Leveraging Online Resources to Prioritize Candidate Genes for Functional Analyses: Using the Fetal Testis as a Test Case

Kathryn S. McClelland · Humphrey H.-C. Yao

Reproductive and Developmental Biology Laboratory, National Institute of Environmental Health Sciences, Research Triangle Park, NC, USA

Keywords

Candidate gene · Databases · Data mining · Enrichment · Expression atlas · Gene prioritization · GO terms · Knowledgebases · Ontology · Ovary · Sex determination · Testis · Transcriptomics

Abstract

With each new microarray or RNA-seq experiment, massive quantities of transcriptomic information are generated with the purpose to produce a list of candidate genes for functional analyses. Yet an effective strategy remains elusive to prioritize the genes on these candidate lists. In this review, we outline a prioritizing strategy by taking a step back from the bench and leveraging the rich range of public databases. This *in silico* approach provides an economical, less biased, and more effective solution. We discuss the publicly available online resources that can be used to answer a range of questions about a gene. Is the gene of interest expressed in the system of interest (using expression databases)? Where else is this gene expressed (using added-value transcriptomic resources)? What pathways and processes is the gene involved in (using enriched gene pathway analysis and mouse knock-out databases)? Is this gene correlated with human diseases (using human disease variant databases)? Using mouse fetal

testis as an example, our strategies identified 298 genes annotated as expressed in the fetal testis. We cross-referenced these genes to existing microarray data and narrowed the list down to cell-type-specific candidates (35 for Sertoli cells, 11 for Leydig cells, and 25 for germ cells). Our strategies can be customized so that they allow researchers to effectively and confidently prioritize genes for functional analysis.

© 2017 S. Karger AG, Basel

The process of sex determination is complex, involving multiple pathways that shape the identity of the gonads. Identifying the genes that make up these pathways is like solving a complex puzzle. We propose that by using public databases investigators can efficiently identify new candidate genes for their involvement in gonadogenesis. While scientists have gathered a number of missing puzzle pieces, unexplained cases of disorders of sexual development (DSD) and infertility clearly indicate that many key pathways have not yet been placed in the picture [Ono and Harley, 2013]. In order to fill in the missing pieces, mouse models were used to generate numerous transcriptomic studies that reveal variations in gene expression between sexes, cell types, developmental stages, and different genotypes [Albrecht and Eicher, 2001; Nef

et al., 2005; Beverdam and Koopman, 2006; Bouma et al., 2007, 2010; Coveney et al., 2008; Rolland et al., 2011; Jameson et al., 2012; Munger et al., 2013; McClelland et al., 2015; Inoue et al., 2016]. Each of these studies had the goal to identify new genes that control key processes in sex determination of the gonads.

Over time the strategies used to generate new candidate genes have changed with the availability of different technologies. Initial cDNA array screens in the 1990s and early 2000s compared male and female gonads to search for genes enriched in either testis or ovary, with the assumption that enriched genes in one sex were important for development of that fate [Bowles et al., 2000; Wertz and Herrmann, 2000; McClive et al., 2002; Menke and Page, 2002]. With the advent of microarrays, comparing gene expression between multiple groups, such as testis and ovary at different developmental time-points, became feasible. This approach allowed the identification of genes that were expressed at key developmental time-points and begun the process of fitting them together in pathways that were initiated by marshaling transcription factors such as SOX9 [Grimmond et al., 2000; Nef et al., 2005; Coveney et al., 2008; Munger et al., 2009; Bouma et al., 2010]. Subsequently, to move beyond male and female comparisons and to gain more cell type-specific information, scientists generated transcriptomic data from enriched cell populations isolated from cell type-specific reporter mice. This strategy was used to isolate somatic cells (Sertoli, Leydig, and others) and germ cells to identify key differences between lineages in addition to sexes [Nef et al., 2005; Beverdam and Koopman, 2006; Bouma et al., 2007; Rolland et al., 2011; McClelland et al., 2015; Inoue et al., 2016]. This approach was used in the GUDMAP (GenitoUrinary Development Molecular Anatomy Project) Consortium-backed microarray of 4 key cell types (supporting, germ, interstitial, and endothelial cells), aiming at differences in cell population gene expression over a developmental time course [Jameson et al., 2012]. Over the past 2 decades, transcriptomic studies identified numerous factors and sex-linked genes that are critical for gonadogenesis. Efforts by other groups continued to fill in the puzzle with the advent of RNA-seq and whole exome sequencing. RNA-seq allows gene expression to be assayed at a greater sensitivity in a more unbiased way than microarrays on whole gonads and sorted cells [Lindeman et al., 2015; McClelland et al., 2015; Inoue et al., 2016]. In tandem, the decreasing cost for next-generation sequencing has made whole exome sequencing of DSD patients economically feasible [Ono and Harley, 2013; for review see Ostrer, 2014; Eggers et al., 2016]. The function

