Conversely, the right-most column shows the PPI score according to the STRING database. The raw values are available by



Cell types	Exp. IDs	GATA2's Colocalization partners		GATA2: Average	SRX070876: HUVEC	SRX070877: HUVEC	SRX150427: HUVEC	GATA2: STRING
Primary endothelial cells	SRX393518	TAL1	↻	🟥	🟥	🟥	🟥	»
Primary endothelial cells	SRX393516	TAL1	↻	🟥	🟥	🟥	🟥	»
HAEC	SRX2355074	JUNB	↻	🟥	🟥	🟥	🟥	»
HUVEC	SRX150427	GATA2	↻	🟥	🟥	🟥	⬛	»
HUVEC	SRX1116232	EP300	↻	🟥	🟥	🟥	🟥	»
HUVEC	SRX070877	GATA2	↻	🟥	🟥	⬛	🟥	»
HUVEC	SRX070876	GATA2	↻	🟥	⬛	🟥	🟥	»
HUVEC	SRX150358	JUN	↻	🟩	🟥	▨	🟥	🟧»
HUVEC	SRX425264	RELA	↻	🟩	🟩	🟩	🟩	»
HUVEC	SRX425254	BRD4	↻	🟩	🟩	🟩	▨	»
HUVEC	SRX070879	FOS	↻	🟩	▨	▨	🟥	🟧»
HUVEC	SRX070878	FOS	↻	🟩	▨	🟥	▨	🟧»
HUVEC	SRX425263	RELA	↻	🟩	▨	🟩	🟩	»
Coronary artery smooth mus	SRX699737	JUN	↻	▨	▨	▨	▨	🟧»
HAEC	SRX2355073	JUN	↻	▨	▨	▨	▨	🟧

Fig. 5.9 Example of the Colocalization function. The similarity of GATA2 ChIP-seq data from HUVECs (three columns on the middle) versus ChIP-seq data such as TAL1, JUNB, and EP300 in cardiovascular cells (rows) is shown with the cell colors indicating the colocalization scores. Note that the right-most column shows PPI scores by the STRING database

downloading a tab-delimited text file from the "Download: TSV (text)" link at the top of the web page (not shown in Fig. 5.9). To learn more about the Colocalization function, see the tutorial movie for Colocalization on the home page of ChIP-Atlas (Fig. 5.1).

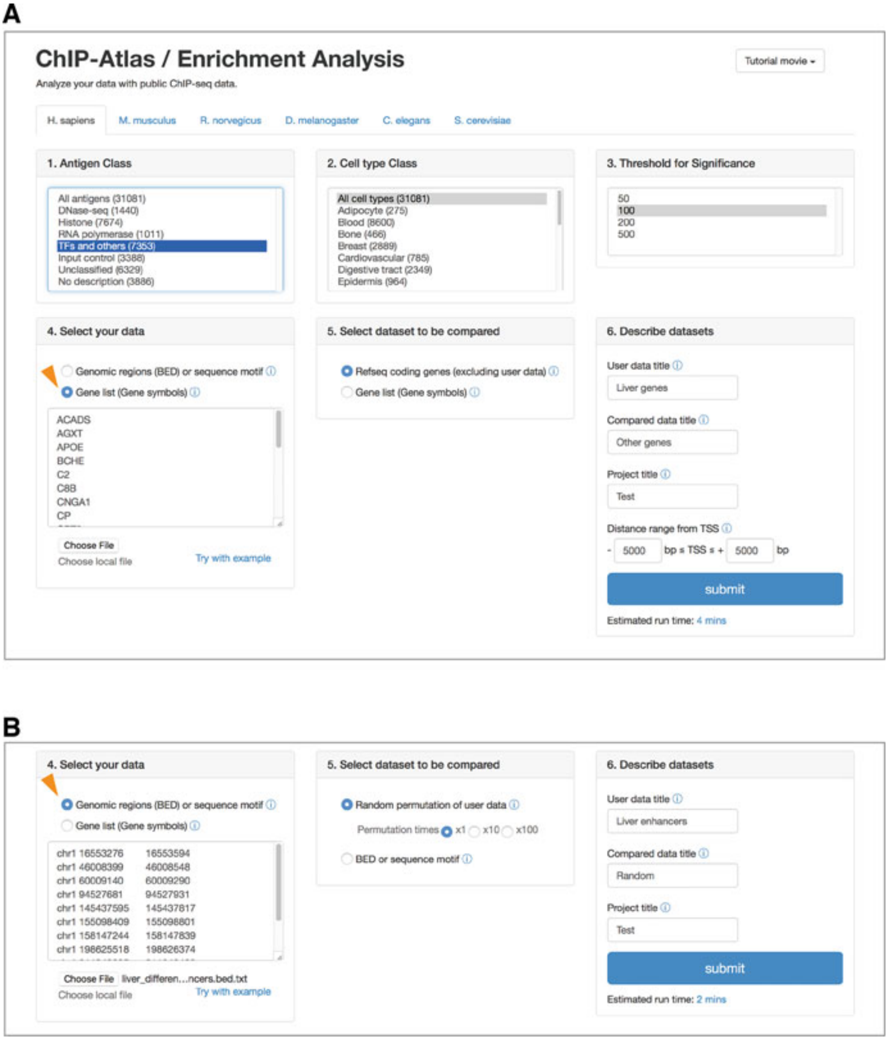
5.4.5 Enrichment Analysis Using a Gene Set as a Query

The Enrichment Analysis function of ChIP-Atlas is a unique tool to search for TFs and histone modifications with enriched binding around a given gene set of interest. Transcriptome analysis for given cell types often yields hundreds of genes upregulated in one cell type. To determine whether the TFs collectively regulate the upregulated genes, motif enrichment analysis is generally performed to search for TF binding motifs significantly enriched around the loci of the gene set [e.g. GATHER (http://changlab.uth.tmc.edu/gather/gather.py; Chang and Nevins 2006) and GSEA (http://www.gsea-msigdb.org; Mootha et al. 2003; Subramanian et al. 2005)]. However, TFs do not necessarily bind to the motif loci, and the binding pattern is context-dependent, and thus different by cell types and tissues. The Enrichment Analysis tool of ChIP-Atlas, in contrast, depends on real binding data, or ChIP-seq data, labeled with cell type and tissue information. Furthermore, one can analyze the enrichment of histone modifications and DNase hypersensitivity sites in specific cell types using the Enrichment Analysis tool. This tutorial demonstrates how to search for TFs enriched around genes specifically expressed in the liver. To conduct this search using the Enrichment Analysis tool, however, you have to learn how to obtain the liver-specific gene set from another database (RefEx; see Chap. 5; Ono et al. 2017) as shown in steps 1–4 below. You may also skip to step 5 and use the Enrichment Analysis tool with an example gene set.

1. In order to obtain genes specifically expressed in the liver, go to the RefEx home page (http://refex.dbcls.jp/index.php?lang=en).
2. Hover over the image of a liver and click on “liver/hepato.”
3. Genes specifically expressed in the liver and hepatocytes are shown. Click on “Downloads” button on the top right of the page.
4. Open the downloaded tab-delimited text file with Microsoft Excel or another spreadsheet application. The second column shows RefSeq IDs of the liver-specific genes, which must be converted to official gene symbols before using the Enrichment Analysis tool: (1) Copy the RefSeq IDs and navigate to the web page “Hyperlink Management System: ID Converter System” (http://biodb.jp/#ids; Imanishi and Nakaoka 2009); (2) paste the RefSeq IDs into the text area; (3) choose “RefSeq ID” from the “Source ID list” and “HUGO gene symbol” from the “Convert to” list before clicking on the “Search” button; (4) download the converted gene symbols by clicking on the “Download” button; (5) open the downloaded tab-delimited text file with Microsoft Excel or another spreadsheet application; the second column will include the symbols of the liver-specific genes, which can be submitted to the Enrichment Analysis tool.
5. Click on “Enrichment Analysis” on the ChIP-Atlas home page (Figs. 5.1 and 5.10a).
6. Choose “TFs and others” for “Antigen Class.”
7. Choose “All cell types” for “Cell type Class.”
8. Choose “100” for “Threshold for Significance.”

9. Select the “Gene list” radio button on the “Select your data” panel, copy the symbols of the liver-specific gene symbols prepared in step 4 above, and paste them into the text area. Alternatively, click on “Try with example,” which will load potential target genes of POU5F1 in the text area.
10. Select the “Refseq coding genes” radio button on the “Select dataset to be compared” panel.
11. Specify the “Distance range from TSS” as TSS ± “5000” bp, and enter arbitrary titles for the data and project.
12. Click on the “Submit” button to transport the gene list to the NIG supercomputer system before performing enrichment analysis. The status of the job will be shown in the following order: “Requesting” (job being entered to NIG supercomputer); “queued” (job in standby queue); “running” (job being executed); “finished” (job has been completed).
13. While the results are processing, it is recommended to note the URL for the result, so that you can revisit the result page again in the future.
14. If the job status is “finished,” click on the result URL to open a table showing ChIP-seq data with enriched binding around liver-specific genes (Fig. 5.11). The first row of the table, for instance, indicates that HNF4A ChIP-seq data for liver-derived Hep G2 cells (SRX100505) has 21,259 peaks, of which 114 peaks overlapped with the TSS ± 5000 bp coordinates of the submitted liver-specific genes ($n = 210$) and 3456 peaks overlapped with those of other RefSeq coding genes ($n = 18{,}335$). This enrichment yielded a P value of $1 \times 10^{-29.2}$ (Fisher’s exact probability test), Q-value of $1 \times 10^{-25.3}$ (Benjamini and Hochberg method), and fold enrichment of 2.88 (=the ratio of 114/210 to 3456/18335). The table is sorted according to P value, which shows significant enrichment of known master regulators of liver development, HNF4A/G and FOXA1/2, that directly reprogram skin fibroblasts into hepatocyte-like cells (Sekiya and Suzuki 2011). Remarkably, the cell type of the top 24 hits are all Hep G2 cells (“Cell” column), even though the number of experiments in this class is relatively small for human data (Fig. 5.4); this suggests that the analysis has high detection power. This example suggests that the Enrichment Analysis feature of ChIP-Atlas can identify master regulators that collectively organize a gene set of interest.
15. Use the search box on the top right to quickly filter with given keywords such as certain TFs, cell type names, or “TRUE” or “FALSE,” which indicates whether the fold enrichment (FE) is greater or less than 1, respectively (see right-most column).

Tips: While RefSeq coding genes, excluding liver-specific genes, were used for comparison in step 10 of this tutorial, other gene lists are also acceptable as comparison data by selecting the “Gene list” radio button in “Select dataset to be compared” (Fig. 5.10a). Thus, if you have a gene list originating from a transcriptome analysis, you can perform the following tests: upregulated genes vs. other RefSeq genes and upregulated genes vs. downregulated genes. The estimated runtime will be shown below the “Submit” button, which is calculated



ChIP-Atlas / Enrichment Analysis

Analyze your data with public ChIP-seq data.

1. Antigen Class
All antigens (31081)
DNase-seq (1440)
Histone (7674)
RNA polymerase (1011)
TFs and others (7353)
Input control (3388)
Unclassified (6329)
No description (3886)

2. Cell type Class
All cell types (31081)
Adipocyte (275)
Blood (8600)
Bone (466)
Breast (2889)
Cardiovascular (785)
Digestive tract (2349)
Epidermis (964)

3. Threshold for Significance
50
100
200
500

4. Select your data
○ Genomic regions (BED) or sequence motif
● Gene list (Gene symbols)
ACADS
AGXT
APOE
BCHE
C2
C8B
CNGA1
CP
Choose File
Choose local file
Try with example

5. Select dataset to be compared
● Refseq coding genes (excluding user data)
○ Gene list (Gene symbols)

6. Describe datasets
User data title: Liver genes
Compared data title: Other genes
Project title: Test
Distance range from TSS: - 5000 bp ≤ TSS ≤ + 5000 bp
submit
Estimated run time: 4 mins

4. Select your data
● Genomic regions (BED) or sequence motif
○ Gene list (Gene symbols)

```
chr1 16553276  16553594
chr1 46008399  46008548
chr1 60009140  60009290
chr1 94527681  94527931
chr1 145437595 145437817
chr1 155098409 155098801
chr1 158147244 158147839
chr1 198625518 198626374
```

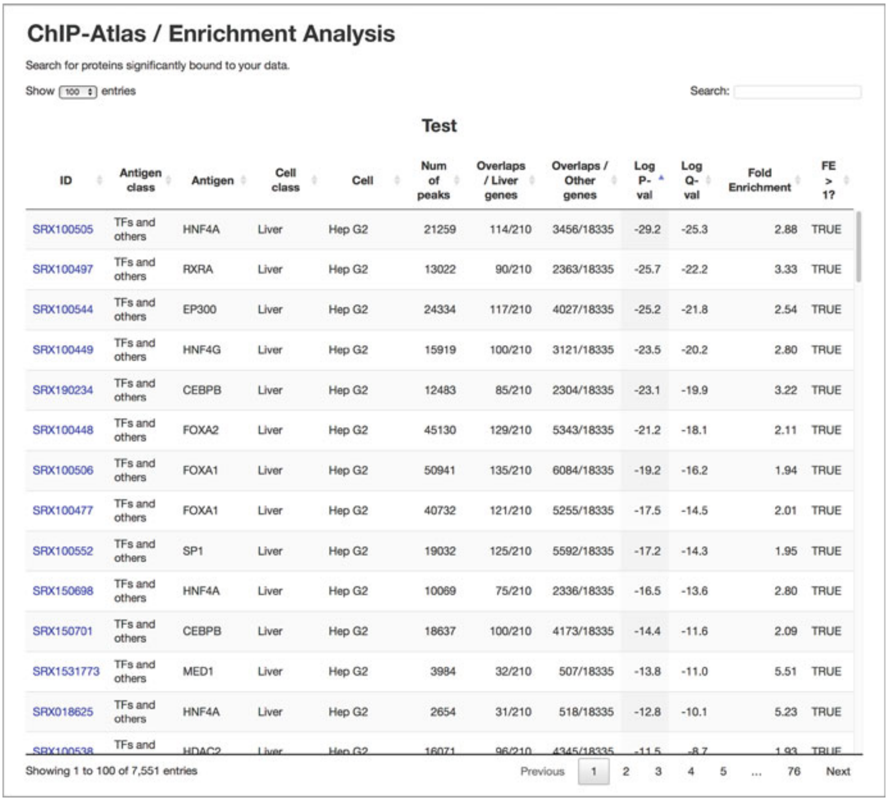
Choose File liver_differen…ncers.bed.txt
Choose local file
Try with example

5. Select dataset to be compared
● Random permutation of user data
Permutation times ● x1 ○ x10 ○ x100
○ BED or sequence motif

6. Describe datasets
User data title: Liver enhancers
Compared data title: Random
Project title: Test
submit
Estimated run time: 2 mins

Fig. 5.10 Settings for Enrichment Analysis function. An Enrichment Analysis web page is shown with the settings to submit genes specifically expressed in the liver (**a**) or enhancer loci specifically activated in the liver (**b**)

with an empirical multivariable function of the amount of submitted data and ChIP-Atlas data to be analyzed. The runtime will be shorter when the submitted data is smaller and ChIP-Atlas data is filtered with antigen and cell type classes. Submitted gene names must be official gene symbols; for example, *Oct4* and *p53* genes in mice must be converted to *Pou5f1* and *Trp53*. PANTHER is a useful tool that offers batch conversion services for gene synonyms (http://www.pantherdb.org/genes/batchIdSearch.jsp; Mi et al. 2017). Human gene lists can also be submitted under



ChIP-Atlas / Enrichment Analysis

Search for proteins significantly bound to your data.

Show 100 entries

Search:

Test

ID	Antigen class	Antigen	Cell class	Cell	Num of peaks	Overlaps / Liver genes	Overlaps / Other genes	Log P-val	Log Q-val	Fold Enrichment	FE > 1?
SRX100505	TFs and others	HNF4A	Liver	Hep G2	21259	114/210	3456/18335	-29.2	-25.3	2.88	TRUE
SRX100497	TFs and others	RXRA	Liver	Hep G2	13022	90/210	2363/18335	-25.7	-22.2	3.33	TRUE
SRX100544	TFs and others	EP300	Liver	Hep G2	24334	117/210	4027/18335	-25.2	-21.8	2.54	TRUE
SRX100449	TFs and others	HNF4G	Liver	Hep G2	15919	100/210	3121/18335	-23.5	-20.2	2.80	TRUE
SRX190234	TFs and others	CEBPB	Liver	Hep G2	12483	85/210	2304/18335	-23.1	-19.9	3.22	TRUE
SRX100448	TFs and others	FOXA2	Liver	Hep G2	45130	129/210	5343/18335	-21.2	-18.1	2.11	TRUE
SRX100506	TFs and others	FOXA1	Liver	Hep G2	50941	135/210	6084/18335	-19.2	-16.2	1.94	TRUE
SRX100477	TFs and others	FOXA1	Liver	Hep G2	40732	121/210	5255/18335	-17.5	-14.5	2.01	TRUE
SRX100552	TFs and others	SP1	Liver	Hep G2	19032	125/210	5592/18335	-17.2	-14.3	1.95	TRUE
SRX150698	TFs and others	HNF4A	Liver	Hep G2	10069	75/210	2336/18335	-16.5	-13.6	2.80	TRUE
SRX150701	TFs and others	CEBPB	Liver	Hep G2	18637	100/210	4173/18335	-14.4	-11.6	2.09	TRUE
SRX1531773	TFs and others	MED1	Liver	Hep G2	3984	32/210	507/18335	-13.8	-11.0	5.51	TRUE
SRX018625	TFs and others	HNF4A	Liver	Hep G2	2654	31/210	518/18335	-12.8	-10.1	5.23	TRUE
SRX100538	TFs and	HDAC2	Liver	Hep G2	16071	96/210	4345/18335	-11.5	-8.7	1.93	TRUE

Showing 1 to 100 of 7,551 entries

Fig. 5.11 Example of an Enrichment Analysis result. The result of the submission from Fig. 5.10a is shown, which lists ChIP-seq data whose peaks preferentially overlapped within TSS ± 5 kb of liver-specific genes

the mouse and rat organism tabs, because many official gene symbols are shared in these three organisms except for upper/lower case differences, which are ignored after submission to Enrichment Analysis. To learn more about the Enrichment Analysis function, see the tutorial movie for Enrichment Analysis on the home page of ChIP-Atlas (Fig. 5.1).

5.4.6 Enrichment Analysis Using Genomic Coordinates as a Query

To analyze enriched binding of TFs and histone modifications, the Enrichment Analysis feature of ChIP-Atlas can be used with a gene list, as shown above (Sect. 5.4.5), as well as with genomic coordinates of interest. This tutorial demonstrates how to search for TFs enriched for liver-specific enhancers, which are identified by the FANTOM5 consortium (Andersson et al. 2014).

1. In order to obtain enhancers specifically activated in the liver, go to the FANTOM5 Human Enhancer Tracks web page (http://slidebase.binf.ku.dk/human_enhancers/presets) and download the "liver" data under the section "3. Enhancers specifically expressed in organs/tissues."
2. Repeat steps 5–8 in Sect. 5.4.5.
3. Select the "Genomic regions" radio button in the "Select your data" panel (Fig. 5.10b), and choose the file downloaded in step 1. The "BED-formatted" coordinates for liver-specific enhancers will be loaded in the text area.
4. Select the "Random permutation of user data" radio button in the "Select dataset to be compared" panel, and select "x1" for permutation time, which means that the genomic coordinates of liver enhancers will be randomly permutated once to generate background genomic regions.
5. Enter arbitrary titles for the data and project before clicking on the "Submit" button.
6. If the job status is "finished," click on the result URL (data not shown). Enriched ChIP-seq data for liver-specific enhancers will be similar to those for liver-specific genes as shown in Sect. 5.4.5, with HNF4A/G and FOXA1/2 in Hep G2 cells highly ranked. Altogether, these data indicate that HNF4 and FOXA proteins are significantly bound not only around the TSS of liver-specific genes but also to liver-specific distal enhancers, which may facilitate the regulation of liver/hepatocyte lineage determination.

Tips: Acceptable genomic regions must be in BED format: the first three columns of the tab-delimited file indicate chromosome (column 1) and the beginning and end of the coordinates (columns 2 and 3, respectively). If the submitted BED file has extra columns (fourth and later) as those obtained in step 1 of this tutorial, they are automatically removed prior to Enrichment Analysis processing. We note that only BED files in the following genome assemblies are acceptable: hg19 (*H. sapiens*), mm9 (*M. musculus*), rn6 (*R. norvegicus*), dm3 (*D. melanogaster*), ce10 (*C. elegans*), and sacCer3 (*S. cerevisiae*). If the BED file was prepared in another genome assembly, convert it to an acceptable one with the UCSC liftOver tool (https://genome.ucsc.edu/cgi-bin/hgLiftOver). Random background (step 4) is generated with random chromosomes and genomic positions with the same length and number as the submitted BED intervals. Increasing the number of random permutations will produce more even random background data, although the runtime will be longer. In addition to the enhancer region search as shown in this tutorial, the Enrichment Analysis service is used to analyze a wide variety of genomic contexts such as a user's own ChIP-seq peaks, SNPs associated with human traits, and evolutionary-relevant genomic regions to find TF enrichment (Oki et al. 2018; Ferris et al. 2018; Anan et al. 2018).

5.5 Database Access and Mining Methods

Data processed and recorded in ChIP-Atlas are all assigned unique URLs and are publicly available; thus, they can be used for subsequent analysis using command-line scripting and for connecting with other biodatabases. For example, the RegulatorTrail web service (https://regulatortrail.bioinf.uni-sb.de; Kehl et al. 2017) uses ChIP-Atlas Target Genes data to show regulator–target gene interactions, and the DeepBlue epigenomic data server (http://deepblue.mpi-inf.mpg.de; Albrecht et al. 2016) has recently imported the entire peak-call data of ChIP-Atlas in order to prepare an integrative analysis platform by combining with other (epi)genomic data. This section briefly introduces how to download the ChIP-Atlas data; further detailed instructions are available in the ChIP-Atlas documentation (https://github.com/inutano/chip-atlas/wiki#downloads_doc).

NOTE: All URLs in this section begin with "http://dbarchive.biosciencedbc.jp/kyushu-u," and variables are shown in bold letters.

5.5.1 Downloading Each SRX Data

- URL for alignment data in BigWig format:
  /**GENOME**/eachData/bw/**SRX_ID**.bw
  - Example (BigWig file for human SRX097088 data):
    /**hg19**/eachData/bw/**SRX097088**.bw
- URL for Peak-call data in BED format:
  /**GENOME**/eachData/bed**THRESHOLD**/**SRX_ID**.**THRESHOLD**.bed
  (**THRESHOLD** = 05, 10, or 20)
  - Example (BED file for human SRX097088 data with MACS2 *Q*-value <1E–05):
    /**hg19**/eachData/bed**05**/**SRX097088**.**05**.bed

5.5.2 Assembled Peak-Call Data Used in "Peak Browser"

- URL for assembled Peak-call data in bed format:
  /**GENOME**/allPeaks_light/allPeaks_light.**GENOME**.**THRESHOLD**.bed.gz
  - Example (assembled mouse peak-call data with MACS2 *Q*-value <1E–10):
    /**mm9**/allPeaks_light/allPeaks_light.**mm9**.**10**.bed.gz

This file is composed of genomic coordinates (columns 1–3), SRX IDs, (column 4), and MACS2 scores (column 5). To restore the antigen and cell type names corresponding to SRX IDs, a corresponding table is available from following URL:

```
/metadata/experimentList.tab
```

5.5.3 Analyzed Data Used in "Target Genes" and "Colocalization"

- URL for directory storing "Target Genes" results in HTML and TSV format: `/GENOME /target/`
- URL for directory storing "Colocalization" results in HTML and TSV format:
  `/GENOME /colo/`
  where **GENOME** is the genome assembly shown in Fig. 5.4b.

Acknowledgements **Funding**: ChIP-Atlas project was supported by the National Bioscience Database Center (NBDC) of the Japan Science and Technology Agency (JST).

References

Albrecht F, List M, Bock C, Lengauer T (2016) DeepBlue epigenomic data server: programmatic data retrieval and analysis of epigenome region sets. Nucleic Acids Res 44:W581–W586. https://doi.org/10.1093/nar/gkw211

Anan K, Hino S, Shimizu N et al (2018) LSD1 mediates metabolic reprogramming by glucocorticoids during myogenic differentiation. Nucleic Acids Res. 46(11):5441–5454. https://doi.org/10.1093/nar/gky234

Andersson R, Gebhard C, Miguel-Escalada I et al (2014) An atlas of active enhancers across human cell types and tissues. Nature 507:455–461. https://doi.org/10.1038/nature12787

Chang JT, Nevins JR (2006) GATHER: a systems approach to interpreting genomic signatures. Bioinformatics 22:2926–2933. https://doi.org/10.1093/bioinformatics/btl483

Ferris E, Abegglen LM, Schiffman JD, Gregg C (2018) Accelerated evolution in distinctive species reveals candidate elements for clinically relevant traits, including mutation and cancer resistance. Cell Rep 22:2742–2755. https://doi.org/10.1016/j.celrep.2018.02.008

Imanishi T, Nakaoka H (2009) Hyperlink management system and ID converter system: enabling maintenance-free hyperlinks among major biological databases. Nucleic Acids Res 37:W17–W22. https://doi.org/10.1093/nar/gkp355

Kehl T, Schneider L, Schmidt F et al (2017) RegulatorTrail: a web service for the identification of key transcriptional regulators. Nucleic Acids Res 45:W146–W153. https://doi.org/10.1093/nar/gkx350

Langmead B, Salzberg SL (2012) Fast gapped-read alignment with Bowtie 2. Nat Methods 9:357–359. https://doi.org/10.1038/nmeth.1923

Linnemann AK, O'Geen H, Keles S et al (2011) Genetic framework for GATA factor function in vascular biology. Proc Natl Acad Sci 108:13641–13646. https://doi.org/10.1073/pnas.1108440108

Mi H, Huang X, Muruganujan A et al (2017) PANTHER version 11: expanded annotation data from Gene Ontology and Reactome pathways, and data analysis tool enhancements. Nucleic Acids Res 45:D183–D189. https://doi.org/10.1093/nar/gkw1138

Mootha VK, Lindgren CM, Eriksson K-F et al (2003) PGC-1α-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. Nat Genet 34: 267–273. https://doi.org/10.1038/ng1180

Oki S, Ohta T, Shioi G, et al (2018) Integrative analysis of transcription factor occupancy at enhancers and disease risk loci in noncoding genomic regions. bioRxiv 262899. https://doi.org/10.1101/262899

Ono H, Ogasawara O, Okubo K, Bono H (2017) RefEx, a reference gene expression dataset as a web tool for the functional analysis of genes. Sci Data 4:170105. https://doi.org/10.1038/sdata.2017.105

Park PJ (2009) ChIP-seq: advantages and challenges of a maturing technology. Nat Rev Genet 10: 669–680. https://doi.org/10.1038/nrg2641

Robinson JT, Thorvaldsdóttir H, Winckler W et al (2011) Integrative genomics viewer. Nat Biotechnol 29:24–26. https://doi.org/10.1038/nbt.1754

Sekiya S, Suzuki A (2011) Direct conversion of mouse fibroblasts to hepatocyte-like cells by defined factors. Nature 475:390–393. https://doi.org/10.1038/nature10263

Subramanian A, Tamayo P, Mootha VK et al (2005) Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc Natl Acad Sci U S A 102:15545–15550. https://doi.org/10.1073/pnas.0506580102

Yu M, Selvaraj SK, Liang-Chu MMY et al (2015) A resource for cell line authentication, annotation and quality control. Nature 520:307–311. https://doi.org/10.1038/nature14397

Zhang Y, Liu T, Meyer CA et al (2008) Model-based analysis of ChIP-Seq (MACS). Genome Biol 9:R137. https://doi.org/10.1186/gb-2008-9-9-r137

Chapter 6
RefEX: Reference Expression Dataset

Hiromasa Ono and Hidemasa Bono

Abstract Reference Expression dataset (RefEx) is a web tool which allows users to search by the gene name, various types of IDs, chromosomal regions in genetic maps, gene family based on InterPro, gene expression patterns, or biological categories based on Gene Ontology. RefEx also provides information about genes with tissue-specific expression, and the relative gene expression values are shown as choropleth maps on 3D human body images from BodyParts3D. Combined with FANTOM dataset, RefEx enables users to draw insights regarding the functional interpretation of unfamiliar genes.

Keywords RNA-SEQ · Probe level · Database · Transcriptome · Update · Atlas · Bioinformatics · Profiles · Archive · Biology

6.1 Introduction

Reference Expression dataset (RefEx) (Ono et al. 2017) is a web tool which allows users to browse gene expression profiles by genes collected from public databases. It can be searched by various types of IDs including gene names, chromosomal regions, gene family based on InterPro (Mitchell et al. 2015), gene expression patterns, or gene annotations based on Gene Ontology (Ashburner et al. 2000). Information about genes with tissue-specific expression are also provided, and the relative gene expression values are shown as choropleth maps on 3D human body images from BodyParts3D (Mitsuhashi et al. 2009). RefEx provides insight regarding the functional interpretation of unfamiliar genes through its web interface.

The purpose of RefEx is to provide a web tool for visualization of reference gene expression pattern of mammalian tissues and cell lines measured using different

H. Ono · H. Bono (✉)
Database Center for Life Science, Joint Support-Center for Data Science Research, Research Organization of Information and Systems, Mishima, Japan
e-mail: bono@dbcls.rois.ac.jp

© The Author(s), under exclusive license to Springer Nature Singapore Pte Ltd. 2021

I. Abugessaisa, T. Kasukawa (eds.), *Practical Guide to Life Science Databases*,
https://doi.org/10.1007/978-981-16-5812-9_6

methods, which can facilitate the reuse of the precious data archived in several public databases.

RefEx provides suitable datasets as a reference for gene expression data from 40 normal tissues from human, mouse, and rat collected from public gene expression databases. The collected gene expression data are classified based on four different measurement strategies (Expressed Sequence Tags (ESTs), GeneChip, Cap Analysis of Gene Expression (CAGE), and RNA-Seq). These four types of data were linked based on the NCBI gene IDs in the dataset in RefEx. In addition to these datasets, RefEx currently includes quantified gene expression data from Functional Annotation of the Mammalian genome 5 (FANTOM5) dataset for human and mouse (The FANTOM Consortium & the RIKEN PMI and CLST (DGT) 2014).

Target user group of RefEx is biologists who wish to reuse public data, but accessing the data remains difficult due to its sheer magnitude and complicated access.

6.2 Database Overview

6.2.1 Importance of Reference Gene Expression Datasets

Gene expression data are exponentially accumulating after the advent of gene expression measurement methods on a genomic scale. Many datasets are now archived in the public gene expression databases [NCBI Gene Expression Omnibus (GEO) (Barrett et al. 2013) and EBI ArrayExpress (Kolesnikov et al. 2015)]. Because the description about datasets is written by different researchers who produced the gene expression data, they are so different that it is not machine readable currently.

Nevertheless, there is strong demand for a comprehensive set of reference gene expression data from huge gene expression data in public. The availability of such data is of benefit to biologists who wish to reuse it, but accessing the data remains difficult due to its sheer magnitude and complicated access. Recently, a meta-analysis of RNA-Seq expression data across various species, tissues, and studies was reported (Sudmant et al. 2015). However, the interpretation of such data is not easy. Biologists are often at a loss because of the sheer number of datasets in public databases provided by numerous researchers. From such situations, reference expression datasets are needed for the inference of functions of genes, and a proper web interface for visualizing such data is essential.

In addition, concerted patterns of gene expression profiles for different quantification methods can strengthen the evidence of these patterns. Also, tissue-specific expression can be a key feature to examine the function of genes of interest, and lists of genes with tissue-specific expression can help biologists to explore unannotated genes with prominent expression patterns. Thus, the functional annotation of genes from meta-analysis and the interface to access the data with graphical visualization are urgently required.

6.2.2 Current Status of Reference Gene Expression Data

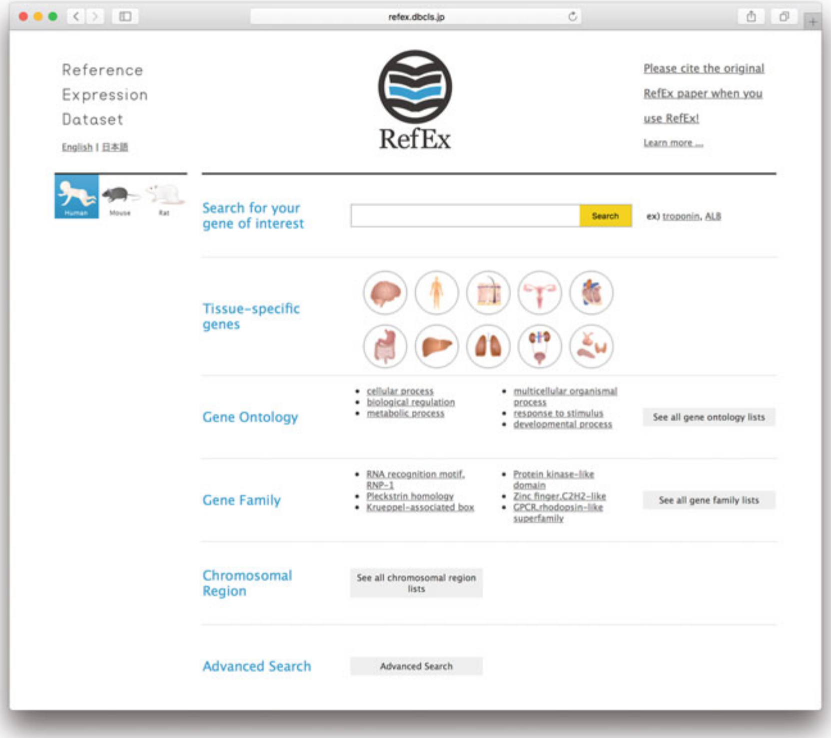
As a reference gene expression data for a genomic scale, expression profiles in normal mammalian tissues by GeneChip were first maintained by researchers at the Genomics Institute of the Novartis Research Foundation (GNF). Users can access the microarray data produced in that project at the GNF Expression Atlas (also known as GNF SymAtlas), now called BioGPS (Wu et al. 2016). EBI also maintains the Expression Atlas which provides gene expression patterns under various biological conditions based on data archived in ArrayExpress (Petryszak et al. 2016). Recently, using RNA sequencing and CAGE (Shiraki et al. 2003), the FANTOM collaboration consortium released terabytes of transcriptome sequencing data from adult and fetal human and mouse tissue primary cell lines that can be used as the reference gene expression data (Lizio et al. 2015).

6.2.3 The Main Feature of RefEx

The main feature of RefEx is a simple web interface, which allows users to compare expression profiles by different methods at a glance (Fig. 6.1). It provides access to curated data from several other public databases, with expression levels in 40 tissues measured by four well-established gene expression quantification technologies (ESTs, Affymetrix GeneChip, CAGE, and RNA-Seq). The web interface allows users to browse the expression profiles by the gene name, various types of IDs, chromosomal regions in genetic maps, gene family based on InterPro, gene expression patterns, or biological categories based on Gene Ontology. The web interface also includes the way to browse the expression profile for adult and fetal human and mouse tissues obtained by the FANTOM5 project using CAGE for gene expression quantification. All the data provided through the RefEx web interface is listed with corresponding digital object identifiers (DOI) in Table 6.1.

6.2.4 Future Update and Availability of the Database

RefEx is planned to be updated when the useful dataset for gene expression is publicly available. Data in RefEx is freely available under a Creative Commons Attribution 4.0 International License (CC BY 4.0). Users can download a concatenated version of all the data at the RefEx download page (https://refex.dbcls.jp/download.php?lang=en), including the log-transformed ratios of the gene expression, the functional annotation of the genes, the list of tissue-specific genes, and the sample information, in a tab-delimited text format (Table 6.1). Data in RefEx is also available at figshare (https://doi.org/10.6084/m9.figshare.c.3812815) and the scripts to make RefEx data are available at GitHub (https://github.com/dbcls/RefEx/).


refex.dbcls.jp
RefEx
English | 日本語
Human
Mouse
Rat
Search
See all gene ontology lists
See all gene family lists
See all chromosomal region lists
Advanced Search

Reference Expression Dataset

Please cite the original RefEx paper when you use RefEx!

Learn more ...

Search for your gene of interest

ex) troponin, ALB

Tissue-specific genes

Gene Ontology

- cellular process
- biological regulation
- metabolic process
- multicellular organismal process
- response to stimulus
- developmental process

Gene Family

- RNA recognition motif, RNP-1
- Pleckstrin homology
- Krueppel-associated box
- Protein kinase-like domain
- Zinc finger,C2H2-like
- GPCR,rhodopsin-like superfamily

Chromosomal Region

Advanced Search

Fig. 6.1 The top page of RefEx web tool (https://refex.dbcls.jp/)

6.3 Content and Architecture of the Database

The type of data stored in RefEx is the processed and quantified gene expression data for human, mouse, and rat. All data in RefEx is originally from the public database. Table 6.2 shows original data sources for RefEx dataset.

The data in RefEx were manually collected by RefEx curators from public databases, including the International Nucleotide Sequence Database (INSD, consisting of GenBank/DDBJ/ENA) (Cochrane et al. 2016), the NCBI Gene Expression Omnibus (GEO), and RNA-Seq data in the Sequence Read Archive (SRA) (Kodama et al. 2012). The raw data from the public databases were re-organized and compared against each other. Four types of data were linked based on the NCBI gene IDs, while the EST data were based on the Unigene IDs, and the GeneChip data were based on the Probe set IDs. Detailed information regarding four data extraction methods are described below. All scripts used to produce the data and additional descriptions are available on the GitHub site at https://github.com/dbcls/RefEx.

Table 6.1 Summary table of the entity stored in the database. Original data is available from RefEx page for download (https://refex.dbcls.jp/download.php?lang=en)

1. Gene expression data		
Human	EST 10 tissues	DOI: https://doi.org/10.6084/m9.figshare.4028625
	EST 40 tissues	DOI: https://doi.org/10.6084/m9.figshare.4028634
	GeneChip 10 tissues	DOI: https://doi.org/10.6084/m9.figshare.4028643
	GeneChip 40 tissues	DOI: https://doi.org/10.6084/m9.figshare.4028652
	CAGE 10 tissues	DOI: https://doi.org/10.6084/m9.figshare.4028619
	CAGE 40 tissues	DOI: https://doi.org/10.6084/m9.figshare.4028622
	CAGE all	DOI: https://doi.org/10.6084/m9.figshare.4028613
	RNA-seq 10 tissues	DOI: https://doi.org/10.6084/m9.figshare.4028661
	RNA-seq 40 tissues	DOI: https://doi.org/10.6084/m9.figshare.4028667
Mouse	EST 10 tissues	DOI: https://doi.org/10.6084/m9.figshare.4028628
	EST 40 tissues	DOI: https://doi.org/10.6084/m9.figshare.4028637
	GeneChip 10 tissues	DOI: https://doi.org/10.6084/m9.figshare.4028646
	GeneChip 40 tissues	DOI: https://doi.org/10.6084/m9.figshare.4028655
	CAGE all	DOI: https://doi.org/10.6084/m9.figshare.4028616
	RNA-seq 10 tissues	DOI: https://doi.org/10.6084/m9.figshare.4028664
	RNA-seq 40 tissues	DOI: https://doi.org/10.6084/m9.figshare.4028670
Rat	EST 10 tissues	DOI: https://doi.org/10.6084/m9.figshare.4028631
	EST 40 tissues	DOI: https://doi.org/10.6084/m9.figshare.4028640
	GeneChip 10 tissues	DOI: https://doi.org/10.6084/m9.figshare.4028649
	GeneChip 40 tissues	DOI: https://doi.org/10.6084/m9.figshare.4028658
2. Tissue specificity [calculated by ROKU (Kadota et al. 2006) method]		
Human	GeneChip	DOI: https://doi.org/10.6084/m9.figshare.4028700
	RNA-seq	DOI: https://doi.org/10.6084/m9.figshare.4028709

(continued)

Table 6.1 (continued)

Mouse	GeneChip	DOI: https://doi.org/10.6084/m9.figshare.4028703
Rat	GeneChip	DOI: https://doi.org/10.6084/m9.figshare.4028706
3. ID relation table		
Human	DOI: https://doi.org/10.6084/m9.figshare.4028676	
Mouse	DOI: https://doi.org/10.6084/m9.figshare.4028679	
Rat	DOI: https://doi.org/10.6084/m9.figshare.4028682	
4. Tissue table		
Common	10 tissue names	DOI: https://doi.org/10.6084/m9.figshare.4028712
	40 tissue names	DOI: https://doi.org/10.6084/m9.figshare.4028718
Human	Sample classifications (GeneChip)	DOI: https://doi.org/10.6084/m9.figshare.4028598
	Sample classifications (RNA-seq)	DOI: https://doi.org/10.6084/m9.figshare.4028607
Mouse	Sample classifications (GeneChip)	DOI: https://doi.org/10.6084/m9.figshare.4028601
	Sample classifications (RNA-seq)	DOI: https://doi.org/10.6084/m9.figshare.4028610
Rat	Sample classifications (GeneChip)	DOI: https://doi.org/10.6084/m9.figshare.4028604
5. Sample annotations		
Human	Sample annotations (GeneChip)	DOI: https://doi.org/10.6084/m9.figshare.4028691
	Sample annotations (CAGE)	DOI: https://doi.org/10.6084/m9.figshare.4028685
Mouse	Sample annotations (GeneChip)	DOI: https://doi.org/10.6084/m9.figshare.4028703
	Sample annotations (CAGE)	DOI: https://doi.org/10.6084/m9.figshare.4028688
Rat	Sample annotations (GeneChip)	DOI: https://doi.org/10.6084/m9.figshare.4028697
6. RDF		
Human	RefEx FANTOM5 RDF	The NBDC RDF Portal (http://integbio.jp/rdf/)

6.3.1 EST

The original EST data were retrieved from the EST division of the INSD. The number of ESTs was counted by source organ based on the BodyMap method (Okubo et al. 1992) according to the cDNA annotation of each EST entry. The EST data in RefEx originated from the BodyMap-Xs database, which contains

Table 6.2 Original data sources for RefEx dataset. EST data is originally from the International Nucleotide Sequence Database (INSD). Original data of GeneChip data and CAGE & RNA-seq data is from the NCBI Gene Expression Omnibus (GEO) and the Sequence Read Archive (SRA), respectively

	Human	Mouse	Rat
EST	INSD	INSD	INSD
GeneChip	GSE7307	GSE10246	GSE952
CAGE	PRJDB3010	PRJDB1100	Not available
RNA-seq	PRJEB2445	PRJNA30467	Not available

previously compiled gene expression data from the INSD EST division for reuse (Ogasawara et al. 2006). After counting the number of ESTs, gene expression data were obtained for the 40 normal tissues stored in the BodyMap-Xs database (https://doi.org/10.6084/m9.figshare.4028721). For visualization purposes, the data were grouped into ten subsets (i.e., brain, blood, connective, reproductive, muscular, alimentary, liver, lung, urinary, and endo/exocrine; https://doi.org/10.6084/m9.figshare.4028715). This categorization of the organs was also applied to the gene expression data that were obtained by the other methods.

6.3.2 GeneChip

The GeneChip data deposited in the NCBI GEO database were selected for the reference dataset (tissue-specific patterns of mRNA expression) (Table 6.2). Those data were analyzed based on a typical microarray data analysis method (Wu et al. 2016). The expression values of the genes were calculated from the original CEL files after robust multi-array averaging (RMA) normalization (Irizarry et al. 2003) by the affy package (Gautier et al. 2004) in R (ver.3.0.3)/BioConductor (ver.2.12) (Gentleman et al. 2004).

6.3.3 CAGE

CAGE is a technique that produces a snapshot of the 5′ end of the mRNA population in a biological sample, and the CAGE data collected in the RIKEN FANTOM5 project were counted by source organ based on the original data, the FANTOM5 CAGE peak expression, and the annotation tables (Lizio et al. 2015). The CAGE tag counts were mapped onto the reference genome sequences (hg19 for human and mm9 for mouse) and reflect the intensity of the gene expression of the corresponding transcripts. The tag counts are normalized by tag per million (TPM). The processed data in RefEx is converted to log 2 for each TPM value of the original FANTOM 5 CAGE data and then organized for each sample classification and the data to which the same GeneID is assigned are added up and averaged. In addition to the 40 normal tissues, the FANTOM5 project collected hundreds of samples from cell lines, primary cells, and adult and fetal tissues of human (https://doi.org/10.6084/m9.figshare.4028685) and mouse (https://doi.org/10.6084/m9.figshare.4028688).

6.3.4 RNA-Seq

For RNA-Seq data, the normal tissue transcriptome sequence data were selected from the SRA. Utilizing human and mouse reference genome sequences (hg19 for human and mm9 for mouse), these data were processed using a typical RNA-Seq data analysis pipeline with TopHat (ver.2.0.7) (Trapnell et al. 2009) and Cufflinks (ver.2.0.2) (Trapnell et al. 2010), and the transcript abundances were calculated and normalized to fragments per kilobase of transcript per million reads (FPKM).

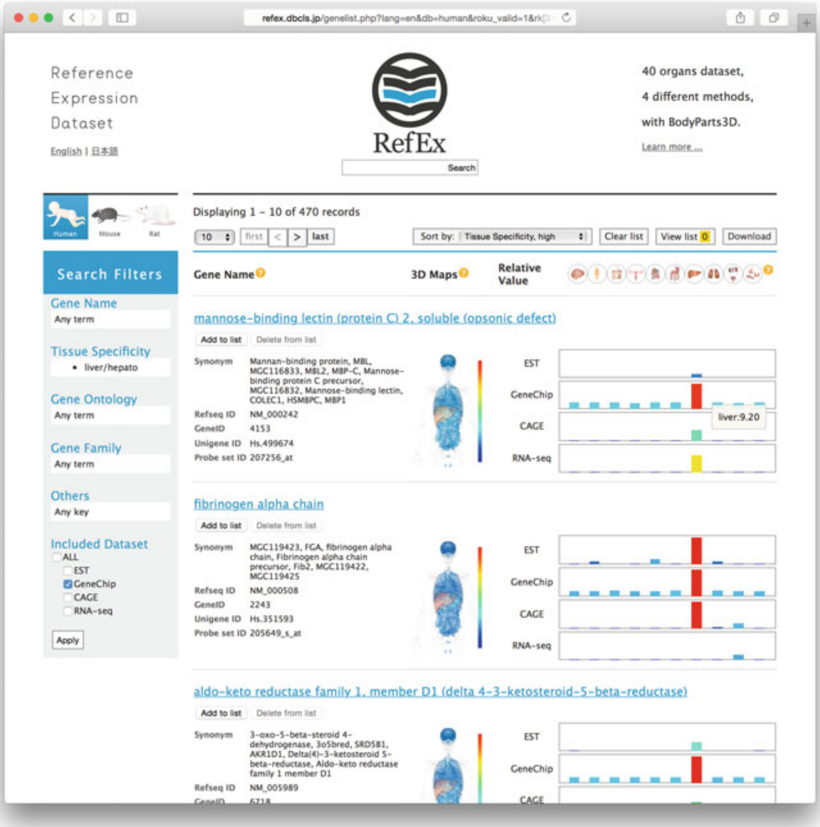
Four types of data above were linked based on the NCBI gene IDs in the dataset in RefEx. The EST data were clustered by sequence similarity, and the NCBI UniGene IDs were added to those clusters. The GeneChip data were based on Affymetrix probe IDs, which were originally designed based on the UniGene database (Wagner and Agarwala 2013). The remaining two methods were based on direct sequencing and were developed after the completion of the human and mouse genome sequencing projects; the data obtained by these methods can be mapped to the reference genomes by the genomic position. Thus, the NCBI Gene IDs were adopted, which are currently widely used to integrate other gene IDs, as a standard. Mapping the various gene IDs (UniGene ID, Affymetrix probe ID, and NCBI Gene ID) onto the various genomes was performed using the Biomart REST API (http://www.biomart.org/martservice.html).

The quality of expression data above is guaranteed by the biological replicates. For example, the RNA-Seq data used were generated from multiple reads (single-read and paired-end read) and those of mouse from triplicate sequence reads were averaged to represent the gene expression value.

6.4 Database Access and Mining Methods

6.4.1 Gene Expression Visualization Tool in RefEx

The relative gene expression values are shown in RefEx as choropleth maps on 3D human body images from BodyParts3D (Mitsuhashi et al. 2009). BodyParts3D has been developed by the Database Center for Life Science (DBCLS) as a dictionary-type anatomy database in which anatomical concepts are represented by 3D structural data that specify the corresponding segments on a 3D whole-body model of an adult human male. Foundational Model of Anatomy (FMA) ontology (https://bioportal.bioontology.org/ontologies/FMA) was used to map the gene expression



refex.dbcls.jp/genelist.php?lang=en&db=human&roku_valid=1&rk

Reference Expression Dataset

English | 日本語

RefEx

40 organs dataset, 4 different methods, with BodyParts3D.

Learn more ...

Displaying 1 – 10 of 470 records

Sort by: Tissue Specificity, high

Search Filters

Gene Name: Any term

Tissue Specificity: liver/hepato

Gene Ontology: Any term

Gene Family: Any term

Others: Any key

Included Dataset

□ ALL
□ EST
☑ GeneChip
□ CAGE
□ RNA-seq

Gene Name	3D Maps	Relative Value

mannose-binding lectin (protein C) 2, soluble (opsonic defect)

Synonym	Mannan-binding protein, MBL, MGC116833, MBL2, MBP-C, Mannose-binding protein C precursor, MGC116832, Mannose-binding lectin, COLEC1, HSMBPC, MBP1
Refseq ID	NM_000242
GeneID	4153
Unigene ID	Hs.499674
Probe set ID	207256_at

fibrinogen alpha chain

Synonym	MGC119423, FGA, fibrinogen alpha chain, Fibrinogen alpha chain precursor, Fib2, MGC119422, MGC119425
Refseq ID	NM_000508
GeneID	2243
Unigene ID	Hs.351593
Probe set ID	205649_s_at

aldo-keto reductase family 1, member D1 (delta 4-3-ketosteroid-5-beta-reductase)

Synonym	3-oxo-5-beta-steroid 4-dehydrogenase, 3o5bred, SRD5B1, AKR1D1, Delta(4)-3-ketosteroid 5-beta-reductase, Aldo-keto reductase family 1 member D1
Refseq ID	NM_005989
GeneID	6718

Fig. 6.2 The search results for liver-specific genes. This view can easily be viewed by clicking the liver icon at the top of the RefEx page (Fig. 6.1)

data onto the corresponding tissues. Because drawing the choropleth maps dynamically on a 3D human body is quite labor-intensive, still images were prepared for only the GeneChip data for the whole entries. Figure 6.2 clearly illustrates that the selected transcript is highly expressed in the liver tissue. On the right (Fig. 6.2), the relative expression levels in 40 types of normal tissues that were more precisely classified are displayed. The visualization can help users to understand the differences in the gene expression patterns among tissues more intuitively.

6.4.2 How to Query RefEx

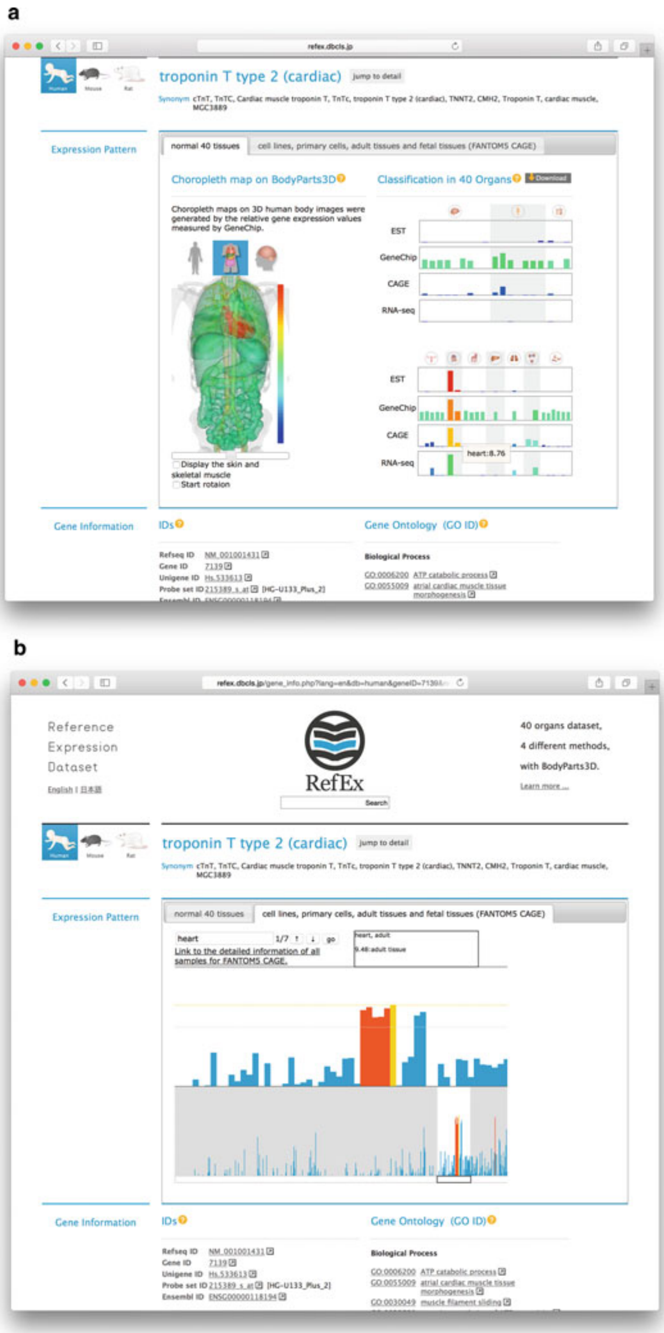
Users can easily query RefEx with an effective filter to extract genes with concerted gene expression profiles. For example, genes with liver-specific gene expression can be retrieved only a single click. Details are available in a video tutorial (https://doi.org/10.7875/togotv.2016.068).

After the publication of the FANTOM5 project, pre-calculated gene expression data from the CAGE data in the FANTOM5 project were incorporated into RefEx. The most important benefit of the FANTOM5 CAGE data is that the search targets are much more abundant. The original version of RefEx only had forty tissue search targets (Fig. 6.3a). However, it is now possible to search more than 500 human samples, encompassing cell lines, primary cells, and adult and fetal tissues (Fig. 6.3b). RefEx also enables users to browse high-resolution gene expression data from approximately 800 samples (human plus mouse).

By clicking the tab on the right-hand side, users can switch to a FANTOM5 CAGE data viewer (Fig. 6.3b). This viewer shows the expression patterns of all samples in the lower portion of the screen and displays an enlarged view of a specific area in the upper portion of the screen. Because this is a representation of the expression profile in humans, 556 samples are shown in a bar chart in the lower portion of the screen. Therefore, a user can observe an overview of expression patterns in all the samples. The area displayed in the enlarged box can be moved freely by dragging. When a user enters a keyword into the search window of the viewer, the sample name containing that keyword is highlighted. The FANTOM5 CAGE data correspond to the tissue classification in the original RefEx and are linked to the original FANTOM5 data. The expression values of the samples obtained in the FANTOM5 project are averaged and listed in RefEx.

6.4.3 How to Download Data from RefEx

While the data shown in RefEx is originally from the public database and the sources for all data records are summarized in Table 6.2, the data used in RefEx including processed gene expression data can be downloaded from RefEx download page (https://refex.dbcls.jp/download.php?lang=en). These data are deposited in figshare (Fig. 6.4), which is a repository where users can make all of their research outputs available in a citable, shareable, and discoverable manner. Forty-one datasets uploaded to figshare can be accessible from figshare collection at https://doi.org/10.6084/m9.figshare.c.3812815.



troponin T type 2 (cardiac)

Synonym cTnT, TnTC, Cardiac muscle troponin T, TnTc, troponin T type 2 (cardiac), TNNT2, CMH2, Troponin T, cardiac muscle, MGC3889

Choropleth map on BodyParts3D

Choropleth maps on 3D human body images were generated by the relative gene expression values measured by GeneChip.

□ Display the skin and skeletal muscle
□ Start rotaion

Classification in 40 Organs

IDs

Refseq ID NM_001001431
Gene ID 7139
Unigene ID Hs.533613
Probe set ID 215389_s_at [HG-U133_Plus_2]

Gene Ontology (GO ID)

Biological Process

GO:0006200 ATP catabolic process
GO:0055009 atrial cardiac muscle tissue morphogenesis

40 organs dataset,
4 different methods,
with BodyParts3D.

troponin T type 2 (cardiac)

Synonym cTnT, TnTC, Cardiac muscle troponin T, TnTc, troponin T type 2 (cardiac), TNNT2, CMH2, Troponin T, cardiac muscle, MGC3889

Link to the detailed information of all samples for FANTOM5 CAGE.

IDs

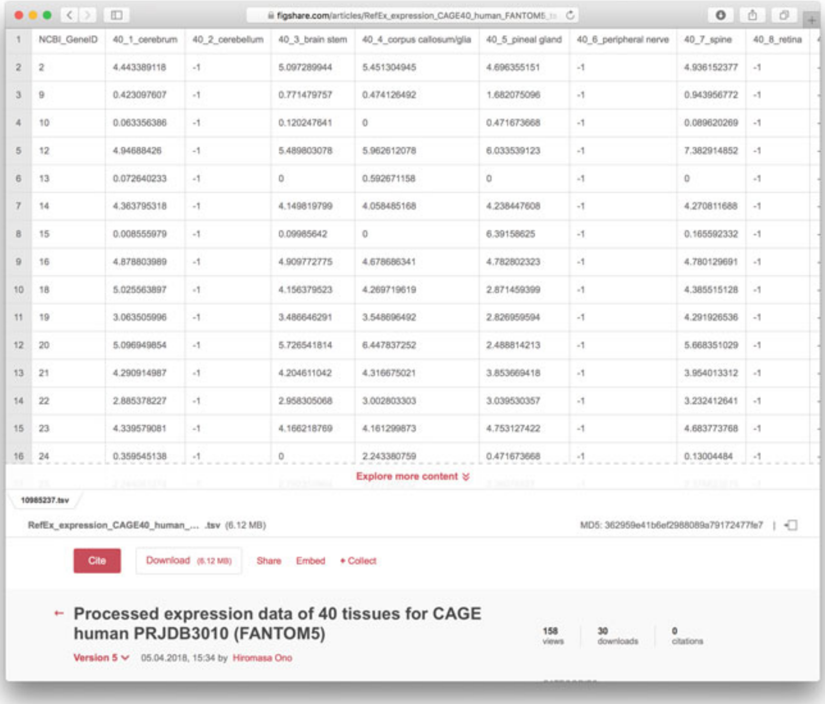
Refseq ID NM_001001431
Gene ID 7139
Unigene ID Hs.533613
Probe set ID 215389_s_at [HG-U133_Plus_2]
Ensembl ID ENSG00000118194

Gene Ontology (GO ID)

Biological Process

GO:0006200 ATP catabolic process
GO:0055009 atrial cardiac muscle tissue morphogenesis
GO:0030049 muscle filament sliding

Fig. 6.3 Detailed expression view by a gene (Troponin T type 2). (**a**) Forty normal tissues. (**b**) FANTOM5 CAGE



figshare.com/articles/RefEx_expression_CAGE40_human_FANTOM5_ts

1	NCBI_GeneID	40_1_cerebrum	40_2_cerebellum	40_3_brain stem	40_4_corpus callosum/glia	40_5_pineal gland	40_6_peripheral nerve	40_7_spine	40_8_retina
2	2	4.443389118	-1	5.097289944	5.451304945	4.696355151	-1	4.936152377	-1
3	9	0.423097607	-1	0.771479757	0.474126492	1.682075096	-1	0.943956772	-1
4	10	0.063356386	-1	0.120247641	0	0.471673668	-1	0.089620269	-1
5	12	4.94688426	-1	5.489803078	5.962612078	6.033539123	-1	7.382914852	-1
6	13	0.072640233	-1	0	0.592671158	0	-1	0	-1
7	14	4.363795318	-1	4.149819799	4.058485168	4.238447608	-1	4.270811688	-1
8	15	0.008555979	-1	0.09985642	0	6.39158625	-1	0.165592332	-1
9	16	4.878803989	-1	4.909772775	4.678686341	4.782802323	-1	4.780129691	-1
10	18	5.025563897	-1	4.156379523	4.269719619	2.871459399	-1	4.385515128	-1
11	19	3.063505996	-1	3.486646291	3.548696492	2.826959594	-1	4.291926536	-1
12	20	5.096949854	-1	5.726541814	6.447837252	2.488814213	-1	5.668351029	-1
13	21	4.290914987	-1	4.204611042	4.316675021	3.853669418	-1	3.954013312	-1
14	22	2.885378227	-1	2.958305068	3.002803303	3.039530357	-1	3.232412641	-1
15	23	4.339579081	-1	4.166218769	4.161299873	4.753127422	-1	4.683773768	-1
16	24	0.359545138	-1	0	2.243380759	0.471673668	-1	0.13004484	-1

RefEx_expression_CAGE40_human_... .tsv (6.12 MB)

MD5: 362959e41b6ef2988089a79172477fe7

Processed expression data of 40 tissues for CAGE human PRJDB3010 (FANTOM5)

Version 5 05.04.2018, 15:34 by Hiromasa Ono

158 views | 30 downloads | 0 citations

Fig. 6.4 Data in RefEx can be downloaded from figshare. https://doi.org/10.6084/m9.figshare.4028622.v5

6.4.4 Programmatic Technique to Access RefEx

As a member of the integrated database project in Japan, the Resource Description Framework (RDF) version of RefEx resides at the National Bioscience Database Center (NBDC) RDF portal, and the RefEx dataset is ready for use at the NBDC RDF Portal (https://integbio.jp/rdf/).

6.5 Use-Cases and Demo to Utilize the Database

As a simple use-case of RefEx, gene expression profiles for specific genes of interest in normal tissues were often used in medical research. The mRNA expression levels of isocitrate dehydrogenase 3 (NAD(+)) alpha (IDH3α) and vascular endothelial growth factor A (VEGF-A) were used and visualized as main indicators in ten major groups of normal tissues (Fig. 6.5) (Zeng et al. 2015) in conjunction with the calculated prognostic values by the PrognoScan database (Mizuno et al. 2009).

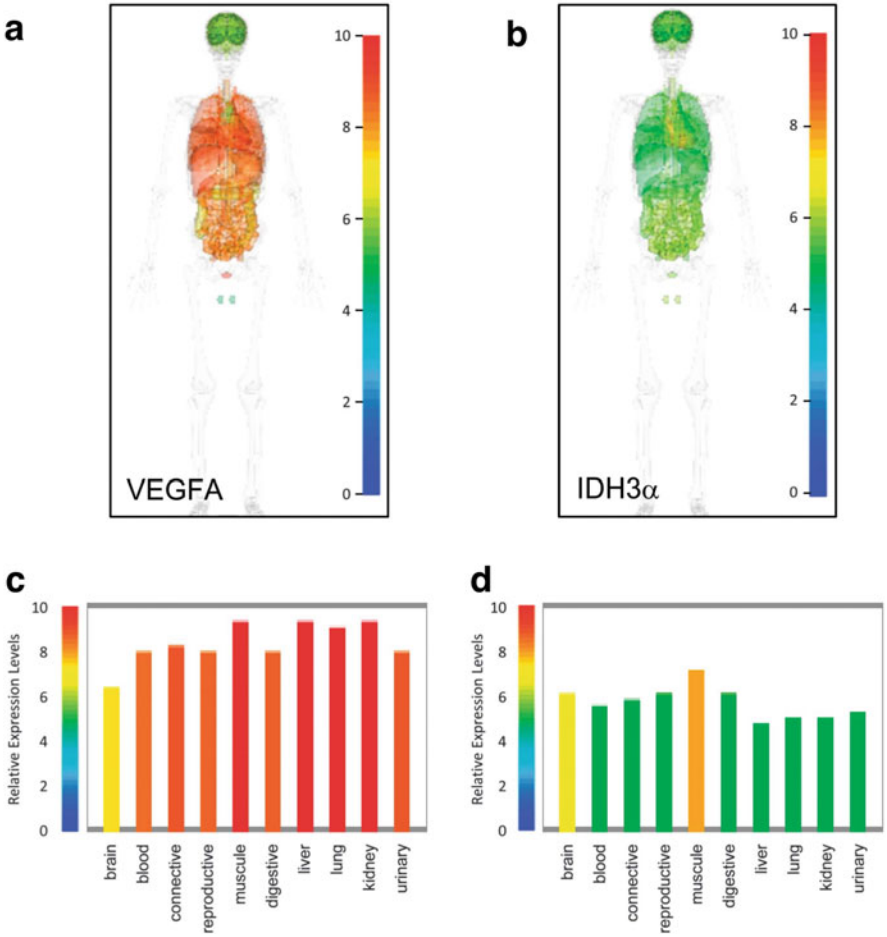


Fig. 6.5 RefEx-based quantification of the mRNA expression levels of VEGF (**a**, **c**) gene and IDH3a (**b**, **d**) gene in the indicated ten major groups of normal tissues (Zeng et al. 2015)

Another application is used as a reference of tissue-specific information. In a study of murine colon proteomes, colon-specific genes in the mouse version of RefEx were compared to a list of genes from murine colon proteomes that was generated by the researchers' own results (Magdeldin et al. 2012). A similar example was a study of liver-specific genes to investigate biomarkers indicating liver injury in humans. The gene expression profiles of albumin (ALB), apolipoprootein H (APOH), group-specific component (GC), and α-1 microglobulin/bikunin precursor (AMBP) were used to confirm the liver-specific expression (Okubo et al. 2016).

The data retrieved from RefEx strengthened the authors' hypotheses without the further confirmation in wet-lab. The gene expression profile of noncardiac myosin, light chain 12A (MRLC) in RefEx was used to confirm the conclusions because noncardiac MRLC was expressed in the heart at the same level as that in the skeletal muscle while it was annotated as “noncardiac” (Mizutani et al. 2016). As an another example, RefEx was used to list the gene expression profiles of all genes previously reported to cause deafness in a review article (Nishio et al. 2015).

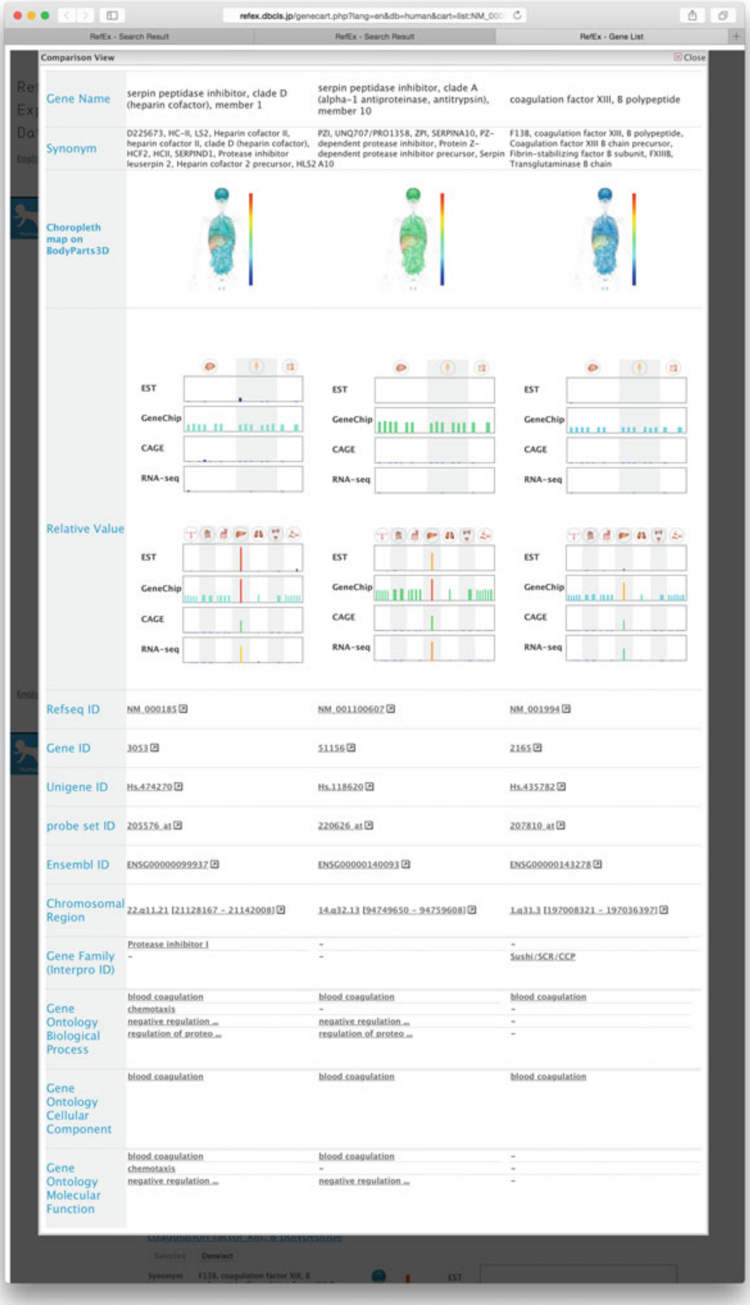
As an intermediate use-case of RefEx using the RefEx web interface, users can add up to three genes to their list and compare these genes simultaneously. Users can compare all the detailed information about the genes in that list, including the expression data. This parallel comparison enables users to easily identify the differences among the genes. Overlapped terms, such as the Gene Ontology and the InterPro gene family terms, are arranged in the same row (Fig. 6.6). Therefore, RefEx is also useful as a tool for investigating the relationships of unknown genes found in gene expression analyses.

For the advanced use-case, RefEx can be used in comparative transcriptomic studies as a reliable reference expression dataset of human normal tissues and cell lines. To compare the characteristics between human small intestine and *Bombyx mori* larval midgut, gene expression profiles from a human colon carcinoma cell line (Caco-2) cells, which is used to estimate human intestinal absorption, human small intestine, and *B. mori* larval midgut were compared to identify common drug transporters in the human intestine and *B. mori* larval midgut (Ichino et al. 2018). While the *B. mori* larval midgut RNA-seq data were originally analyzed, the gene expression data for Caco-2 cells and human small intestine were reused from RefEx [“Processed expression data of all samples for CAGE human PRJDB3010 (FANTOM5)” downloaded from https://doi.org/10.6084/m9.figshare.4028613.v4]. As a result, 26 drug transporter homologs were found and those were common in the *B. mori* larval midgut and human intestine (Fig. 6.7).

6.6 Summary and Future Development of the Database

RefEx is a tool for an interactive analysis of gene expression patterns on the web via the latest version of web browsers such as Firefox, Safari, and Chrome. RefEx has three main applications. First, users can examine the expression profiles of unfamiliar genes in normal body tissues, cells, and cell lines based on actual measurement data rather than only from a description in a journal article. Second, a search for tissue-specific genes can be performed simply by clicking on the appropriate tissue icon at the top of the RefEx page. Third, users can compare differences in gene expression levels related to the use of different experimental methods.

Currently, transcripts, which are based on RefSeq mRNA records, are used to integrate different types of measurement methods for gene expression. However, according to a high-throughput sequencing data analysis, over 90% of human genes undergo alternative splicing (Pan et al. 2008), and many of these are not yet included in RefSeq. To address this limitation, the definitions of the transcripts need to be redefined to include noncoding RNA in tight collaboration with the FANTOM project. In the upcoming version of RefEx, it is planned to use personalized gene



refex.dbcls.jp/genecart.php?lang=en&db=human&cart=list:NM_000

RefEx - Search Result | RefEx - Search Result | RefEx - Gene List

Comparison View ☒Close

Gene Name	serpin peptidase inhibitor, clade D (heparin cofactor), member 1	serpin peptidase inhibitor, clade A (alpha-1 antiproteinase, antitrypsin), member 10	coagulation factor XIII, B polypeptide
Synonym	D22S673, HC-II, LS2, Heparin cofactor II, heparin cofactor II, clade D (heparin cofactor), HCF2, HCII, SERPIND1, Protease inhibitor leuserpin 2, Heparin cofactor 2 precursor, HLS2	PZI, UNQ707/PRO1358, ZPI, SERPINA10, PZ-dependent protease inhibitor, Protein Z-dependent protease inhibitor precursor, Serpin A10	F13B, coagulation factor XIII, B polypeptide, Coagulation factor XIII B chain precursor, Fibrin-stabilizing factor B subunit, FXIIIB, Transglutaminase B chain
Choropleth map on BodyParts3D			
Relative Value			
Refseq ID	NM_000185	NM_001100607	NM_001994
Gene ID	3053	51156	2165
Unigene ID	Hs.474270	Hs.118620	Hs.435782
probe set ID	205576_at	220626_at	207810_at
Ensembl ID	ENSG00000099937	ENSG00000140093	ENSG00000143278
Chromosomal Region	22.q11.21 [21128167 - 21142008]	14.q32.13 [94749650 - 94759608]	1.q31.3 [197008321 - 197036397]
Gene Family (Interpro ID)	Protease inhibitor I -	- -	- Sushi/SCR/CCP
Gene Ontology Biological Process	blood coagulation chemotaxis negative regulation ... regulation of proteo ...	blood coagulation - negative regulation ... regulation of proteo ...	blood coagulation - - -
Gene Ontology Cellular Component	blood coagulation	blood coagulation	blood coagulation
Gene Ontology Molecular Function	blood coagulation chemotaxis negative regulation ...	blood coagulation - negative regulation ...	- - -

Fig. 6.6 Intermediate use-case. Comparison view. Up to three genes can be compared simultaneously. Users can compare all detailed information in parallel. The expression data and the overlapped annotated terms from Gene Ontology and the InterPro gene family are arranged in the same row

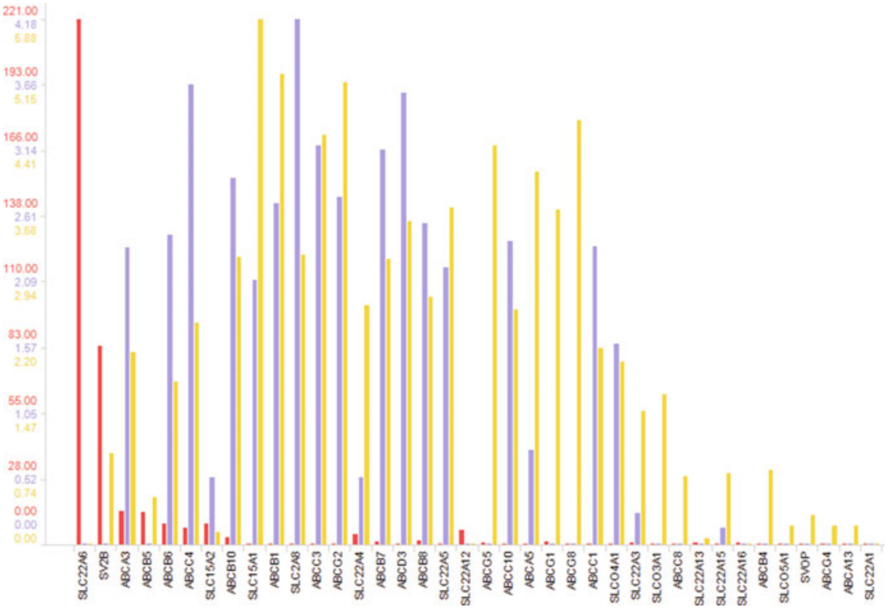


Fig. 6.7 Advanced use-case. The use-case in the comparison of *B. mori* midgut (red), Human small intestine (yellow), and Caco-2 cells (purple)

expression data from the Genotype-Tissue Expression database (GTEx) (GTEx Consortium et al. 2015).

Funding This work was supported by the Integrated Database Project of the Ministry of Education, Culture, Sports, Science and Technology (MEXT) of Japan and the National Bioscience Database Center (NBDC) of the Japan Science and Technology Agency (JST). Computations were partially performed on the NIG supercomputer at the ROIS National Institute of Genetics.

References

Ashburner M et al (2000) Gene ontology: tool for the unification of biology. Nat Genet 25:25–29

Barrett T et al (2013) NCBI GEO: archive for functional genomics data sets—update. Nucleic Acids Res 41:D991–D995

Cochrane G, Karsch-Mizrachi I, Takagi T (2016) The International Nucleotide Sequence Database Collaboration. Nucleic Acids Res 44:D48–D50

Gautier L et al (2004) Affy—analysis of Affymetrix GeneChip data at the probe level. Bioinformatics 20:307–315

Gentleman RC et al (2004) Bioconductor: open software development for computational biology and bioinformatics. Genome Biol 5:R80

GTEx Consortium et al (2015) Human genomics. The Genotype-Tissue Expression (GTEx) pilot analysis: multitissue gene regulation in humans. Science 348:648–660

Ichino F et al (2018) Construction of a simple evaluation system for the intestinal absorption of an orally administered medicine using Bombyx mori larvae. Drug Discov Ther 12(1):7–15

Irizarry RA et al (2003) Exploration, normalization, and summaries of high density oligonucleotide array probe level data. Biostatistics 4:249–264

Kadota K et al (2006) ROKU: a novel method for identification of tissue-specific genes. BMC Bioinformatics 7:294

Kodama Y, Shumway M, Leinonen R (2012) International Nucleotide Sequence Database Collaboration. The Sequence Read Archive: explosive growth of sequencing data. Nucleic Acids Res 40:D54–D56

Kolesnikov N et al (2015) ArrayExpress update—simplifying data submissions. Nucleic Acids Res 43:D1113–D1116

Lizio M et al (2015) Gateways to the FANTOM5 promoter level mammalian expression atlas. Genome Biol 16:22

Magdeldin S et al (2012) Murine colon proteome and characterization of the protein pathways. BioData Mining 5:11

Mitchell A et al (2015) The InterPro protein families database: the classification resource after 15 years. Nucleic Acids Res 43:D213–D221

Mitsuhashi N et al (2009) BodyParts3D: 3D structure database for anatomical concepts. Nucleic Acids Res 37:D782–D785

Mizuno H, Kitada K, Nakai K, Sarai A (2009) PrognoScan: a new database for meta-analysis of the prognostic value of genes. BMC Med Genet 2:18

Mizutani T et al (2016) Heterogeneous filament network formation by myosin light chain isoforms effects on contractile energy output of single cardiomyocytes derived from human induced pluripotent stem cells. Regen Ther 3:90–96

Nishio S et al (2015) Gene expression profiles of the cochlea and vestibular endorgans: localization and function of genes causing deafness. Ann Otol Rhinol Laryngol 124:6S–48S

Ogasawara O et al (2006) BodyMap-Xs: anatomical breakdown of 17 million animal ESTs for cross-species comparison of gene expression. Nucleic Acids Res 34:D628–D631

Okubo K et al (1992) Large scale cDNA sequencing for analysis of quantitative and qualitative aspects of gene expression. Nat Genet 2:173–179

Okubo S et al (2016) Albumin and apolipoprotein H mRNAs in human plasma as potential clinical biomarkers of liver injury: analyses of plasma liver-specific mRNAs in patients with liver injury. Biomarkers 21:353–362

Ono H, Ogasawara O, Okubo K, Bono H (2017) RefEx, a reference gene expression dataset as a web tool for the functional analysis of genes. Sci Data 4:170105

Pan Q et al (2008) Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing. Nat Genet 40:1413–1415

Petryszak R et al (2016) Expression Atlas update—an integrated database of gene and protein expression in humans, animals and plants. Nucleic Acids Res 44:D746–D752

Shiraki T et al (2003) Cap analysis gene expression for high-throughput analysis of transcriptional starting point and identification of promoter usage. Proc Natl Acad Sci U S A 100:15776–15781

Sudmant PH, Alexis MS, Burge CB (2015) Meta-analysis of RNA-seq expression data across species, tissues and studies. Genome Biol 16:287

The FANTOM Consortium & the RIKEN PMI and CLST (DGT) (2014) A promoter-level mammalian expression atlas. Nature 507:462–470

Trapnell C, Pachter L, Salzberg SL (2009) TopHat: discovering splice junctions with RNA-Seq. Bioinformatics 25:1105–1111

Trapnell C et al (2010) Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. Nat Biotechnol 28:511–515

Wagner L, Agarwala R (2013) The NCBI handbook, 2nd edn. National Center for Biotechnology Information

Wu C et al (2016) BioGPS: building your own mash-up of gene annotations and expression profiles. Nucleic Acids Res 44:D313–D316

Zeng L et al (2015) Aberrant IDH3α expression promotes malignant tumor growth by inducing HIF-1-mediated metabolic reprogramming and angiogenesis. Oncogene 34:4758–4766

Chapter 7
The Mouse Gene Expression Database (GXD)

Martin Ringwald, James A. Kadin, and Joel E. Richardson

Keywords Bioinformatics · Database · Development · Gene expression · Mouse

7.1 Preamble

The Gene Expression Database (GXD) is an extensive, highly curated, and freely available community resource of mouse gene expression information. Its primary emphasis is on endogenous gene expression during mouse development. GXD integrates data from RNA in situ hybridization, immunohistochemistry, knock-in reporter, RT-PCR, northern blot, and western blot experiments for wild-type and mutant mice. These data are generated by the research community worldwide and are systematically collected through curation of the literature, electronic data submissions, and collaborations with projects that produce these types of data at a large scale. GXD curators annotate all these expression data in standardized ways using official genetic nomenclature, controlled vocabularies, and an extensive anatomical ontology. As a major component of the larger Mouse Genome Informatics (MGI) resource, GXD integrates its expression data closely with genotype, functional, phenotype, and disease-oriented data, thus enabling users to search for expression data and images using a large variety of biologically and biomedically relevant parameters. A critical bridge between genotype and phenotype data, the expression information provided by GXD fosters insights into the molecular mechanism of mammalian development, differentiation, and disease. Target user groups include scientists pursuing basis research and/or clinical/translational research, as well as computational biologists and bioinformaticians.

M. Ringwald (✉) · J. A. Kadin · J. E. Richardson
The Jackson Laboratory, Bar Harbor, ME, USA
e-mail: Martin.Ringwald@jax.org; James.Kadin@jax.org; Joel.Richardson@jax.org

© The Author(s), under exclusive license to Springer Nature Singapore Pte Ltd. 2021

I. Abugessaisa, T. Kasukawa (eds.), *Practical Guide to Life Science Databases*,
https://doi.org/10.1007/978-981-16-5812-9_7

7.2 Database Overview

Recent technological advances have made it possible to rapidly determine the sequence of individual human genomes and to correlate genetic mutations with human diseases. This has made it more important than ever to gain insights into the molecular mechanisms that lead from genomic mutations to diseases. Expression analysis of all developmental stages in the mouse, and in mouse mutants modeling human mutations, has been and is being used to address this critical issue. However, these expression data are voluminous, complex, and heterogeneous. They are generated by many different laboratories and scattered through tens of thousands of publications. Without the help of centralized databases, it is impossible to keep abreast of all this information, much less to access and search these data in an integrated way. Since its first release in 1998, the Gene Expression Database for Mouse Development (GXD) has provided this crucial function to the research community focusing on classical types of expression data. These include: RNA in situ hybridization, knock-in reporter, and immunohistochemistry data that provide detailed spatial information about the expression of genes at the RNA and protein level; RT-PCR experiments that can detect small amounts and small differences in transcripts; and northern blot and western blot experiments which reveal the number and length of transcripts and proteins, respectively, made from a given gene in specific tissues.

Different types of expression assays provide different but complementary insights into what transcripts and proteins are made from a given gene, and where and when these gene products are expressed. Therefore, GXD is designed as an open-ended system that can integrate different types of expression data and dynamically represent novel insights based on new data (Ringwald et al. 1994). Expression patterns are described using an extensive, hierarchically structured anatomical ontology (Hayamizu et al. 2013, 2015). In this way, expression results from assays with differing spatial resolution are recorded in a standardized and integrated manner and expression patterns can be queried at different levels of detail. Importantly, the expression data are fully integrated with the genetic, functional, phenotypic, and disease-oriented data in MGI (http://www.informatics.jax.org) to place the data in the larger biological context and to enable the combined analysis of these data (Drabkin et al. 2015; Smith et al. 2015; Smith et al. 2018a, b).

GXD is freely available at http://www.informatics.jax.org/expression.shtml. New data are added daily and made available to the public on a weekly basis. There are several software releases per year featuring new web utilities. Over the years, GXD has grown tremendously, both in terms of data content and search and display functions (Ringwald et al. 1999; Smith et al. 2014; Finger et al. 2015, 2017; Smith et al. 2018b). The following sections describe the current content and search features of GXD in more detail.

7.3 Data Structure, Data Curation, and Content of the Database

GXD records the data at the level of individual expression assays. The expression experiments captured by GXD can be divided into in situ and blot assays and into RNA and protein assays (Table 7.1a). For each assay, GXD records the gene studied, the assay type, the molecular probes and experimental conditions used, the age and genetic background of specimens, and the expression results obtained for each of these specimens, such as the time and tissue of expression, and the number and sizes of bands detected in blot assays (see examples shown in Figs. 7.1 and 7.2). By storing the expression data at this elemental level, data from different assays can be combined to provide increasingly complete information about the expression patterns of RNA and protein products made from a given gene.

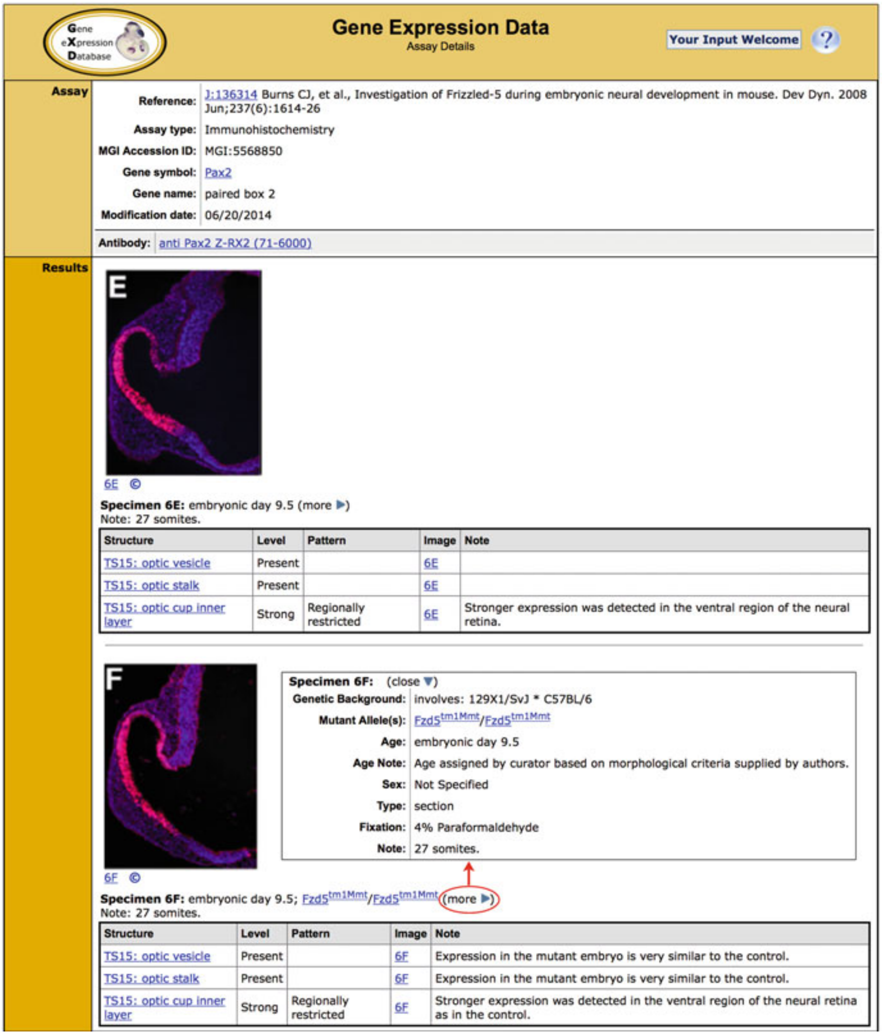
GXD's focus is on endogenous gene expression in mouse strains and targeted/well-defined mutants. Not included are: (1) expression data from mutants generated through random insertions of transgenes because the genotype of these mutants is often ill-defined; (2) expression data from transgenes under ectopic promoters because they do not reflect the endogenous expression; and (3) expression data from animals that have been treated with drugs or other substances or exposed to environmental challenges.

Data are acquired from the literature, from electronic data submissions of individual laboratories, and by collaborations with projects that generate the data GXD collects at a large scale.

GXD systematically surveys scientific journals to index all publications that include data on endogenous gene expression during mouse development. As a first curation step for each paper, the genes and ages analyzed and the expression assay types used are recorded. Annotations are based on the entire publication, including supplemental information, and use official nomenclature for genes and controlled vocabularies for age and assay types. This index, combined with bibliographic information from PubMed, is made available for searches via the Gene Expression Literature Search (see Sect. 7.4). This is a powerful search tool because GXD's

Table 7.1 GXD assay types and current data content

(a) Assay types		
Gene product	*In situ assays*	*Blot assays*
RNA	RNA in situ hybridization In situ reporter (Knock-in)	Northern blot RT-PCR
Protein	Immunohistochemistry	Western blot
(b) Current data content		
Amount	*Data type*	
341,905	Expression images	
1,659,173	Annotated expression results	
87,645	Expression assays	
14,723	Genes	
4085	Mouse mutants with expression data	



Gene Expression Data

Assay Details

Assay

Reference:	J:136314 Burns CJ, et al., Investigation of Frizzled-5 during embryonic neural development in mouse. Dev Dyn. 2008 Jun;237(6):1614-26
Assay type:	Immunohistochemistry
MGI Accession ID:	MGI:5568850
Gene symbol:	Pax2
Gene name:	paired box 2
Modification date:	06/20/2014

Antibody: anti Pax2 Z-RX2 (71-6000)

Results

6E ©

Specimen 6E: embryonic day 9.5 (more ▶)
Note: 27 somites.

Structure	Level	Pattern	Image	Note
TS15: optic vesicle	Present		6E	
TS15: optic stalk	Present		6E	
TS15: optic cup inner layer	Strong	Regionally restricted	6E	Stronger expression was detected in the ventral region of the neural retina.