of genes, identified from mutations in DSD patients, can then be confirmed in the mouse model in complete knockouts or exact recapitulations of the mutated human gene. One common feature of these approaches is that they produce an extensive list of candidate genes. The key question then becomes: how do we prioritize the list of candidate genes and determine which ones to pursue for further functional analyses?

Although the way we identify candidate genes continues to evolve, the way in which we prioritize them for further functional analysis remains rudimentary; the most obvious candidates were selected based on whether they fit into existing pathways with known roles in gonadogenesis. From experience we know that transcription factors or sex chromosome-linked genes are good candidates for functional analyses. Traditionally, scientists tried to prioritize genes by targeting them with a sexually dimorphic expression pattern, profiling gene expression over a developmental time course, and then performing *in situ* hybridization or immunohistochemistry on the cell type in which the gene is expressed [Menke and Page, 2002; McClive et al., 2003; Bouma et al., 2004; Rolland et al., 2011]. The gene with the most robust sex-specific expression, the best working *in situ* hybridization probe (or antibody), and some background information in PubMed comes forward as the candidate gene. This prioritization problem is not new: genes identified in early cDNA screens were present in subsequent microarrays and were found to be critical for gonadogenesis 20–25 years after their identification. The problem was not the identification of these genes, the issue was to prioritize them or identify them as good candidates for further functional analysis. Confronted by a list that contains hundreds of genes, the ‘gene prioritization problem’ looms bigger than ever. We argue that current methods of selecting a candidate gene are unnecessarily expensive, ineffective, and biased in today’s information rich world. We propose that the community could better prioritize genes for functional validation by utilizing publically available gene expression and transcriptomic databases. These resources represent a valuable pool of information that can be used freely by any scientist to move projects and hypotheses forward.

In this review, we discuss the development and use of gene ontology (GO) terms by databases and the key features of a ‘good’ gene list before describing 4 key resources: expression atlases and databases, added-value transcriptomic databases, pathway or GO term enrichment analysis, and mouse knockout or human disease variant databases. We describe how to utilize these databases to

prioritize genes and discuss the strengths and limitations of each resource. Finally, we demonstrate how to use these resources as a discovery tool. As an example, we analyze 'testis' annotated genes in the Eurexpress expression database. Although the reanalysis of existing data sets and in silico discovery is still a fairly new concept in the field of sex determination, we argue that by capitalizing on the wide range of publically available databases, scientists can prioritize genes in a candidate list without exhausting research dollars and time chasing ghosts.

Online Tools for Generating and Analyzing the Genes-Of-Interest List

While enormous amounts of functional data are available in the published literature, each paper typically reports on a limited number of genes and developmental stages. However, there are numerous publically available databases that host a wide range of expression and transcriptomic data covering murine embryonic development. To increase the utility of individual databases, most core databases are linked with specialist databases to create ways to springboard between resources. For example, Mouse Genome Informatics (MGI; <http://www.informatics.jax.org/>) (Table 1) provides a comprehensive data page linking to most major resources for each gene. However, it is not yet possible to perform a bulk query and retrieve a set of metrics or a list of genes from a core database. How the user implements the strategies outlined in this paper will depend on the nature of the question and what types of data are used. If the investigator has already had data from a patient or mouse model, prioritizing genes by focusing on key biological processes or signaling pathways might be the first step. Conversely, if investigators are searching for new candidates, then re-analyzing published transcriptomic data may provide new genes of interest, followed by validating the expression patterns in the existing expression databases. In this section, we aim to guide investigators through how each type of database works and what can be gained from each one so they can customize their search to solve their 'gene prioritization problem'.