Specimen 6F: (close ▼)

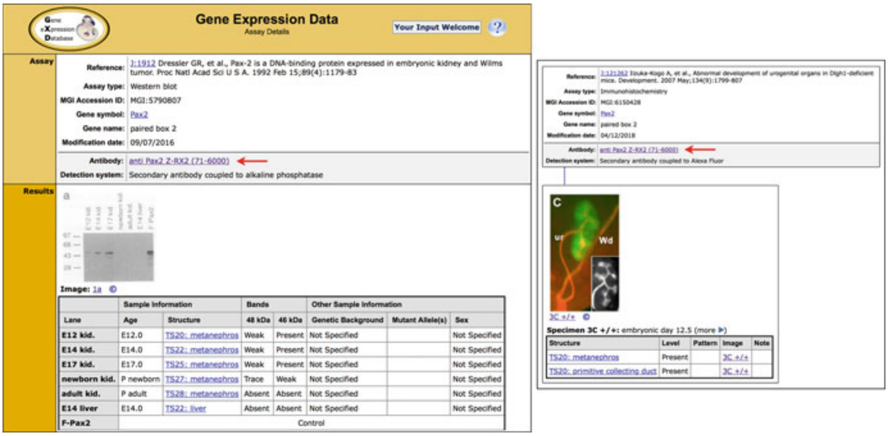
Genetic Background:	involves: 129X1/SvJ * C57BL/6
Mutant Allele(s):	$Fzd5^{tm1Mmt}/Fzd5^{tm1Mmt}$
Age:	embryonic day 9.5
Age Note:	Age assigned by curator based on morphological criteria supplied by authors.
Sex:	Not Specified
Type:	section
Fixation:	4% Paraformaldehyde
Note:	27 somites.

6F ©

Specimen 6F: embryonic day 9.5; $Fzd5^{tm1Mmt}/Fzd5^{tm1Mmt}$ (more ▶)
Note: 27 somites.

Structure	Level	Pattern	Image	Note
TS15: optic vesicle	Present		6F	Expression in the mutant embryo is very similar to the control.
TS15: optic stalk	Present		6F	Expression in the mutant embryo is very similar to the control.
TS15: optic cup inner layer	Strong	Regionally restricted	6F	Stronger expression was detected in the ventral region of the neural retina as in the control.

Fig. 7.1 Example of a detailed data record. An Assay Details Page for an immunohistochemistry (in situ type) assay is shown as an example. The Assay section lists the reference from which the data were derived, the assay type, the gene studied, and the antibody used, with links to more information about the antibody. The Results section describes the specimens used in the assay. Specimen information includes the age and, in the case of mutants, mutant allele information using standard genetic nomenclature. Further specimen details can be viewed by expanding the "more" toggle (red circle and arrow). Expression results are annotated to each specimen: the developmental stage (Theiler stage) and anatomical structure that was analyzed, using terms from the anatomical ontology; the level and patterns of expression, as described by the authors; a reference to the image; and additional notes when applicable. Images are displayed together with the annotations whenever possible



Gene Expression Data
Assay Details

Assay

Reference: J:1912 Dressler GR, et al., Pax-2 is a DNA-binding protein expressed in embryonic kidney and Wilms tumor. Proc Natl Acad Sci U S A. 1992 Feb 15;89(4):1179-83
Assay type: Western blot
MGI Accession ID: MGI:5790807
Gene symbol: Pax2
Gene name: paired box 2
Modification date: 09/07/2016

Antibody: anti Pax2 Z-RX2 (71-6000)
Detection system: Secondary antibody coupled to alkaline phosphatase

Results

Image: 1a ©

	Sample Information		Bands		Other Sample Information		
Lane	Age	Structure	48 kDa	46 kDa	Genetic Background	Mutant Allele(s)	Sex
E12 kid.	E12.0	TS20: metanephros	Weak	Present	Not Specified		Not Specified
E14 kid.	E14.0	TS22: metanephros	Weak	Present	Not Specified		Not Specified
E17 kid.	E17.0	TS25: metanephros	Weak	Present	Not Specified		Not Specified
newborn kid.	P newborn	TS27: metanephros	Trace	Weak	Not Specified		Not Specified
adult kid.	P adult	TS28: metanephros	Absent	Absent	Not Specified		Not Specified
E14 liver	E14.0	TS22: liver	Absent	Absent	Not Specified		Not Specified
F-Pax2	Control						

Reference: J:121262 Iizuka-Kogo A, et al., Abnormal development of urogenital organs in Dlgh1-deficient mice. Development. 2007 May;134(9):1799-807
Assay type: Immunohistochemistry
MGI Accession ID: MGI:6150428
Gene symbol: Pax2
Gene name: paired box 2
Modification date: 04/12/2018

Antibody: anti Pax2 Z-RX2 (71-6000)
Detection system: Secondary antibody coupled to Alexa Fluor

3C +/+ ©

Specimen 3C +/+: embryonic day 12.5 (more ▶)

Structure	Level	Pattern	Image	Note
TS20: metanephros	Present		3C +/+	
TS20: primitive collecting duct	Present		3C +/+	

Fig. 7.2 Integration of data. Left side: An Assay Details Page for a western (blot type) experiment is shown as an example. The Assay section provides the same information as entries for in situ studies (compare Fig. 7.1). The Results section describes the samples used and shows the annotations for each lane in the blot. The sample information includes the age, the genetic background, mutant alleles, and sex of the sample if applicable and available. The expression results include the anatomical structure of the sample and the number and sizes of detected bands and their expression levels, as described by the authors. Right side: Part of an Immunohistochemistry assay using the same antibody (red arrow) and analyzing the same developmental stage and anatomical structure (TS20: metanephros) as lane 1 of the western blot shown on the left side. This demonstrates just some of the points of data integration in GXD: the gene, the probe (antibody), and the anatomy. Although the three experiments in Figs. 7.1 and 7.2 are from different publications, they examine expression of the same gene and use the same antibody. These examples illustrate how GXD can provide increasingly complete expression information for a given gene by integrating different types of expression data from many different sources

literature index records are comprehensive and up-to-date from 1990 to the present. GXD has records for more than 26,600 references and nearly 16,000 genes.

The Gene Expression Literature Index described above helps GXD Curators to prioritize papers for detailed expression annotation. In this second curation step, the expression data are annotated in detailed, standardized ways, as illustrated in Figs. 7.1 and 7.2, by employing standard nomenclature for genes, mouse strains, and alleles; controlled vocabularies; and an extensive hierarchically structured anatomical ontology to describe the time and space of gene expression. The annotations are complemented with the original image data whenever possible. If GXD does not have the permission to include an image, a reference to the corresponding figure in the publication is provided to guide users to the original data. The annotation of expression patterns relies on the authors' description within the paper. Because the authors are in a much better position to interpret their own data, GXD Curators do not interpret the images presented within the paper.

Data acquired via electronic data submissions and through collaboration with large-scale projects are annotated using the same standards as employed for literature curation. These data are usually bulk-loaded into a staging database for computational and manual review. Computational checks are performed, for example, to make sure that molecular probes are assigned to the correct genes on the latest genome assembly. Issues of missing or ambiguous data are resolved in collaborations with these laboratories, and annotations provided by them are mapped to controlled vocabularies and ontologies. Once the data have been cleaned and reviewed, they are imported into GXD and made publicly available.

Due to these data acquisition and curation efforts, GXD's content has increased tremendously over the years. As of October 2018, the database contains detailed expression data for over 14,700 genes, covering data from numerous strains of wild-type mice and from more than 4000 mouse mutants (Table 7.1b). GXD now holds more than 342,000 images and over 1.66 million expression result annotations (as defined in Figs. 7.1 and 7.2).

The extensive and standardized data annotation enables close data integration within GXD and the larger MGI system. Expression Assays, References, Genes, Probes, Specimens, Mutant Alleles, Anatomical Terms, and Images are all accessioned objects (with MGI IDs) that serve as data integration points. For example, the gene object ties the data for an expression assay to all the other data for that gene in MGI (such as genetic, functional, and phenotypic data); the probe object connects all the experiments that use the same probe; a given anatomy term links all expression data annotated to that term; and a mutant allele object ties the data to all the other data for this specific mutation. These connections within GXD and MGI result in powerful search capabilities.

7.4 Database Access and Mining Methods

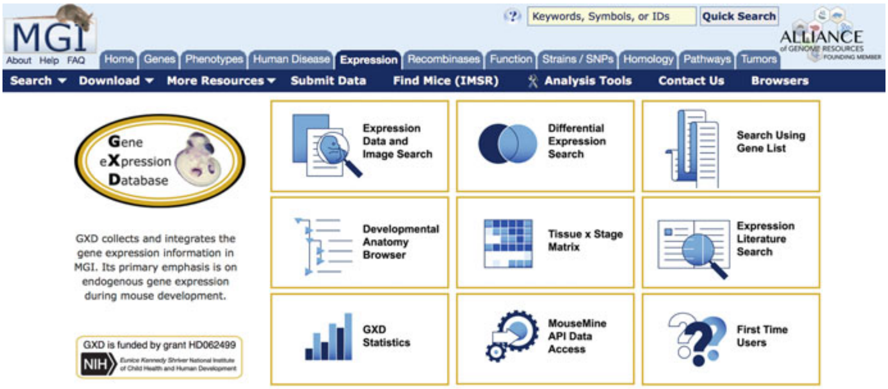
GXD's data are accessible via web-based search interfaces, web services APIs, and ftp. As the majority of our users are biomedical researchers who search the database via the web interface, we will focus on this mode of access first.

The GXD Home Page at http://www.informatics.jax.org/expression.shtml is the best starting point to explore GXD (Fig. 7.3). Graphical tiles provide a quick overview of and access to GXD's search functions. A Highlights section alerts users of newly added features and data. The "About GXD" button provides access to general information about the database. "Fast Track Your Data" leads to guidelines for electronic data submissions.

A quick and effective way to become familiar with GXD's utilities is the following:

Click on the *First Time Users* tile to see a one-page flow chart that illustrates how the overall search interface works (http://www.informatics.jax.org/mgihome/GXD/FirstTimeUsers.shtml).

Then explore the different search functions by clicking on the following tiles:



GXD collects and integrates the gene expression information in MGI. Its primary emphasis is on endogenous gene expression during mouse development.

GXD is funded by grant HD062499

Fig. 7.3 Navigation tiles on the GXD Home Page. The First Time Users tile gives a quick one-page overview of GXD's web interface. Other tiles provide access to GXD's search functions. The GXD Home Page is accessible at http://www.informatics.jax.org/expression.shtml or via the Expression tab at MGI

Expression Data and Image Search: leads to the Standard Search of the Gene Expression Data Query Form (http://www.informatics.jax.org/gxd; Fig. 7.6) which allows you to search for expression data and images using many different parameters (as discussed in Sect. 7.5.2).

Differential Expression Search (http://www.informatics.jax.org/gxd/differential): allows you to search for genes expressed in some anatomical structures and/or developmental stages but not in others; or to search for genes that have been shown to be expressed in a specific anatomical structure (or its substructures) and nowhere else.

Search Using Gene List: leads to the Batch Search of the Gene Expression Data Query Form (http://www.informatics.jax.org/gxd/batchSearch) which lets you use lists of genes to retrieve expression data.

Developmental Anatomy Browser (http://www.informatics.jax.org/vocab/gxd/anatomy/; Fig. 7.5): allows you to navigate the anatomical ontology, locate a specific anatomical structure, and to obtain the expression data and phenotype data associated with that structure and its substructures (as discussed in Sect. 7.5.1).

Tissue x Stage Matrix (http://www.informatics.jax.org/gxd/tissue_matrix): displays all of GXD's expression data in a tissue-by-developmental stage matrix. Starting with a high-level overview of all of GXD's data, you can interactively view and select expression data for specific tissues and/or developmental stages.

All the search and browse functions described above return the search results in a common format, as a multi-tabbed summary. Each tab displays the results in a different view: at the level of genes, assays, assay results, and images, and in interactive tissue-by-stage and tissue-by-gene matrix views (see Figs. 7.7 and 7.8). The summaries can be filtered and sorted to refine the data sets. Links provided in the summaries lead to detailed expression entries such as the ones shown in Figs. 7.1 and 7.2.

The *Expression Literature Search* (http://www.informatics.jax.org/gxdlit) provides access to the GXD's Literature content records. It allows you to quickly find all publications that contain expression data for specific mouse genes, assays, ages, and authors. Searches are more complete and more specific than, for example, PubMed searches because the full text of publications, including supplemental data, is indexed by GXD using standard gene nomenclature and controlled vocabularies.

Via the *MouseMine API Data Access* tile (http://www.mousemine.org/mousemine/begin.do#Expression) users can get programmatic access to GXD data via MouseMines's web services API, as described in Sect. 7.5.3.

FTP Access: Several expression database reports are generated weekly and are available at http://www.informatics.jax.org/downloads/reports/index.html#expression.

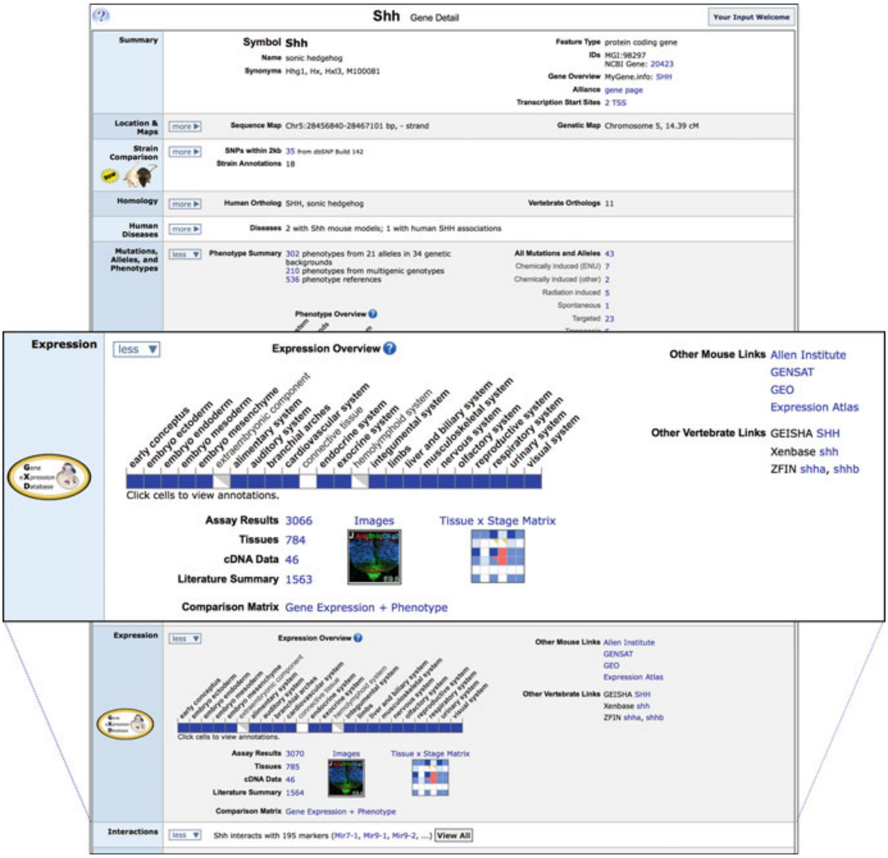
User Support: Dedicated User Support is provided. User Support can be contacted via email at mgi-help@jax.org or by clicking the "Contact Us" link in the navigation bar of all MGI web pages.

7.5 Use Cases and Demo to Use the Database

7.5.1 Simple Use Cases

7.5.1.1 Where and When Is a Given Gene Expressed?

The quickest way to find the expression data for a gene of interest is to review the Expression Section of its MGI Gene Detail Page (Fig. 7.4). To get to the desired Gene Detail page, use the quick search box (available on all MGI pages) and follow the gene link on the summary return. On the Gene Detail page, scroll down to the expression section. This section provides an overview of the gene's expression pattern via a grid of high-level anatomical structures and access to a Tissue x Stage Matrix to interactively explore the expression pattern of the gene. Links to all expression images for the gene and to other expression-related summaries are also available. In addition, there is a link to the Gene Expression + Phenotype Comparison Matrix. This matrix visually juxtaposes tissues where the gene is expressed against tissues where mutations in the gene cause abnormalities and thus enables the anatomical comparison of expression and phenotype data for the gene in MGI. Gene-based links to expression data in external resources are also provided. These include links to mouse data at the Allen Institute (Sunkin et al. 2013), GENSAT (Schmidt et al. 2013), GEO (Barrett et al. 2013), and the EBI Expression Atlas (Papatheodorou et al. 2018). There are also links to expression information for the orthologous genes in zebrafish at Zfin (Howe et al. 2017), xenopus at Xenbase (Karimi et al. 2018), and chicken at GEISHA (Antin et al. 2014), other important vertebrate model organisms used in developmental research.



Shh Gene Detail

Your Input Welcome

Summary

Symbol Shh
Name sonic hedgehog
Synonyms Hhg1, Hx, Hxl3, M100081

Feature Type protein coding gene
IDs MGI:98297
NCBI Gene: 20423
Gene Overview MyGene.info: SHH
Alliance gene page
Transcription Start Sites 2 TSS

Location & Maps — more ▶ — Sequence Map Chr5:28456840-28467101 bp, - strand — Genetic Map Chromosome 5, 14.39 cM

Strain Comparison — more ▶ — SNPs within 2kb 35 from dbSNP Build 142
Strain Annotations 18

Homology — more ▶ — Human Ortholog SHH, sonic hedgehog — Vertebrate Orthologs 11

Human Diseases — more ▶ — Diseases 2 with Shh mouse models; 1 with human SHH associations

Mutations, Alleles, and Phenotypes — less ▼ — Phenotype Summary 302 phenotypes from 21 alleles in 34 genetic backgrounds
210 phenotypes from multigenic genotypes
536 phenotype references

Phenotype Overview

All Mutations and Alleles 43
Chemically induced (ENU) 7
Chemically induced (other) 2
Radiation induced 5
Spontaneous 1
Targeted 23

Expression — less ▼

Expression Overview

early conceptus
embryo ectoderm
embryo endoderm
embryo mesoderm
embryo mesenchyme
extraembryonic component
alimentary system
auditory system
branchial arches
cardiovascular system
connective tissue
endocrine system
exocrine system
hemolymphoid system
integumental system
limbs
liver and biliary system
musculoskeletal system
nervous system
olfactory system
reproductive system
respiratory system
urinary system
visual system

Click cells to view annotations.

Assay Results 3066
Tissues 784
cDNA Data 46
Literature Summary 1563

Images

Tissue x Stage Matrix

Comparison Matrix Gene Expression + Phenotype

Other Mouse Links Allen Institute
GENSAT
GEO
Expression Atlas

Other Vertebrate Links GEISHA SHH
Xenbase shh
ZFIN shha, shhb

Expression — less ▼

Expression Overview

Click cells to view annotations.

Assay Results 3070
Tissues 785
cDNA Data 46
Literature Summary 1564

Images

Tissue x Stage Matrix

Comparison Matrix Gene Expression + Phenotype

Other Mouse Links Allen Institute
GENSAT
GEO
Expression Atlas

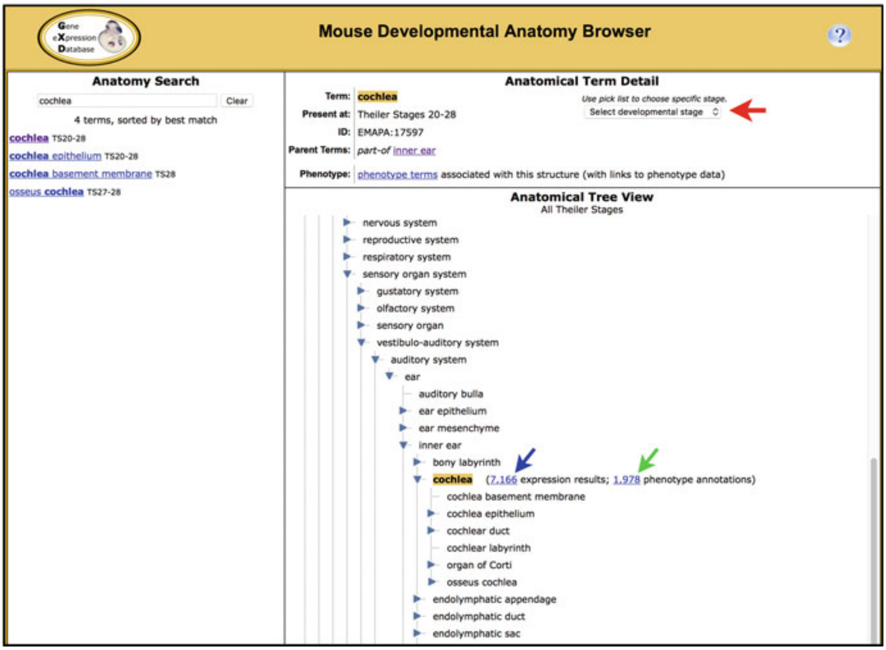
Other Vertebrate Links GEISHA SHH
Xenbase shh
ZFIN shha, shhb

Interactions — less ▼ — Shh interacts with 195 markers (Mir7-1, Mir9-1, Mir9-2, ...) View All

Fig. 7.4 The Expression Section of the MGI Gene Detail Page summarizes the available expression data. The summary for the gene *Shh* is shown. A grid of high-level anatomical terms gives an overview of the expression pattern of the gene. Blue cells indicate that expression was detected in wild-type embryos. Gray triangles indicate either absence of expression in wild-type embryos or expression data from mutant embryos. Clicking on an individual cell in the grid leads to the Tissue-by-Stage matrix for the corresponding anatomical structure (such as the one shown in Fig. 7.8). Clicking on the Tissue-by-Stage Matrix icon leads to an interactive matrix display for all high-level anatomical structures. Links to all the expression images for the gene, to all assay results, and to other summaries are also provided. In addition, the expression section features gene-based links to external resources (right side). The upper part of the entire Gene Detail page is shown in the background. Gene Detail pages are central data hubs, summarizing and providing access to all the data for a given gene in MGI, as well as extensive gene-based links to external resources

7.5.1.2 What Genes Are Expressed in a Given Tissue/Anatomical Structure?

A good approach for answering this question is to use the Mouse Developmental Anatomy Browser (Fig. 7.5). By searching or browsing the mouse developmental anatomy ontology, one can locate the anatomical structure of interest and look up the



Anatomy Search

cochlea [Clear]

4 terms, sorted by best match

- cochlea TS20-28
- cochlea epithelium TS20-28
- cochlea basement membrane TS28
- osseus cochlea TS27-28

Anatomical Term Detail

Term: cochlea

Present at: Theiler Stages 20-28

ID: EMAPA:17597

Parent Terms: *part-of* inner ear

Use pick list to choose specific stage.

Select developmental stage

Phenotype: phenotype terms associated with this structure (with links to phenotype data)

Anatomical Tree View

All Theiler Stages

```
▶ nervous system
▶ reproductive system
▶ respiratory system
▼ sensory organ system
  ▶ gustatory system
  ▶ olfactory system
  ▶ sensory organ
  ▼ vestibulo-auditory system
    ▼ auditory system
      ▼ ear
          auditory bulla
        ▶ ear epithelium
        ▶ ear mesenchyme
        ▼ inner ear
          ▶ bony labyrinth
          ▼ cochlea   (7,166 expression results; 1,978 phenotype annotations)
              cochlea basement membrane
            ▶ cochlea epithelium
            ▶ cochlear duct
              cochlear labyrinth
            ▶ organ of Corti
            ▶ osseus cochlea
          ▶ endolymphatic appendage
          ▶ endolymphatic duct
          ▶ endolymphatic sac
```

Fig. 7.5 Using the Mouse Developmental Anatomy Browser. A search for the anatomical structure “cochlea” is illustrated. The Anatomical Tree View displays the search term “cochlea,” highlighted, in the hierarchical context of the anatomy. Links to associated expression data (blue arrow) lead to summaries like the ones shown in Figs. 7.7 and 7.8. Links to mouse phenotype data associated with cochlea (green arrow) are provided as well. Users can explore the anatomy ontology further by collapsing and expanding branches of the tree view. Clicking on another anatomical term in the tree view will select and highlight that term and links to associated expression and phenotype data will be displayed. The initial tree view shows the “abstract” version of the anatomy (Hayamizu et al. 2015) that covers all developmental stages. The Anatomical Term Detail section indicates at which developmental stages a selected anatomical structure exists. Accordingly, the associated expression results will include the annotations for all the stages at which the selected structure is present. By using the drop-down to select a specific developmental stage (in the Anatomical Term Detail section, red arrow) one can obtain the corresponding stage-specific tree view and associated stage-specific expression data

expression data (and phenotype data) for this structure and its substructures. The stages of interest can also be specified. It can be the entire developmental stage range, during which the selected anatomical structure exists, or a specific developmental stage within that range.

7.5.2 Intermediate Use Cases

7.5.2.1 Combining Search Parameters to Formulate Complex Queries

The Standard Search mode of the Gene Expression Data query form enables expression searches using one or many parameters. Thus, it can be used for the basic gene- and anatomy-based searches described above. However, it can also be used for more complex and more specific searches. An example of a multi-parameter query is provided in Fig. 7.6, illustrating a search for the expression data of "DNA-binding transcription factors" "detected" in the "eye" of "*Pax6* mutants."

Search results are represented as six tabbed summaries displaying the data at different levels of detail: at the gene, assay, assay results, and image level, and as interactive tissue-by-stage and tissue-by-gene matrix views (Figs. 7.7 and 7.8). These summaries can be refined by various sorting and filtering options and provide access to the detailed expression entries, such as those shown in Figs. 7.1 and 7.2.

7.5.2.2 Differential Expression Search and Batch Search

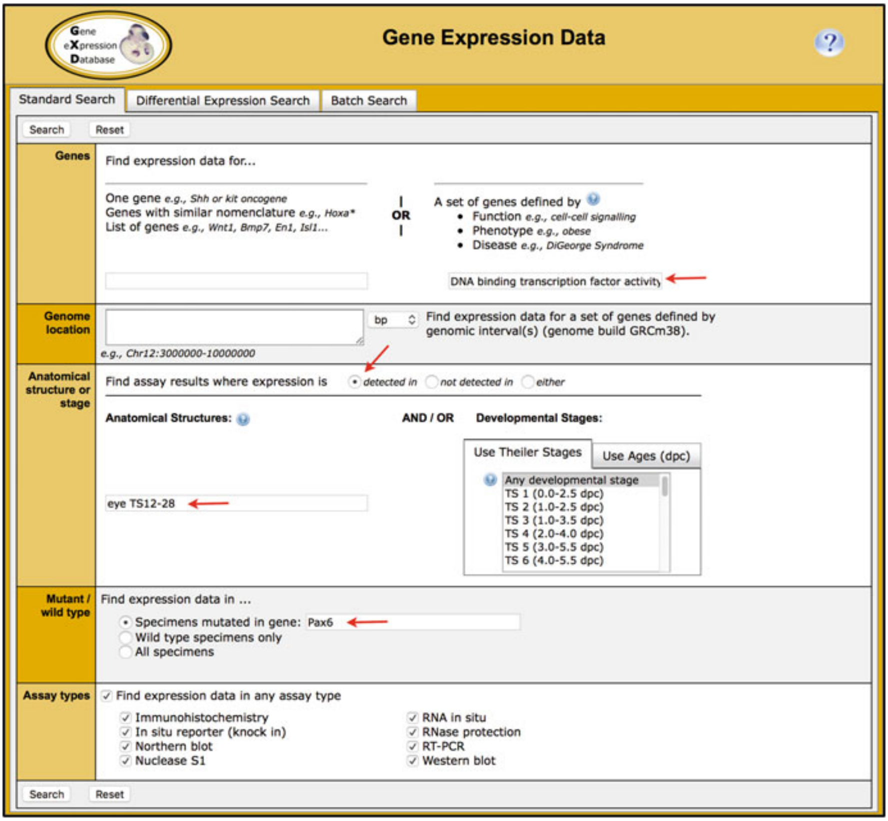
Other intermediate use cases are supported by the Differential Expression Search and Batch Search modes of the Gene Expression Data Query form. Using the Differential Expression Search, one can search for genes that are expressed in a specific anatomical structure and nowhere else. This can be very helpful, for example, for development of tissue-specific recombinase driver lines. The Batch Search is useful if one has identified a list of interesting genes, for example, through RNA-seq experiments. In this case, the Batch Search provides effective means to look up the expression information for all these genes in GXD.

7.5.3 Advanced Use Cases

Web Services Access to GXD Data via MouseMine (http://www.mousemine.org/mousemine/begin.do#Expression).

MouseMine is a data warehouse established by MGI to support access to mouse data by computational biologists and bioinformatics programmers (Motenko et al. 2015). Based on the InterMine system (Smith et al. 2012; Lyne et al. 2015), MouseMine offers a powerful query system with an interactive query builder; the ability to encapsulate queries as form-based templates; the ability to save results as lists, to combine lists using set operations, and to "plug" a list into a query; the ability to download results in various formats; and more.

MouseMine contains most of the data types in MGI, including GXD expression data, as well as other data not in MGI (e.g. interaction data from BioGrid; https://thebiogrid.org). MouseMine is updated weekly. As in MGI, the data in MouseMine



Gene eXpression Database

Gene Expression Data

Standard Search | Differential Expression Search | Batch Search

Search Reset

Genes Find expression data for...

One gene *e.g., Shh or kit oncogene*
Genes with similar nomenclature *e.g., Hoxa**
List of genes *e.g., Wnt1, Bmp7, En1, Isl1...*

OR

A set of genes defined by
- Function *e.g., cell-cell signalling*
- Phenotype *e.g., obese*
- Disease *e.g., DiGeorge Syndrome*

DNA binding transcription factor activity

Genome location bp Find expression data for a set of genes defined by genomic interval(s) (genome build GRCm38).
e.g., Chr12:3000000-10000000

Anatomical structure or stage Find assay results where expression is ● *detected in* ○ *not detected in* ○ *either*

Anatomical Structures: **AND / OR** **Developmental Stages:**

eye TS12-28

Use Theiler Stages | Use Ages (dpc)

Any developmental stage
TS 1 (0.0-2.5 dpc)
TS 2 (1.0-2.5 dpc)
TS 3 (1.0-3.5 dpc)
TS 4 (2.0-4.0 dpc)
TS 5 (3.0-5.5 dpc)
TS 6 (4.0-5.5 dpc)

Mutant / wild type Find expression data in ...
● Specimens mutated in gene: Pax6
○ Wild type specimens only
○ All specimens

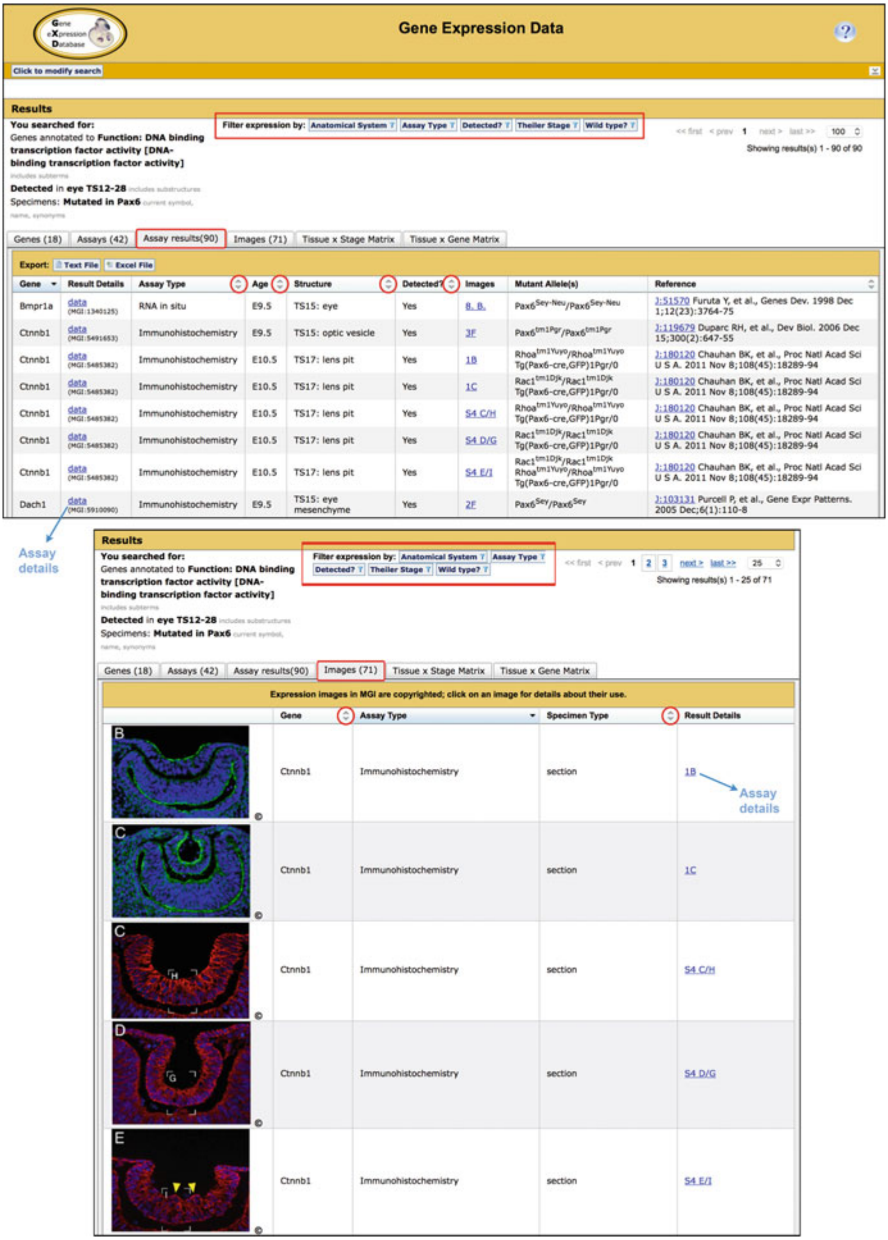
Assay types ☑ Find expression data in any assay type

☑ Immunohistochemistry	☑ RNA in situ
☑ In situ reporter (knock in)	☑ RNase protection
☑ Northern blot	☑ RT-PCR
☑ Nuclease S1	☑ Western blot

Search Reset

Fig. 7.6 The Gene Expression Data—Standard Search enables multi-parameter queries. The Genes section permits users to search by a specific gene or sets of genes based on their function [as defined by Gene Ontology terms (Drabkin et al. 2015)], their association with mouse phenotypes [as defined by Mammalian Phenotype Ontology terms (Smith and Eppig 2012)], or their association with human diseases [as defined by Disease Ontology terms (Schriml and Mitraka 2015)]. The Genome location section allows users to restrict expression searches to genes located in a specific chromosomal region, thus supporting candidate gene searches. Via the Anatomical Structure or stage section, users can search for expression data in specific anatomical structures and/or developmental stages and they can specify whether all results should be returned or only those where expression was detected (present) or not detected (absent). Anatomical searches include substructures, as defined by the anatomy ontology hierarchies. Using the Mutant/wild-type section, one can search for expression in wild-type embryos or in specific mutants. The Assay Types section allows for the selection of specific expression assay types. The illustrated search uses four parameters, asking for the expression data of genes with "DNA-binding transcription factor activity" that are "detected" in the "eye" of "*Pax6* mutant mice" (red arrows). The search results returned by this query are shown in Figs. 7.7 and 7.8

are fully integrated, enabling powerful queries that combine multiple types and sources of data. The easiest way to see an example is to run one of the predefined template queries. The most frequently used templates are organized into tabs at the bottom of the MouseMine home page. Figure 7.9 shows the home page, with the



Gene Expression Data

Results

You searched for:
Genes annotated to **Function: DNA binding transcription factor activity [DNA-binding transcription factor activity]** includes subterms
Detected in **eye TS12-28** includes substructures
Specimens: **Mutated in Pax6** current symbol, name, synonyms

Filter expression by: Anatomical System | Assay Type | Detected? | Theiler Stage | Wild type?

<< first < prev 1 next > last >> 100
Showing results(s) 1 - 90 of 90

Genes (18) | Assays (42) | Assay results(90) | Images (71) | Tissue x Stage Matrix | Tissue x Gene Matrix

Export: Text File | Excel File

Gene	Result Details	Assay Type	Age	Structure	Detected?	Images	Mutant Allele(s)	Reference
Bmpr1a	data (MGI:1340125)	RNA in situ	E9.5	TS15: eye	Yes	8, B,	$Pax6^{Sey-Neu}/Pax6^{Sey-Neu}$	J:51570 Furuta Y, et al., Genes Dev. 1998 Dec 1;12(23):3764-75
Ctnnb1	data (MGI:5491653)	Immunohistochemistry	E9.5	TS15: optic vesicle	Yes	3F	$Pax6^{tm1Pgr}/Pax6^{tm1Pgr}$	J:119679 Duparc RH, et al., Dev Biol. 2006 Dec 15;300(2):647-55
Ctnnb1	data (MGI:5485382)	Immunohistochemistry	E10.5	TS17: lens pit	Yes	1B	$Rhoa^{tm1Yuyo}/Rhoa^{tm1Yuyo}$ Tg(Pax6-cre,GFP)1Pgr/0	J:180120 Chauhan BK, et al., Proc Natl Acad Sci U S A. 2011 Nov 8;108(45):18289-94
Ctnnb1	data (MGI:5485382)	Immunohistochemistry	E10.5	TS17: lens pit	Yes	1C	$Rac1^{tm1Djk}/Rac1^{tm1Djk}$ Tg(Pax6-cre,GFP)1Pgr/0	J:180120 Chauhan BK, et al., Proc Natl Acad Sci U S A. 2011 Nov 8;108(45):18289-94
Ctnnb1	data (MGI:5485382)	Immunohistochemistry	E10.5	TS17: lens pit	Yes	S4 C/H	$Rhoa^{tm1Yuyo}/Rhoa^{tm1Yuyo}$ Tg(Pax6-cre,GFP)1Pgr/0	J:180120 Chauhan BK, et al., Proc Natl Acad Sci U S A. 2011 Nov 8;108(45):18289-94
Ctnnb1	data (MGI:5485382)	Immunohistochemistry	E10.5	TS17: lens pit	Yes	S4 D/G	$Rac1^{tm1Djk}/Rac1^{tm1Djk}$ Tg(Pax6-cre,GFP)1Pgr/0	J:180120 Chauhan BK, et al., Proc Natl Acad Sci U S A. 2011 Nov 8;108(45):18289-94
Ctnnb1	data (MGI:5485382)	Immunohistochemistry	E10.5	TS17: lens pit	Yes	S4 E/I	$Rac1^{tm1Djk}/Rac1^{tm1Djk}$ $Rhoa^{tm1Yuyo}/Rhoa^{tm1Yuyo}$ Tg(Pax6-cre,GFP)1Pgr/0	J:180120 Chauhan BK, et al., Proc Natl Acad Sci U S A. 2011 Nov 8;108(45):18289-94
Dach1	data (MGI:5910090)	Immunohistochemistry	E9.5	TS15: eye mesenchyme	Yes	2F	$Pax6^{Sey}/Pax6^{Sey}$	J:103131 Purcell P, et al., Gene Expr Patterns. 2005 Dec;6(1):110-8

Results

You searched for:
Genes annotated to **Function: DNA binding transcription factor activity [DNA-binding transcription factor activity]** includes subterms
Detected in **eye TS12-28** includes substructures
Specimens: **Mutated in Pax6** current symbol, name, synonyms

Filter expression by: Anatomical System | Assay Type | Detected? | Theiler Stage | Wild type?

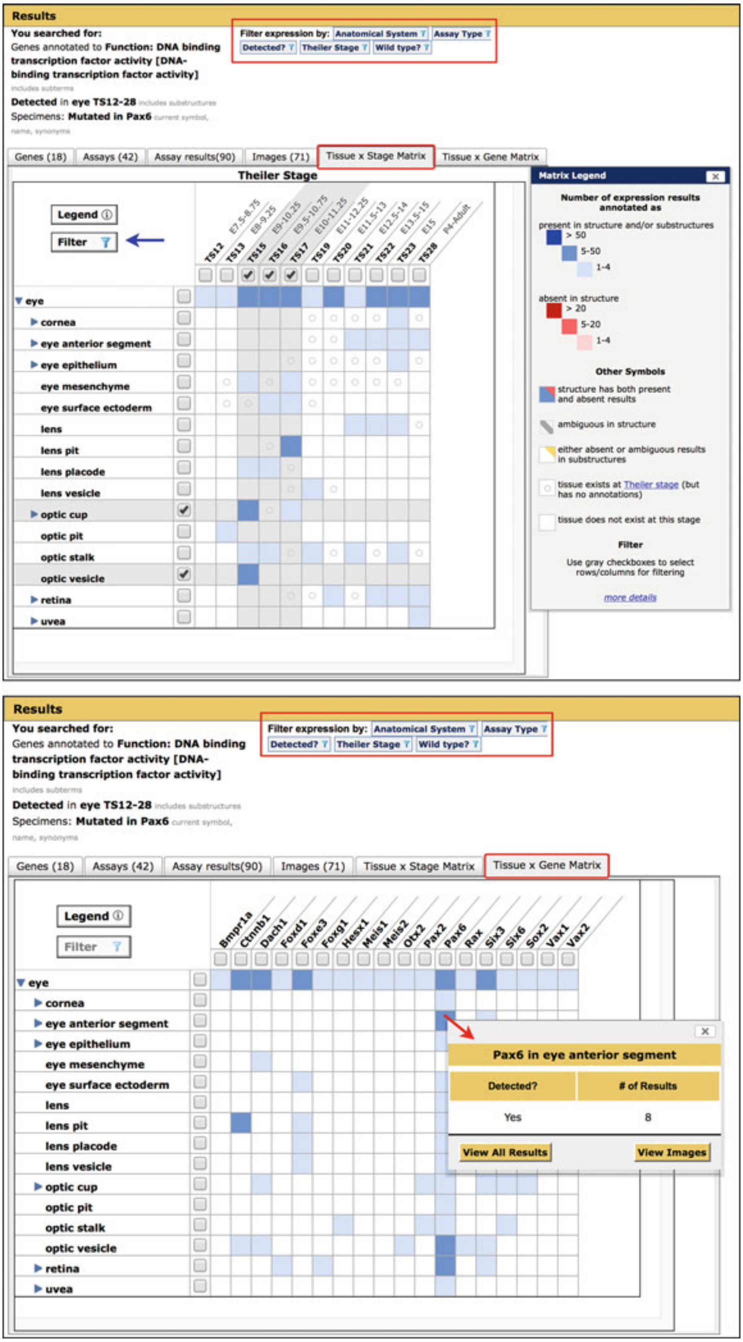
<< first < prev 1 2 3 next > last >> 25
Showing results(s) 1 - 25 of 71

Genes (18) | Assays (42) | Assay results(90) | Images (71) | Tissue x Stage Matrix | Tissue x Gene Matrix

Expression images in MGI are copyrighted; click on an image for details about their use.

	Gene	Assay Type	Specimen Type	Result Details
B ©	Ctnnb1	Immunohistochemistry	section	1B
C ©	Ctnnb1	Immunohistochemistry	section	1C
C ©	Ctnnb1	Immunohistochemistry	section	S4 C/H
D ©	Ctnnb1	Immunohistochemistry	section	S4 D/G
E ©	Ctnnb1	Immunohistochemistry	section	S4 E/I

Fig. 7.7 Search Results are returned in a multi-tabbed summary. The Assay Results summary is shown at the top, the Images summary at the bottom. Summaries can be refined further using various filtering options (red box: "Filter expression by:"). The filters are applied to the content of all the tabbed summary views. Both summaries can also be sorted as indicated by red circles. Further, the Assay Results summary provides options to export the data as text or Excel files. Both summaries link to detailed expression entries (such as the ones shown in Figs. 7.1 and 7.2); links are indicated by blue arrows



Results

You searched for:
Genes annotated to **Function: DNA binding transcription factor activity [DNA-binding transcription factor activity]** includes subterms
Detected in **eye TS12-28** includes substructures
Specimens: **Mutated in Pax6** current symbol, name, synonyms

Matrix Legend

Number of expression results annotated as

present in structure and/or substructures
- > 50
- 5-50
- 1-4

absent in structure
- > 20
- 5-20
- 1-4

Other Symbols
- structure has both present and absent results
- ambiguous in structure
- either absent or ambiguous results in substructures
- ○ tissue exists at Theiler stage (but has no annotations)
- tissue does not exist at this stage

Filter

Use gray checkboxes to select rows/columns for filtering

more details

Results

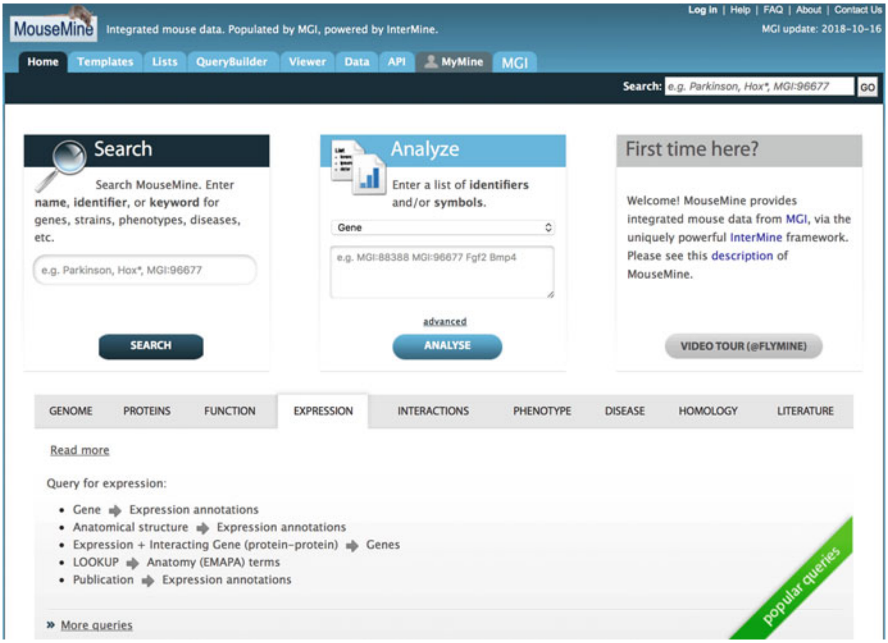
You searched for:
Genes annotated to **Function: DNA binding transcription factor activity [DNA-binding transcription factor activity]** includes subterms
Detected in **eye TS12-28** includes substructures
Specimens: **Mutated in Pax6** current symbol, name, synonyms

Pax6 in eye anterior segment

Detected?	# of Results
Yes	8

View All Results View Images

Fig. 7.8 Two search result tabs are interactive matrix views. The Tissue-by-Stage matrix is shown at the top, the Tissue-by-Gene matrix at the bottom. The anatomy axis on the left side of each can be expanded and collapsed based on the hierarchy of the anatomy ontology. Filtering can be done by selecting specific rows and columns (as indicated by check marks in top matrix) and applying the



MouseMine Integrated mouse data. Populated by MGI, powered by InterMine.