The Construction of a GO Framework

The GO framework [Ashburner et al., 2000] is an example of a flexible and rapidly developing textual framework for the molecular functions of gene products, including their subcellular localization, the biological processes in which they function, and the processes in which

they participate during embryogenesis. The GO Consortium (<http://geneontology.org/>) (Table 1), initially a collaboration between the Mouse Genome Informatics, FlyBase, and *Saccharomyces* Genome databases, was formed to create descriptions of biological processes and establish standards for the community-wide organization of a set of frameworks in order to enable the integration of data from different model organisms [FlyBase Consortium, 1999; Ashburner et al., 2000; Ball et al., 2000; Blake et al., 2000; Ringwald et al., 2000]. The structured language laid out by the Gene Ontology Consortium to describe the properties of gene products are called GO terms. GO terms are catalogued in a central thesaurus hosted by the GO Consortium (<http://geneontology.org/page/ontology-documentation>) (Table 1). Databases using GO terms have these terms as set outputs (or answers). Therefore, it is important to understand how GO frameworks are constructed, what the limitations of GO terms are, and how GO terms are utilized by different databases.

There are 3 central themes defined by GO terms based on the function of the protein: (1) terms that describe the *Cellular Component*, the parts of a cell or its extracellular environment; (2) terms that describe the *Molecular Function*, the basic functions of a gene product at the molecular level; and (3) terms that describe *Biological Process*, which is defined as the molecular events necessary for the function of cells, tissues, organs, and organisms [Ashburner et al., 2000]. Within the ontology framework, each term has defined relationships to one or more other terms in the same classification and sometimes to terms in other classifications. Some terms are parent terms to a host of more specific terms like a branching tree to describe things in greater and greater detail. For an example of GO annotation, the gene product 'Sox9' can be described by the Molecular Function term 'transcription', the Biological Process terms 'cell differentiation' and 'system development', and the Cellular Component term 'nucleus'. To look up terms for specific genes, a brief overview can be found in the Gene Ontology Classifications section of MGI and more detail can be obtained by querying AmiGO2 (<http://amigo.geneontology.org/amigo/landing>) (Table 1).

The consistent use of GO terms for cellular and biological processes and for anatomical structures throughout development is essential for the success, development, and maintenance of biological databases. Creating and maintaining ontologies seem straightforward; however, the creation and maintenance of ontologies in the context of developmental biology poses some interesting problems. The subtleties between different ontology