Log in | Help | FAQ | About | Contact Us

MGI update: 2018-10-16

e.g. Parkinson, Hox*, MGI:96677

Search

Search MouseMine. Enter **name, identifier,** or **keyword** for genes, strains, phenotypes, diseases, etc.

e.g. Parkinson, Hox*, MGI:96677

Analyze

Enter a list of **identifiers** and/or **symbols.**

Gene

e.g. MGI:88388 MGI:96677 Fgf2 Bmp4

First time here?

Welcome! MouseMine provides integrated mouse data from MGI, via the uniquely powerful InterMine framework. Please see this description of MouseMine.

Query for expression:

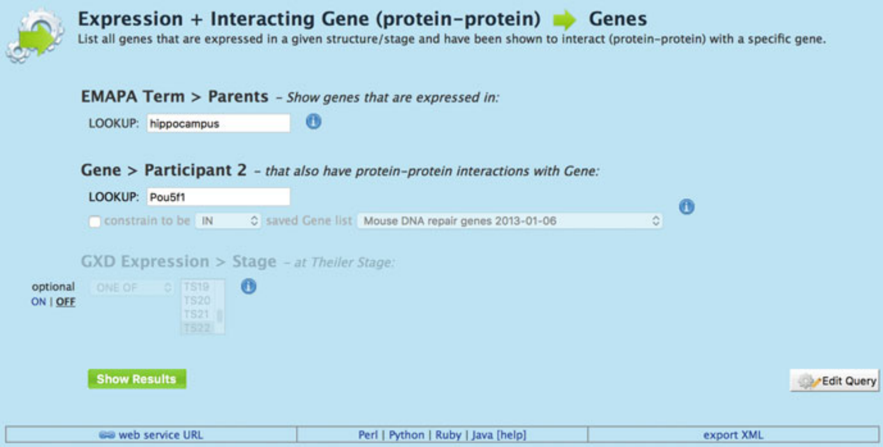
- Gene ➔ Expression annotations
- Anatomical structure ➔ Expression annotations
- Expression + Interacting Gene (protein-protein) ➔ Genes
- LOOKUP ➔ Anatomy (EMAPA) terms
- Publication ➔ Expression annotations

Fig. 7.9 The MouseMine home page, showing the expression tab. Each tab section shows the most frequently used templates for that area

Expression tab open, and Fig. 7.10 shows one of the templates, "Expression + Interacting Gene (protein-protein) → Genes," which returns genes expressed in a specified tissue that have protein–protein interactions with a specified gene. As with all template forms, this one is preloaded to run a sample query, in this case, to return genes expressed in the hippocampus that interact with POU5F1. Figure 7.11 shows the result of clicking the "Show Results" button—a paginated table of results combining expression information from GXD and interaction data from BioGrid. The genes themselves (columns 1 and 2) can be exported or saved as a list (of distinct items) to drive further queries. Other options allow the user to change the sort order, add or remove columns, download in various formats, and more.

While MouseMine offers a full-featured interactive User Interface (UI), the more significant feature is that all UI functionality is available via (indeed, depends upon) a comprehensive web services interface. Using MouseMine web services one can easily write tools and applications that can run any query, access and modify lists,

Fig. 7.8 (continued) Filter button (blue arrow), or by using the general filter options (red box: "Filter expression by:"). As indicated in the bottom matrix, clicking on an individual cell opens a small summary window of the data represented in that cell with links to the corresponding Assay Results and Images summaries that in turn link to the detailed expression entries



Expression + Interacting Gene (protein-protein) ➡ Genes

List all genes that are expressed in a given structure/stage and have been shown to interact (protein-protein) with a specific gene.

EMAPA Term > Parents – *Show genes that are expressed in:*

LOOKUP: hippocampus

Gene > Participant 2 – *that also have protein-protein interactions with Gene:*

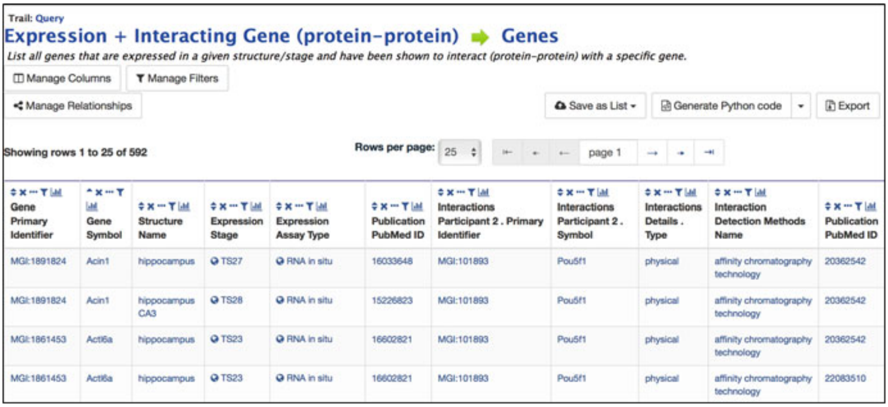
LOOKUP: Pou5f1

□ constrain to be IN saved Gene list Mouse DNA repair genes 2013-01-06

GXD Expression > Stage – *at Theiler Stage:*

optional ON | OFF ONE OF TS19 TS20 TS21 TS22

Fig. 7.10 Example template form. This template finds genes with products expressed in the specified location that interact with the specified gene's product. Every form is prepopulated to run an example query (in this case with the location "hippocampus" and the gene "*Pou5f1*"), so one can simply click "Show Results" to see how it works



Trail: Query

Expression + Interacting Gene (protein-protein) ➡ Genes

List all genes that are expressed in a given structure/stage and have been shown to interact (protein-protein) with a specific gene.

Showing rows 1 to 25 of 592

Rows per page: 25 page 1

Gene Primary Identifier	Gene Symbol	Structure Name	Expression Stage	Expression Assay Type	Publication PubMed ID	Interactions Participant 2 . Primary Identifier	Interactions Participant 2 . Symbol	Interactions Details . Type	Interaction Detection Methods Name	Publication PubMed ID
MGI:1891824	Acin1	hippocampus	TS27	RNA in situ	16033648	MGI:101893	Pou5f1	physical	affinity chromatography technology	20362542
MGI:1891824	Acin1	hippocampus CA3	TS28	RNA in situ	15226823	MGI:101893	Pou5f1	physical	affinity chromatography technology	20362542
MGI:1861453	Actl6a	hippocampus	TS23	RNA in situ	16602821	MGI:101893	Pou5f1	physical	affinity chromatography technology	20362542
MGI:1861453	Actl6a	hippocampus	TS23	RNA in situ	16602821	MGI:101893	Pou5f1	physical	affinity chromatography technology	22083510

Fig. 7.11 The results from running the query are shown in Fig. 7.10. Query results are presented in tables that offer many options for dynamically adjusting the query, saving lists, downloading, etc

export data in numerous formats, etc. The full set of endpoints is documented and available for experimentation at http://iodocs.apps.intermine.org/mgi/docs#/. For this chapter, we will give a small example. For more details on InterMine web services API, see Kalderimis et al. (2014). For help using MouseMine's API, contact mgi-help@jax.org.

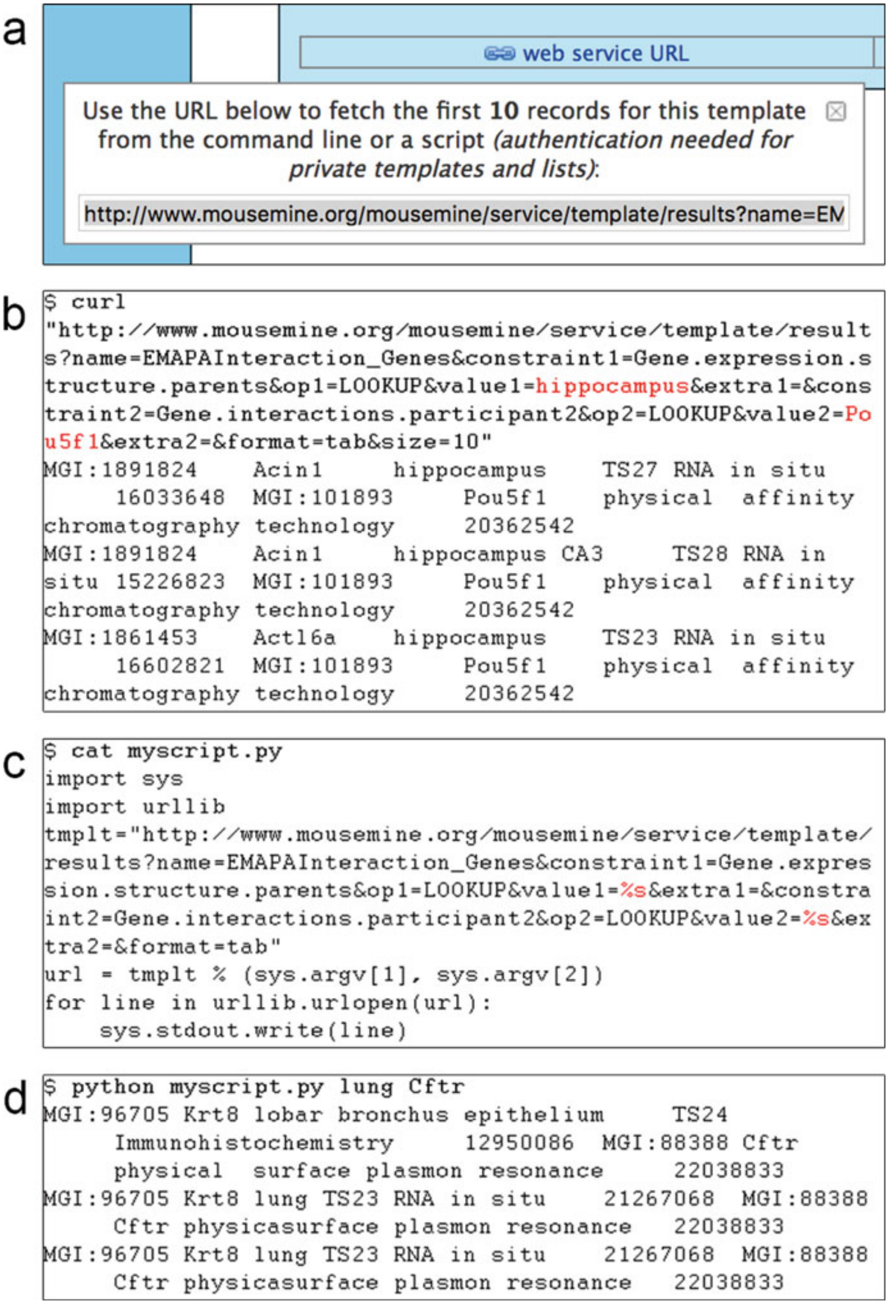
Rather than create an application from scratch, one can take advantage of a particularly developer-friendly feature of the UI: its ability to generate working code. For example, at the bottom of every template form is a series of links. Figure 7.12a shows one: it provides a URL to run the current template with the current parameters and return the first 10 results in tab-delimited format. As shown in Fig. 7.12b, one can simply pass this URL to curl or other similar tools to download the results on the command line. By adjusting the parameters (highlighted in red), one can retrieve data for a different gene, return data in a different format (e.g., json), or adjust the number of results (to return all results, remove the size parameter). With a tiny bit of work, one can wrap the generated URL in a python script that runs this template with parameters the user enters on the command line (Fig. 7.12c, d).

Users may access the API directly, by constructing appropriate URLs as in this example, or via client libraries in one of the several languages. Other links at the bottom center of the form in Fig. 7.10 generate working applications that use these libraries. For example, clicking the "Python" link generates the output in Fig. 7.13, which runs the same query as the URL from Fig. 7.12a, but using the Python client library.

7.6 Summary and Future Developments of the Database

GXD collects and integrates different types of mouse expression data generated by the research community worldwide and makes these data freely and widely available. The expression data can be interrogated from many biological and biomedical perspectives. Basic and complex query capabilities are provided via web-based search forms and browsers and through programmatic access. So far, GXD has focused on classical types of expression data, such as in situ and blot data. GXD will continue to acquire and curate these important data. In addition, the database will be expanded to capture array and RNA-seq expression data. By integrating different types of expression data that provide complementary expression information, GXD will continue to provide increasingly complete information about what RNA and protein products are made from a given gene, where and when these products are expressed, and how their expression changes in different mouse strains and mutants. GXD will continue to integrate the expression data with genetic, functional, and phenotypic information to support research into the molecular mechanisms of mammalian development, differentiation, and disease.

Funding GXD is supported by the Eunice Kennedy Shriver National Institute of Child Health and Human Development (NICHD) of the National Institutes of Health (NIH); Grant number: HD062499.



Use the URL below to fetch the first **10** records for this template from the command line or a script *(authentication needed for private templates and lists)*:

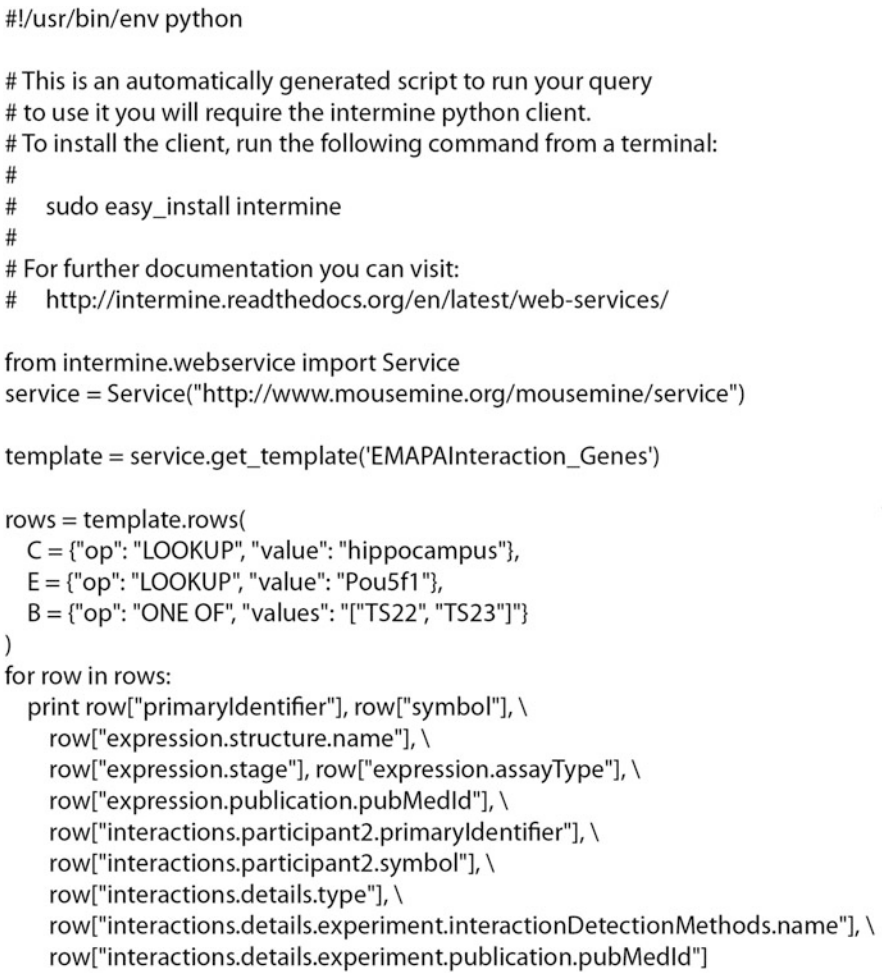
```
http://www.mousemine.org/mousemine/service/template/results?name=EM
```

```
$ curl
"http://www.mousemine.org/mousemine/service/template/result
s?name=EMAPAInteraction_Genes&constraint1=Gene.expression.s
tructure.parents&op1=LOOKUP&value1=hippocampus&extra1=&cons
traint2=Gene.interactions.participant2&op2=LOOKUP&value2=Po
u5f1&extra2=&format=tab&size=10"
MGI:1891824    Acin1    hippocampus    TS27 RNA in situ
     16033648  MGI:101893    Pou5f1    physical  affinity
chromatography technology    20362542
MGI:1891824    Acin1    hippocampus CA3    TS28 RNA in
situ 15226823  MGI:101893    Pou5f1    physical  affinity
chromatography technology    20362542
MGI:1861453    Actl6a   hippocampus    TS23 RNA in situ
     16602821  MGI:101893    Pou5f1    physical  affinity
chromatography technology    20362542
```

```
$ cat myscript.py
import sys
import urllib
tmplt="http://www.mousemine.org/mousemine/service/template/
results?name=EMAPAInteraction_Genes&constraint1=Gene.expres
sion.structure.parents&op1=LOOKUP&value1=%s&extra1=&constra
int2=Gene.interactions.participant2&op2=LOOKUP&value2=%s&ex
tra2=&format=tab"
url = tmplt % (sys.argv[1], sys.argv[2])
for line in urllib.urlopen(url):
    sys.stdout.write(line)
```

```
$ python myscript.py lung Cftr
MGI:96705 Krt8 lobar bronchus epithelium     TS24
     Immunohistochemistry     12950086  MGI:88388 Cftr
     physical  surface plasmon resonance     22038833
MGI:96705 Krt8 lung TS23 RNA in situ     21267068  MGI:88388
     Cftr physicasurface plasmon resonance    22038833
MGI:96705 Krt8 lung TS23 RNA in situ     21267068  MGI:88388
     Cftr physicasurface plasmon resonance    22038833
```

Fig. 7.12 Simple Web Service example. (**a**) The MouseMine UI has many features to help the developer write applications. At the bottom of every form is a link that generates a URL to run the current template with the current inputs. (**b**) One can run the query on the command line by simply



```
#!/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/

from intermine.webservice import Service
service = Service("http://www.mousemine.org/mousemine/service")

template = service.get_template('EMAPAInteraction_Genes')

rows = template.rows(
  C = {"op": "LOOKUP", "value": "hippocampus"},
  E = {"op": "LOOKUP", "value": "Pou5f1"},
  B = {"op": "ONE OF", "values": "["TS22", "TS23"]"}
)
for row in rows:
  print row["primaryIdentifier"], row["symbol"], \
    row["expression.structure.name"], \
    row["expression.stage"], row["expression.assayType"], \
    row["expression.publication.pubMedId"], \
    row["interactions.participant2.primaryIdentifier"], \
    row["interactions.participant2.symbol"], \
    row["interactions.details.type"], \
    row["interactions.details.experiment.interactionDetectionMethods.name"], \
    row["interactions.details.experiment.publication.pubMedId"]
```

Fig. 7.13 MouseMine client libraries. An alternative to accessing the MouseMine API directly (via URLs) is to use one of the client libraries available in several languages. UI assists are available to generate working programs. Clicking the Python link at the bottom of the form in Fig. 7.10 generates the code in this figure (the comments have been removed to save space)

Fig. 7.12 (continued) pasting the URL as an argument to curl (or other similar tool). (**c**) With a little more effort, the URL can be wrapped in a Python script that plugs in the user's parameters. (**d**) myscript.py is a command-line substitute for the web-based query form. The user can download results for any tissue and gene combination

References

Antin PB, Yatskievych TA, Davey S, Darnell DK (2014) GEISHA: an evolving gene expression resource for the chicken embryo. Nucleic Acids Res 42:D933–D937

Barrett T, Wilhite SE, Ledoux P, Evangelista C, Kim IF, Tomashevsky M, Marshall KA, Phillippy KH, Sherman PM, Holko M, Yefanov A, Lee H, Zhang N, Robertson CL, Serova N, Davis S, Soboleva A (2013) NCBI GEO: archive for functional genomics data sets—update. Nucleic Acids Res 41:D991–D995

Drabkin HJ, Christie KR, Dolan ME, Hill DP, Ni L, Sitnikov D, Blake JA (2015) Application of comparative biology in GO functional annotation: the mouse model. Mamm Genome 26:574–583

Finger JH, Smith CM, Hayamizu TF, McCright IJ, Xu J, Eppig JT, Kadin JA, Richardson JE, Ringwald M (2015) The mouse gene expression database: new features and how to use them effectively. Genesis 53:510–522

Finger JH, Smith CM, Hayamizu TF, McCright IJ, Xu J, Law M, Shaw DR, Baldarelli RM, Beal JS, Blodgett O et al (2017) The mouse Gene Expression Database (GXD): 2017 update. Nucleic Acids Res 45:D730–D736

Hayamizu TF, Wicks MN, Davidson DR, Burger A, Ringwald M, Baldock RA (2013) EMAP/EMAPA ontology of mouse developmental anatomy: 2013 update. J Biomed Semantics 4:15

Hayamizu TF, Baldock RA, Ringwald M (2015) Mouse anatomy ontologies: enhancements and tools for exploring and integrating biomedical data. Mamm Genome 26:422–430

Howe DG, Bradford YM, Eagle A, Fashena D, Frazer K, Kalita P, Mani P, Martin R, Moxon ST, Paddock H, Pich C, Ramachandran S, Ruzicka L, Schaper K, Shao X, Singer A, Toro S, Van Slyke C, Westerfield M (2017) The Zebrafish Model Organism Database: new support for human disease models, mutation details, gene expression phenotypes and searching. Nucleic Acids Res 45:D758–D768

Kalderimis A, Lyne R, Butano D, Contrino S, Lyne M, Heimbach J, Hu F, Smith R, Stěpán R, Sullivan J, Micklem G (2014) InterMine: extensive web services for modern biology. Nucleic Acids Res 42:W468–W472

Karimi K, Fortriede JD, Lotay VS, Burns KA, Wang DZ, Fisher ME, Pells TJ, James-Zorn C, Wang Y, Ponferrada VG, Chu S, Chaturvedi P, Zorn AM, Vize PD (2018) Xenbase: a genomic, epigenomic and transcriptomic model organism database. Nucleic Acids Res 46:D861–D868

Lyne R, Sullivan J, Butano D, Contrino S, Heimbach J, Hu F, Kalderimis A, Lyne M, Smith RN, Štěpán R, Balakrishnan R, Binkley G, Harris T, Karra K, Moxon SA, Motenko H, Neuhauser S, Ruzicka L, Cherry M, Richardson J, Stein L, Westerfield M, Worthey E, Micklem G (2015) Cross-organism analysis using InterMine. Genesis 53:547–560

Motenko H, Neuhauser SB, O'Keefe M, Richardson JE (2015) MouseMine: a new data warehouse for MGI. Mamm Genome 26:325–330

Papatheodorou I, Fonseca NA, Keays M, Tang YA, Barrera E, Bazant W, Burke M, Füllgrabe A, Fuentes AM, George N, Huerta L, Koskinen S, Mohammed S, Geniza M, Preece J, Jaiswal P, Jarnuczak AF, Huber W, Stegle O, Vizcaino JA, Brazma A, Petryszak R (2018) Expression Atlas: gene and protein expression across multiple studies and organisms. Nucleic Acids Res 46: D246–D251

Ringwald M, Baldock R, Bard J, Kaufman M, Eppig JT, Richardson JE, Nadeau JH, Davidson D (1994) A database for mouse development. Science 265:2033–2034

Ringwald M, Mangan ME, Eppig JT, Kadin JA, Richardson JE, the Gene Expression Database Group (1999) GXD: a gene expression database for the laboratory mouse. Nucleic Acids Res 27:106–112

Schmidt EF, Kus L, Gong S, Heintz N (2013) BAC transgenic mice and the GENSAT database of engineered mouse strains. Cold Spring Harb Protoc 2013:3

Schriml LM, Mitraka E (2015) The Disease Ontology: fostering interoperability between biological and clinical human disease-related data. Mamm Genome 26:584–589

Smith CL, Eppig JT (2012) The Mammalian Phenotype Ontology as a unifying standard for experimental and high-throughput phenotyping data. Mamm Genome 23:653–668

Smith RN, Aleksic J, Butano D, Carr A, Contrino S, Hu F, Lyne M, Lyne R, Kalderimis A, Rutherford K, Stepan R, Sullivan J, Wakeling M, Watkins X, Micklem G (2012) InterMine: a flexible data warehouse system for the integration and analysis of heterogeneous biological data. Bioinformatics 28:3163–3165

Smith CM, Finger JH, Kadin JA, Richardson JE, Ringwald M (2014) The gene expression database for mouse development (GXD): putting developmental expression information at your fingertips. Dev Dyn 243:1176–1186

Smith CM, Finger JH, Hayamizu TF, McCright IJ, Xu J, Eppig JT, Kadin JA, Richardson JE, Ringwald M (2015) GXD: a community resource of mouse Gene Expression Data. Mamm Genome 26:314–324

Smith CL, Blake JA, Kadin JA, Richardson JE, Bult CJ, the Mouse Genome Database Group (2018a) Mouse Genome Database (MGD)-2018: knowledgebase for the laboratory mouse. Nucleic Acids Res 46:D836–D842

Smith CM, Hayamizu TF, Finger JH, Bello SM, McCright IJ, Xu J, Baldarelli RM, Beal JS et al (2018b) The mouse gene expression database (GXD): 2019 update. Nucleic Acids Res. https://doi.org/10.1093/nar/gky922

Sunkin SM, Ng L, Lau C, Dolbeare T, Gilbert TL, Thompson CL, Hawrylycz M, Dang C (2013) Allen Brain Atlas: an integrated spatio-temporal portal for exploring the central nervous system. Nucleic Acids Res 41:D996–D1008

Chapter 8
Protein Structural Changes Based on Structural Comparison

Ryotaro Koike and Motonori Ota

Abstract Advances in structural biology have provided a wealth of information on protein structures. In many proteins, multiple structures under distinct functional states are available. The comparison of such structures reveals structural changes during the transition between the states, for example, those from ligand-free to -bound states. These structural changes are important for understanding the molecular mechanism of protein function. Currently, a number of computational methods have been developed to compare distinct structural states of the same protein and describe protein structural changes. The resulting structural changes are stored in databases. After a brief introduction of pre-existing methods and databases, we introduce Motion Tree, which illustrates various structural changes using a tree diagram and provides an explanation of how to use Motion Tree. We also introduce PSCDB, which presents structural changes for 837 proteins including homodimers. Structural changes are classified into seven categories based on the types of motions and bound ligands. PSCDB is available via the Internet.

Keywords Molecular visualization · Conformational change · Molecular structure and function

8.1 Introduction

Proteins are flexible molecules that often change their structure in response to external stimuli, such as ligand binding and post-translational modification (Amemiya et al. 2011; Brylinski and Skolnick 2008; Hayward 2004; Xin and Radivojac 2012). These structural changes are closely linked to their function. For example, structural changes to the calcium pump are required for ion transport (Kobayashi et al. 2015; Toyoshima and Nomura 2002), whereas the actin monomer

R. Koike · M. Ota (✉)
Graduate School of Informatics, Nagoya University, Nagoya, Japan
e-mail: mota@i.nagoya-u.ac.jp

© The Author(s), under exclusive license to Springer Nature Singapore Pte Ltd. 2021
I. Abugessaisa, T. Kasukawa (eds.), *Practical Guide to Life Science Databases*,
https://doi.org/10.1007/978-981-16-5812-9_8

changes shape to form a long fiber (Oda et al. 2009). Particular transferases including kinases undergo large structural changes to shield the ligand from water molecules (Koike et al. 2008). Systematic examination of available enzymatic structures reveals a strong correlation between structural changes and their function (Kanematsu et al. 2013; Koike et al. 2008, 2014). These observations highlight that structural changes are important in uncovering how proteins function.

Comparison of distinct structures of the same protein in the Protein Data Bank (PDB) that has been determined under different conditions or functional states (Smiley et al. 1971) provides a powerful approach to analyze structural changes (Kinjo et al. 2017). Here, comparison of the structures allows identification of structural changes. For this purpose, we commonly identify rigid bodies in the structure and denote motions using rigid bodies. This operation is considered to be reasonable because all atoms in the structure do not move independently and in many cases they move collectively. Note that rigid bodies in this sense are not exactly rigid bodies that are strictly fixed without any internal motion but are rigid bodies that allow fluctuation of atoms under a given threshold value.

A number of computational methods have been developed to detect rigid bodies from two structures of a protein. In Table 8.1, we present some of these methods that are currently available (Abyzov et al. 2010; Hayward and Berendsen 1998; Hayward et al. 1997; Koike et al. 2014; Ponzoni et al. 2015; Poornam et al. 2009; Sim et al. 2015; Wriggers and Schulten 1997). The most direct and intuitively simplest approach involves identification of common substructures in two structures. For example, in the Hingefind method (Wriggers and Schulten 1997), the common substructures are identified using iterative superposition of protein structures. Other approaches focus on the deviation of each atom because atoms in the rigid bodies collectively move together in a concerted manner (Hayward and Berendsen 1998; Hayward et al. 1997; Hinsen et al. 1999). Movements of Cα atoms are represented as rotation and translation vectors and screw axis, in which the Cα atoms in one structure are superimposed with the corresponding Cα atoms in the other structure. According to the similarity of the movements, Cα atoms are grouped and defined as rigid bodies. Other methods focus on the distance difference matrices of two structures (Abyzov et al. 2010; Nichols et al. 1995). In a rigid body, distances between Cα atoms are almost unchanged in the two structures. Consequently, rigid bodies are defined by the sets of Cα atoms where all the distance difference matrices are under a given threshold. Motion Tree employs the distance difference matrix and identifies various types of rigid bodies, i.e., from small loop motions to large domain motions (Koike et al. 2014). We introduce Motion Tree in the following section.

Structural changes of proteins are stored in databases, and we can access and browse them via the Internet (see Table 8.2) (Amemiya et al. 2012; Chang et al. 2016, 2012; Gerstein and Krebs 1998; Hrabe et al. 2016; Juritz et al. 2011; Monzon et al. 2016; Qi et al. 2005). Some databases provide pairs of structures that are apparently distinct and exhibit drastic structural changes (Gerstein and Krebs 1998; Lee et al. 2003; Qi et al. 2005). Such large structural changes should be attributed to a cause, e.g., ligand binding, chemical modification, thermal fluctuation, crystal packing, but are often not considered in much detail in such databases. Some other

Table 8.1 Tools to detect rigid-body motions

Name	URL	Description	References
DynDom	http://fizz.cmp.uea.ac.uk/dyndom/	This is the most widely used software that focuses on domain motion. DynDom3D, applicable to protein complexes, has also been developed. The source code is distributed at the website. The server is also available online	Hayward and Berendsen (1998), Hayward et al. (1997), Poornam et al. (2009)
Hingefind	http://biomachina.org/disseminate/hingefind/hingefind.html	Rigid bodies are detected by iterative superposition. This is available as a plug-in for VMD and also included in X-PLOR	Wriggers and Schulten (1997)
RigidFinder	http://rigidfinder.molmovdb.org/	Quasi-dynamic programming is applied to the distance difference matrix and rigid bodies with various sizes are detected. This program can be downloaded from the website	Abyzov et al. (2010)
DAGR	http://dna.ssu.ac.kr/index.php?pid=program	Clique detection is used to detect rigid domains. The server is also available online	Sim et al. (2015)
SPECTRUS	http://spectrus.sissa.it/	Distance fluctuation is used to detect rigid bodies. The server accepts a pair or multiple structures to calculate distance fluctuations. Additionally, a single structure can be used to conduct elastic network analysis. Users can download the source code from the website. The server is also available online	Ponzoni et al. (2015)
Motion Tree	http://idp1.force.cs.i.nagoya-u.ac.jp/MotionTree/	Hierarchical clustering using a distance difference matrix enables hierarchical description of various rigid-body motions with various sizes. The binary codes can be downloaded from the website. The server is also available online	Koike et al. (2014)

databases focus on structural changes related to function (Amemiya et al. 2012; Chang et al. 2012; Qi and Hayward 2009). Only pairs of structures under distinct functional states, e.g., ligand-free and bound states, are stored. From the latter type of databases, the Protein Structural Change DataBase (PSCDB) is introduced below.

Table 8.2 Databases of protein structural changes

Name	URL	Description	References
MolMovDB	http://www.molmovdb.org/	This is the oldest database that presents protein structural changes. The structural changes are classified according to the size of mobile parts. MolMovDB also focuses on whether the motion is hinge-like or shear-like	Gerstein and Krebs (1998)
NRDPDM	http://www.cmp.uea.ac.uk/dyndom/	This is a non-redundant and comprehensive database of protein domain motions. A subset is also available as a database of domain motions in enzymes	Qi et al. (2005)
CCProf	http://zoro.ee.ncku.edu.tw/ccprof/	Conformational change profiles (CCPs) of >3000 proteins are presented in the database. The profile represents the deviation of each residue during a structural change. (The search engine on the website looks unavailable as of May, 2018)	Chang et al. (2016)
CoDNaS	http://ufq.unq.edu.ar/codnas/	Conformational ensembles of proteins are presented in the database. The ensemble, called “conformational diversity” in the database, means multiple structures of the same protein obtained from the PDB	Monzon et al. (2016)
PDBFlex	http://pdbflex.org/	This database shows structural flexibility of nearly 30,000 proteins. The structural ensemble provides an estimate of local and global structural flexibility	Hrabe et al. (2016)
PCDB	http://www.pcdb.unq.edu.ar/	Conformational diversities of protein domains are provided in this database. (The Internet connection to the website is unstable as of May, 2018)	Juritz et al. (2011)
AH-DB	http://ahdb.ee.ncku.edu.tw/	This database provides pairs of structures in ligand-free and -bound forms. There are nearly 750,000 pairs for ~3600 proteins. (The search engine on the website looks unavailable as of May, 2018)	Chang et al. (2012)
PSCDB	http://idp1.force.cs.i.nagoya-u.ac.jp/pscdb/	This database presents ligand-free and -bound pairs for 839 proteins, including homodimers. The structural changes are classified into seven categories by types of motions and bound ligands	Amemiya et al. (2012)

8.2 Motion Tree

8.2.1 Overview

Motion Tree is a computational method that identifies and describes structural changes between two distinct structures of an identical protein (Koike et al. 2014). Significant features of Motion Tree are summarized in the following three points.

(1) Motion Tree can identify rigid bodies of various sizes from small to large ones. In a protein structural change, a rigid body can be defined as a small region like a loop or a secondary structure element, or a large region such as a protein domain. In contrast, most other methods detect either small or large rigid regions. (2) Motion Tree can detect motions with different magnitudes and describe them in a hierarchical manner. The magnitude of motions indicates how large a rigid body moves: some rigid bodies move slightly, whereas others move drastically. The magnitude is also an essential factor to describe structural changes. For example, in adenylate kinase (ADK), sub-domains, ATP and AMP lids, move drastically upon substrate binding, whereas a helix within the AMP lid moves only slightly (Koike et al. 2014). The motion of the helix is disregarded when focusing only on large movements. Conversely, if we focus on small movements, the motion of the AMP lid is decomposed into several motions of rigid bodies, one of which might be the helix. (3) Motion Tree describes structural changes based on the distance difference matrix of two structures rather than on their superposition. The superposition of two structures is a standard approach. However, it is not straightforward to decide which optimal parts should be superposed. In other words, methods relying on structural superposition provide various results that are dependent on the superimposed regions. In contrast, Motion Tree is uniquely determined because two structures provide only a unique distance difference map.

8.2.2 Illustration of Structural Changes with Motion Tree

We explain how Motion Tree illustrates protein structural changes using dUTPase as an example. The two distinct structures of dUTPase in free and dUDP-bound forms were compared (Harkiolaki et al. 2004) (Fig. 8.1a). To describe the structural change upon dUDP binding, a tree diagram, called as Motion Tree, was determined (Fig. 8.1b). A node in the Motion Tree (black circle) represents a relative motion between two rigid bodies indicated by two child branches at the node. The height of the node indicates the magnitude of motion and a node closer to the root indicates a more drastic motion. In this case, node A shows the (relative) motion of domain B (and the rest), which is the largest structural change upon ligand binding (Fig. 8.1b). Node B shows the motion of the binding loop within domain B. The magnitude of the motion for node B is significantly smaller than that of node A. Each residue of the dUTPase corresponds to each leaf of the tree. The smaller rigid parts, i.e., domain B and binding loop, are represented by the leaves indicated by the child branches at the nodes and highlighted by red bars on the left side of the tree. The remaining parts are also rigid bodies (e.g., domain A in Fig. 8.1a) and shown by blue bars.

The Motion Tree illustrates both small loop motion and large domain motion using a tree diagram, in which the heights of the nodes indicate the magnitudes of motions. This representation also clarifies the "nesting" relationship, i.e., the binding loop is a part of domain B. This hierarchical diagram enables us to understand easily the whole structural change between two structures at a glance, in terms of the size of

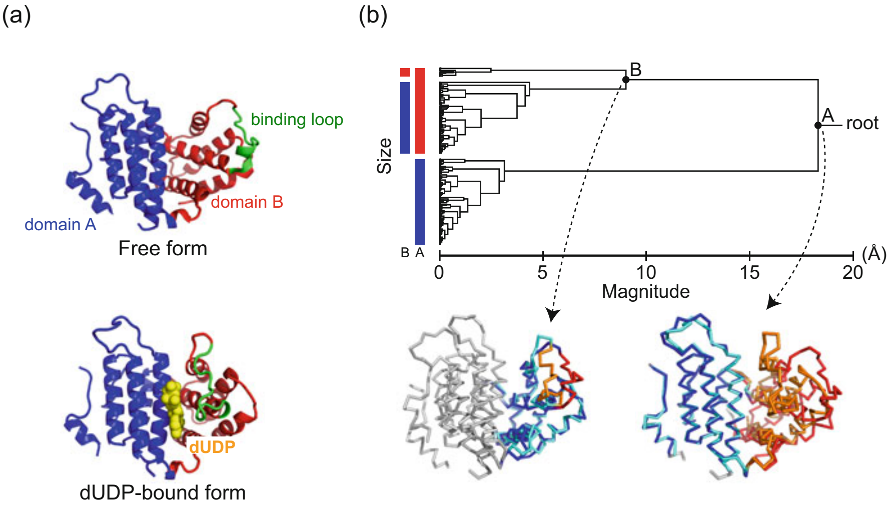


Fig. 8.1 Structures and Motion Tree of dUTPase. (**a**) Two structures compared with Motion Tree. The free (PDB ID: 1ogl, upper panel) and ligand-bound structures (PDB ID: 1ogk; D chain, lower) are represented in cartoon model. The domains A and B and the binding loop are colored in blue, red, and green, respectively. The bound ligand is shown as yellow spheres. (**b**) Motion Tree describing structural changes to dUTPase upon ligand binding. The nodes A and B (black circles) indicate the motion between domains A and B and the motion of the binding loop in domain B, respectively. Structural superposition at the bottom illustrates the structural changes detected at nodes A (right) and B (left). The two structures are superposed using Cα atoms in larger rigid bodies (blue parts in the free form and cyan in the bound form) and the motion of smaller rigid bodies (red and orange) is highlighted. The bars on the left indicate the sizes (total residues) of the two rigid bodies at each node. The blue and red bars correspond to larger and smaller rigid bodies, respectively

the rigid bodies, the magnitude of motions, and the nesting relationships among rigid bodies.

8.2.3 *Availability of Motion Tree*

Motion Tree is accessible through the web server (http://idp1.force.cs.i.nagoya-u.ac.jp/MotionTree/). At the web page, the input boxes at the bottom are used to specify the two structures to be compared. The structures are described by the combination of the PDB identifier and chain identifier such as 4akeA. Motion Tree starts upon clicking the "Execute" button near the input boxes. The window of the web browser is automatically updated, and the Motion Tree and structures appear after completion (Fig. 8.2a). The compared structures are superimposed using the Cα atoms of the largest rigid bodies and displayed in the right panel. The identified rigid bodies are shown by different colors.

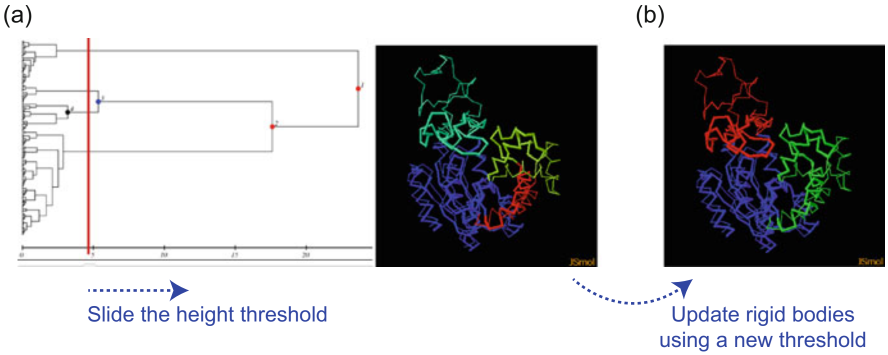

(a)
(b)
Slide the height threshold
Update rigid bodies using a new threshold

Fig. 8.2 The Motion Tree of ADK on the Motion Tree server (URL: http://idp1.force.cs.i.nagoya-u.ac.jp/MotionTree/). (**a**) Motion Tree and structures of ADK. Motion Tree (left panel) describes the structural change between ligand-free (PDB ID: 4ake; A chain) and -bound (PDB ID: 2eck; A chain) forms. The red line indicates the height threshold used to define rigid bodies. In the case of the 5.0 Å threshold, structures are divided into four clusters, corresponding to four rigid bodies with different colors shown in the right panel. The red line is adjustable using a slide box shown at the bottom. (**b**) Three rigid bodies defined by a 10.0 Å threshold

The colored rigid bodies are determined according to a threshold tree height, which is represented as a vertical red line (default value is 5 Å) in the tree diagram. In Fig. 8.2a, the threshold tree height is ~5 Å, and the red line and a tree have four points of intersection, indicating that the structure was divided into four rigid bodies (Fig. 8.2a, right panel). The red line can be adjusted using the slide bar under the tree. The rigid body representation is updated when the threshold height increases so that the red line and a tree have three points of intersection (Fig. 8.2b).

The Motion Tree program is downloadable from the website (http://idp1.force.cs.i.nagoya-u.ac.jp/rk1/mtntr/classification/). Two binary codes for Linux and Mac OS are distributed on the site. The program (mtntr) receives two PDB files (pdbfile1 (chain-id1) and pdbfile2(chain-id2)) and produces a Motion Tree (in postscript), and a set of PDB files, each of which shows a rigid-body motion at each node (see bottom panels in Fig. 8.1b). The standard command is:

```
> ./mtntr pdbfile1 chain-id1 pdbfile2 chain-id2
```

The script files to highlight the motions in Rasmol are also produced.

8.3 PSCDB

8.3.1 Overview

PSCDB is a database of protein structural changes that occur upon ligand binding (Amemiya et al. 2012). Structural changes are defined by using a pair of structures from a particular protein under different conditions, i.e., ligand-free (apo) and -bound (holo) forms. The pairs of apo- and holo-forms are the fundamental resource of the database. Most data are structural changes for monomeric proteins but those for homodimeric proteins, the most abundant protein complexes in the PDB (Koike et al. 2018), are also stored. Currently, PSCDB has collected structural changes for 839 distinct proteins, each of which was selected from a SCOP family. The structural changes are classified into seven categories (Fig. 8.3): (1) coupled domain motion

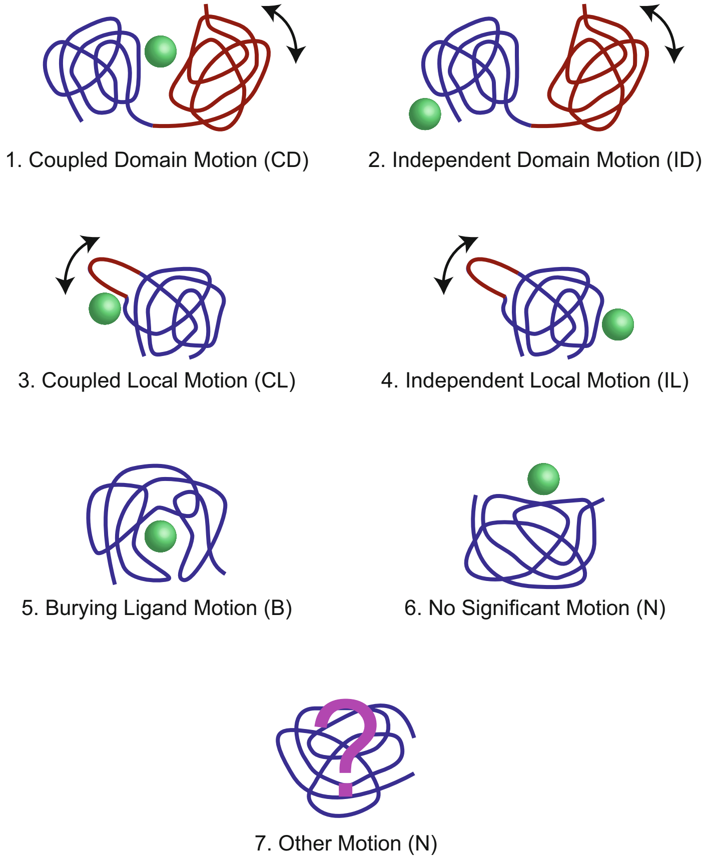

1. Coupled Domain Motion (CD)
2. Independent Domain Motion (ID)
3. Coupled Local Motion (CL)
4. Independent Local Motion (IL)
5. Burying Ligand Motion (B)
6. No Significant Motion (N)
?
7. Other Motion (N)

Fig. 8.3 Schematic diagrams for the seven categories of structural changes. A protein molecule is illustrated as a string and the mobile portion is highlighted in red. The ligand molecule is shown as a green sphere

(CD); (2) independent domain motion (ID); (3) coupled local motion (CL); (4) independent local motion (IL); (5) burying ligand motion (B); (6) no significant motion (N); and (7) other motion (O). On the PSCDB web page, each entry is presented with its classification, rigid parts, function, and ligands. In addition, PSCDB provides morphing movies of motions.