Table 1. List of resources that can be used to prioritize candidate genes

Resource	Functional description	URL
MGI (Mouse Genome Informatics)	international database resource for the laboratory mouse, providing integrated genetic, genomic, and biological data to facilitate the study of human health and disease	http://www.informatics.jax.org
Gene Ontology (GO) Consortium	develops up-to-date, comprehensive, computational models of biological systems, from the molecular level to larger pathways, cellular- and organism-level systems	http://geneontology.org/Descriptors http://geneontology.org/page/ontology-documentation http://geneontology.org/page/development
AmiGO2	searchable interface of GO Consortium	http://amigo.geneontology.org/amigo/landing
DAVID (Database for Annotation, Visualization and Integrated Discovery)	comprehensive set of functional annotation tools for investigators to understand the biological meaning behind a large list of genes	https://david.ncifcrf.gov/
GSEA (Gene Set Enrichment Analysis)	computational tool that determines whether a set of genes shows statistically significant, concordant differences between 2 biological states	http://software.broadinstitute.org/gsea/index.jsp
Entrez Gene	provides a record including nomenclature, reference sequences (RefSeqs), maps, pathways, variations, phenotypes, and links to genome-, phenotype-, and locus-specific resources	https://www.ncbi.nlm.nih.gov/gene/
ENSEMBL	genome browser for vertebrate genomes that supports research in comparative genomics, evolution, sequence variation, and transcriptional regulation	http://useast.ensembl.org/index.html
UniProt	comprehensive, high-quality and freely accessible resource of protein sequence and functional information	http://www.uniprot.org/
TRANSFAC	houses data on eukaryotic transcription factors, their experimentally proven binding sites, consensus binding sequences (positional weight matrices), and regulated genes	http://www.gene-regulation.com/pub/databases.html (last updated 2005)
EMBL-EBI Gene Expression Atlas (GEA)	gene expression data are reanalyzed in-house to detect genes showing interesting baseline and differential expression patterns in different cell types and organs, in addition to different developmental stages, disease states, and biological/experimental conditions	https://www.ebi.ac.uk/gxa/home
Online Mendelian Inheritance in Man (OMIM)	comprehensive online collection of records on human disease genes and genetic phenotypes	https://omim.org/
UK10K Project	example of one of the many smaller efforts to understand the link between low-frequency and rare genetic changes and human disease	http://www.uk10k.org
Human Gene Mutation Database (HGMD, from Cardiff University)	collection of germline mutations in nuclear genes that underlie, or are associated with, human inherited disease (a) HGMD Cardiff: unlicensed and limited access (no access to mutations identified in the last 2.5 years) (b) HGMD Professional: licensed through Qiagen and complete access to entire collection	http://www.hgmd.cf.ac.uk/ac/index.php https://www.qiagenbioinformatics.com/products/human-gene-mutation-database/
IKMC (International Knockout Mouse Consortium)	generates targeted ES cells of all known protein-coding mouse genes and companion Cre driver lines	access through IKMC/IMPC hub
IMPC (International Mouse Phenotyping Consortium)	generates the mouse strains and performs standardized phenotyping all gene trap alleles are housed at Jax (Jackson Laboratories)	www.knockoutmouse.org IKMC/IMPC web portals were merged to create a central hub: http://www.mousephenotype.org Jax: https://www.jax.org
IMSR (International Mouse Strain Resource)	searchable catalogue of over 2,000 strains	www.findmice.org ;
DMDD (Deciphering the Mechanisms of Developmental Disorders)	results from the Consortium to study knock-out lines where embryonic development is compromised	https://dmdd.org.uk
HGNC (HUGO Gene Nomenclature Committee) Multi-Symbol Checker	checks submitted gene names in a list against HUGO verified names and their known synonyms	http://www.genenames.org/cgi-bin/symbol_checker

terms used by the Biological Process Ontology Guidelines for Development to describe development and differentiation of lineages and tissues can be found in the GO Consortium resources (<http://geneontology.org/page/development>). The goal of ontology is to represent knowledge in a computer-interpretable manner so that it can (hopefully) be used to cross-reference data among different databases. The success of a database and the ontology depends greatly on the needs of the end-user of the database, the developmental biologist who wants to know about the genes and proteins involved and who wants to map the expression of the genes associated with a tissue and cell population at a different developmental stage. For example for the process of differentiation of one cell type into a different cell type, there are multiple ways this process could be represented by GO terms. Consider these 3 similar yet distinct terms that could be used to describe the differentiation process (GO term in *italics*):

[cell type] cell fate commitment

The process whereby the developmental fate of a cell becomes restricted such that it will develop into a **[cell type]** cell.

[cell type] cell fate specification

The process whereby a cell becomes capable of differentiating autonomously into a **[cell type]** cell in an environment that is neutral with respect to the developmental pathway. Upon specification, the cell fate can be reversed.

[cell type] cell fate determination

The process whereby a cell becomes capable of differentiating autonomously into a **[cell type]** cell regardless of its environment. Upon determination, the cell fate cannot be reversed.

Each of these terms, although similar on first glance, encapsulates important information about the cell and the process that takes place, and they are not interchangeable. Knowing what pertinent GO terms mean in detail and how those GO terms are nested within other GO terms is critical to extract important information about the function of a gene product.

GO terms can be a powerful tool to describe biology and make findings accessible across data formats and platforms. The core item the investigator needs to begin is a ‘good’ candidate gene list, but what defines a ‘good’ list?

Key Features of a ‘Good’ Candidate Gene List

Lists of ‘interesting’ genes, which may range in size from hundreds to thousands of genes, can reveal patterns, signaling cascades, and processes important for organogenesis. The analysis of these data sets has gradually be-

come the responsibility of the biologist rather than the bioinformatician, as more biologist-friendly tools become available. In addition to analyzing new primary data, reanalysis and inclusion of published data should be considered to bolster a new analysis. Before considering how to analyze the genes-of-interest list, the quality of the list itself should be assessed. Realistically, any correlations and directions indicated by a list should be confirmed by bench work and functional analysis. Regardless of this, as a general rule a ‘good’ gene list has a few key features [modified from Huang et al., 2009a]: (1) a reasonable number of biological (not technical) replicates and the reproducibility of the list before analysis, either from independent experiments or by statistical testing (appropriate *p* values and test parameter should be chosen for each experiment), are important considerations, especially when re-analyzing data; (2) the genes-of-interest list contains a series (not just 1 or 2 bona fide marker genes (i.e., *Sox9*, *Amh*, and *Dhh* for the fetal testis) that would be expected as a result for the given experiment or analysis. When looking at reprocessed data, it is important to remember that the statistical approaches for microarray and RNA-seq analysis have changed over time. This is especially important when processing data from pre-2012. Simple reprocessing of experiments using newer analysis methodology may result in subtle or more dramatic changes in the final genes-of-interest list compared to the original analysis. If using reprocessed data, checking for known marker genes remains a simple, yet effective metric to determine if the analysis and data is clean; and (3) the number of genes on the list is big enough for pathway analysis (around 100 genes) but not so big that it is not comparable (around 2,000–3,000 genes). The size of a list can affect the ontologies that are selected for, especially for smaller lists. As a test, the enriched terms should appear in the queried list and not in a random list of approximately the same size (number of genes).

Expression Atlases and Databases

The first question the investigator often asks is in which cell type/s a gene is expressed during gonadogenesis. In the past, in situ hybridization on embryonic gonads was used to validate the expression of candidate genes from mRNA expression studies and to determine in which cell type the gene was expressed. This first pass analysis should now be considered redundant for many developmental biologists with the completion of large-scale in situ hybridization atlases that aim to capture gene expression throughout the murine embryo on a gene-by-gene basis. Several databases used high-throughput ro-

Table 2. Expression databases and resources that can be used to prioritize candidate genes

Expression databases	Stage/tissue	Aim and scope of the database	URL
ABA (Allen Brain Atlas)	embryonic to adult brain and spinal cord	database of gene expression patterns in the mouse brain and spinal cord	http://brain-map.org/ for mouse: http://mouse.brain-map.org/
GENSAT (Gene Expression Nervous System Atlas)	10.5/15.5 dpc heads, P7/adult brain	mapping the expression of genes in the central nervous system of the mouse, using both in situ hybridization and transgenic mouse techniques; the EGFP BAC-transgenic mice created by this project are available to the scientific community (278 new Cre driver lines)	http://www.gensat.org/
GUDMAP (GenitoUrinary Development Molecular Anatomy Project)	mostly 14.5 dpc whole embryos	consortium of laboratories working to provide the scientific and medical community with tools to facilitate research into the GU system	http://www.gudmap.org/
FaceBase	various mouse, human and zebrafish	consortium generating data in support of advancing research into craniofacial development and malformation; includes data from high-throughput genetic, molecular, biological, imaging, and computational techniques	https://www.facebase.org
Eurexpress	14.5 dpc whole embryos	transcriptome atlas database for mouse embryo	http://www.eurexpress.org/ee/
GenePaint	10.5/14.5 dpc whole embryos; 15.5, P7, adult (P57) brain	digital atlas of gene expression profiles in the mouse	http://www.genepaint.org/
The e-Mouse Atlas Project (EMAP) which includes: EMA (The e-Mouse Atlas) EMAGE (Edinburgh Mouse Atlas of Gene Expression) eHistology (Online Atlas of Mouse Development)	1 cell to 18.5 dpc embryo; P0–P57	3D anatomical atlas of mouse embryo development including detailed histology	http://www.emouseatlas.org/emap/ema/
		anatomy is reconstructed from serial sections of single embryos at each representative developmental stage enabling 3D graphical display and analysis of in situ expression data	http://www.emouseatlas.org/emage/
		allows interactive exploration of cellular-resolution color images detailing mouse development with annotations from <i>The Atlas of Mouse Development</i> [Kaufman, 1994]	http://www.emouseatlas.org/emap/eHistology/
MGI-GXD (Mouse Genome Informatics Gene Expression Database)	1 cell to 18.5 dpc embryo; adult	collects and integrates the gene expression information in MGI, its primary emphasis is on endogenous gene expression during mouse development; integrates different types of data and provides links to other resources to place the data into the larger biological and analytical context	http://www.informatics.jax.org/expression.shtml
MGEIR (Mouse Gene Expression Information Resource)	pulls data from GXD and EMAP	collaboration between the GXD project and the EMAP project	under development
3D Atlas of Human Embryo Development	morphological data from Carnegie Stage 7–23	morphological (not an expression) atlas with 3D fully reconstructed human embryos covering the phase of organogenesis between Carnegie Stage 7 (15–17 days) and 23 (56–60 days)	http://www.3dembryoatlas.com/