8.3.2 Data Construction

PSCDB selects pairs of apo- and holo-forms as follows: First, protein structures are obtained from the PDB. The structures are paired according to sequence identities. If the identity is more than 95% with more than 90% coverage of the entire sequence, the proteins are assumed to be the same. Among them, the pairs of apo- and holo-structures are selected as potential contents of PSCDB. Next, the pairs satisfying the following criteria are retained in the PSCDB. The pairs are crystal structures with resolution better than 3.0 Å. Oligomeric states of a structural pair coincide: both are monomers or homodimers. The oligomeric states are basically judged by PiQSi (Levy 2007) but in some cases, PQS (Henrick and Thornton 1998) and the biological units of the PDB entry are referenced (Amemiya et al. 2011). A SCOP family (for single domain proteins) (Murzin et al. 1995) or a combination of SCOP families (multiple domain proteins) can be assigned to proteins based on SCOP entries and homology (40% sequence identity). A representative pair is selected from a set of proteins labeled by the same SCOP annotation described above. A structural pair, which exhibits a large structural change and has a small number of missing residues, is favorable as the representative. The RMSD of Cα atoms in the superposition is used as a measure of magnitude (Amemiya et al. 2011). After applying these selection criteria 839 structural pairs were selected.

Structural changes are classified into the seven categories (CD, ID, CL, IL, B, N, O) mentioned above. Classification principles are summarized as follows: The structural changes are basically divided into large and slight. There are two types of large structural changes. Domain motions (D) are detected by DynDom (Hayward and Berendsen 1998; Hayward et al. 1997) and local motions (L) are detected by our original procedure (Amemiya et al. 2011). When the ligand positions are close to the boundary of the rigid bodies (within 4.0 Å from both rigid bodies), we assume the motion is coupled with ligand binding (C). Otherwise, the structural change is independent of ligand binding (I). Combinations of D/L and C/I result in four categories of motions. When many domain and local motions are detected between a pair of structures, the structure is decomposed into a pair of a large fixed part and a small moving part (namely a component), and a category is assigned to each component. The coexistence of D and L in a protein is treated as D, and the coexistence of C and I is treated as C. Slight structural changes (Cα RMSD <1 Å) are further divided according to the accessible surface area (ASA) of ligands: a low ASA ($<10\%$ relative ASA) of the ligand defines burying ligand motion (B). In this category, as the apo-form is almost identical to the holo-form and the binding site is

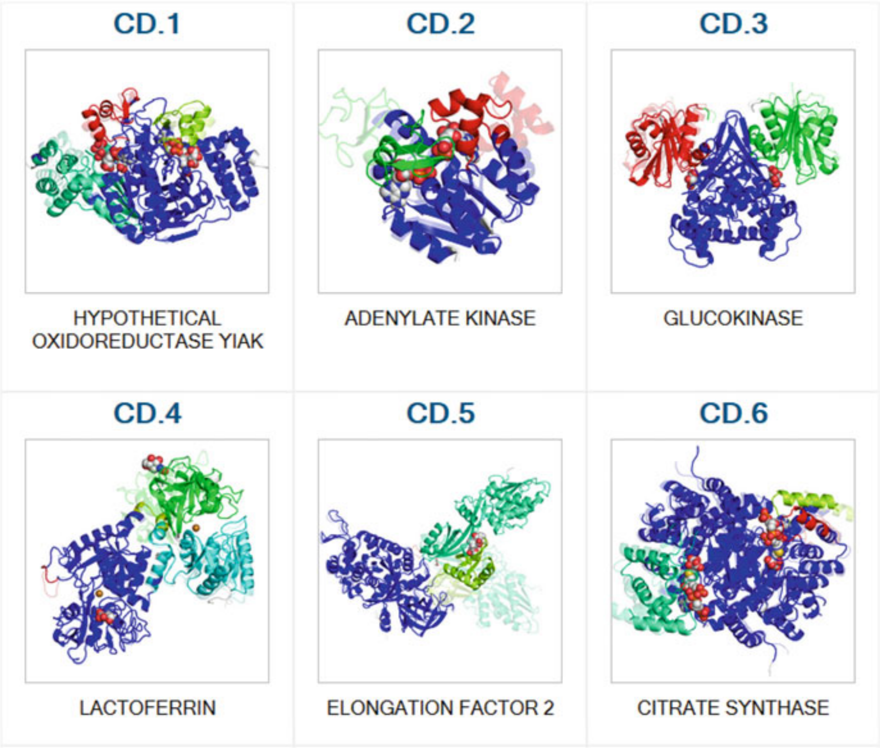


Fig. 8.4 Gallery page showing each protein with the PSCID and an image. This page is linked from "Gallery" button in the left side of the top page (URL: http://idp1.force.cs.i.nagoya-u.ac.jp/pscdb/). Each PSCID or image is a link to the summary page of the protein

significantly shielded from water molecules, the ligand can never access the binding site from the outside. This suggests that some structural change, e.g., open and close motion, is required to enable the ligand to bind. When the ASA of the ligand is high (>10% relative ASA), the motion is no significant (N). The rest of the pairs are categorized to "other motion" (O) and are considered outliers. A unique PSCID, the combination of the category and an identification number in the category is assigned to each of the 839 proteins (Fig. 8.4).

Note that a number of slight structural changes (B and N) are stored in PSCDB (Fig. 8.3). This means that drastic structural changes are not always required upon ligand binding. Our previous studies revealed that the magnitude of the structural changes correlates with the function of proteins (Koike et al. 2008) and the reaction mechanism (Kanematsu et al. 2013). Slight structural changes in some specific proteins have enabled deduction of their molecular functions.

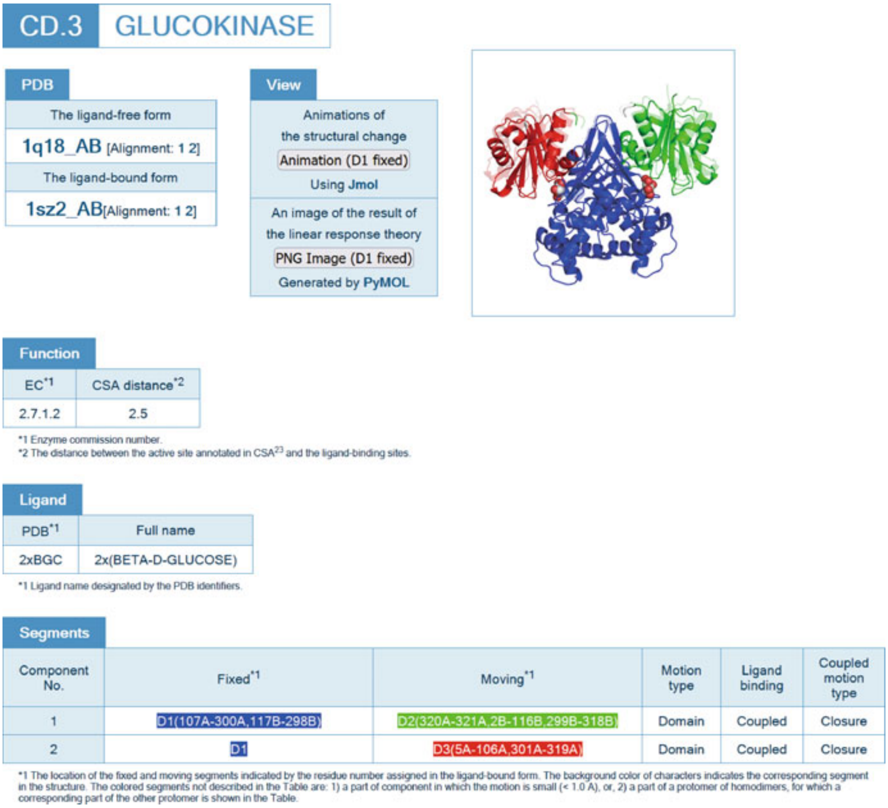
8.3.3 Browsing PSCDB

The top web page of PSCDB (http://idp1.force.cs.is.nagoya-u.ac.jp/pscdb/) shows schematic diagrams of the seven categories (Fig. 8.3). Clicking a particular diagram provides a list of the proteins in the selected category. The list is shown as a table with PSCID, protein name, PDB codes and chain identifiers of apo- and holo-forms, ligands, and a string of six digits. Digits one to five represent the numbers of components for coupled domain, independent domain, coupled local, independent local, and burying ligand motions in the protein, respectively. The last digit shows the number of ligand molecules on the protein surface and frequently implies "no significant motion" for the protein. The search box is available in the upper right corner of every page. PDB identifiers or protein names can be used as queries. The gallery page, accessible from the link button on the left side, is also useful. An image of the change in the protein structure is shown in a panel with the PSCID and protein name (Fig. 8.4).

From the list or the gallery page, users can access a summary of the structural changes to a protein. As an example, structural changes to glucokinase (Lunin et al. 2004) are shown in Fig. 8.5. The PSCID is CD.3, and the PDB codes of ligand-free and bound forms are given in the "PDB" box (upper left). Based on the structures, two components showing coupled domain motions are detected (the "Segments" box). The rigid parts of the protein are illustrated as a cartoon model (blue, green, and red), and ligand molecules are shown as spheres in the upper right panel. The rigid regions are explicitly shown in the "Segments" box. Users can view domain motions (i.e., morphing) by clicking the "Animation" button. The predicted domain motions using linear response theory (Ikeguchi et al. 2005) can also be viewed using the "PNG image" button. The enzyme commission (EC) number and information about the active sites are provided in the "Function" box. The bound ligand molecules are in the "Ligand" box.

8.4 Future Work

We are confident that Motion Tree is an ideal, universal tool to describe protein structural changes because Motion Tree can detect both small and large rigid bodies with any magnitude of structural change and illustrates such changes systematically using a tree diagram. Motion Tree has already been applied to identify structural changes to monomers (Koike et al. 2014) and homodimers (Koike et al. 2018). We are planning to extend this analysis to higher-order homo-oligomers and hetero-oligomers. We anticipate that structural changes to such oligomers are more complex than those observed for monomers and homodimers, and our analysis will highlight the relationships between structural changes and functions more clearly. Results obtained from Motion Tree can be stored in PSCDB, thereby increasing the size of the database and thus making PSCDB a more comprehensive database. Advances in



CD.3 GLUCOKINASE

PDB

The ligand-free form
1q18_AB [Alignment: 1 2]
The ligand-bound form
1sz2_AB[Alignment: 1 2]

View

Animations of the structural change

Animation (D1 fixed)

Using Jmol

An image of the result of the linear response theory

PNG Image (D1 fixed)

Generated by PyMOL

Function

EC[*1]	CSA distance[*2]
2.7.1.2	2.5

[*1] Enzyme commission number.
[*2] The distance between the active site annotated in CSA[23] and the ligand-binding sites.

Ligand

PDB[*1]	Full name
2xBGC	2x(BETA-D-GLUCOSE)

[*1] Ligand name designated by the PDB identifiers.

Segments

Component No.	Fixed[*1]	Moving[*1]	Motion type	Ligand binding	Coupled motion type
1	D1(107A-300A,117B-298B)	D2(320A-321A,2B-116B,299B-318B)	Domain	Coupled	Closure
2	D1	D3(5A-106A,301A-319A)	Domain	Coupled	Closure

[*1] The location of the fixed and moving segments indicated by the residue number assigned in the ligand-bound form. The background color of characters indicates the corresponding segment in the structure. The colored segments not described in the Table are: 1) a part of component in which the motion is small (< 1.0 A), or, 2) a part of a protomer of homodimers, for which a corresponding part of the other protomer is shown in the Table.

Fig. 8.5 A summary page of glucokinase. The page shows the PSCID, protein name, PDB codes and chain identifiers, link buttons to morphing animation and png image, function, ligands, and the rigid bodies. Information is arranged as "PDB," "View," "Function," "Ligand," and "Segments" boxes

structural biology have generated a large number of structural complexes, which provide not only a static image but also dynamic information. The described method and database containing structural changes to proteins are valuable resources for extracting biological knowledge from protein structures.

Acknowledgements We are grateful to Akinori Kidera and Takayuki Amemiya for their contributions to Motion Tree and PSCDB.

References

Abyzov A, Bjornson R, Felipe M, Gerstein M (2010) RigidFinder: a fast and sensitive method to detect rigid blocks in large macromolecular complexes. Proteins 78(2):309–324. https://doi.org/10.1002/prot.22544

Amemiya T, Koike R, Fuchigami S, Ikeguchi M, Kidera A (2011) Classification and annotation of the relationship between protein structural change and ligand binding. J Mol Biol 408 (3):568–584. https://doi.org/10.1016/j.jmb.2011.02.058

Amemiya T, Koike R, Kidera A, Ota M (2012) PSCDB: a database for protein structural change upon ligand binding. Nucleic Acids Res 40(Database issue):D554–D558. https://doi.org/10.1093/nar/gkr966

Brylinski M, Skolnick J (2008) What is the relationship between the global structures of apo and holo proteins? Proteins 70(2):363–377. https://doi.org/10.1002/prot.21510

Chang DT, Yao TJ, Fan CY, Chiang CY, Bai YH (2012) AH-DB: collecting protein structure pairs before and after binding. Nucleic Acids Res 40(Database issue):D472–D478. https://doi.org/10.1093/nar/gkr940

Chang CW, Chou CW, Chang DT (2016) CCProf: exploring conformational change profile of proteins. Database 2016. https://doi.org/10.1093/database/baw029

Gerstein M, Krebs W (1998) A database of macromolecular motions. Nucleic Acids Res 26 (18):4280–4290

Harkiolaki M, Dodson EJ, Bernier-Villamor V, Turkenburg JP, Gonzalez-Pacanowska D, Wilson KS (2004) The crystal structure of Trypanosoma cruzi dUTPase reveals a novel dUTP/dUDP binding fold. Structure 12(1):41–53

Hayward S (2004) Identification of specific interactions that drive ligand-induced closure in five enzymes with classic domain movements. J Mol Biol 339(4):1001–1021. https://doi.org/10.1016/j.jmb.2004.04.004. S0022283604004061 [pii]

Hayward S, Berendsen HJ (1998) Systematic analysis of domain motions in proteins from conformational change: new results on citrate synthase and T4 lysozyme. Proteins 30(2):144–154. https://doi.org/10.1002/(SICI)1097-0134(19980201)30:2<144::AID-PROT4>3.0.CO;2-N. [pii]

Hayward S, Kitao A, Berendsen HJ (1997) Model-free methods of analyzing domain motions in proteins from simulation: a comparison of normal mode analysis and molecular dynamics simulation of lysozyme. Proteins 27(3):425–437. https://doi.org/10.1002/(SICI)1097-0134(199703)27:3<425::AID-PROT10>3.0.CO;2-N. [pii]

Henrick K, Thornton JM (1998) PQS: a protein quaternary structure file server. Trends Biochem Sci 23(9):358–361

Hinsen K, Thomas A, Field MJ (1999) Analysis of domain motions in large proteins. Proteins 34 (3):369–382. https://doi.org/10.1002/(SICI)1097-0134(19990215)34:3<369::AID-PROT9>3.0.CO;2-F. [pii]

Hrabe T, Li Z, Sedova M, Rotkiewicz P, Jaroszewski L, Godzik A (2016) PDBFlex: exploring flexibility in protein structures. Nucleic Acids Res 44(D1):D423–D428. https://doi.org/10.1093/nar/gkv1316

Ikeguchi M, Ueno J, Sato M, Kidera A (2005) Protein structural change upon ligand binding: linear response theory. Phys Rev Lett 94(7):078102. https://doi.org/10.1103/PhysRevLett.94.078102

Juritz EI, Alberti SF, Parisi GD (2011) PCDB: a database of protein conformational diversity. Nucleic Acids Res 39(Database issue):D475–D479. https://doi.org/10.1093/nar/gkq1181

Kanematsu Y, Koike R, Amemiya T, Ota M (2013) Substrate-shielding and hydrolytic reaction in hydrolases. Proteins 81(6):926–932. https://doi.org/10.1002/prot.24253

Kinjo AR, Bekker GJ, Suzuki H, Tsuchiya Y, Kawabata T, Ikegawa Y, Nakamura H (2017) Protein Data Bank Japan (PDBj): updated user interfaces, resource description framework, analysis tools for large structures. Nucleic Acids Res 45(D1):D282–D288. https://doi.org/10.1093/nar/gkw962

Kobayashi C, Koike R, Ota M, Sugita Y (2015) Hierarchical domain-motion analysis of conformational changes in sarcoplasmic reticulum Ca(2)(+)-ATPase. Proteins 83(4):746–756. https://doi.org/10.1002/prot.24763

Koike R, Amemiya T, Ota M, Kidera A (2008) Protein structural change upon ligand binding correlates with enzymatic reaction mechanism. J Mol Biol 379(3):397–401. https://doi.org/10.1016/j.jmb.2008.04.019. S0022-2836(08)00448-8 [pii]

Koike R, Ota M, Kidera A (2014) Hierarchical description and extensive classification of protein structural changes by Motion Tree. J Mol Biol 426(3):752–762. https://doi.org/10.1016/j.jmb.2013.10.034

Koike R, Amemiya T, Horii T, Ota M (2018) Structural changes of homodimers in the PDB. J Struct Biol 202(1):42–50. https://doi.org/10.1016/j.jsb.2017.12.004

Lee RA, Razaz M, Hayward S (2003) The DynDom database of protein domain motions. Bioinformatics 19(10):1290–1291

Levy ED (2007) PiQSi: protein quaternary structure investigation. Structure 15(11):1364–1367. https://doi.org/10.1016/j.str.2007.09.019

Lunin VV, Li Y, Schrag JD, Iannuzzi P, Cygler M, Matte A (2004) Crystal structures of Escherichia coli ATP-dependent glucokinase and its complex with glucose. J Bacteriol 186(20):6915–6927. https://doi.org/10.1128/JB.186.20.6915-6927.2004

Monzon AM, Rohr CO, Fornasari MS, Parisi G (2016) CoDNaS 2.0: a comprehensive database of protein conformational diversity in the native state. Database 2016. https://doi.org/10.1093/database/baw038

Murzin AG, Brenner SE, Hubbard T, Chothia C (1995) SCOP: a structural classification of proteins database for the investigation of sequences and structures. J Mol Biol 247(4):536–540. https://doi.org/10.1006/jmbi.1995.0159

Nichols WL, Rose GD, Ten Eyck LF, Zimm BH (1995) Rigid domains in proteins: an algorithmic approach to their identification. Proteins 23(1):38–48. https://doi.org/10.1002/prot.340230106

Oda T, Iwasa M, Aihara T, Maeda Y, Narita A (2009) The nature of the globular- to fibrous-actin transition. Nature 457(7228):441–445. https://doi.org/10.1038/nature07685

Ponzoni L, Polles G, Carnevale V, Micheletti C (2015) SPECTRUS: a dimensionality reduction approach for identifying dynamical domains in protein complexes from limited structural datasets. Structure 23(8):1516–1525. https://doi.org/10.1016/j.str.2015.05.022

Poornam GP, Matsumoto A, Ishida H, Hayward S (2009) A method for the analysis of domain movements in large biomolecular complexes. Proteins 76(1):201–212. https://doi.org/10.1002/prot.22339

Qi G, Hayward S (2009) Database of ligand-induced domain movements in enzymes. BMC Struct Biol 9:13. https://doi.org/10.1186/1472-6807-9-13. 1472-6807-9-13 [pii]

Qi G, Lee R, Hayward S (2005) A comprehensive and non-redundant database of protein domain movements. Bioinformatics 21(12):2832–2838. https://doi.org/10.1093/bioinformatics/bti420

Sim J, Sim J, Park E, Lee J (2015) Method for identification of rigid domains and hinge residues in proteins based on exhaustive enumeration. Proteins 83(6):1054–1067. https://doi.org/10.1002/prot.24799

Smiley IE, Koekoek R, Adams MJ, Rossmann MG (1971) The 5 A resolution structure of an abortive ternary complex of lactate dehydrogenase and its comparison with the apo-enzyme. J Mol Biol 55(3):467–475

Toyoshima C, Nomura H (2002) Structural changes in the calcium pump accompanying the dissociation of calcium. Nature 418(6898):605–611. https://doi.org/10.1038/nature00944

Wriggers W, Schulten K (1997) Protein domain movements: detection of rigid domains and visualization of hinges in comparisons of atomic coordinates. Proteins 29(1):1–14. https://doi.org/10.1002/(SICI)1097-0134(199709)29:1<1::AID-PROT1>3.0.CO;2-J. [pii]

Xin F, Radivojac P (2012) Post-translational modifications induce significant yet not extreme changes to protein structure. Bioinformatics 28(22):2905–2913. https://doi.org/10.1093/bioinformatics/bts541

Chapter 9
Single Cell Databases: An Emerging and Essential Tool

Scott Walker, Imad Abugessaisa, and Takeya Kasukawa

Abstract As single-cell studies become commonplace, the need for reuse, effective curation, and downstream analysis of data becomes a necessity. The development of single-cell protocols has presented a challenge for database design and construction. Newer droplet-based or gel beads in emulsion (GEM) techniques such as Drop-seq or 10× Genomics protocols can produce from 100,000 to millions of cells. To then be able to display vital metadata or secondary analyses in a coherent way will be essential moving forward. Here the current state of single-cell databases (SCDB) is explored, focusing on the needs of researchers, types of data stored and fields of study covered, common elements of SCDBs, secondary analyses, and use-case examples for some of the DBs mentioned.

Keywords Single cell · Gene expression · Development · Database · Cell type differential expression

9.1 Introduction

9.1.1 Summary of the Databases/Data Repositories

The rise of single cell transcriptomics, driven by the growing demand and necessity for studies to have a more focused lens on their subject matter, has subsequently necessitated the need for dedicated databases (DBs). Consistent improvement in technologies for cell isolation, capture, and analysis has grown a niche subject area to a foundational resource for most biological frontiers. These DBs are often backed by international consortia (Human Cell Atlas Data Portal), government backed/funded institutions [EMBL-EBI's Single Cell Expression Atlas (Papatheodorou et al. 2020)], research institution backed [RIKEN IMS's SCPortalen (Abugessaisa

S. Walker · I. Abugessaisa · T. Kasukawa (✉)
Laboratory for Large-Scale Biomedical Data Technology, RIKEN Center for Integrative Medical Sciences (IMS), Yokohama, Kanagawa, Japan
e-mail: scott.walker@riken.jp; imad.abugessaisa@riken.jp; takeya.kasukawa@riken.jp

© The Author(s), under exclusive license to Springer Nature Singapore Pte Ltd. 2021

I. Abugessaisa, T. Kasukawa (eds.), *Practical Guide to Life Science Databases*,
https://doi.org/10.1007/978-981-16-5812-9_9

et al. 2019), Karolinska Institutet's PanglaoDB (Franzén et al. 2019)], university-run [UCSC Cell Browser, Hong Kong University's SCDevDB (Wang et al. 2019), etc.].

9.1.2 Purpose of the Databases/Data Repositories

As well as providing structured secure storage of large datasets, including human data that requires strict protocols to avoid accidental mismanagement of personal data, these databases provide platforms and a resource for downstream analyses. The purpose of the databases is twofold, they are data repositories and in some cases they are analysis tools with attributed datasets. The data is publicly available for viewing and download to researchers and students, requiring in some cases to sign up by providing an email address; this is a necessity when handling human data, for prevention of data mismanagement. More niche databases offer a one-stop-shop for researchers of that field to access relevant data without trawling larger databases; e.g. Vascular Single Cell Database, from Tianjin Neurological Institute (China) and Karolinska Institutet (Sweden) (He et al. 2018) that focuses solely on brain vascular cell types. These datasets also offer pipelines for quality control (QC) for researchers own data as well as the data curated within.

9.1.3 The Source and Type of the Dataset Stored

Consortia (like the Human Cell Atlas group) source the datasets for their DB from consortium members (laboratories in universities or research institutions). Likewise, university or institution-run databases host data collected "in-house." Datasets attributed to studies that are publicly available are often included (with the lead authors' permission and proper accreditation) and if they fit the purpose of the database. The type of data collected includes detailed metadata about the dataset study, original sample data, and also data of each individual cell; this metadata is either curated manually or by automated pipelines. In-depth QC information in the form of standard metrics and cell image data (as in SCPortalen) is accompanied by cell-specific gene expression levels. The type of datasets stored, sourced, and curated is dependent upon the method, or protocol, by which they are collected. The protocol selected is constrained by the budget of the study as well as the biological objective. Full-length protocols like Smart-seq have a greater sequencing depth; the tradeoff being a higher cost compared with other methods that sequence from the $3'$ end (MARS-seq, CEL-seq or Drop-seq). However, Smart-seq lacks the use of unique molecular identifiers (UMIs) which is a feature that is now common as standard due to its perceived effect in reducing PCR bias and improving accuracy (Islam et al. 2014). Cell isolation is achieved by many protocols with flow-activated cell sorting (FACS). In this method cells are first tagged with a fluorescent monoclonal antibody to subsequently recognize specific surface markers, thus enabling sorting of cells into distinct populations; MARS-seq is a protocol that utilizes this method (Jaitin et al. 2014). More recently, microfluidic methods [sometimes referred to as droplet based or (Gel bead in EMulsion (GEM) methods] are ubiquitous due to their extremely high throughput, scalability, relatively lower cost, and also require only a small initial sample volume, compared to FACS. An example is the widely used Drop-seq (Macosko et al. 2015) and the Chromium Chemistry protocol from 10× Genomics (Hwang et al. 2018; Zheng et al. 2017).

9.1.4 Target Users

Use of these DBs is targeted at researchers or students in all fields of Medical science like Oncology, Neurology, and Dermatology, etc. As well as, researchers in evolutionary developmental (Evo-devo) biology with growing numbers of datasets for classical model animals like *Drosophila melanogaster* and *Caenorhabditis elegans*. Conceivably, there are endless applications, e.g. oncologists could collaborate together to outline the cellular idiosyncrasies of rare cancer types in order to establish clearer methods of how to detect said cancers; a consortium of neurobiologists could utilize data collected from patients globally to establish more diverse datasets for more nuanced degenerative brain illnesses; or an entomologist interested in developmental processes surrounding imaginal disks in a non-model organism could find a community of like-minded researchers to build a dataset for this purpose. In all relevant fields of biology, the usage and desire for single cell data will increase in tandem with its necessity.

9.2 Database Overview

9.2.1 The Importance of the Type of the Dataset Stored

Commonality and normalization are important for datasets in single cell databases (SCDBs), the information stored is intended for use in downstream processes (as indicated below in the use-case scenarios). To enable comparison between datasets and samples, the correct format and type of dataset are valued. At the surface level databases are constructed with a purpose or aim in mind and therefore usually contain datasets that are comparable. However, not all databases are part of a consortium group, so database managers must consider the appropriate data, file type, and format that would be of most use to the intended audience, that is, the wider scientific community. Common downstream processes include clustering analysis, e.g. principle component analysis (PCA), t-distributed stochastic neighbor embedding (tSNE) or, the more in-vogue, uniform manifold approximation and projection (UMAP), whose popularity comes from its maintaining of global and local clustering. Often then, tools for performing such analyses are provided in embedded panels, chiefly Plotly interactive plots, and matrices of the analysis are often provided as standard. Quality control is another vital step in ensuring the dataset's integrity and is therefore a key component of SCDBs.

9.2.2 The Current Status and What Has Been Done

Single cell databases are relatively new, establishing within the last 5 years, and they are growing rapidly. As the numbers grow, each DB is beginning to find its own niche. Many DBs focus on human and mouse cells, as there is a relatively larger pool of datasets to curate. Intuitively, many DBs are oriented towards providing data for the biomedical field. The Human Cell Atlas (HCA) Data Portal (Regev et al. 2017), run by the HCA Consortium, was established in order to create an atlas of all human cell types (as the name suggests); so far they have acquired data on 4.5 million cells that encompass 33 different organs. For mouse cells, the *Tabula muris* Consortium database (Schaum et al. 2018) has compiled 100,000 cells belonging to 20 different organs and tissues. However, not all DBs are focused on these model organisms, The Single Cell Expression Atlas (Papatheodorou et al. 2020), that is operated by The European Molecular Biology Laboratory's (EMBL) European Bioinformatics Institute (EBI), includes single cell data of four million cells from 14 different species of model animals (*Homo sapiens*, *Mus musculus*, *Danio rerio*, *Gallus gallus*, etc.), plants (*Arabidopsis thaliana*) as well as fungi and protists. Other DBs look to create expression atlases for cells related to specific diseases [e.g. scREAD (Jiang et al. 2020), SC2Diseases (Zhao et al. 2021)], or organs [e.g. HCA's Bone Marrow Single Cell Interactive Web Portal (Hay et al. 2018)].

9.2.3 The Main Feature(s) of the Databases/Data Repositories

As mentioned already, single cell databases can be classified in two ways, either as a pure repository or as an analysis tool. This is not to say that these features are mutually exclusive, there are many databases with features of both; these are, however, the two main functions. To navigate the datasets at the user's disposal, many DBs will initially present the user with a quick search feature on the landing page. Along with navigation features, many DBs have pipelines that allow for curation of metadata to coincide with metadata obtained by the dataset authors. This is all deposited alongside the more vital metrics like cell-specific gene regulation and expression, all of which are curated and are publicly accessible. Naturally, databases will include download (and bulk download) functionality, providing the user with standard file types for downstream analysis, i.e. zipped folders, BAM files, or FASTQ files all depending on the target data. These are common features of a public data repository, many DBs offer extra functionality in the form of analysis tools. These tools are integrated in the database, for example, clustering analysis and gene expression search capabilities of The Broad Institutes Single Cell Portal. Alternatively these tools come as separate entities linked to the data repository, for example, UCSC's Xena tool that is linked to the UCSC Cell Browser (Table 9.1 and Fig. 9.1).

Table 9.1 Summary table of the entities stored in a single cell database

Entity type	Description
Single cell data	Tabulated data related to the cell, in terms of function or gene expression
Metadata	Metadata relates to the study or source of the dataset as well as quality control information about the cell itself
Links	Many SCDBs contain links to external DBs that relate to further downstream analysis or reference data for the cells
Images	Some SCDBs include cell image data to enhance cell QC data, visualizations, e.g. gene body coverage or expression heat maps are also commonplace
Tools	Analysis tools are imbedded or are linked to externally for analysis of the user's own data in tandem with the database's or to compare only the data stored in the DB

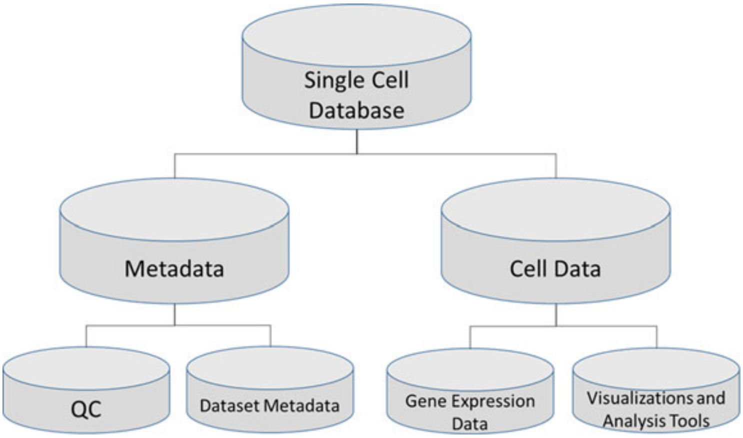


Fig. 9.1 Single cell database overview

9.2.4 Future Update and Availability of the Database

Single cell DBs are fluid and constantly changing. Emerging technologies present new opportunities for researchers coinciding with new challenges for DB managers. 10× genomics' Chromium protocol has allowed for the capture and sampling of 10,000 or more than one million cells in a single study. This creates a boundary for manually curated cell-specific metadata and also limits the attainable depth of sequencing in some cases. As technologies improve and evolve the data collected will increase steadily. Future database design must then adapt to these larger datasets. This is to ensure the size of the datasets does not introduce computing power as a limiting factor and also that the data is comparable with other datasets.

9.3 Content and Architecture of the Database

9.3.1 Type of the Data Stored

The data stored in SCDBs can be separated into categories of quality control, dataset metrics, single cell metrics and related visualizations. As shown in Table 9.2, quality

Table 9.2 Description of single cell database data types

Data category	Data type	Description	Data download format
Quality control metrics	Cell centric	Cell image data—qualitative QC data as well as quantitative, e.g. cDNA concentration, % mitochondrial genes, no. of uniquely mapped reads, no. of reads mapped to too many loci, no. of unmapped reads, and no. of sequenced tags	Jpeg images BAM files, fastq files, and tabulated, csv metadata
	Sample centric	Number of runs, gene body coverage, FastQC Reports	Tabulated csv/tsv files, html FastQC report or in zipped (.gz) folder
Dataset metrics	Technical details	Experiment protocol, cell capture and sequencing protocols, library preparation, instrument specifications, barcode sequence(s)	Tabulated metadata in csv/tsv format
		Source organism and organ(s), disease state, and developmental stage	Tabulated metadata in csv/tsv format
	General details	Dataset title, objective, description, Biomaterial provider, paper citation	Tabulated metadata in csv/tsv format
		Cell count, accession number, and links to INSD gene databases (see Sect. 9.3.5)	Tabulated metadata in csv/tsv format
Single cell data	Cell metrics	Inferred cell type, cell cycle phase, cell ontology, cell line, sequence data	Tabulated metadata in csv/tsv format, fastq, BAM files
	Gene expression	Differential gene expression of all genes or a selection of highly expressed genes or genes of interest, gene descriptions	BAM files, fastq file csv/tsv metadata
		Normalization measures: Transcripts per million reads (TPM) or fragments per kilobase of transcripts per million (FPKM)	Tabulated metadata in csv/tsv format
		Accession number and links to INSD gene databases	URL links
Visualizations	Quality control	Gene body coverage plot, gene expression correlation diagram	Image (jpeg/png)
	Downstream analysis	Cell clustering (PCA, tSNE, or UMAP), differential gene expression heatmaps	Image (jpeg/png) and matrices (csv)

control data are broken down into cell-specific and sample specific metrics. Dataset details include the technical metadata relating to the datasets creation and also the general details of the study of origin and the dataset owners. The single cell metrics are the meat of the data, what is most desirable for users, these are the details of the cells themselves and their gene expression profiles. There are many visualizations offered by SCDBs that display quality control features and also downstream analyses.

9.3.2 Data Collection Methods

The initial forays into single cell transcriptomics (Tang et al. 2009) used existing next generation sequencing platforms, some with an automated and massively parallel setup, e.g. MARS-seq (Jaitin et al. 2014). However, the approach is laborious and technically limiting in some cases. Now, there is the advent of droplet-based approaches as mentioned previously with Drop-seq and 10× Genomics' platform; the process is more streamlined and cell capture is increased by orders of magnitude. This approach involves the disaggregation of tissues and the individual cells are passed through microfluidic devices whereby single cells are captured in oil droplets that also contain beads affixed with primers and barcoded oligonucleotides; these barcoded help to disseminate the cell of origin of each sequence in downstream bioinformatics pipelines. The result is a library of cDNA formed from captured RNA of each cell providing a snapshot of expressed genes. The datasets formed from such an approach are extremely large, often with high percentages of undamaged or "good" cells; the depth of sequencing is sometimes limited with these data collection methods however.

9.3.3 Curation Approaches

Databases working exclusively with datasets originating "in-house," i.e. within their own institution or consortium, have a limited barrier to the curation of datasets for their DB. Before the curation process, there can be communication with the lab producing the dataset so that the formatting and metadata collection is smooth and without gaps or error. In this case, it is a matter of a simple transfer of data, also taking into account training and security protocols to ensure the safeguarding of human data. Databases that rely on datasets from non-affiliated labs may have to jump through a few more hurdles in order to acquire the datasets of interest from the owner of the data. The data targeted may already be in the public domain and both the data and surrounding metadata are readily available. However, to acquire some detailed metadata, communication between both parties may be required. In many cases, data curation is manual, whereas there are ways of automating data processing. There are automated tools that can be utilized for curation of metadata.

Chiefly, the E-Utilities tool from NIH and its accompanying UNIX/LINUX command line tool Entrez Direct (EDirect). This tool can query multiple databases such as Gene Expression Omnibus (GEO), Sequence Read Archive (SRA), and PubMed, for associated dataset and study metadata or more specific queries for sample or run metadata.

9.3.4 Processing Strategy

In SCDBs there is distinct primary and secondary analysis. In primary analysis, raw sequence read .fastq files are accessed from International Nucleotide Sequencing Database Consortium (INSDC) databases. These reads are then processed for QC and are also aligned to reference genomes using sequence aligners such as STAR (Dobin et al. 2013). The resulting BAM files and accompanying mapping log files are used for secondary analysis. The secondary analyses include gene expression quantification, i.e. FPKM and TPM as well as gene expression correlation of highly expressed genes, finally further downstream analysis, e.g. dimensionality reduction (PCA, tSNE, or UMAP).

9.3.5 Dataset Indexing and Accession Numbers

Datasets are indexed in SCDBs by accession numbers using the INSD accession numbers (AN) (Table 9.3). As an example, consider a fictitious dataset's study with the study AN of DRP123456, the "D" identifies this study as curated by the DDBJ database. Within this study is the sample of interest for the SCDB, concerning mouse retinal epithelial cells, with the sample AN of DRS123456. Within this sample the DB would contain all experiments, identified with DRX followed by six unique numbers. And finally within this experiment the user would be able to identify all runs performed with the AN DRX again followed by six unique numbers. In most cases the run AN would uniquely index the cells. However, where this is not the case, the DB may assign a unique identifier, e.g. 1502-A07 using the run number and the well-plate coordinates of the experiment, respectively.

9.3.6 Quality Control Method

Cell and sequence quality control is essential for data integrity. Sequence QC can be inserted into DB pipelines by collecting metrics from alignment, for instance, the percentage of reads mapped to multiple loci; percentage of reads mapped to too many loci; calculation of the percentage possibility of genomic contamination; also quantifying unmapped reads. There are some tools for high throughput QC of

Table 9.3 Description of International Nucleotide Sequence Database Consortium (INSDC) accession numbers

Accession number (AN) description	DNA Data Bank of Japan (DDBJ)	European Nucleotide Archive (ENA)	National Center for Biotechnology Information's (NCBI) Sequence Read Archive (SRA)
Submission AN—details of the organization that submitted the data	DRA123456	ERA123456	SRA123456
Project AN—details of the project, e.g. objective and scope	PRJD[A-Z][0-9]+	PRJE[A-Z][0-9]+	PRJN[A-Z][0-9]+
Study AN—details of the study, e.g. title and abstract	DRP123456	ERP123456	SRP123456
Sample AN—details of the sample, i.e. organism or disease	DRS123456	ERS123456	SRS123456
Biosample AN—descriptive information about the sample, i.e. cell line or tissue biopsy	SAMD[A-Z][0-9]+	SAME[A-Z][0-9]+	SAMN[A-Z][0-9]+
Analysis Object AN—used for internal processes	DRZ123456	ERZ123456	SRZ123456
Experiment AN—the specific experiment details, condition, and protocol info	DRX123456	ERX123456	SRX123456
Run AN—the run number of the experiment	DRR123456	ERR1234567	SRR12345678

sequencing data, for example, FASTQC (Andrews 2010). FASTQC accepts BAM/SAM or fastq files and outputs quality statistics like per-base sequence quality; per-base sequence, N and GC content; per sequence quality scores and per-sequence GC content; sequence length distribution; sequence duplication levels; overrepresented sequences; and kmer content.

9.3.7 Database Update and Maintenance Strategy

New methods of data collection or processing are constantly being made which provides new challenges in data storage and secondary analysis. To account for this, SCDBs are usually created with some wiggle room to account for potential future upgrades, e.g. by allowing for future storage of large datasets with a million or more cells. In some cases, however, a new technology, protocol, or processing program provides new data metrics which were not previously considered by the DB manager. In this case an update of the target data tables is conducted. New data is added to DBs regularly as new studies are published, then the data is processed through the DB pipeline and is curated into the DB.

9.4 Database Access and Data Tools

9.4.1 Accessing and Browsing the Content of Single Cell Databases (Fig. 9.2)

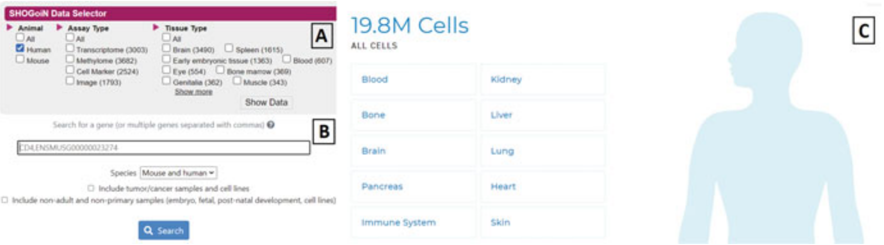
With most SCDBs, the user will initially be presented with a "quick search" feature. This allows for easy navigation to the cell data for the organism, organ, or disease of interest; accessed through a series of radio buttons, drop-down menus, or simply a search bar using keywords (Fig. 9.2). Further options will be available for narrowing searches after an initial quick search in most cases.

9.4.2 How to Query the Database

Data query in SCDB can be performed in order to locate a specific study, to narrow down a quick search further or also to perform some downstream comparison. The DB structure permits for querying of metadata information as a way of locating a study, e.g. to show studies from certain donor groups (Fig. 9.3a). Once a study (or studies) of interest is located, many DBs provide users with downstream comparison tools, for instance, the ability to cluster-specific cell types or compare expression of specific genes between cell types; the example shown is from scREAD (Jiang et al. 2020), an Alzheimer's disease-specific SCDB (Fig. 9.3b).

9.4.3 How to Upload/Download Data to the Database and Fig. 9.4

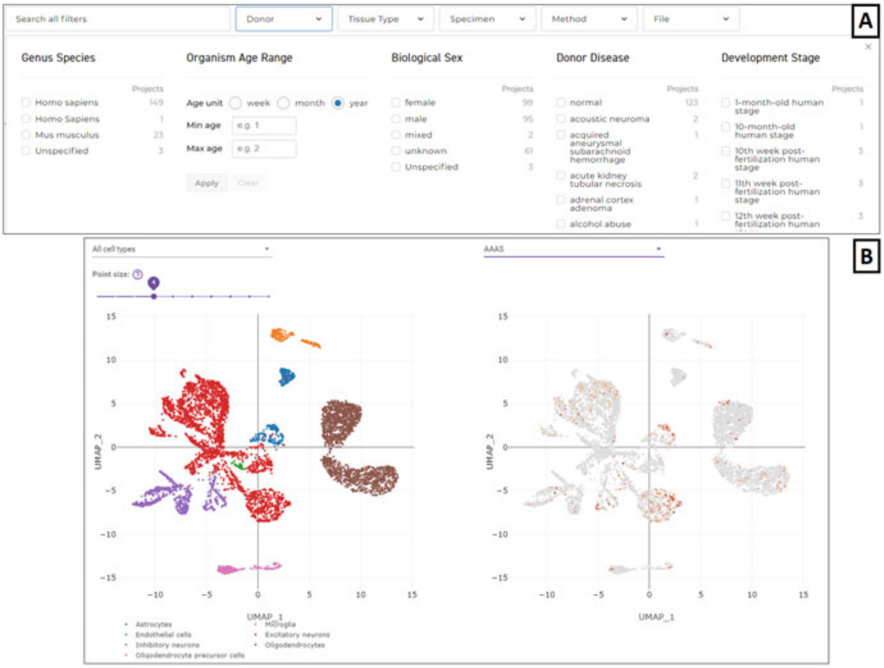
Data download from SCDBs is performed on a dataset-by-dataset basis via links on the dataset summary page. Bulk download capabilities are commonplace, to download all datasets, or datasets related to a certain query/selection criteria. Data



CD4,ENSMUSG00000023274

Species: Mouse and human

Fig. 9.2 Quick search functionality of: (**a**) SHOGoiN DB, (**b**) HCA Data Portal, and (**c**) PanglaoDB



Genus Species	Projects
□ Homo sapiens	149
□ Homo Sapiens	1
□ Mus musculus	23
□ Unspecified	3

Organism Age Range

Age unit ○ week ○ month ● year

Min age e.g. 1

Max age e.g. 2

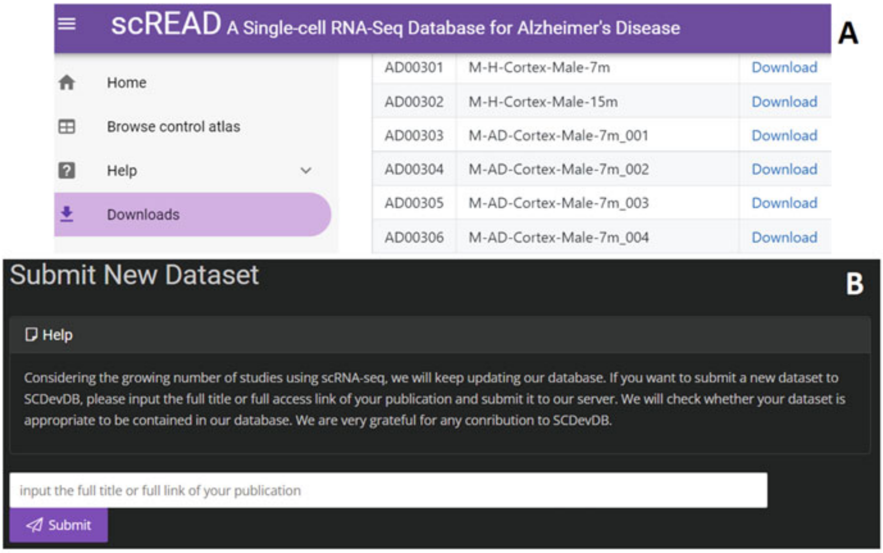
Apply Clear

Biological Sex	Projects
□ female	98
□ male	95
□ mixed	2
□ unknown	61
□ Unspecified	3

Donor Disease	Projects
□ normal	123
□ acoustic neuroma	2
□ acquired aneurysmal subarachnoid hemorrhage	1
□ acute kidney tubular necrosis	2
□ adrenal cortex adenoma	1
□ alcohol abuse	1

Development Stage	Projects
□ 1-month-old human stage	1
□ 10-month-old human stage	1
□ 10th week post-fertilization human stage	3
□ 11th week post-fertilization human stage	3
□ 12th week post-fertilization human	3

Fig. 9.3 Querying capabilities of (**a**) HCA Data Portal and (**b**) downstream querying and comparison in scREAD


≡
scREAD A Single-cell RNA-Seq Database for Alzheimer's Disease
A
Home
Browse control atlas
Help
Downloads
B
Submit

AD00301	M-H-Cortex-Male-7m	Download
AD00302	M-H-Cortex-Male-15m	Download
AD00303	M-AD-Cortex-Male-7m_001	Download
AD00304	M-AD-Cortex-Male-7m_002	Download
AD00305	M-AD-Cortex-Male-7m_003	Download
AD00306	M-AD-Cortex-Male-7m_004	Download

Submit New Dataset

Help

Considering the growing number of studies using scRNA-seq, we will keep updating our database. If you want to submit a new dataset to SCDevDB, please input the full title or full access link of your publication and submit it to our server. We will check whether your dataset is appropriate to be contained in our database. We are very grateful for any conribution to SCDevDB.

input the full title or full link of your publication

Fig. 9.4 (**a**) Example of dataset download links from scREAD. (**b**) The dataset upload input for SCdevDB available for download include processed reads (BAM or FASTQ files), clustering matrices, or metadata tables (Fig. 9.4a). Some DBs also allow for data upload. Here data can be requested to be curated to the database itself (Fig. 9.4b). Another capability is that data can be uploaded temporarily in a sandbox in order to use the DB's analysis tools. The user can then compare gene expression levels with existing datasets or cell clustering with other related data.