botic technology to conduct RNA in situ hybridization on sectioned or whole murine tissues [Geffers et al., 2012]. The consortiums then manually mapped and annotated gene expression patterns on standardized images using anatomical GO terms (Fig. 1A) [Gitton et al., 2002; Raymond et al., 2002; Visel et al., 2004]. These databases can play a critical role in validation of the quality of a gene list, taking an in-depth look at many genes in a list with a focus on a specific process or pathway, or can be used as a discovery tool to uncover new candidate genes and expres-

sion patterns (see Using Databases as a Discovery Tool to Construct a Genes-Of-Interest List for Testis Development).

In the race to map the expression pattern to the transcriptome throughout development, specialized knowledgebases, such as the Allen Brain Atlas (ABA; <http://brain-map.org/>) [Lein et al., 2007; Jones et al., 2009], which catalogues gene expression in brain sections throughout development and adulthood, have developed alongside with more general whole embryo atlases, such

as Eurexpress (mainly whole mid-gestation murine embryos; <http://www.eurexpress.org/ee/>) [Diez-Roux et al., 2011] and the fledgling 3D Atlas of Human Embryo Development developed by the Academic Medical Centre in the Netherlands (Carnegie Stage 7–23 (15–60 days) (<http://www.3dembryoatlas.com/>) [de Bakker et al., 2016] (Table 2). Specialized databases cataloguing expression of genes in specific murine tissues such as the ABA, the Gene Expression Nervous System Atlas (GENSAT; <http://www.gensat.org/>; the nervous system) [Gong et al., 2003], the GenitoUrinary Development Molecular Anatomy Project (GUDMAP; the gonads, reproductive tract, kidney and urinary tract) [Brunskill et al., 2008], and FaceBase (<https://www.facebase.org>; curated craniofacial data from mouse, human, and zebrafish) [Van Otterloo et al., 2016] have now been integrated into broader databases. These collaborations bring together specialist expression data with that from projects on the whole embryo such as the Edinburgh Mouse Atlas of Gene Expression (EMAGE; <http://www.emouseatlas.org/emage/>) [Hill et al., 2004] and the Mouse Genome Informatics (MGI) Gene Expression Database (GXD; <http://www.informatics.jax.org/expression.shtml>) [Christiansen et al., 2006]. Generally, each database uses a set of standard anatomical ontology to annotate gene expression in each image [for details, see Hayamizu et al., 2015].