9.4.4 Programming and Automated Technique to Access the Database

1. *Web Services*: These are application-to-application interaction tools performed over the web. They directly interact with the SQL, Javascript, etc. code within the database to access the data within. The requirements for this process are: Extensible Markup Language (XML), Simple Object Access Protocol (SOAP), Web Services Description Language (WSDL), and Universal Description, Discovery, and Integration (UDDI).
2. *File Transfer Protocol* (*FTP*): FTP allows for client to server communication and file transfer. Authentication can be used with simple text login, i.e. username and password. If the database is set up as such, authentication is sometimes not necessary.
3. *Application Programming Interface* (*API*): APIs are embedded within the database for connecting to other bioinformatics tools. They are also used by SCDBs to connect to related tools, an example of this is the DAVID API (david.ncifcrf.gov/content.jsp?file=DAVID_API.html), used for functional annotation processes.
4. *Bioinformatics tools (R/Python) packages*: There are also R or Python specific tools to achieve the same functionality. An example is MySQLdb module, a Python module that utilizes the Python API to connect to MySQL databases. There are similar R packages, e.g. the DBI package.

9.4.5 Database Integration Strategy

Another key feature of the SCDB is smooth integration with external DBs, or connected tools; this is seem most clearly with the INSDC accession number system. Here, smooth integration is essential for data reliability and universal understanding. Other integration forms are the use of external tools that connect with and utilize a SCDB. UCSC's Cell Browser acts as the data repository that connects seamlessly with its partner tool called UCSC Xena (www.xena.ucsc.edu); they are not mutually exclusive and can be used separately. The final integration method involves consortia. A consortium's data repository can be linked with many databases that serve differing roles or occupy a specific SCDB niche. The obvious example here is the HCA Data Portal, which acts as the data repository for other DBs like UCSC cell browser, Broad Institute's Single Cell Portal (currently Beta), and also the Single Cell Expression Atlas; analysis tools like cellxgene (Li et al. 2020) are also integrated.

9.5 Use-Cases and Capabilities of Single Cell Database

9.5.1 *Simple Use-Case Example*

Consider a researcher in neurological development, to illustrate a point in the article or a presentation they wish to have a figure and data relating to the expression of genes of interest in the brain during development. The researcher visits SCdevDB (Wang et al. 2019) and searches for their gene(s) of interest (Fig. 9.5).

In this case, the researcher searches for the expression of CBLN2 (*cerebellin 2 precursor*) and specifies the frontal lobe. The result is outputted immediately (Fig. 9.6). The user can now see stage-by-stage expression values for CBLN2 in an interactive Plotly plot, simply and efficiently.

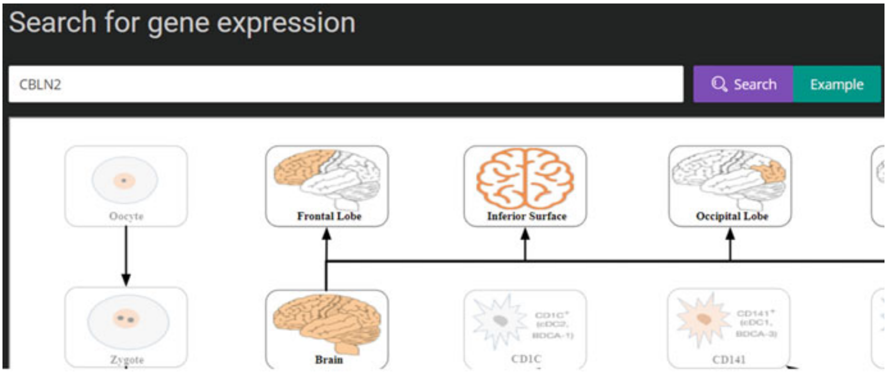


Fig. 9.5 From SCdevDB showing the search feature, the user can search for a gene symbol or accession number, then can select the tissue of interest to see gene expression at each developmental stage

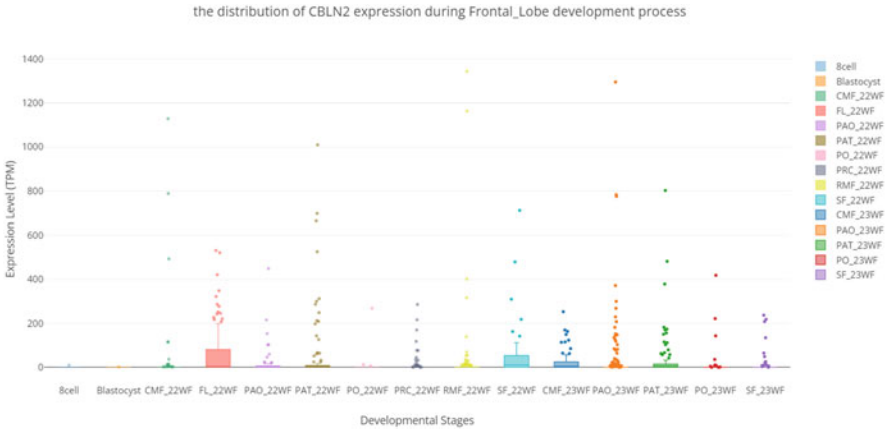


Fig. 9.6 The expression values of CLBN2 in the frontal lobe at different developmental stages, produced by SCdevDB

9.5.2 Intermediate Use-Case Example

Consider a researcher in Evo-devo biology, using *Drosophila melanogaster* as a model organism to study ocular development; this researcher is interested in gene expression during development. This researcher knows there have been studies utilizing scRNA-seq in this area so they access EMBL-EBI's Single Cell Expression Atlas and select studies related to *D. melanogaster* (Fig. 9.7).

As there are only seven *D. melanogaster* datasets (currently), the researcher can easily locate the dataset they are interested in from the list (Fig. 9.8).

Accessing the dataset summary, the researcher can view a dimensionality reduction plot (tSNE) and can then navigate to the "Marker Gene" tab to view the top five expressed genes for each cell cluster. In no time at all the researcher can view potential genes of interest for their own study; they have the ability to download the study metadata, cluster matrix, and marker gene heat map plot (Fig. 9.9).

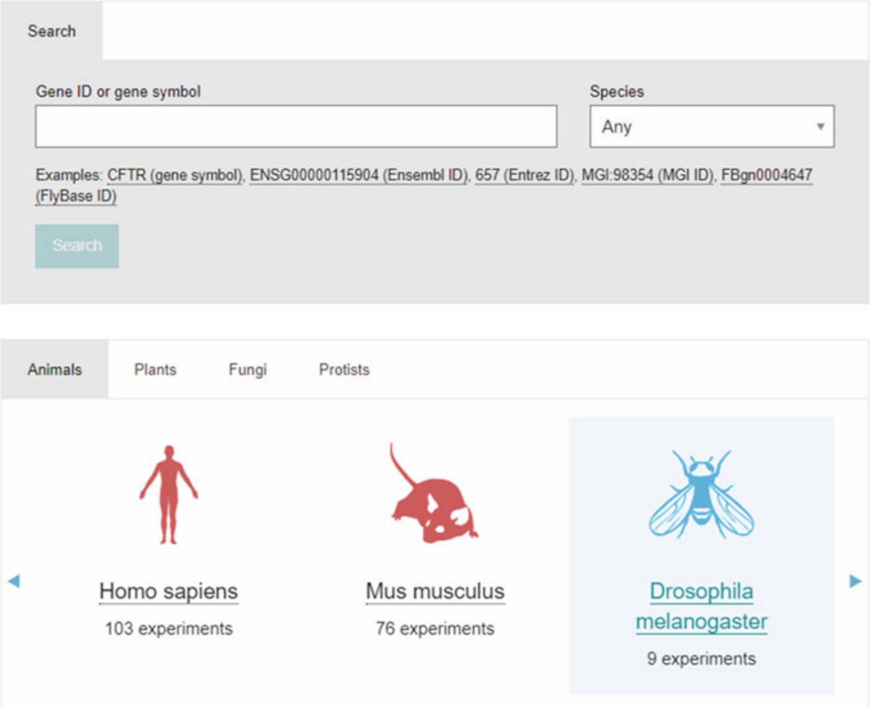
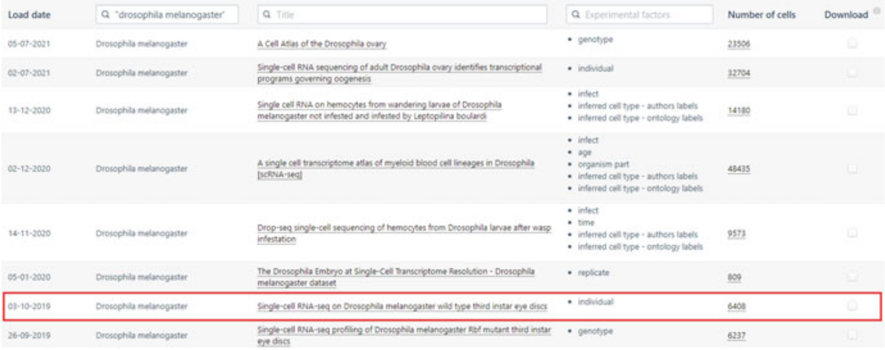
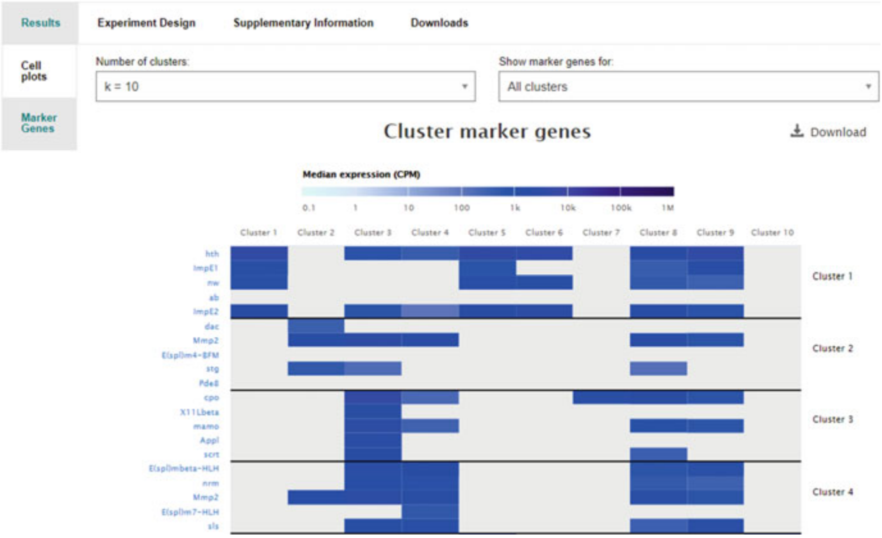


Fig. 9.7 Landing page of Single Cell Expression Atlas



Load date	"drosophila melanogaster"	Title	Experimental factors	Number of cells	Download
05-07-2021	Drosophila melanogaster	A Cell Atlas of the Drosophila ovary	• genotype	23506	□
02-07-2021	Drosophila melanogaster	Single-cell RNA sequencing of adult Drosophila ovary identifies transcriptional programs governing oogenesis	• individual	32704	□
13-12-2020	Drosophila melanogaster	Single cell RNA on hemocytes from wandering larvae of Drosophila melanogaster not infested and infested by Leptopilina boulardi	• infect • inferred cell type - authors labels • inferred cell type - ontology labels	14180	□
02-12-2020	Drosophila melanogaster	A single cell transcriptome atlas of myeloid blood cell lineages in Drosophila [scRNA-seq]	• infect • age • organism part • inferred cell type - authors labels • inferred cell type - ontology labels	48435	□
14-11-2020	Drosophila melanogaster	Drop-seq single-cell sequencing of hemocytes from Drosophila larvae after wasp infestation	• infect • time • inferred cell type - authors labels • inferred cell type - ontology labels	9573	□
05-01-2020	Drosophila melanogaster	The Drosophila Embryo at Single-Cell Transcriptome Resolution - Drosophila melanogaster dataset	• replicate	809	□
03-10-2019	Drosophila melanogaster	Single-cell RNA-seq on Drosophila melanogaster wild type third instar eye discs	• individual	6408	□
26-09-2019	Drosophila melanogaster	Single-cell RNA-seq profiling of Drosophila melanogaster Rbf mutant third instar eye discs	• genotype	6237	□

Fig. 9.8 *Drosophila melanogaster* single cell RNA-seq datasets available on EMBL-EBI Single Cell Expression Atlas



Number of clusters: k = 10

Show marker genes for: All clusters

Fig. 9.9 Cluster-specific differential gene expression for *Drosophila melanogaster* instar eye disks from EMBL-EBI Single Cell Expression Atlas

9.5.3 Advanced Use-Case and Fig. 9.7 with Panels

Consider another user who is researching stem cells, they want to see expression patterns of common cell type marker genes in bone marrow samples for a study they are planning in the future. The user accesses PanglaoDB (Franzén et al. 2019) and accesses the "Cell Type Marker" feature (Fig. 9.10). The user then filters for the type of cell they are interested in.

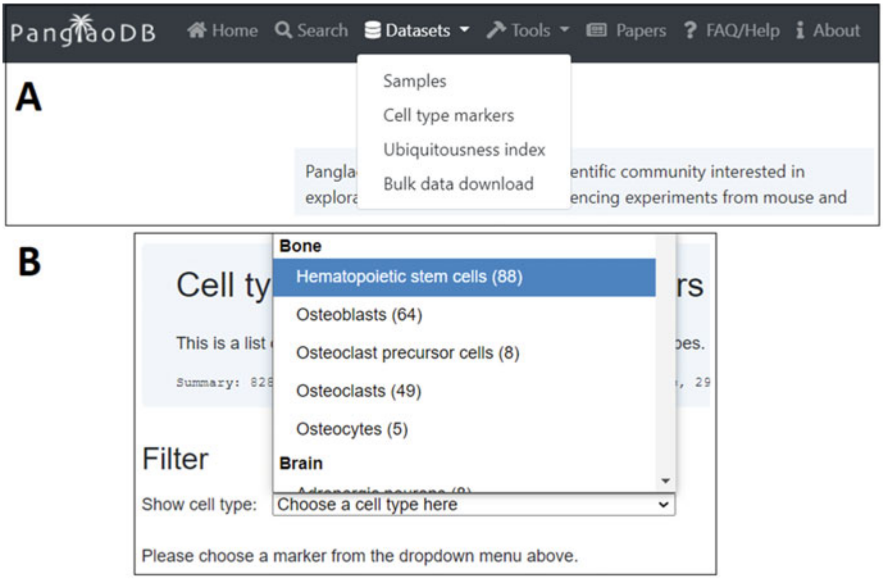
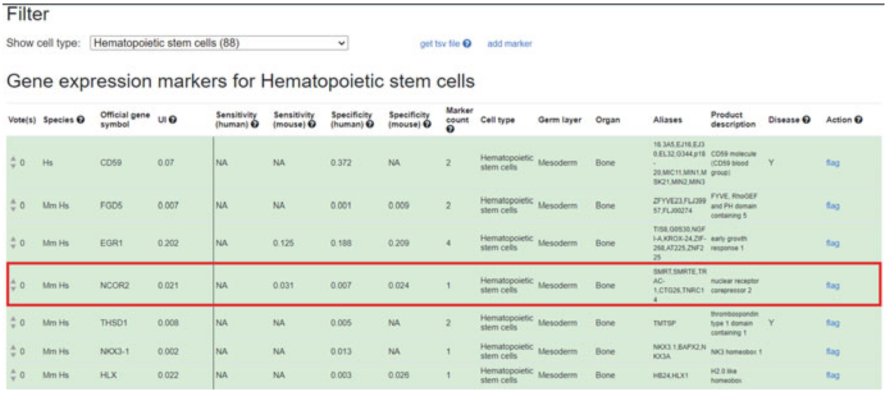


Fig. 9.10 (**a**) The landing page of PanglaoDB, the user can select a feature for cell type markers from the top dropdown menu. (**b**) From the "cell type markers" page the user can then filter for the cells of interest

The user can then browse the list of cell type markers and select the genes of interest using metrics like mouse or human specificity and sensitivity as well as ubiquitousness (Fig. 9.11).

Once the genes of interest are noted, the user can then return to the samples available on the database and can filter for the type of sample they are interested in. The user wishes to see the available human bone marrow samples all using the 10× Chromium protocol; this is because the user wants to normalize their data with the same protocol for downstream analysis (Fig. 9.12a). From the filtering parameters the user can easily browse the samples of interest for them (Fig. 9.12b).

After selecting the samples available, the user can see metadata metrics and dimensionality reduction plots. An interactive UMAP plot is selected (Fig. 9.13a). The user can now overlay the gene expression of the gene of interest that was noted earlier (Fig. 9.13b). In this way, the user can see cluster-specific gene expression of genes of interest; this precision is not possible with bulk RNA sequencing data and thus demonstrates the power of scRNA-seq.



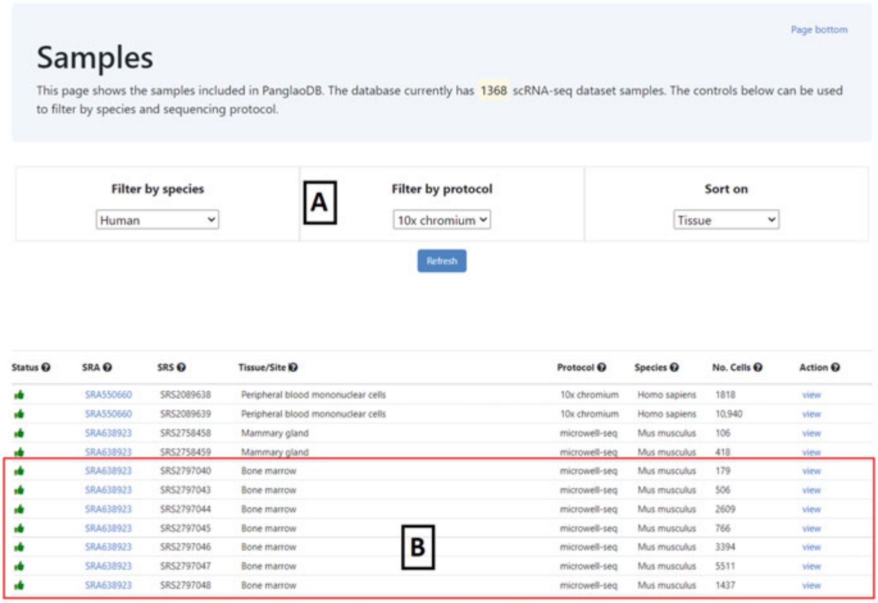
Filter

Show cell type: Hematopoietic stem cells (88)

Gene expression markers for Hematopoietic stem cells

Vote(s)	Species	Official gene symbol	UI	Sensitivity (human)	Sensitivity (mouse)	Specificity (human)	Specificity (mouse)	Marker count	Cell type	Germ layer	Organ	Aliases	Product description	Disease	Action
0	Hs	CD59	0.07	NA	NA	0.372	NA	2	Hematopoietic stem cells	Mesoderm	Bone	16.3A5,EJ16,EJ30,EL32,G344,p18-20,MIC11,MIN1,MSK21,MIN2,MIN3	CD59 molecule (CD59 blood group)	Y	flag
0	Mm Hs	FGD5	0.007	NA	NA	0.001	0.009	2	Hematopoietic stem cells	Mesoderm	Bone	ZFYVE23,FLJ39957,FLJ00274	FYVE, RhoGEF and PH domain containing 5		flag
0	Mm Hs	EGR1	0.202	NA	0.125	0.188	0.209	4	Hematopoietic stem cells	Mesoderm	Bone	TIS8,G0S30,NGFI-A,KROX-24,ZIF-268,AT225,ZNF225	early growth response 1		flag
0	Mm Hs	NCOR2	0.021	NA	0.031	0.007	0.024	1	Hematopoietic stem cells	Mesoderm	Bone	SMRT,SMRTE,TRAC-1,CTG26,TNRC14	nuclear receptor corepressor 2		flag
0	Mm Hs	THSD1	0.008	NA	NA	0.005	NA	2	Hematopoietic stem cells	Mesoderm	Bone	TMTSP	thrombospondin type 1 domain containing 1	Y	flag
0	Mm Hs	NKX3-1	0.002	NA	NA	0.013	NA	1	Hematopoietic stem cells	Mesoderm	Bone	NKX3.1,BAPX2,NKX3A	NK3 homeobox 1		flag
0	Mm Hs	HLX	0.022	NA	NA	0.003	0.026	1	Hematopoietic stem cells	Mesoderm	Bone	HB24,HLX1	H2.0 like homeobox		flag

Fig. 9.11 List of cell type marker genes and relating metrics for Bone Hematopoietic stem cells on PanglaoDB



Samples

This page shows the samples included in PanglaoDB. The database currently has 1368 scRNA-seq dataset samples. The controls below can be used to filter by species and sequencing protocol.

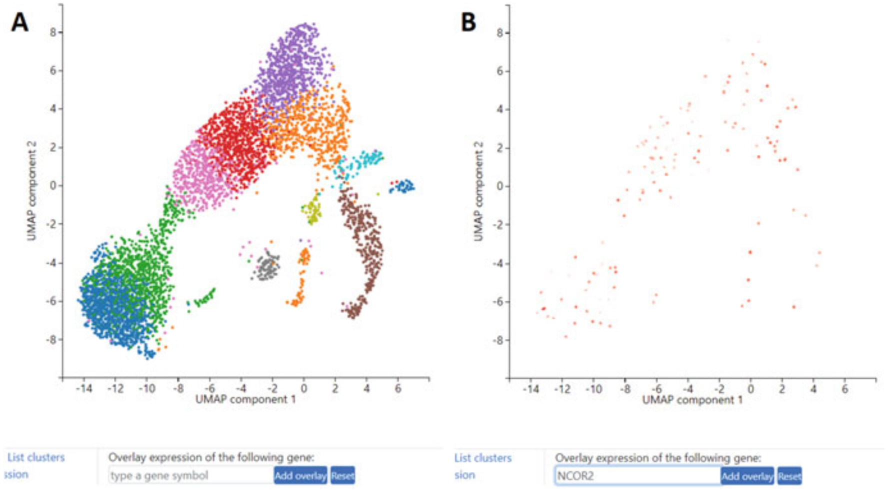
Filter by species: Human

Filter by protocol: 10x chromium

Sort on: Tissue

Status	SRA	SRS	Tissue/Site	Protocol	Species	No. Cells	Action
👍	SRA550660	SRS2089638	Peripheral blood mononuclear cells	10x chromium	Homo sapiens	1818	view
👍	SRA550660	SRS2089639	Peripheral blood mononuclear cells	10x chromium	Homo sapiens	10,940	view
👍	SRA638923	SRS2758458	Mammary gland	microwell-seq	Mus musculus	106	view
👍	SRA638923	SRS2758459	Mammary gland	microwell-seq	Mus musculus	418	view
👍	SRA638923	SRS2797040	Bone marrow	microwell-seq	Mus musculus	179	view
👍	SRA638923	SRS2797043	Bone marrow	microwell-seq	Mus musculus	506	view
👍	SRA638923	SRS2797044	Bone marrow	microwell-seq	Mus musculus	2609	view
👍	SRA638923	SRS2797045	Bone marrow	microwell-seq	Mus musculus	766	view
👍	SRA638923	SRS2797046	Bone marrow	microwell-seq	Mus musculus	3394	view
👍	SRA638923	SRS2797047	Bone marrow	microwell-seq	Mus musculus	5511	view
👍	SRA638923	SRS2797048	Bone marrow	microwell-seq	Mus musculus	1437	view

Fig. 9.12 Sample list of PanglaoDB with filter parameters (**a**) means the user can easily select the datasets they are interested in (**b**)



Overlay expression of the following gene: type a gene symbol

Overlay expression of the following gene: NCOR2

Fig. 9.13 (**a**) UMAP projection of a bone marrow dataset on PanglaoDB. (**b**) Gene expression overlay for the CTNNAL1 gene

9.6 Summary and Future Development of the Database

As is with any emerging technologies, particularly those that have a large data production, new systems for data storage, analysis, and sharing are required. Single cell RNA sequencing is no different. The basis for single cell databases is established with blend of well-trodden database conventions and emerging technologies and tools. Common elements of the databases include a normalizing primary analysis pipeline; customizable, integrated secondary analyses tools with interactive visualizations; integration with INSDC databases; integration with downstream functional analysis or ontology tools; download and bulk download tools for cell metadata and genomic data. There are exceptional efforts being undertaken to create complete atlases for human and mouse cells in all tissues and organs. Such efforts are progressing to other model organisms, e.g. *Drosophila melanogaster.* In the face of public health crises, to tackle neglected tropical diseases or to understand the interplay of dividing cells at the initiation of an organism's life, single cell studies will dominate this era of biotechnology.

So, as single cell transcriptomics becomes routine, smaller boutique databases are being formed to accommodate those more niche fields. Future developments of these databases must involve analysis and storage capabilities for larger datasets. Cell-specific layouts will require adjustments to accommodate these mega datasets with millions of cells. Also, current technologies will continue strives to increase the depth of sequencing for these larger datasets. The data that is curated and processed must be kept available to the public scientific body without technological or financial barriers; pre-processing of computationally exhaustive actions can abate this. Larger single cell databases will look to become a one-stop-shop for researchers to upload, download, view, and analyze their own data among a library of compatible datasets.

Database name	Institution/consortium (country)	URL
SCPortalen	RIKEN IMS (Japan)	www.single-cell.riken.jp/
Human Cell Atlas Data Portal	HCA Consortium	www.data.humancellatlas.org/
Single Cell Expression Atlas	EMBL-EBI (Cambridge, UK)	www.ebi.ac.uk/gxa/sc/home
Single Cell Portal (beta)	Broad Institute (MIT/Harvard, USA)	www.singlecell.broadinstitute.org/single_cell
UCSC Cell Browser	UCSC	www.cells.ucsc.edu/
UCSC Xena	UCSC	www.xena.ucsc.edu/
PanglaoDB	Karolinska Institutet (Sweden)	www.panglaodb.se/index.html
SCDevDB	University of Hong Kong	www.scdevdb.deepomics.org/
scRNASeqDB	University of Texas (USA)	www.bioinfo.uth.edu/scrnaseqdb/
scQuery	Carnegie Mellon University (USA)	www.scquery.cs.cmu.edu/
Vascular Single Cell Database	Tianjin Neurological Institute (China) and karolinska Inst. (Sweden)	www.betsholtzlab.org/VascularSingleCells/database.html
SHOGoiN	CiRA(Kyoto University, Japan)	www.stemcellinformatics.org/
ASAP (Automated Single Cell Analysis Portal)	Swiss Institute of Bioinformatics	www.asap.epfl.ch/
Human Cell Atlas Bone Marrow Single Cell Interactive Web Portal	HCA Consortium	www.altanalyze.org/ICGS/HCA/splash.php
scREAD	Ohio State University (USA)	www.bmbls.bmi.osumc.edu/scread/
SC2Disease	Xi'an University (China)	www.easybioai.com/sc2disease/
Tabula Muris	Tabula Muris Consortium	www.tabula-muris.ds.czbiohub.org/

No external funding

References

Abugessaisa I, Noguchi S, Hasegawa A, Kondo A, Kawaji H, Carninci P, Kasukawa T (2019) refTSS: a reference data set for human and mouse transcription start sites. J Mol Biol 431(13):2407–2422. https://doi.org/10.1016/j.jmb.2019.04.045

Andrews S (2010) FastQC: a quality control tool for high throughput sequence data. http://www.bioinformatics.babraham.ac.uk/projects/fastqc/

Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, Batut P, Chaisson M, Gingeras TR (2013) STAR: ultrafast universal RNA-seq aligner. Bioinformatics 29(1):15–21. https://doi.org/10.1093/bioinformatics/bts635

Franzén O, Gan LM, Björkegren JLM (2019) PanglaoDB: a web server for exploration of mouse and human single-cell RNA sequencing data. Database 2019(1):46. https://doi.org/10.1093/database/baz046

Hay SB, Ferchen K, Chetal K, Grimes HL, Salomonis N (2018) The Human Cell Atlas bone marrow single-cell interactive web portal. Exp Hematol 68:51–61. https://doi.org/10.1016/j.exphem.2018.09.004

He L, Vanlandewijck M, Mäe MA, Andrae J, Ando K, Del Gaudio F, Nahar K, Lebouvier T, Laviña B, Gouveia L, Sun Y, Raschperger E, Segerstolpe Å, Liu J, Gustafsson S, Räsänen M, Zarb Y, Mochizuki N, Keller A et al (2018) Data descriptor: single-cell RNA sequencing of mouse brain and lung vascular and vessel-associated cell types. Sci Data 5(1):1–11. https://doi.org/10.1038/sdata.2018.160

Hwang B, Lee JH, Bang D (2018) Single-cell RNA sequencing technologies and bioinformatics pipelines. In: Experimental and molecular medicine, volume 50, issue 8. Nature Publishing Group. https://doi.org/10.1038/s12276-018-0071-8

Islam S, Zeisel A, Joost S, La Manno G, Zajac P, Kasper M, Lönnerberg P, Linnarsson S (2014) Quantitative single-cell RNA-seq with unique molecular identifiers. Nat Methods 11(2):163–166. https://doi.org/10.1038/nmeth.2772

Jaitin DA, Kenigsberg E, Keren-Shaul H, Elefant N, Paul F, Zaretsky I, Mildner A, Cohen N, Jung S, Tanay A, Amit I (2014) Massively parallel single-cell RNA-seq for marker-free decomposition of tissues into cell types. Science 343(6172):776–779. https://doi.org/10.1126/science.1247651

Jiang J, Wang C, Qi R, Fu H, Ma Q (2020) scREAD: a single-cell RNA-seq database for Alzheimer's disease. IScience 23(11):101769. https://doi.org/10.1016/j.isci.2020.101769

Li K, Ouyang Z, Lin D, Mingueneau M, Chen W, Sexton D, Zhang B (2020) Cellxgene VIP unleashes full power of interactive visualization, plotting and analysis of scRNA-seq data in the scale of millions of cells. BioRxiv:2020.08.28.270652. https://doi.org/10.1101/2020.08.28.270652

Macosko EZ, Basu A, Satija R, Nemesh J, Shekhar K, Goldman M, Tirosh I, Bialas AR, Kamitaki N, Martersteck EM, Trombetta JJ, Weitz DA, Sanes JR, Shalek AK, Regev A, McCarroll SA (2015) Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets. Cell 161(5):1202–1214. https://doi.org/10.1016/j.cell.2015.05.002

Papatheodorou I, Moreno P, Manning J, Fuentes AMP, George N, Fexova S, Fonseca NA, Füllgrabe A, Green M, Huang N, Huerta L, Iqbal H, Jianu M, Mohammed S, Zhao L, Jarnuczak AF, Jupp S, Marioni J, Meyer K et al (2020) Expression Atlas update: from tissues to single cells. Nucleic Acids Res 48(D1):D77–D83. https://doi.org/10.1093/nar/gkz947

Regev A, Teichmann SA, Lander ES, Amit I, Benoist C, Birney E (2017) The human cell atlas. eLife 6

Schaum N, Karkanias J, Neff NF, May AP, Quake SR, Wyss-Coray T, Darmanis S, Batson J, Botvinnik O, Chen MB, Chen S, Green F, Jones RC, Maynard A, Penland L, Pisco AO, Sit RV, Stanley GM, Webber JT et al (2018) Single-cell transcriptomics of 20 mouse organs creates a Tabula muris. Nature 562(7727):367–372. https://doi.org/10.1038/s41586-018-0590-4

Tang F, Barbacioru C, Wang Y, Nordman E, Lee C, Xu N, Wang X, Bodeau J, Tuch BB, Siddiqui A, Lao K, Surani MA (2009) mRNA-Seq whole-transcriptome analysis of a single cell. Nat Methods 6(5):377–382. https://doi.org/10.1038/nmeth.1315

Wang Z, Feng X, Li SC (2019) SCDevDB: a database for insights into single-cell gene expression profiles during human developmental processes. Front Genet 10(SEP):903. https://doi.org/10.3389/fgene.2019.00903

Zhao T, Lyu S, Lu G, Juan L, Zeng X, Wei Z, Hao J, Peng J (2021) SC2disease: a manually curated database of single-cell transcriptome for human diseases. Nucleic Acids Res 49(D1):D1413–D1419. https://doi.org/10.1093/nar/gkaa838

Zheng GXY, Terry JM, Belgrader P, Ryvkin P, Bent ZW, Wilson R, Ziraldo SB, Wheeler TD, McDermott GP, Zhu J, Gregory MT, Shuga J, Montesclaros L, Underwood JG, Masquelier DA, Nishimura SY, Schnall-Levin M, Wyatt PW, Hindson CM et al (2017) Massively parallel digital transcriptional profiling of single cells. Nat Commun 8(1):1–12. https://doi.org/10.1038/ncomms14049

Chapter 10
scIVA: Single Cell Database and Tools for Interactive Visualisation and Analysis

Liam M. Crowhurst, Onkar Mulay, Nathan Palpant, and Quan H. Nguyen

Abstract Single cell RNA sequencing (scRNA-seq) data is advancing the understanding of complex biology by enabling the investigation of cell-specific biological processes. A scRNA-seq gene count matrix is often thousand times larger than that from a traditional (bulk) RNA sequencing dataset. While the technology to generate data for hundreds of thousands of cells is now established, the large data poses challenges for experimental biologists, who are experts in their biological domain, but are not computationally ready to analyse and understand the data. Here we present scIVA, an interactive visualisation and analysis pipeline available as a web application, allowing biologists to easily mine their scRNA-seq data. We show that scIVA provides a lightweight and highly accessible framework to build interactive databases for data mining of scRNA-seq data. We present the use of scIVA to build two sample interactive datasets. These include a single cell dataset for in vitro pluripotent stem cells (a simple use case) and a cardiomyocyte differentiation time course dataset (a more complex use case). scIVA is publicly accessible, and we expect that it can be used as a streamlined, interactive, and intuitive single cell data mining platform for experimentalists.

Keywords Genomics · DNA variants · Gene expression · Patiotemporal data · Multidimensional sequencing · Imaging data · Biological regulatory networks

10.1 Introduction

Summary: single cell RNA sequencing (scRNA-seq) data is advancing the understanding of complex biology by enabling the investigation of cell-specific biological processes. A scRNA-seq gene count matrix is often thousand times larger than a traditional (bulk) RNA sequencing dataset. While technology to generate data for

L. M. Crowhurst · O. Mulay · N. Palpant · Q. H. Nguyen (✉)
Institute for Molecular Bioscience, University of Queensland, Brisbane, QLD, Australia
e-mail: quan.nguyen@uq.edu.au

© The Author(s), under exclusive license to Springer Nature Singapore Pte Ltd. 2021

I. Abugessaisa, T. Kasukawa (eds.), *Practical Guide to Life Science Databases*,
https://doi.org/10.1007/978-981-16-5812-9_10

hundreds of thousands of cells is now established, the large data poses challenges for experimental biologists, who are experts in their biological domain, but are not computationally ready to analyse and understand the data. Here we present *scIVA*, an interactive visualisation and analysis pipeline available as a web application, which allows biologists to easily mine their scRNA-seq data. We show that *scIVA* provides a light-weight and highly accessible framework to build interactive databases for data mining of scRNA-seq data. We present the use of *scIVA* to build two sample interactive datasets. These include a single cell dataset for in vitro pluripotent stem cells (a simple use case) and a cardiomyocyte differentiation time course dataset (a more complex use case).

Purpose of the database: *scIVA* provides a streamlined, interactive, and intuitive data mining platform for those with little or no coding experience to gain insights into large scRNA-seq datasets. Users, such as experimental biologists, can use point-and-click control to upload data, execute exploratory data analysis (EDA), visualise the analysis results, generate new hypotheses and analysis questions, perform statistical tests, produce summary tables and high quality plots that can be exported.

Data input for scIVA: scRNA-seq raw count data can be uploaded and analysed by *scIVA* workflow. In the case that partial analyses have been done prior to *scIVA*, users can also upload cell-type information and gene list of interest for more targeted analysis.

Availability: The *scIVA* web application is freely available at http://computationalgenomics.com.au/shiny/scIVA/ and the open-source code is maintained at https://github.com/BiomedicalMachineLearning/scIVA/. The two interactive datasets shown as use cases of *scIVA* are maintained in our Linux server and are publicly accessible at http://computationalgenomics.com.au/shiny/hipsc/ and http://computationalgenomics.com.au/shiny/hipsc2cm/.

10.2 Database Overview

10.2.1 scRNA-seq Technology and Single Cell Data

Knowledge of how gene expression affects phenotypes has mainly been derived using data generated from bulk samples consisting of millions of cells; meaning cellular differences are averaged out. Although the ensemble approaches can successfully characterise dominant patterns shared by the majority of cells, important changes occurring on subsets of cells likely remain hidden. Examples of such subtle but critical changes include early evolution of tumours from original single cancerous stem cells (Schmidt and Efferth 2016) or cell-type specific responses/resistance to drugs (Schmidt and Efferth 2016; Miyamoto et al. 2015). Due to genotypic or phenotypic heterogeneity between cells, the unambiguous inference of the causal relationship between the differences in genome sequences or gene expression and changes in phenotypes can be achieved most accurately at single cell level (Macaulay et al. 2017). Recent technological advances have allowed the measurement of gene expression at single cell resolution, enabling specific, sensitive, and

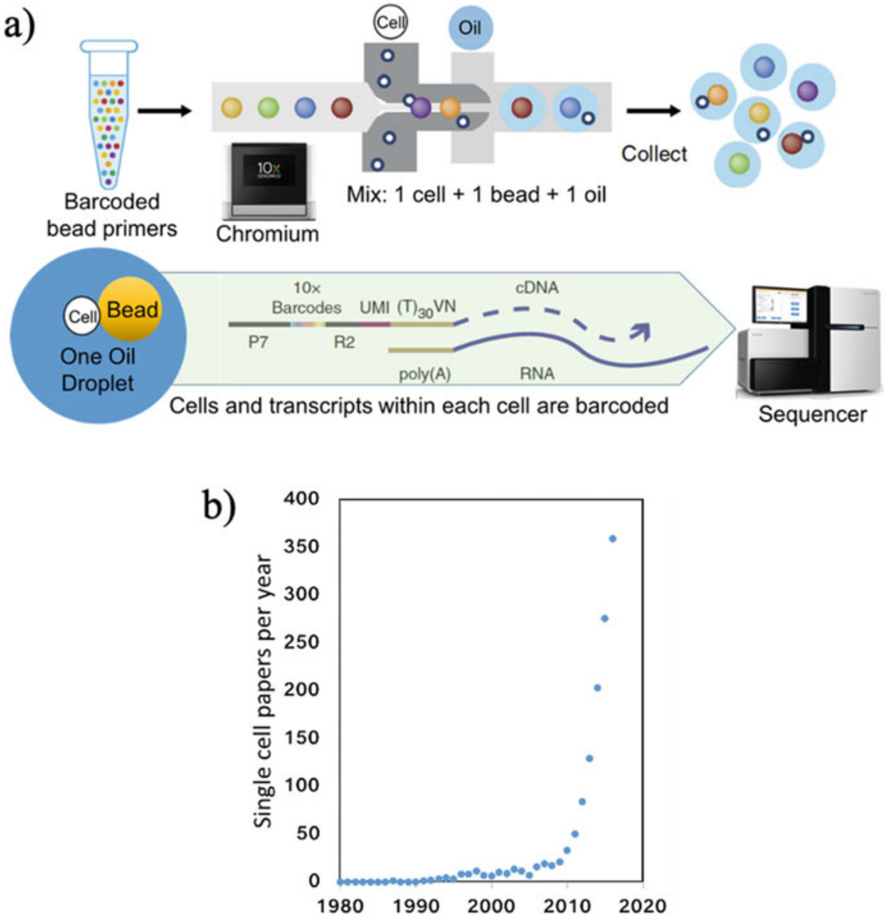


Fig. 10.1 Large-scale single-cell transcriptome sequencing. (**a**) Thousands of droplets can be separated by the Chromium platform. Cell lysis and cDNA synthesis reactions occur inside a droplet (adapted from Zheng et al. 2017). Each cDNA molecule is linked to a cell barcode (10× Barcode) and a unique identifier (UMI) for the molecule. Barcoded cDNAs were pooled and sequenced. Output reads can be computationally assigned to the original transcripts and cells based on the barcodes and identifiers. (**b**) Exponential growth in scRNA-seq research publications (Data from PubMed till 2018)

novel discovery of biological processes, which were undetectable by standard whole-tissue sequencing technologies (Regev et al. 2017; Giladi and Amit 2018).

Single cell transcriptome sequencing is an ultra-sensitive technology that can compartmentalise serial biochemical reactions to index individual RNA molecules to the original cell and amplify indexed molecules to create libraries ready for next generation sequencing (Wang and Navin 2015) (Fig. 10.1a). By using microfluidics device like the 10× Genomics Chromium platform, each single cell is separated into one oil droplet. Serial reactions occur in each droplet to add barcodes to sequencing reads that allow to find which cells the reads belong to, and thus each cell has an expression profile measured. scRNA-seq created unprecedented resolution (individual cells) and sample size (each cell is a sample) to decompose gene expression data of a mixed population into that of different cell subtypes (Altschuler and Wu 2010). scRNA-seq technology has rapidly evolved (Fig. 10.1b), creating a valuable opportunity for devising innovative analysis approaches to exploit the unprecedented high-resolution and large-scale single cell data. Single cell gene expression data is commonly represented as a matrix consisting of tens of thousands of genes and thousands to millions of cells. These data have created a demand among biologists for quick and interactive visualisation and exploratory analysis.

10.2.2 Landscape of scRNA-seq Data Analysis Tools

Although increasing scRNA-seq data analysis tools have been developed in the past few years, few interactive tools are available for users with little or no programming experience (Fig. 10.2). Such 'user-friendly' tools are increasingly in demand, as data is exponentially produced and more experimental groups have access to single cell sequencing technologies. As of July 2021, the Human Cell Atlas (HCA) data portal has made available 13.8 M cells from 70 organs of 1.3 thousand donors, with 125 projects by 281 labs (https://data.humancellatlas.org/).

Although data visualisation is the most common analysis task, most tools do not have interactive visualisation options. Often experimental biologists are able to raise numerous intuitive biologically relevant questions. Interactive tools are especially powerful to facilitate the hypothesis forming process by these researchers. Thus, point-and-click, web-based interfaces that can be applied interactively are useful in single cell biology research. Interactive tools also enable experimental researchers and readers of single cell research publications to explore the huge amount of data that is often not exhaustedly explored by any single investigator. Recently, several interactive tools have been developed to serve such a demand. One of the first and most well-established webtools is *ASAP* (Gardeux et al. 2017), which has now been expanded to the fifth version. Commercial platform like BioTuring is also available for users with a sign-up account and a paid option for the full version (Le et al. 2020) Most notable development would be the genexcell tool developed to facilitate the common analyses in the HCA consortium (Megill et al. 2021).

10.2.2.1 Overview of *scIVA*

scIVA is a light-weight, user-friendly web-based tool freely available for experimental biologists to query, analyse, and visualise their single cell datasets of interest. *scIVA* also serves as a framework for building interactive single cell databases. The *scIVA* GUI (Graphical User Interface) enables biologists to quickly perform four main categories of scRNA-seq data analysis tasks and consists of four modules (Fig. 10.3). These tasks are for exploratory data analysis and for testing gene expression across cells and different clusters within a mixed, heterogenous dataset.

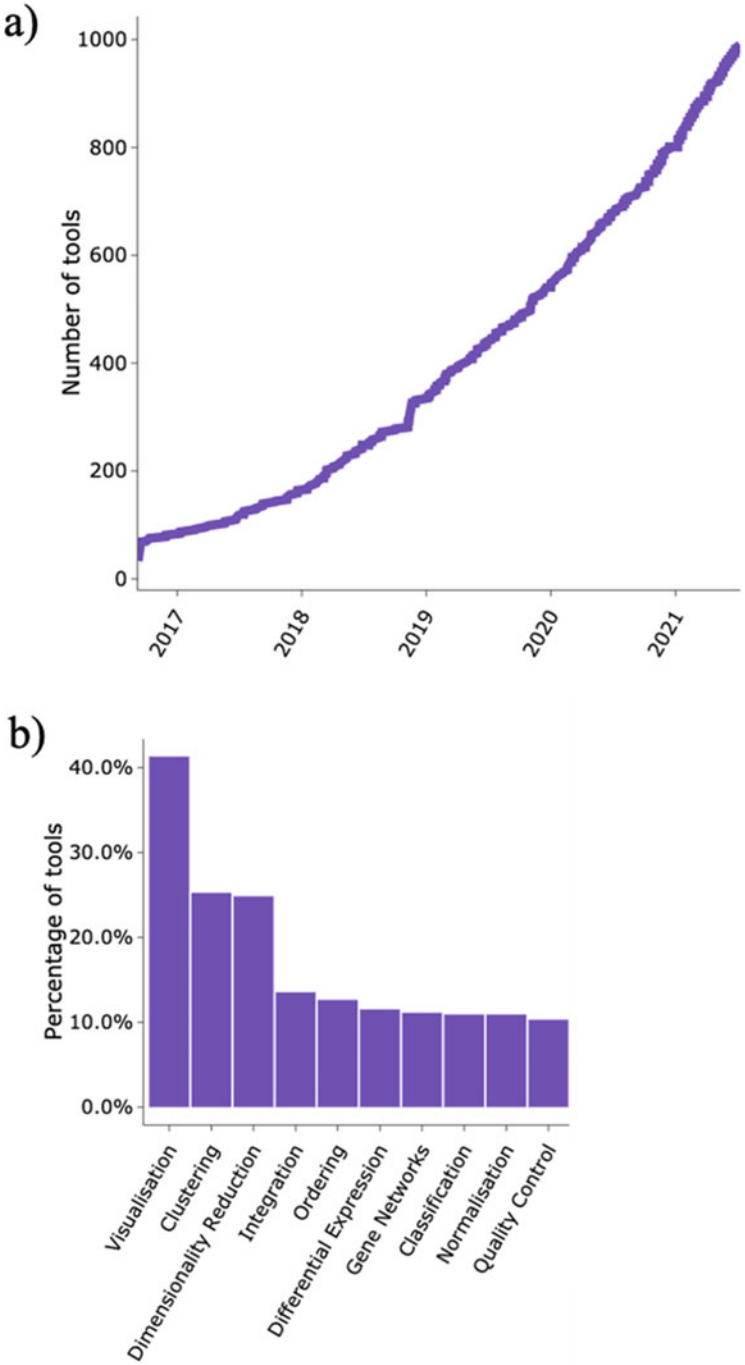


Fig. 10.2 Landscape of analysis tools available for scRNA-seq data. (**a**) A rapid increase in the number of analysis tools (software). (**b**) The top ten most common analysis tasks for single cell data. The updated data was generated by the scRNA-tools website (Zappia et al. 2018)

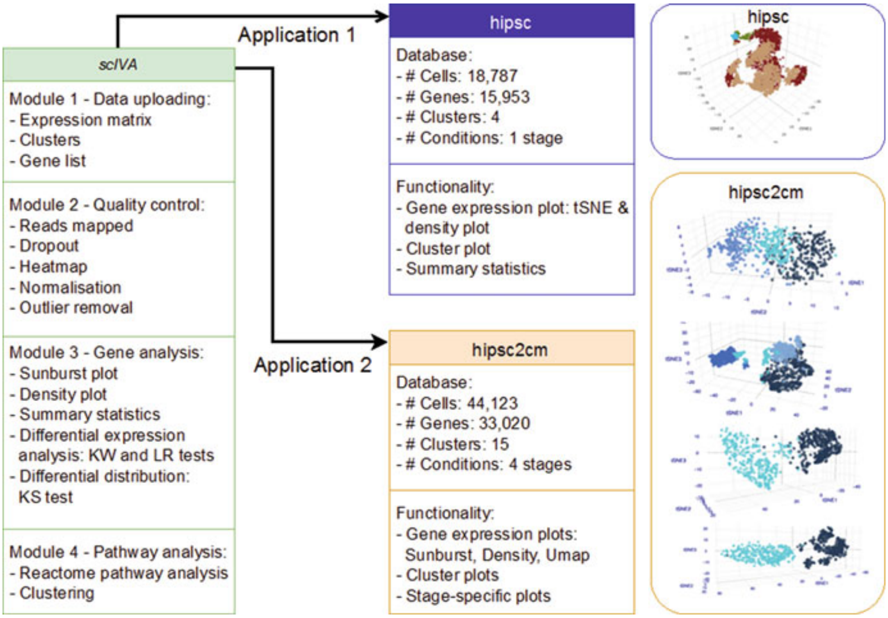
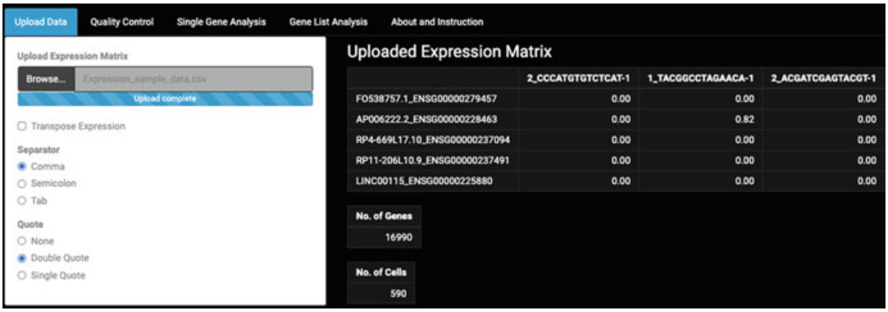


Fig. 10.3 Overview of the scIVA and two interactive scRNA-seq databases for mining. Four modules of the scIVA interactive software and their components are shown. An overview of the two databases (interactive datasets), *hipsc* and *hispsc2cm*, with information about sizes, scales, key interactive functionalities, and visualisation of single cell clustering for each differentiation stages on tSNE (t-distributed stochastic neighbour embedding) space is shown

Specific analyses include: quality control, normalisation, visualising single gene expression in different clusters, statistical tests to compare gene expression in single cell clusters, visualisation of gene expression by clusters, and functional analysis of a group of genes (which can be selected from gene-centric analysis tools, Fig. 10.3.

Two single cell datasets, namely hipsc (human induced pluripotent stem cell) and hipsc2cm (hipsc cells differententiated to cardiomyocytes), hereby interchangeably referred to as databases in the sense that the data was organised to enable the exploration of thousands of genes and cells across multiple clusters/cell types at single cell level. The two interactive datasets are presented as examples of applying *scIVA* applications for development of single cell resource for data mining (Fig. 10.3). These interactive, publicly available datasets are: *hipsc* (a human induced pluripotent stem cells with 18,787 cells) and *hipsc2cm* (a differentiation dataset with 43,168 cells). Each of these datasets has multiple functions in scIVA that allow users to fully mine the data, exploring expression changes at single cell level across cell types within one stage of pluripotency (*hipsc*) or multiple stages from pluripotent cells to differentiated cell types (*hipsc2cm*). The generation of hipsc and hipsc2cm datasets is described in our previous publications (Nguyen et al. 2018; Friedman et al. 2018). Here we show *scIVA* as an important development to utilise single cell data for either reanalysis or evaluating new questions, hypotheses by experimental biologist without the need for coding.



Upload Expression Matrix

Expression_sample_data.csv

□ Transpose Expression

Separator
● Comma
○ Semicolon
○ Tab

Quote
○ None
● Double Quote
○ Single Quote

Uploaded Expression Matrix

	2_CCCATGTGTCTCAT-1	1_TACGGCCTAGAACA-1	2_ACGATCGAGTACGT-1
FO538757.1_ENSG00000279457	0.00	0.00	0.00
AP006222.2_ENSG00000228463	0.00	0.82	0.00
RP4-669L17.10_ENSG00000237094	0.00	0.00	0.00
RP11-206L10.9_ENSG00000237491	0.00	0.00	0.00
LINC00115_ENSG00000225880	0.00	0.00	0.00

No. of Genes
16990

No. of Cells
590

Fig. 10.4 Data input. User can browse and upload a gene expression dataset from their local computer to *scIVA* server. The server can be flexible and several options to transpose data (so cells in columns and genes in rows) recognise different formats. *scIVA* can also take pre-computed clustering information and a set of cell IDs to be analysed (with the default to analyse all genes and all cells)

10.3 *scIVA* Data Input and Preprocessing

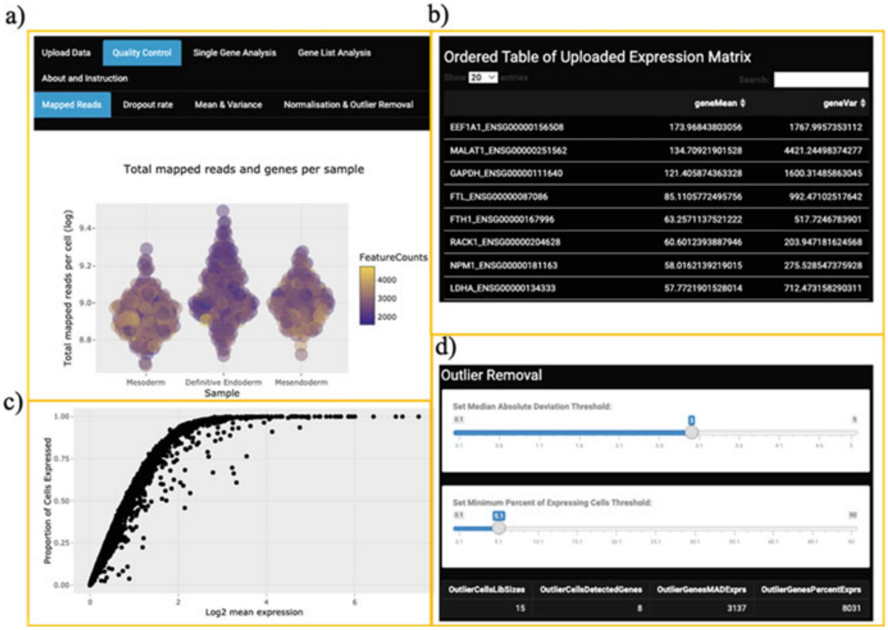
10.3.1 Data Collection Methods and Types of the Data Stored

scIVA allows for flexible data input and formatting, compatible to different single cell sequencing technology platforms. *scIVA* allows data uploading with multiple formats accepted, including csv, tsv, and xlsx, with or without quotations (Fig. 10.4). It also provides options to transpose the data matrices and choose headers from the uploaded dataset. Following the initial upload, users can subset data by clusters or upload a gene list to subset by genes or specify a subset of cells to focus the analysis on. Immediately post uploading, a preview of the uploaded data is automatically generated, with cell, gene, and cluster counts.

Datasets are stored in our local Linux server and made publicly accessible via the domain computationalgenomics.com.au, for example, in the cases of *hipsc* and *hipsc2cm*. The *scIVA* most common usage type is direct uploading and on-the-fly analysis of an scRNA-seq dataset without storing the data on a server. *scIVA* framework is ready and flexible to be developed to host external datasets in any server. The authors also plan to add new datasets to the local server, including new types of single cell transcriptomics data, like spatial transcriptomics.

10.3.2 Quality Control Methods and Curation Approaches

scIVA provides a range of quality control measures for cells and genes in a single cell dataset (Fig. 10.5). First, an interactive beeswarm plot is automatically generated to provide an overview of the uploaded dataset, with stratification by clusters or experimental design conditions. The beeswarm plot shows number of genes per



Ordered Table of Uploaded Expression Matrix

Show 20 entries Search:

	geneMean	geneVar
EEF1A1_ENSG00000156508	173.96843803056	1767.9957353112
MALAT1_ENSG00000251562	134.70921901528	4421.24498374277
GAPDH_ENSG00000111640	121.405874363328	1600.31485863045
FTL_ENSG00000087086	85.1105772495756	992.47102517642
FTH1_ENSG00000167996	63.2571137521222	517.7246783901
RACK1_ENSG00000204628	60.6012393887946	203.947181624568
NPM1_ENSG00000181163	58.0162139219015	275.528547375928
LDHA_ENSG00000134333	57.7721901528014	712.473158290311

Outlier Removal

Set Median Absolute Deviation Threshold: 3

Set Minimum Percent of Expressing Cells Threshold: 5.1

OutlierCellsLibSizes	OutlierCellsDetectedGenes	OutlierGenesMADExprs	OutlierGenesPercentExprs
15	8	3137	8031

Fig. 10.5 Quality control and data preprocessing. Quality matrices are generated automatically for each of the data uploaded. The general quality of the total reads per cells and genes per cells is shown by beeswarm plot. Numbers of genes with zero expression values are shown relative to the mean expression. Users can get mean expression values and variance across cells for every gene. After these visual observations, users can filter our genes and cells by specifying cutoffs

cells and total RNA sequencing reads that are mapped to each cell. From this plot, users can visually assess if cell outliers exist in the data, based on data distribution.

Additional quality control includes assessing ‘dropout’ rate, a pattern of seeing more cells with zero expression values for genes with lower average gene expression. A scatter plot is generated, showing mean expression with proportions of cells expressing the gene. Also included is a ranking system of genes by mean and variance across the whole dataset and across clusters. This option facilitates gene selection for downstream analysis based on gene expression pattern. For example, those genes that are most variable across the dataset are likely genes of interest. *scIVA* also allows for data to be filtered, so that gene and cell outliers are removed from the downstream analysis of the dataset. The resulting data can be downloaded and re-uploaded for subsequent analysis. For interactive view, user can zoom in and out as well as export plots and table generated by *scIVA*.

With respect to data indexing, there are two identification levels, namely at cell level and dataset level. In a dataset, cell indexes can be uniquely identified based on the barcode sequences (Fig. 10.4). All of the analysis plots incorporate interactive mouse-hover functionality to identify each dot (representing each cell, with cell ID, cell expression level, or cell label). Each dataset has its own storage location in our local server and has its unique URL for public access.

10.4 Database Access and Mining Methods

scIVA offers a comprehensive analysis toolkit that does not require any coding by the user and enables fully exploration of the data from quality control to gene-centric analysis and network analysis at cluster or whole dataset level. Below we describe key functionalities in *scIVA*. Applications of *scIVA* in two interactive datasets are described in the next section.

10.4.1 Single Gene Visualisation

A common analysis question is to find how gene expression changes between different clusters. *scIVA* has comprehensive range of interactive visualisation and statistical tests for examining changes of any selected gene between clusters (Figs. 10.6 and 10.7). The key feature of *scIVA* is to provide the ability to visualise the data in an intuitive and interactive way. The Sunburst plot generates a layered graph of the proportion of cells with/without expression of the genes, displaying counts, and percentage of cells expressing the gene of interest. The beeswarm plot displays the proportion of expressed cells and cell-specific expression level by clusters. Multiple density plots are generated to show the distribution of gene expression between clusters for all and for zero-filtered cells. A summary of expression table is generated, which displays counts, percentages, and means for each cluster across all cells and positively expressed cells.

10.4.1.1 Single Gene Analysis

Testing differences in single cell gene expression requires careful consideration of expression distribution, as the majority of the genes express in a small proportion of the cells. *scIVA* provides summary statistics in data browsing table by cells and clusters. To first identify genes with potential expression changes, a Kruskal–Wallis test (Hollander and Wolfe 1973) is performed for all clusters, followed by a Kolmogorov–Smirnov test (Conover 1971) for pair-wise statistical differences in the distributions of gene expression between clusters. Importantly, a modified form of likelihood ratio test, taking into account zero-inflated distribution, is provided to quantitatively perform differential expression analysis at gene level between clusters and to estimate fold change (McDavid et al. 2013).

10.4.1.2 Gene List Analysis

scIVA provides analysis tools for a user uploaded list of multiple genes. Gene list analysis features include Reactome pathway analysis, results of which are displayed

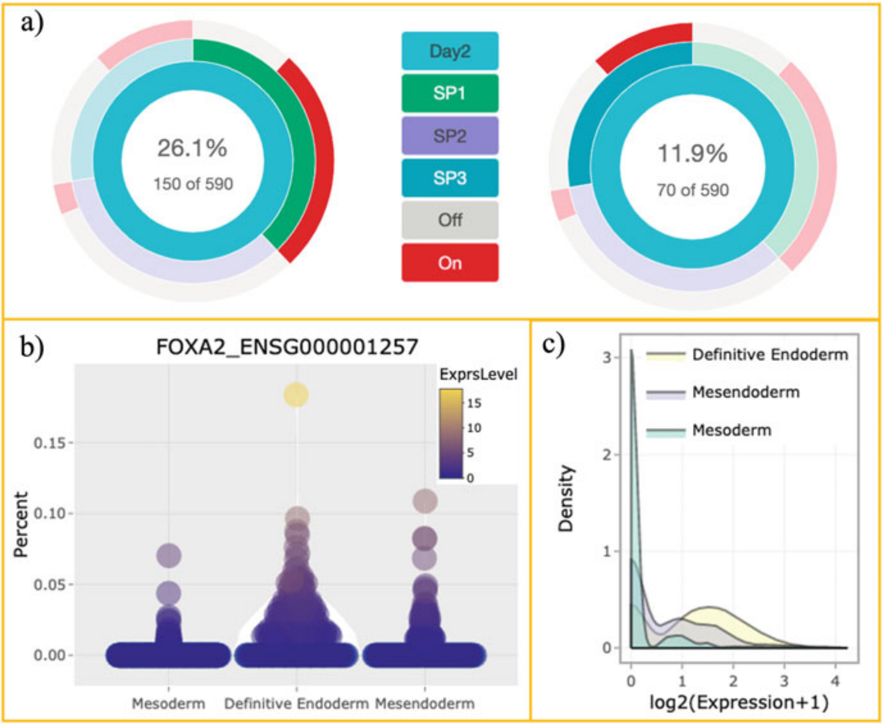
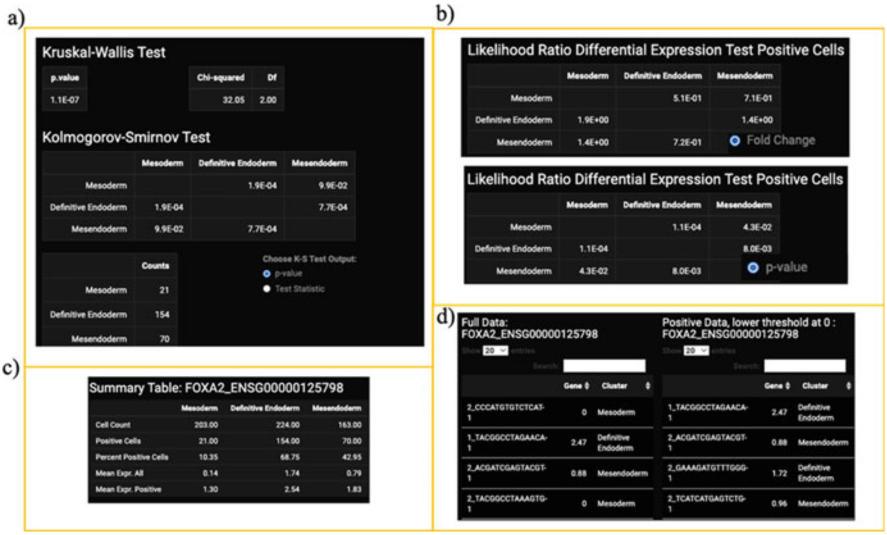


Fig. 10.6 Interactive visualisation of single cell gene expression across subpopulations (clusters). (**a**) Sunburst plot shows proportion of cells expressing a selected gene (FOXA2 in this case). The proportion expressed for two clusters is shown by red bar (definitive endoderm as yellow bar and mesendoderm as purple bars). (**b**) Beeswarm plot displays FOXA2 expression across three clusters. (**c**) Violin plot shows distribution of expression values across all single cells with positive detection of FOXA2, stratified by the three clusters

as interactive genetic pathways using networkD3. The user can also choose to cluster by selected genes, with options to choose the number of clusters, and whether to scale by row. Results will be displayed as a heatmap with dendrogram to illustrate the arrangements of clustering (Fig. 10.8).

10.5 Use Cases and Demo to Utilise the *scIVA* Database Framework

We expect that *scIVA* will be especially useful for interactive analysis of most scRNA-seq research projects, with data from one or more conditions. We present two examples, one simple dataset and another one with four times larger and more conditions (Figs. 10.3, 10.9, 10.10).



Kruskal-Wallis Test

p.value
1.1E-07

Chi-squared	Df
32.05	2.00

Kolmogorov-Smirnov Test

	Mesoderm	Definitive Endoderm	Mesendoderm
Mesoderm		1.9E-04	9.9E-02
Definitive Endoderm	1.9E-04		7.7E-04
Mesendoderm	9.9E-02	7.7E-04	

	Counts
Mesoderm	21
Definitive Endoderm	154
Mesendoderm	70

Choose K-S Test Output:
◉ p-value
○ Test Statistic

Likelihood Ratio Differential Expression Test Positive Cells

	Mesoderm	Definitive Endoderm	Mesendoderm
Mesoderm		5.1E-01	7.1E-01
Definitive Endoderm	1.9E+00		1.4E+00
Mesendoderm	1.4E+00	7.2E-01	

◉ Fold Change

Likelihood Ratio Differential Expression Test Positive Cells

	Mesoderm	Definitive Endoderm	Mesendoderm
Mesoderm		1.1E-04	4.3E-02
Definitive Endoderm	1.1E-04		8.0E-03
Mesendoderm	4.3E-02	8.0E-03	

◉ p-value

Summary Table: FOXA2_ENSG00000125798

	Mesoderm	Definitive Endoderm	Mesendoderm
Cell Count	203.00	224.00	163.00
Positive Cells	21.00	154.00	70.00
Percent Positive Cells	10.35	68.75	42.95
Mean Expr. All	0.14	1.74	0.79
Mean Expr. Positive	1.30	2.54	1.83

Full Data:
FOXA2_ENSG00000125798

Show 20 entries

Search:

	Gene	Cluster
2_CCCATGTGTCTCAT-1	0	Mesoderm
1_TACGGCCTAGAACA-1	2.47	Definitive Endoderm
2_ACGATCGAGTACGT-1	0.88	Mesendoderm
2_TACGGCCTAAAGTG-1	0	Mesoderm

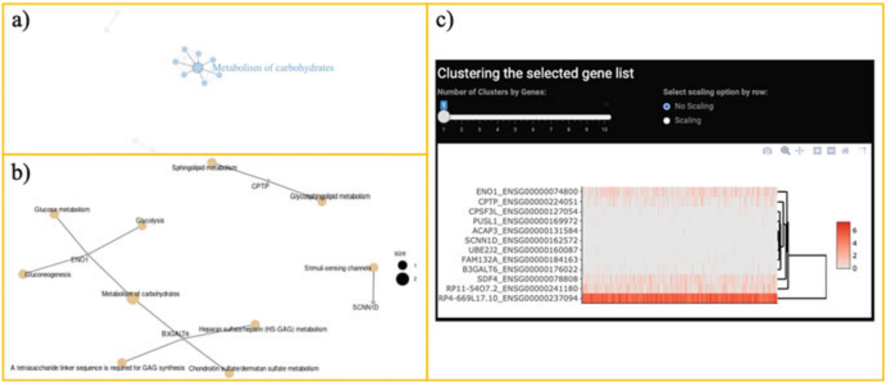
Positive Data, lower threshold at 0 :
FOXA2_ENSG00000125798

Show 20 entries

Search:

	Gene	Cluster
1_TACGGCCTAGAACA-1	2.47	Definitive Endoderm
2_ACGATCGAGTACGT-1	0.88	Mesendoderm
2_GAAAGATGTTTGGG-1	1.72	Definitive Endoderm
2_TCATCATGAGTCTG-1	0.96	Mesendoderm

Fig. 10.7 Statistical tests for any individual gene across cells and clusters. A gene of interest is selected by the user. (**a**) Non-parametric distribution tests include Kruskal–Wallis (test if the gene has differential distribution across clusters) and Kolmogorov–Smirnov test (pair-wise tests for differential distribution between two clusters). *P*-values and KS test statistics can be displayed. (**b**) Likelihood ratio tests for the differential expression of a gene between two cell types (two clusters). (**c**) Summary statistics for the selected gene across each cluster. (**d**) Display of cell statistics (expression and cluster label)

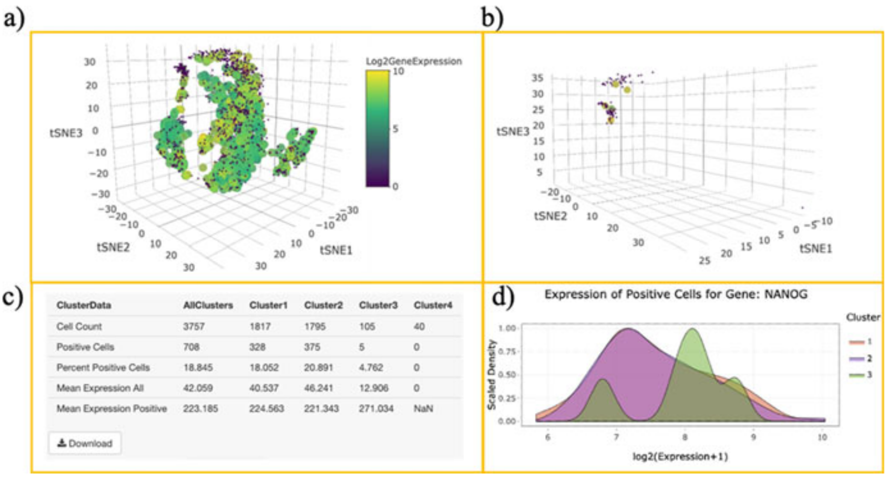


Clustering the selected gene list

Number of Clusters by Genes: 1

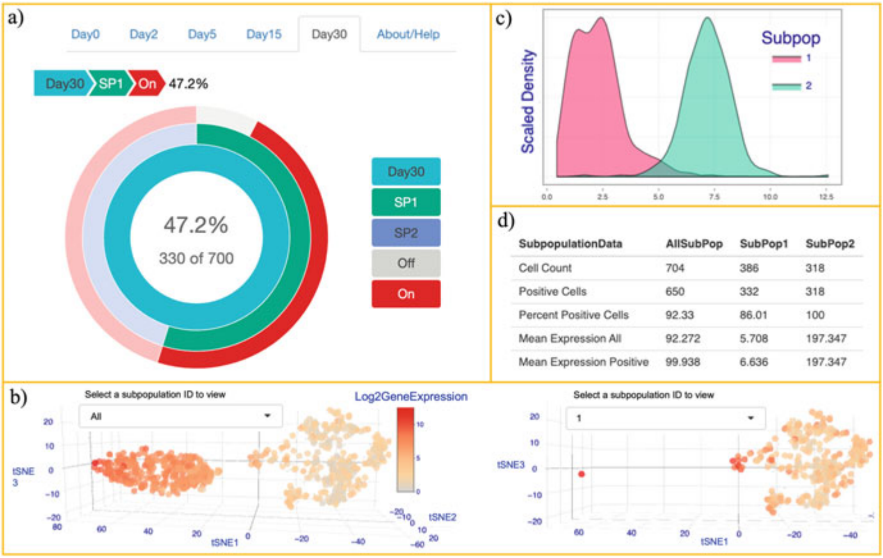
Select scaling option by row:
◉ No Scaling
○ Scaling

Fig. 10.8 Network analysis for a list of multiple genes selected by users. (**a**) and (**b**) are two different representations of networks for the same gene list (**a** is interactive and **b** is static). (**c**) Classifying genes by similarity in expression level across all single cells



ClusterData	AllClusters	Cluster1	Cluster2	Cluster3	Cluster4
Cell Count	3757	1817	1795	105	40
Positive Cells	708	328	375	5	0
Percent Positive Cells	18.845	18.052	20.891	4.762	0
Mean Expression All	42.059	40.537	46.241	12.906	0
Mean Expression Positive	223.185	224.563	221.343	271.034	NaN

Fig. 10.9 *hipsc* interactive dataset. (**a**) Expression of NANOG across all four clusters in the whole dataset. Smaller, purple dots represent cells without NANOG expression. (**b**) An option to display gene expression for each cluster is available. NANOG expression for cells in cluster 3 only is shown as an example. (**c**) Downloadable descriptive statistics of gene expression at cluster level. (**d**) Density plot for NANOG expression across all cells



SubpopulationData	AllSubPop	SubPop1	SubPop2
Cell Count	704	386	318
Positive Cells	650	332	318
Percent Positive Cells	92.33	86.01	100
Mean Expression All	92.272	5.708	197.347
Mean Expression Positive	99.938	6.636	197.347

Fig. 10.10 *hipsc2cm* interactive dataset. An example is shown for one differentiation stage (day 30) and a gene (TNNI1). This dataset has five stages from day 0 to day 30. (**a**) Sunburst plot shows percent of TNNI1 expression for subpopulation 1 (SP1). (**b**) Gene expression plot across all clusters (left) or just cluster 1 (right). (**c**) Density plot shows TNNI1 expression in clusters 1 and 2. (**d**) Summary statistics of TNNI1 across the two subpopulations

For the simple use case, the *hipsc* has 18,845 cells, all supposed to be at a pluripotent stage. This interactive dataset is publicly accessible at http://computationalgenomics.com.au/shiny/hipsc/. Without coding, any user can explore the expression of each of the over 15,000 genes across every individual cell in an interactive Uniform Manifold Approximation and Projection (UMAP) plot. With interactive mouse over control, the expression and ID of the cell will be displayed live. The seemingly homogenous population consists of four subpopulations. These subpopulations display heterogenous gene expression. Selecting a pluripotency marker NANOG as an example, *scIVA* shows that this gene is present and expressed at a range of log2expression from 0 to 10 (Fig. 10.9a). Using subpopulation 3 as an example, most cells had zero expression (small, dark-purple dots), but some cells expressed a high level of NANOG (large, yellowish dots). For a quantitative overview, *scIVA* calculates a summary table, displaying the exact number of cells in each of the four clusters and the proportions of cells with expression higher than 0 (Fig. 10.9c). The density plots in Fig. 10.9d show similar expression patterns between clusters 1 and 2, but variable expression level within cluster 3 and no expression in cluster 4.

scIVA can also be used to build a more complex interactive dataset, for example, the *hipsc2cm*, with five times more cells and more subpopulations (15). This differentiation time course dataset is accessible at http://computationalgenomics.com.au/shiny/hipsc2cm/. The large dataset is split into each of the five conditions (Day 0, Day 2, Day 5, Day 15, and Day 30), with a separate analysis present in each tab of the web interface (Fig. 10.10a). For the same gene selected, users can view the expression across all cells and clusters by switching between different tabs. This configuration both simplifies the display of analysis results and connects the analysis among the timepoints simply by switching tabs. The expression of the proportion of positive cells for each subpopulation is shown in Fig. 10.10a with sunburst plot, Fig. 10.10b with UMAP plots, Fig. 10.10c with density plot, and a quantitative summary table shown in Fig. 10.10d.

10.6 Future Update and Availability of scIVA

The *scIVA* code is open-source and is maintained at the GitHub site https://github.com/BiomedicalMachineLearning/scIVA/. The *scIVA* webtool and the two interactive datasets *hipsc* and *hipsc2cm* are publicly available at the domain, http://computationalgenomics.com.au/. These applications are based on a Shiny server hosted locally in a Centos 7.7, Red Hat Linux system and we are actively keeping the software updated and our platform allows for conveniently updating more single cell datasets. *scIVA* was built as an R-Shiny (Chang et al. 2021) server, hosted in our CentOS 7 Linux system and made publicly available via a secure DMZ Network. *scIVA* implements a range of interactive Javascript applications in *R* like *d3r* (Bostock et al. 2020a), *plotly* (Sievert 2020), *networkD3* (Allaire et al. 2017), *sunburstR* (Bostock et al. 2020b).

Our database integration strategy is that for each new dataset, we apply *scIVA* analysis framework and code to build a separate application in the master server, to be made available at the domain computationalgnomics.com.au. Each new database then is allocated with its new, unique URL, similar to *hipsc* and *hipsc2cm*. This approach allows for flexibility in adding new datasets that follow the consistent framework of *scIVA* and to be hosted at the same domain.

10.7 Summary and Future Development of the Database

scIVA presents itself as an accessible web interface for those with little or no coding experience to facilitate the process for visualisation, hypothesis generation, and testing. Given the increasing use of single cell sequencing techniques, *scIVA* can be widely applied to assist researchers in utilising the unprecedented resolution of the single cell data to reveal important biological processes. Notably, every interactive step is well documented in the open-source code, allowing users to technically understand what is used to generate the visual and GUI they are using. This feature is in contrast to many databases and webtools, which do not provide source codes and thus analyses would at least partially need to go through a computation 'black box'.

To date, nearly 1000 scRNA-seq analysis tools are available, but most of these tools still require scriptings (commonly in R and Python) for analyses and thus are less accessible to experimental biologists.

Consortium-scale webtools like cellxgene (Megill et al. 2021) are robust, but are of limited use for small scale projects (e.g. those with fewer than 100,000 cells and 10 conditions), especially those with data not hosted by the webserver. We expect that *scIVA* provides a light-weight, easy-to-use option for most small scale scRNA-seq projects, without a requirement for depositing the data to the server.

Acknowledgement We thank Professor Joseph Powell for helpful discussions and advice about the scIVA software and the manuscript.

Funding This work was supported by the National Health and Medical Research Council grant GNT2001514 and the Australian Research Council DECRA Fellowship DE190100116.

References

Allaire JJ, Gandrud C, Russell K, Yetman C (2017) networkD3: D3 JavaScript Network Graphs from R. R package version 0.4

Altschuler SJ, Wu LF (2010) Cellular heterogeneity: do differences make a difference? Cell 141 (4):559–563

Bostock M, Russell K, Aisch G, Pearce A (2020a) d3r: 'd3.js' utilities for R. R package version 0.9.1

Bostock M, Rodden K, Warne K, Russell K, (2020b) sunburstR: Sunburst ‘Htmlwidget’. R package version 2.1.5

Chang W, Cheng J, Allaire J, Sievert C, Schloerke B, Xie Y, Allen J, McPherson J, Dipert A, Borges B: (2021) shiny: Web Application Framework for R. R package version 1.6.0

Conover WJ (1971) Practical nonparametric statistics. Wiley, New York

Friedman CE, Nguyen Q, Lukowski SW, Helfer A, Chiu HS, Miklas J, Levy S, Suo S, Han J-DJ, Osteil P et al (2018) Single-cell transcriptomic analysis of cardiac differentiation from human PSCs reveals HOPX-dependent cardiomyocyte maturation. Cell Stem Cell 23(4):586–598.e588

Gardeux V, David FPA, Shajkofci A, Schwalie PC, Deplancke B (2017) ASAP: a web-based platform for the analysis and interactive visualization of single-cell RNA-seq data. Bioinformatics 33(19):3123–3125

Giladi A, Amit I (2018) Single-cell genomics: a stepping stone for future immunology discoveries. Cell 172(1):14–21

Hollander M, Wolfe DA (1973) Nonparametric statistical methods. Wiley, New York

Le T, Phan T, Pham M, Tran D, Lam L, Nguyen T, Truong T, Vuong H, Luu T, Phung N et al (2020) BBrowser: making single-cell data easily accessible. bioRxiv 2020:2020.2012.2011.414136

Macaulay IC, Ponting CP, Voet T (2017) Single-cell multiomics: multiple measurements from single cells. Trends Genet 33(2):155–168

McDavid A, Finak G, Chattopadyay PK, Dominguez M, Lamoreaux L, Ma SS, Roederer M, Gottardo R (2013) Data exploration, quality control and testing in single-cell qPCR-based gene expression experiments. Bioinformatics 29(4):461–467

Megill C, Martin B, Weaver C, Bell S, Prins L, Badajoz S, McCandless B, Pisco AO, Kinsella M, Griffin F et al (2021) cellxgene: a performant, scalable exploration platform for high dimensional sparse matrices. bioRxiv 2021:2021.2004.2005.438318

Miyamoto DT, Zheng Y, Wittner BS, Lee RJ, Zhu H, Broderick KT, Desai R, Fox DB, Brannigan BW, Trautwein J et al (2015) RNA-Seq of single prostate CTCs implicates noncanonical Wnt signaling in antiandrogen resistance. Science 349(6254):1351–1356

Nguyen Q, Lukowski S, Chiu H, Senabouth A, Bruxner T, Christ A, Palpant N, Powell J (2018) Single-cell RNA-seq of human induced pluripotent stem cells reveals cellular heterogeneity and cell state transitions between subpopulations. Genome Res 28(7):1053

Regev A, Teichmann SA, Lander ES, Amit I, Benoist C, Birney E, Bodenmiller B, Campbell PJ, Carninci P, Clatworthy M et al (2017) Science forum: the human cell atlas. elife 6

Schmidt F, Efferth T (2016) Tumor heterogeneity, single-cell sequencing, and drug resistance. Pharmaceuticals 9(2):33

Sievert C (2020) Interactive web-based data visualization with R, plotly, and shiny. Chapman & Hall/CRC

Wang Y, Navin Nicholas E (2015) Advances and applications of single-cell sequencing technologies. Mol Cell 58(4):598–609

Zappia L, Phipson B, Oshlack A (2018) Exploring the single-cell RNA-seq analysis landscape with the scRNA-tools database. PLoS Comput Biol 14(6):e1006245

Zheng GXY, Terry JM, Belgrader P, Ryvkin P, Bent ZW, Wilson R, Ziraldo SB, Wheeler TD, McDermott GP, Zhu J et al (2017) Massively parallel digital transcriptional profiling of single cells. Nat Commun 8:14049

Chapter 11
Access and Visualise High Quality Gene Expression Data with Stemformatics

Jarny Choi and Christine A. Wells

Abstract Stemformatics is an established online gene expression data portal containing hundreds of carefully curated and annotated datasets. The website has recently been completely re-designed using the latest full stack technologies. It contains a suite of visualisation tools which enable easy exploration of the data, as well as an API server which can be used to access the data directly. One of Stemformatics' key features is the integrated data atlas where the user can project their own data and classify their cell types against a reference. In addition to providing gene expression plots for each dataset, there are also tools for finding interesting genes across all datasets. Stemformatics is a fully open-source resource, available at www.stemformatics.org.

Keywords Gene expression · Stem cells · Data integration · Data visualisation · Online data portal · Reference datasets

11.1 Introduction to Stemformatics

Stemformatics is an established gene expression data portal designed for researchers working in stem cell and related fields. It currently contains over 450 datasets, mainly from microarray and bulk RNA sequencing technologies performed on human samples (Choi et al. 2019). These samples span a broad range of cell types, with a major focus on stem cells and their progeny at multiple stages of differentiation and derivation methods. Examples of commonly found cell types include induced pluripotent stem cells (iPSC), monocytes and macrophages, and acute myeloid leukaemia (AML) cells (see Table 11.1). Stemformatics is freely available at stemformatics.org.

J. Choi (✉) · C. A. Wells
The Centre for Stem Cell Systems, Faculty of Medicine, Dentistry and Health Sciences, The University of Melbourne, Parkville, VIC, Australia
e-mail: jarnyc@unimelb.edu.au

© The Author(s), under exclusive license to Springer Nature Singapore Pte Ltd. 2021

I. Abugessaisa, T. Kasukawa (eds.), *Practical Guide to Life Science Databases*,
https://doi.org/10.1007/978-981-16-5812-9_11

Table 11.1 Examples of common cell types in Stemformatics

Cell type	Number of samples
Monocyte	2144
iPSC	1323
ESC	948
Fibroblast	812
AML	728
PBMC	683
MSC	678
Macrophage	568
Mononuclear cell of bone marrow	343
Haematopoietic multipotent progenitor	321
T cell	315
Dendritic cell	190

Like many gene expression data portals, Stemformatics serves multiple purposes for the research community. Some of these can be highlighted by listing key features of Stemformatics:

1. The hosted datasets are sourced from public domains and manually selected for relevance in research context. The sample metadata are then carefully annotated to emphasise biological properties which are reproducible under many conditions. This process adds tremendous value to the datasets since domain specific knowledge is applied to them and the relevant cell types are highlighted and linked together across datasets.
2. It re-processes all data from raw files, rather than just rehosting datasets as they appear in public repositories such as GEO (Edgar 2002) or ArrayExpress (Kolesnikov et al. 2015). It uses quality control checkpoints (Fig. 11.8) to reject any datasets which fail certain criteria such as insufficient primary data available. This results in about 30% of datasets never making it to the portal, ensuring that Stemformatics data are of high quality and have been processed in a consistent manner.
3. It hosts multiple integrated data atlases which combine many datasets into one, enabling users to explore gene expression across multiple datasets at once and benchmark their own cell types against a robust reference.
4. It provides easy-to-use and intuitive tools for biologists to visually explore the data, including interactive gene expression profiles, principal component analysis plots, and more while also providing an application programming interface (API) for easy computational access to the data (Fig. 11.1).

11.1.1 The Integrated Data Atlases

Stemformatics provides a way to view biological patterns across many datasets at once through its novel integrated data atlases. Each atlas is a collection of datasets which represent particular biology and uses a novel gene filtering strategy based on

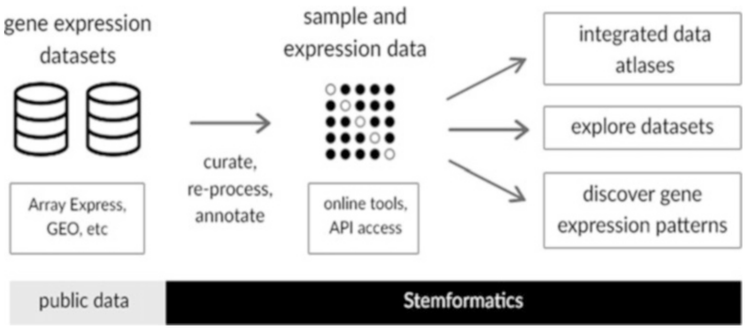


Fig. 11.1 A simplified schematic of Stemformatics which shows public datasets that act as input into Steminformatics, and a uniform data processing pipeline is applied before a dataset is hosted in the system. Various tools provided by the Steminformatics can then be used to answer research questions

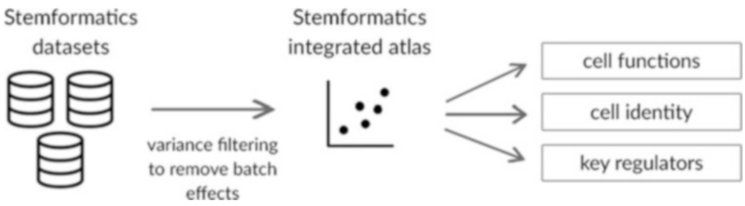

Stemformatics datasets
variance filtering to remove batch effects
Stemformatics integrated atlas
cell functions
cell identity
key regulators

Fig. 11.2 A schematic description of the Stemformatics integrated atlas, showing its derivation by removing batch effects as well as its application in discovering cell identity and key regulators of differentiation

batch variance to remove batch effects which are usually present when many datasets are integrated (Angel et al. 2020) (Fig. 11.2).

Each integrated atlas provides a unique view of gene expression data for several reasons: Most online gene expression data portals (including GEO, ArrayExpress, HCA, Haemosphere, etc.) provide access to each hosted dataset separately, hence gene expression can be viewed per dataset only, for example. Stemformatics integrated atlases in contrast provide an integrated view of tens of datasets and 900+ samples so that gene expression can be viewed across all these datasets, for example. Samples are selected for inclusion in an atlas based on their relevance to the biology being described, as well as meeting the criteria for the computational method used for data integration. These samples are then carefully annotated to ensure that they make biological sense—that they work in the context of the biology being described.

11.1.2 The Myeloid Atlas

The Stemformatics myeloid atlas integrates expression data assayed on a range of human myeloid cells which originate from ex vivo, in vivo, and in vitro sources

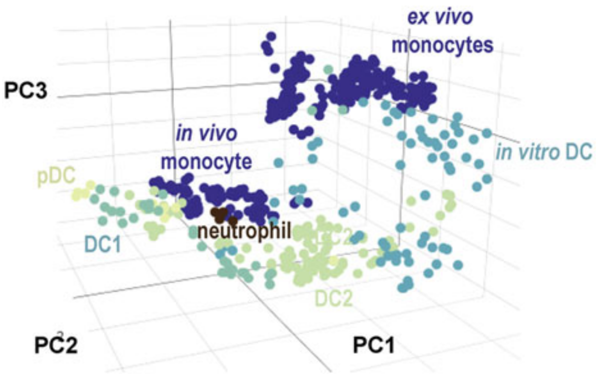


Fig. 11.3 Monocytes and dendritic cell (DC) subsets displayed in the atlas PCA plot show how they differ from each other based on source of derivation

(Rajab et al. 2021). By integrating data from 44 studies spanning over 900 samples, it can be used to identify myeloid subpopulations as well as benchmark one's own data against this reference. The myeloid atlas is available at stemformatics.org/atlas/myeloid.

One of the interesting observations coming from this atlas is the separation of myeloid samples based on their source of derivation (Fig. 11.3): in vitro differentiated myeloid cells do not readily resemble their in vivo counterparts, with ex vivo samples lying somewhere in between in this transcriptional landscape (Rajab et al. 2021). This may have significant implications for models of myeloid cells as well as derivation of these cell types through iPSC or similarly in vitro based methods of differentiation.

Users of the Stemformatics myeloid atlas can project their own data onto the atlas to view the transcriptional position of their own samples within the atlas or download all the atlas data to perform their own analysis (see general features Sect. 11.1.3 below).

11.1.3 General Features of the Atlas Page

The Stemformatics atlas page provides a wealth of tools to intuitively access the data and perform analyses, such as looking up gene expression and projecting other datasets. Key features currently available on the atlas page include:

- Gene expression lookup: where expression of a gene is visualised as a colour gradient on the PCA plot. This plot can also be viewed side by side with the plot coloured by sample group category so that the user can easily see which samples show higher expression of the gene (Fig. 11.4).
- Download all data and plots: all the data used in constructing the atlas can be downloaded as text files, including PCA coordinates, expression values, and even colours used in the plots (Fig. 11.5).
- Project other data: other datasets can be projected onto the atlas (either another Stemformatics dataset or user's own dataset), to visualise the transcriptional distance between the projected samples to those in the atlas (Fig. 11.6).

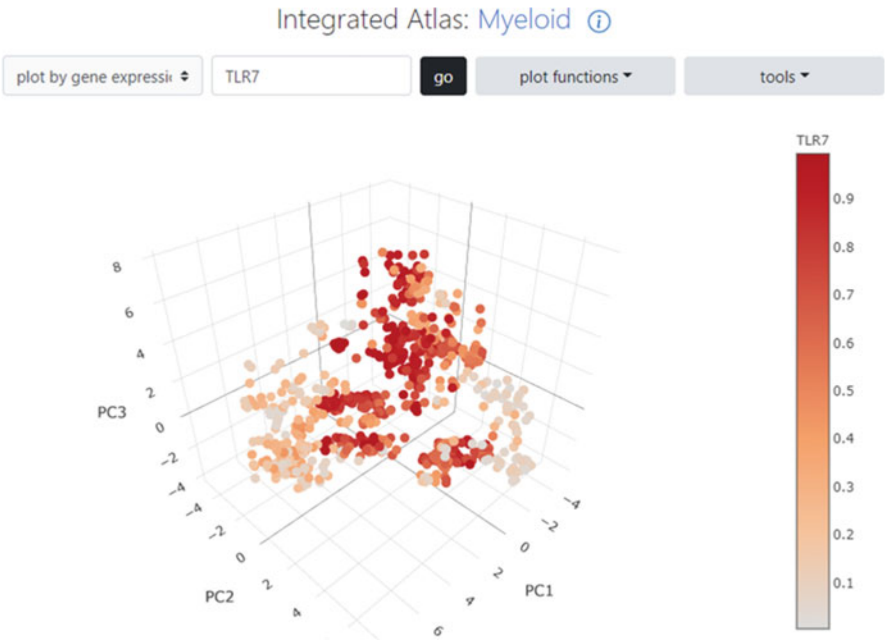


Fig. 11.4 Screenshot of gene expression on the myeloid atlas page

- Combine sample groups: sample groups such as cell type and activation status can be combined dynamically to create a new group that will make a finer distinction between these properties (Fig. 11.7).

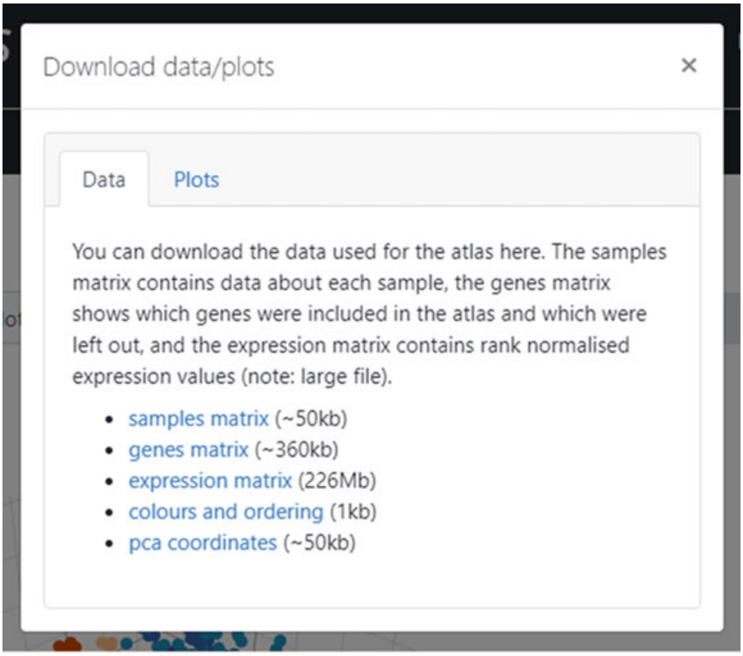
Some of these features are illustrated in more detail under the use-cases section below.

11.2 Data Processing and Sample Annotation

11.2.1 Data Selection

Data selection in Stemformatics is driven primarily by biological questions and research themes across various projects. One of the current research topics is elucidating the identity of iPSC derived cells—macrophages and dendritic cells in particular. How closely do they resemble their in vivo counterparts? What can we learn about the current methods of reprogramming cells if they do differ and can we find ways to improve this process? By having a reference atlas, such as the Stemformatics myeloid atlas, it is possible to ask these questions at the transcriptional level and obtain hypotheses for follow-ups in the lab (Rajab et al. 2021).

For construction of an integrated atlas, the selection of a dataset for inclusion into the atlas and hence into Stemformatics may also depend on technical issues. Since



You can download the data used for the atlas here. The samples matrix contains data about each sample, the genes matrix shows which genes were included in the atlas and which were left out, and the expression matrix contains rank normalised expression values (note: large file).

- samples matrix (~50kb)
- genes matrix (~360kb)
- expression matrix (226Mb)
- colours and ordering (1kb)
- pca coordinates (~50kb)

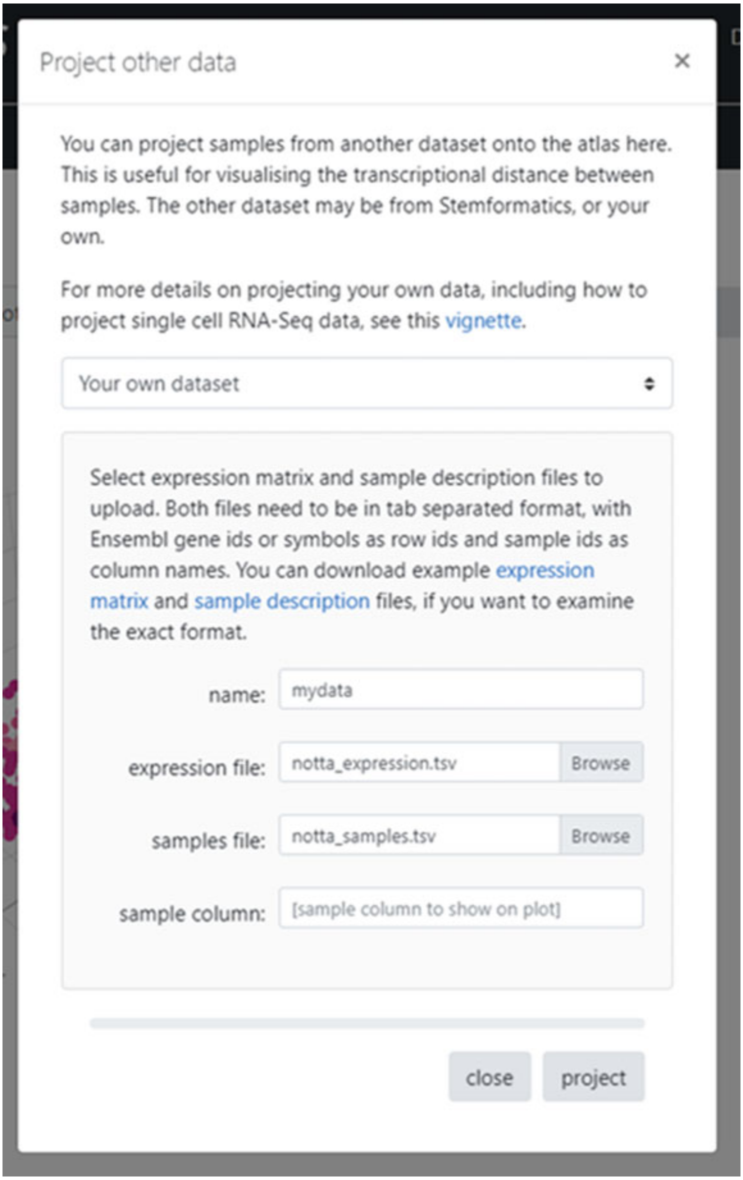
Fig. 11.5 Screenshot of download data/plots dialogue from an atlas page

the integration relies on variance filtering (Angel et al. 2020), it is ideal to have multiple cell types represented across multiple platforms (it is best to have T-cells from both microarray and RNAseq datasets, for example). Where this is difficult to achieve and the inclusion of a dataset leads to platform effect overtaking the biological effect in the atlas, that dataset may not be included in the system.

11.2.2 Data Processing

Once a dataset has been selected for inclusion, the data processing pipeline is applied (Fig. 11.8). Broadly there are three steps each dataset will go through before being included in the system:

1. Quality control checkpoint 1: Evaluate experimental design and that platform is supported by our system. Check that all source data are available. Check sample metadata values against our data dictionary.
2. Get source files and begin expression data processing. For microarray data, use R packages corresponding to the platform to create a normalised matrix at the probe level. For bulk RNAseq data, get fastq files and align against the appropriate genome version and create a counts matrix summarised at the gene level (usually performed on a high performance computer). Note that this step is not necessarily applied for single cell RNAseq datasets due to their complexity, as this is outside



Project other data

You can project samples from another dataset onto the atlas here. This is useful for visualising the transcriptional distance between samples. The other dataset may be from Stemformatics, or your own.

For more details on projecting your own data, including how to project single cell RNA-Seq data, see this vignette.

Your own dataset

Select expression matrix and sample description files to upload. Both files need to be in tab separated format, with Ensembl gene ids or symbols as row ids and sample ids as column names. You can download example expression matrix and sample description files, if you want to examine the exact format.

name: mydata

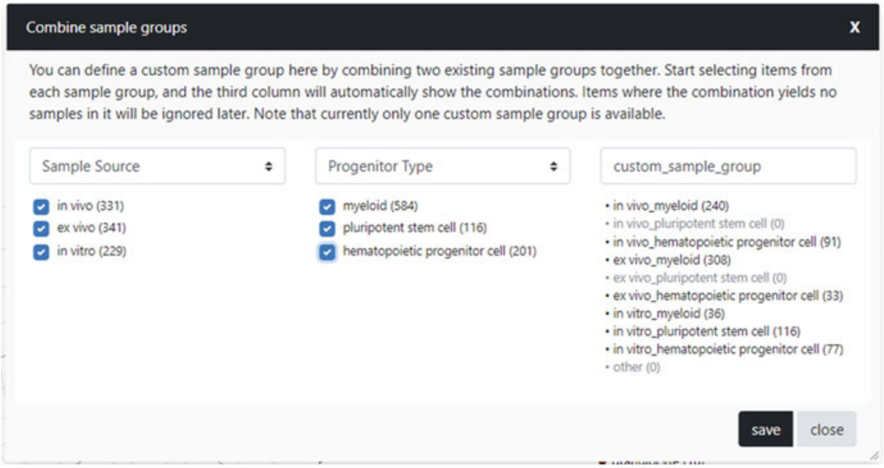
expression file: notta_expression.tsv

samples file: notta_samples.tsv

sample column: [sample column to show on plot]

Fig. 11.6 Screenshot of project other data dialogue from an atlas page

the scope of Stemformatics. However, Stemformatics hosts some Fluidigm C1 platform datasets (Durruthy-Durruthy and Ray 2018), which were processed in this manner. Note also that other projects exist where single cell RNAseq data are processed from raw files, with a uniform pipeline, such as the Human Cell Atlas data portal (humancellatlas n.d.).



Combine sample groups

You can define a custom sample group here by combining two existing sample groups together. Start selecting items from each sample group, and the third column will automatically show the combinations. Items where the combination yields no samples in it will be ignored later. Note that currently only one custom sample group is available.

Sample Source

☑ in vivo (331)
☑ ex vivo (341)
☑ in vitro (229)

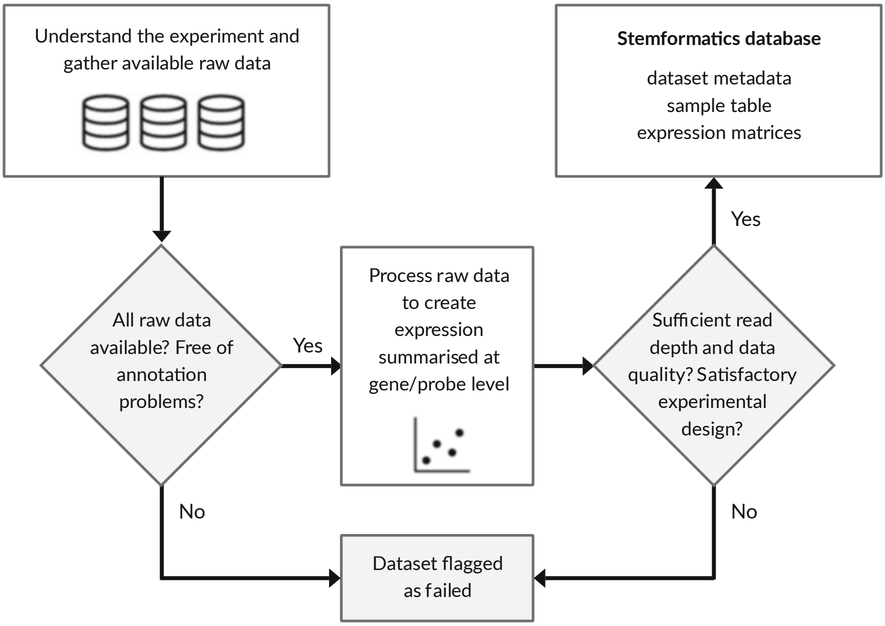
Progenitor Type

☑ myeloid (584)
☑ pluripotent stem cell (116)
☑ hematopoietic progenitor cell (201)

custom_sample_group

- in vivo_myeloid (240)
- in vivo_pluripotent stem cell (0)
- in vivo_hematopoietic progenitor cell (91)
- ex vivo_myeloid (308)
- ex vivo_pluripotent stem cell (0)
- ex vivo_hematopoietic progenitor cell (33)
- in vitro_myeloid (36)
- in vitro_pluripotent stem cell (116)
- in vitro_hematopoietic progenitor cell (77)
- other (0)

Fig. 11.7 Screenshot of combined sample groups dialogue from the myeloid atlas page



Understand the experiment and gather available raw data
All raw data available? Free of annotation problems?
Yes
No
Process raw data to create expression summarised at gene/probe level
Sufficient read depth and data quality? Satisfactory experimental design?
Yes
No
Stemformatics database
dataset metadata
sample table
expression matrices
Dataset flagged as failed

Fig. 11.8 Stemformatics data processing workflow shows two quality control checkpoints that a dataset must pass before being uploaded into the system

3. Quality control checkpoint 2: Evaluate technical aspects of the data, such as read depth and alignment statistics, using the QC metrics produced in step 2. Investigate other possible data issues such as sample swaps. Go back to a previous step if necessary.

This pipeline means that a dataset may fail to process at either quality control checkpoints. Over a decade of processing datasets has shown that roughly 30% of all datasets fail. Common reasons for failure include a lack of primary data, poor experimental design (where batch and biology are confounded, for example), and bad data quality (read depth is too low for RNAseq data) (Choi et al. 2019). This rather high rate of failure suggests that journals should do better in vetting these datasets at review stages.

11.2.3 Sample Annotation

As the collection of datasets grows, it becomes increasingly challenging to find effective ways to annotate sample data, so that they retain as much of the original biological information as possible while working harmoniously with other datasets to accelerate discovery. This is a common problem faced by many medium to large data portals such as Stemformatics, as they wrestle with multiple sources of data, coming from all around the world and having been annotated differently. Let us take an example of a cell type designation of "macrophage" to a particular sample. This simple designation alone is not enough to show how it may have been derived (in vivo or iPSC programmed?) or what kind of stimuli may have been applied, or what the tissue of origin may be. Should there be fields created in the database to capture all these properties then? The problem with this approach is that it may be impossible to fill this information for every dataset, as they may not be present from primary sources or are not relevant within the dataset. Furthermore, deciding on a consistent nomenclature is also difficult when there is no standard way to label many of the entities that appear in sample annotations.

In earlier versions of Stemformatics, a fully flexible and permissive data structure was used to tackle this problem, by allowing key-value pair based annotation tables. This approach is suitable for cases where each dataset can capture all the information about itself in a self-contained way, without the need to correlate with other datasets. However as Stemformatics became more focused on data integration, it has moved away from this approach into a more structured one, where each field is pre-defined, and uniform nomenclature is applied across datasets.

In any system, there is a compromise between these two modalities of data: less structured but more flexible vs more structured but less flexible. In moving from a less structured to a more structured system, Stemformatics is applying domain specific knowledge to the data and highlighting key similarities and differences in the context of the biological questions being asked, rather than simply applying database rules. An example of this is how each integrated atlas has been designed to contain a separate set of sample annotations which complement the core sample annotations attached to the datasets, thus enabling the system to highlight key biological features relevant for each atlas.

11.3 System Architecture

11.3.1 The Challenge of System Design in Research Environments

Stemformatics as a system consists of code which runs various servers and interfaces with the underlying data attached to them. Designing such a system sensibly in a research environment can be challenging (Björn et al. 2019), often due to a lack of resources, compared to a similar sized project in the industry. Hence when developing a reasonably complex data driven system within a research environment, it is important to be aware of certain commonly found constraints:

- The system is built to answer research questions.
- Teams may be very small for both development and maintenance of the system (sometimes one person team for all tasks). Team members also may turn over very quickly, as researchers move labs and professional personnel may look for other opportunities.
- It is not easy to attract funding for database systems from traditional sources, hence the lifecycle of the system may be short, and there may be limited and varying amounts of resources available consistently.

These constraints create some pitfalls for system development within the research environment. A common issue is that the system may be either under- or over-engineered (or both in different parts of the system). An under-engineered system is one where there has not been enough thought put into modifications later on, which perhaps may have to be performed by new personnel. Poor documentation, a lack of structure or reasoning behind the design, or code bloating from simple copy and paste are common, as systems may move from prototype stage to implementation without sufficient thought into the design. This problem is exacerbated by the fact that researchers without the necessary background and knowledge in system architecture and design often begin these projects.

An over-engineered system is one where there are too many steps to find the key functions which perform a task, through excessive wrappers, class definitions, or data mappings. This can happen when experienced software engineers work in the project and translate the levels of engineering found in the industry settings to the research environment. Without specialised teams that can understand the parts of the system or having personnel with enough technical knowledge, an over-engineered system is just as difficult to navigate as an under-engineered one.

In both of these cases, it is very difficult for such systems to be maintained effectively with small teams with high turnover and varying levels of skills. The new Stemformatics system has been built with these constraints in mind and has employed several ways to make it easier to update and maintain the system:

- Document any smallest issues which may take time to resolve, with real examples of commands which were run. This means less prior knowledge of the next maintainer is assumed.
- Design a system with lots of transparency: it should be easy to figure out the base command that performs a function, for example, without having to go through layers and layers of abstraction. The same principle applies to infrastructure on which the system is built, including the virtual machine and the environments used.
- Take a balanced approach between abstraction and direct data access.
- Re-design and refactor parts of the system if necessary as more complexity is added. A guiding principle may be: "How difficult would it be for another person to make a change in this code in 6 months' time?"

11.3.2 Stemformatics System Architecture

Stemformatics uses Australia's national computing infrastructure, NECTAR (ardc n.d.), which provides virtual machines and disk spaces, as well as a set of tools to manage the resources. Centos OS 8 virtual machines are created to host each instance of Stemformatics, where an instance may work as one of development, test, production and backup servers.

On each virtual machine, data are stored in different formats: mongdb (mongodb n.d.) is used for dataset and sample metadata, whereas hdf5 or text files are used for expression and atlas data. These data are then interfaced with an API server, which is built on flask (Palletsprojects n.d.). A separate user interface (UI) server, which is built on nuxtjs (nuxtjs n.d.), communicates with the API server to retrieve data and render plots and tables.

This separation of API and UI servers is a new design that moves away from the previous versions where one python pyramid (trypyramid n.d.) server was responsible for both data and web pages (Stephenson et al. 2018). The new design makes it easier to maintain the system as each server is based on completely different technologies and also empowers others to build their own applications using the API, as well as bioinformaticians to access the data directly without using the website (Fig. 11.9).

Versioning has also been implemented for datasets and atlas data, so that changes can be tracked and any previous version of the dataset accessed. A dataset may change version for various reasons, such as an update in sample annotation or its expression matrices based on new gene annotations or normalisation methods.

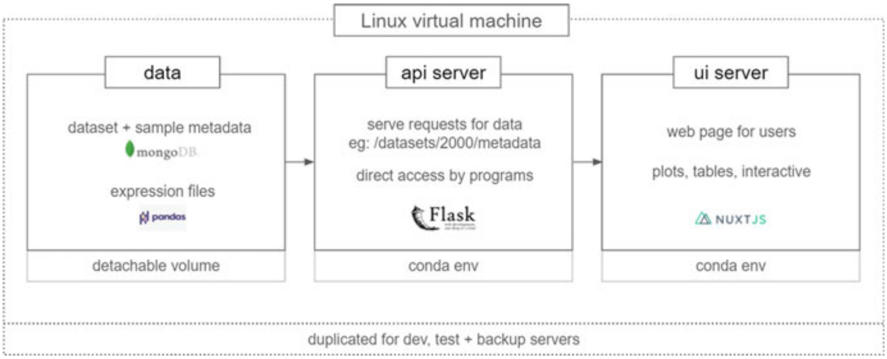


Fig. 11.9 A schematic of the Stemformatics infrastructure, which shows the connection amongst data, the API server and the UI server. Three types of data are stored in the system in different formats: dataset and sample metadata are stored in mongoDB, expression matrices are in hdf5 format, and atlas data are stored as text files

11.4 User Interfaces and Use-Cases

The main means of accessing the Stemformatics data for most users would be through the website, stemformatics.org. This site has recently been completely redesigned, to take advantage of modern browser features as such responsive layouts. Broadly, the site has the following main categories and features (example use-cases follow):

- Find datasets and view details, including PCA plots, sample table, and gene expression. Novel visualisation tools such as a sunburst plot are employed to invite the user to explore the data easily here.
- Explore the integrated data atlases, such as the blood and myeloid atlases, and use its features such as data projection and dynamic sample group combinations.
- View gene and gene set expression across datasets, to suggest key pathways for differentiation.

The other user interface available is the API server, which can be accessed to fetch all the data stored in Stemformatics while bypassing the website altogether. This is designed for bioinformaticians and computational biologists to incorporate the data access directly into their code and for systems developers to build custom user interfaces. All the data available through the website are also available through the API. See an example use-case below.

11.4.1 Use-Case 1 (Basic Level): Find Datasets of Relevance and View Details

A typical workflow for many biologists may be to search for datasets based on cell types and view their details, as well as plot gene expression. By going to Datasets > View collections from the menu (Fig. 11.10), the user can enter a search term or choose from a pre-defined collection of datasets. In this example, the Atlas Datasets are chosen to view all the datasets which have been used in the atlases. Once on this page, filtering can be performed to further subset the datasets, and clicking on the name of an individual dataset will go to the page showing the details of that dataset. In this example the (Silvin et al. 2017) dataset has been selected to view in detail, and expression of RAB1A gene shows differential expression across the three dendritic cell subpopulations within the dataset.

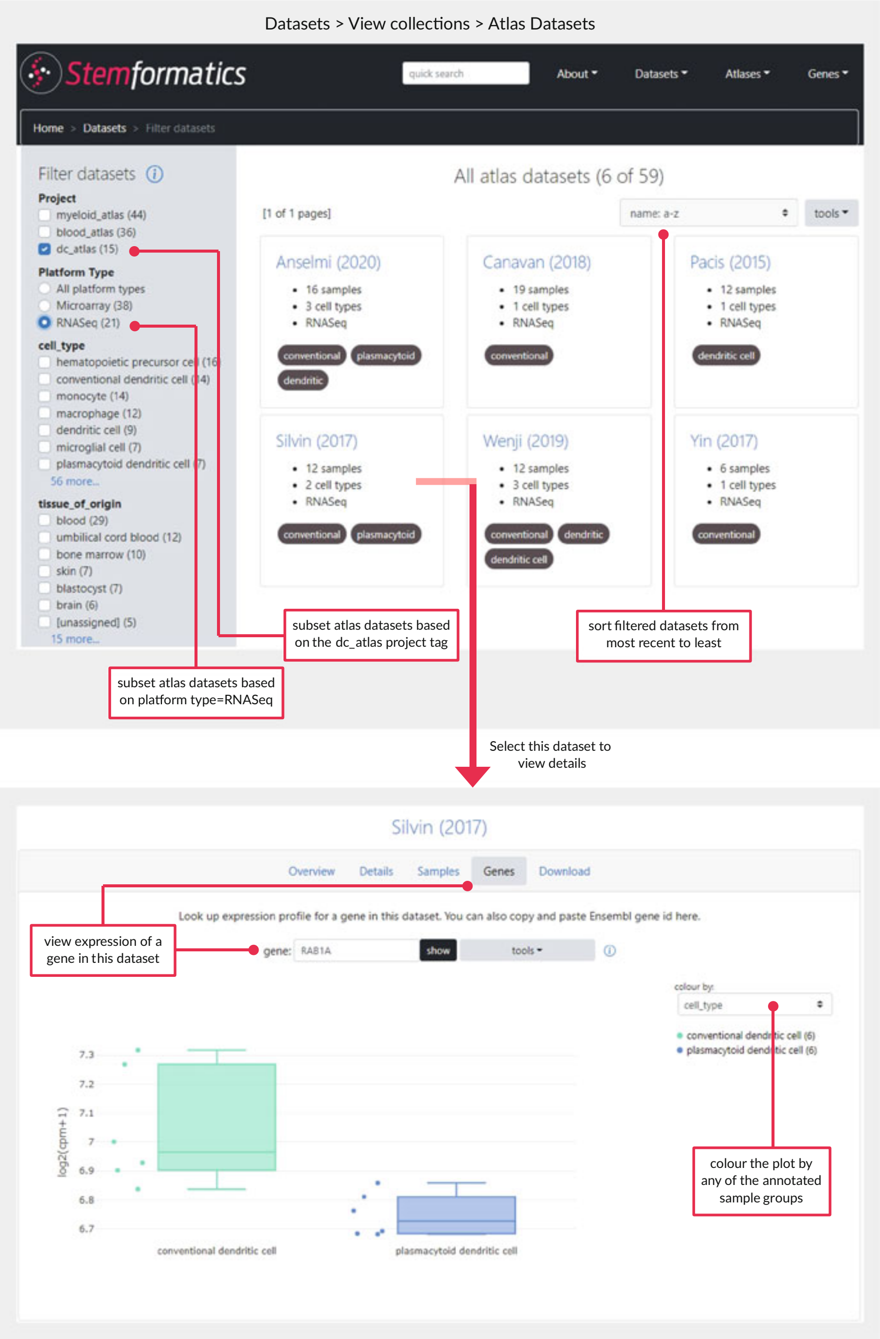
11.4.2 Use-Case 2 (Intermediate): Combine Sample Groups on the Myeloid Atlas

In this example, the user is exploring the myeloid atlas which shows clustering of samples according to “Sample Source” or “Progenitor Type”, where Sample Source labels samples according to “in vivo”, “ex vivo”, and “in vitro” labels, while Progenitor Type labels them according to “Myeloid”, “Pluripotent stem cell”, and “haematopoietic progenitor cell”. This means that the user can see a separation of in vivo sample from in vitro, for example.

Now the user is interested in seeing if there is a separation between myeloid and stem cells within each of the sample sources. To do this, a combined sample group can be created on the atlas page (Tools > Combine sample groups), which will dynamically fill in the combinations as the user makes selections on each sample group (Fig. 11.11).

11.4.3 Use-Case 3 (Intermediate/Advanced): Project One’s Own Data Onto the Blood Atlas

A more advanced usage example is to project one’s own expression data onto an atlas. This generally requires some level of pre-processing of the data prior to projection, to prepare the files in the correct format first. Two files are required for projection: gene expression matrix in tab separated format, with Ensembl gene ids as row index, and sample table also in tab separated format where the row indices should match the columns of the expression matrix. Once the files have been prepared, performing the projection is straightforward through tools > project other data on the atlas page (Fig. 11.12).



Datasets > View collections > Atlas Datasets

subset atlas datasets based on the dc_atlas project tag

sort filtered datasets from most recent to least

subset atlas datasets based on platform type=RNASeq

Select this dataset to view details

Look up expression profile for a gene in this dataset. You can also copy and paste Ensembl gene id here.

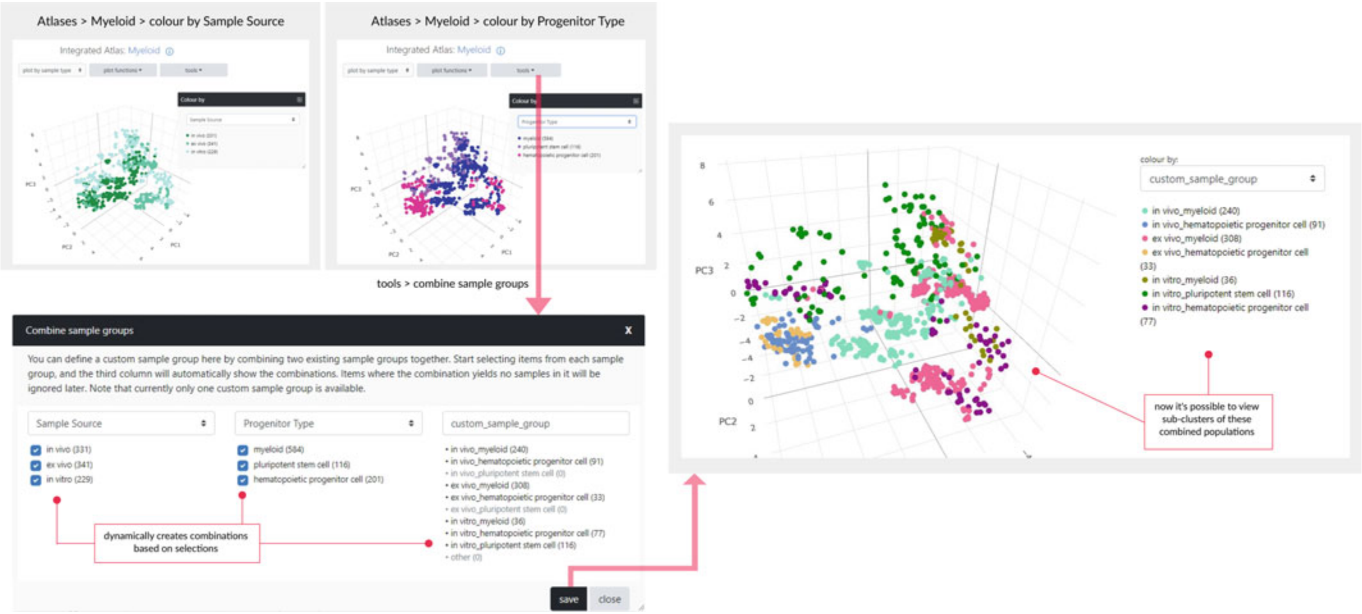
gene: RAB1A

show
tools
colour by:
cell_type
● conventional dendritic cell (6)
● plasmacytoid dendritic cell (6)
log2(cpm+1)
7.3
7.2
7.1
7
6.9
6.8
6.7
conventional dendritic cell
plasmacytoid dendritic cell

view expression of a gene in this dataset

colour the plot by any of the annotated sample groups

Fig. 11.10 Screenshots showing find dataset and view its details workflow


Atlases > Myeloid > colour by Sample Source
Integrated Atlas: Myeloid
plot by sample type
plot functions
tools
Colour by
Sample Source
in vivo (331)
ex vivo (341)
in vitro (229)
PC3
PC2
PC1
Atlases > Myeloid > colour by Progenitor Type
Integrated Atlas: Myeloid
Colour by
Progenitor Type
myeloid (584)
pluripotent stem cell (116)
hematopoietic progenitor cell (201)
tools > combine sample groups
save
close
colour by:
custom_sample_group
PC3
PC2
dynamically creates combinations based on selections
now it's possible to view sub-clusters of these combined populations

Combine sample groups

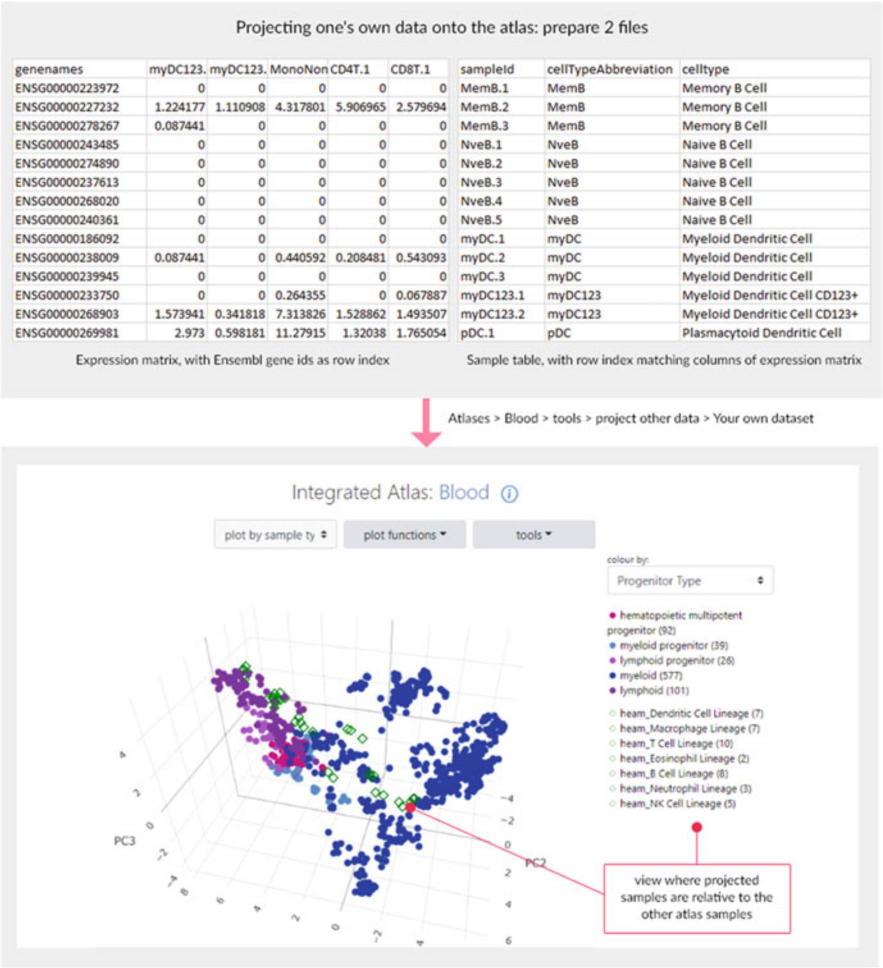
You can define a custom sample group here by combining two existing sample groups together. Start selecting items from each sample group, and the third column will automatically show the combinations. Items where the combination yields no samples in it will be ignored later. Note that currently only one custom sample group is available.

Sample Source	Progenitor Type	custom_sample_group
☑ in vivo (331)	☑ myeloid (584)	• in vivo_myeloid (240)
☑ ex vivo (341)	☑ pluripotent stem cell (116)	• in vivo_hematopoietic progenitor cell (91)
☑ in vitro (229)	☑ hematopoietic progenitor cell (201)	• in vivo_pluripotent stem cell (0)
		• ex vivo_myeloid (308)
		• ex vivo_hematopoietic progenitor cell (33)
		• ex vivo_pluripotent stem cell (0)
		• in vitro_myeloid (36)
		• in vitro_hematopoietic progenitor cell (77)
		• in vitro_pluripotent stem cell (116)
		• other (0)

colour by: custom_sample_group

- ● in vivo_myeloid (240)
- ● in vivo_hematopoietic progenitor cell (91)
- ● ex vivo_myeloid (308)
- ● ex vivo_hematopoietic progenitor cell (33)
- ● in vitro_myeloid (36)
- ● in vitro_pluripotent stem cell (116)
- ● in vitro_hematopoietic progenitor cell (77)

Fig. 11.11 Screenshots showing combine sample groups in the myeloid atlas workflow



Projecting one's own data onto the atlas: prepare 2 files

genenames	myDC123.	myDC123.	MonoNon	CD4T.1	CD8T.1
ENSG00000223972	0	0	0	0	0
ENSG00000227232	1.224177	1.110908	4.317801	5.906965	2.579694
ENSG00000278267	0.087441	0	0	0	0
ENSG00000243485	0	0	0	0	0
ENSG00000274890	0	0	0	0	0
ENSG00000237613	0	0	0	0	0
ENSG00000268020	0	0	0	0	0
ENSG00000240361	0	0	0	0	0
ENSG00000186092	0	0	0	0	0
ENSG00000238009	0.087441	0	0.440592	0.208481	0.543093
ENSG00000239945	0	0	0	0	0
ENSG00000233750	0	0	0.264355	0	0.067887
ENSG00000268903	1.573941	0.341818	7.313826	1.528862	1.493507
ENSG00000269981	2.973	0.598181	11.27915	1.32038	1.765054

Expression matrix, with Ensembl gene ids as row index

sampleId	cellTypeAbbreviation	celltype
MemB.1	MemB	Memory B Cell
MemB.2	MemB	Memory B Cell
MemB.3	MemB	Memory B Cell
NveB.1	NveB	Naive B Cell
NveB.2	NveB	Naive B Cell
NveB.3	NveB	Naive B Cell
NveB.4	NveB	Naive B Cell
NveB.5	NveB	Naive B Cell
myDC.1	myDC	Myeloid Dendritic Cell
myDC.2	myDC	Myeloid Dendritic Cell
myDC.3	myDC	Myeloid Dendritic Cell
myDC123.1	myDC123	Myeloid Dendritic Cell CD123+
myDC123.2	myDC123	Myeloid Dendritic Cell CD123+
pDC.1	pDC	Plasmacytoid Dendritic Cell

Sample table, with row index matching columns of expression matrix

Atlases > Blood > tools > project other data > Your own dataset

Integrated Atlas: Blood

plot by sample ty
plot functions
tools
colour by:
Progenitor Type
● hematopoietic multipotent progenitor (92)
● myeloid progenitor (39)
● lymphoid progenitor (26)
● myeloid (577)
● lymphoid (101)
◇ heam_Dendritic Cell Lineage (7)
◇ heam_Macrophage Lineage (7)
◇ heam_T Cell Lineage (10)
◇ heam_Eosinophil Lineage (2)
◇ heam_B Cell Lineage (8)
◇ heam_Neutrophil Lineage (3)
◇ heam_NK Cell Lineage (5)
PC3
PC2
−4
−2
0
2
4
6

view where projected samples are relative to the other atlas samples

Fig. 11.12 Screenshots showing projection of one's own data onto the blood atlas workflow

11.4.4 *Use-Case 4 (Advanced): Use the API to Download All Sample Metadata for Datasets Containing Blood Samples*

For more computationally savvy users, the API server provides access to all the functions used by the website to query the data and these are outlined on the website under Datasets > API access, which also provides example output for each query. An example workflow may be to search for all RNAseq datasets which contain the term "blood" (python API example here) and save sample table for each dataset in a file:

```
import requests
# First we find out what values are available for platform_type:
r = requests.get(‘https://api.stemformatics.org/values/datasets/
platform_type”)
print(r.json())
>> [ "Microarray","RNASeq","other","scRNASeq"]
# Now fetch all datasets for the query which is a list of dictionary
containing dataset metadata:
r = requests.get(“https://api.stemformatics.org/search/datasets?
platform_type=RNASeq&query_string=blood”)
datasets = r.json()
print(datasets[0])
>> {"dataset_id": 6601, "platform_type": "RNASeq.", ...}
# Loop through these datasets and fetch sample table, saving them as files
locally
for dataset in datasets:
 r = requests.get(“https://api.stemformatics.org/datasets/%s/
samples?as_file=true” % dataset[“dataset_id”])
 with open(“%s_samples.txt” % dataset[“dataset_id”], “w”,
encoding=“utf-8”) as f:
    f.write(r.text)
```

11.5 Summary

Stemformatics provides a valuable resource to the life science research community through its curated approach to data collection and processing as well as its novel integrated atlases. It can be used to discover emerging properties in the data, particularly in the field of transcriptional cell identity, where the user can benchmark their own differentiation or programming protocols, for example, and generate hypotheses. This will continue to be a focus for Stemformatics in the future, to provide high quality datasets with additional annotations, and to provide cutting edge visualisation tools to access the data easily.

The process of building a research based data portal such as Stemformatics involves overcoming many key issues commonly found in similar systems: how to best format, annotate, and store the data, how to design the system to be sustainable without an army of software engineers, and how to add real research value to the community. The new Stemformatics system implements many of the ideas learnt from years of experience and hence provides a valuable resource for other groups wanting to build such systems in future.

Stemformatics website is available at stemformatics.org. Its API server is available at api.stemformatics.org and source code is available at github.com/wellslab.

Funding The current list of funders includes:

- NHMRC Synergy grant 1186371.
- ARC Future Fellowship (FT150100330).
- JEM Research Foundation philanthropic funding.
- University of Melbourne Centre for Stem Cell Systems.

Past funders include:

• Stemformatics was established as part of the ARC Special Research Initiative to Stem Cells Australia (SR1101002).

• QLD Government Smart Futures Fellowship.

References

Angel PW, Rajab N, Deng Y et al (2020) A simple, scalable approach to building a cross-platform transcriptome atlas. PLoS Comput Biol 16(9):e1008219. https://doi.org/10.1371/journal.pcbi.1008219.

ardc (n.d.). ardc.edu.au/services/nectar-research-cloud/

Choi J, Pacheco CM, Mosbergen R et al (2019) Stemformatics: visualize and download curated stem cell data. Nucleic Acids Res 47(D1):D841–D846. https://doi.org/10.1093/nar/gky1064

Durruthy-Durruthy R, Ray M (2018) Using fluidigm C1 to generate single-cell full-length cDNA libraries for mRNA sequencing. Methods Mol Biol 2018:199–221. https://doi.org/10.1007/978-1-4939-7471-9_11

Edgar R (2002) Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. Nucleic Acids Res 30:207–210

Grüning BA, Lampa S, Vaudel M, Blankenberg D (2019) Software engineering for scientific big data analysis. GigaScience 8(5):giz054. https://doi.org/10.1093/gigascience/giz054

humancellatlas (n.d.). data.humancellatlas.org/

Kolesnikov N, Hastings E, Keays M et al (2015) ArrayExpress update—simplifying data submissions. Nucleic Acids Res 43(Database issue):D1113–D1116. https://doi.org/10.1093/nar/gku1057

mongodb (n.d.). www.mongodb.com

nuxtjs (n.d.). nuxtjs.org

Palletsprojects (n.d.). flask.palletsprojects.com

Rajab N, Angel PW, Deng Y et al (2021) An integrated analysis of human myeloid cells identifies gaps in in vitro models of in vivo biology. Stem Cell Rep 16(6):1629–1643. https://doi.org/10.1016/j.stemcr.2021.04.010. [pii] S2213-6711(21)00205-8

Silvin A, Yu CI, Lahaye X, Imperatore F, Brault J-B, Cardinaud S, Becker C, Kwan W-H, Conrad C, Maurin M, Goudot C, Marques-Ladeira S, Wang Y, Pascual V, Anguiano E, Albrecht RA, Iannacone M, García-Sastre A, Goud B, Dalod M, Moris A, Merad M, Palucka AK, Manel N (2017) Constitutive resistance to viral infection in human CD141+ dendritic cells. Sci Immunol 2(13):eaai8071. https://doi.org/10.1126/sciimmunol.aai8071

Stephenson L, Wakeham Y, Seidenman N, Choi J (2018) Building online genomics applications using BioPyramid. Bioinformatics 34(17):3055–3057. https://doi.org/10.1093/bioinformatics/bty207/

trypyramid (n.d.). trypyramid.com