As an example database, the Eurexpress transcriptome atlas catalogues annotated RNA in situ hybridization expression patterns for approximately 18,000 *Mus musculus* protein-coding genes at 14.5 days post coitus (dpc). Using Eurexpress sections taken throughout the entire embryo, annotated gene expression patterns can be visualized in an online viewer. This atlas aims to achieve complete representation of all embryonic tissues throughout the 24 representative sagittal sections. For example, the gene *Star* is expressed in the adrenal and fetal Leydig cells at 14.5 dpc; in the Eurexpress assay it is annotated as expressed in the testis on section 5–8 and 14–16. Flipping through the online slide deck and zooming in, expression of *Star* is clearly detected in the adrenal and testis in section 6 (Fig. 1B). The quality of the in situ hybridization in Eurexpress is high but not always consistent. The Eurexpress Project reports that 18% of genes tested were not detected at 14.5 dpc. Whether this is due to ineffective probe design, the experimental conditions, or the gene not being expressed in the 14.5 dpc embryo is unknown. Therefore, a negative result does not necessarily mean that a gene is not expressed in vivo; false negatives cannot be excluded. In our experience, among all the genes annotated in the testis, 83% of the in situ hybridization results with annotated

testis expression were considered ‘publication quality’ (see Using Databases as a Discovery Tool to Construct a Genes-Of-Interest List for Testis Development). However, in some cases we found that expression was often detected even though it was not annotated. If the structure of interest is not annotated as ‘expressed’ or the assay is classified as ‘not detected’, we recommend the readers to look through the entire slide deck.

Eurexpress and other expression databases allow basic and advanced queries based on annotated anatomy, gene name, gene symbol, template, and gene sequence. Entries in the Eurexpress database are linked to other databases, such as ABA [Lein et al., 2007; Jones et al., 2009], EMAGE [Ringwald et al., 1994; Christiansen et al., 2006], and GENSAT [Gong et al., 2003], and to informational resources such as Entrez Gene (<https://www.ncbi.nlm.nih.gov/gene/>), ENSEMBL (<http://useast.ensembl.org/index.html>), and MGI. The key drawback of these atlases is the lack of the ability to process a batch of genes generated by approaches such as microarrays and RNA-seq. The lack of an option to do bulk queries on a gene list, instead of on individual genes, in many of these databases means that the biologist has to manually go through digital slide decks for each gene to find the structures of interest. Although structures typically occur in the same 3–5 slice window of the 24 slice slide deck, each slice image must then be manually downloaded at high resolution and be processed. Currently, the most efficient way of examining batches of genes within most expression databases is to automatically extract information from the webpage rather than displaying it to the investigator using a third party program. This way the queried image files for each gene can be downloaded as a batch, and the images can then be scrolled through on a host computer. The introduction of a tool that allows the users to submit lists of genes and return batch searches within the database will make accessing the information stored within expression databases far easier and increase the utility of expression databases for biologists. Until this kind of tool is implemented, the investigator has 2 choices: (1) harvest/extract the image files from the database website by writing an automation script, or (2) go through the browser manually. Even if the investigator chooses the more laborious manual approach, there is great value to be gained from trawling through expression atlases.

Transcriptomic Databases (for Non-Bioinformaticians)

The abundance of transcriptomic data stored in the Gene Expression Omnibus (GEO, a public functional ge-

(For legend see next page.)

nomics data repository) provides many opportunities for the investigator. However, many biologists do not have the programming skills to exploit this resource. Luckily, there are some useful databases that can assist the biologist querying transcriptomic data. Transcriptomic microarray data covering various stages of gonad development and a variety of enriched cell types has been generated by a number of groups [Jameson et al., 2012; McClelland et al., 2015; Inoue et al., 2016]. These data complement the spatiotemporal expression data in databases like Eurexpress. It is now standard for all the raw data for these types of studies to be deposited in GEO, so they are available for reanalysis. However, mining published transcriptomic data (such as microarray and RNA-seq) can be challenging without appropriate analysis frameworks. Many ‘at the bench researchers’ lack the expertise to extensively mine the RNA-seq and microarray data stored in GEO. This means researchers are not able to fully utilize published data in their field.

To use data from primary archives, a certain level of expertise is needed. Raw or processed data must be downloaded, and then the data can be analyzed independently or in combination with other data. For microarray data many biologists have some expertise in using software programs such as Partek to analyze and reanalyze data from CEL files; however, RNA-seq data are more complicated to analyze. Added-value databases make the biological content of the expression data more accessible to non-bioinformaticians. These tools aggregate data stored in repositories such as GEO by extracting relevant information from the raw primary data and therefore allow the user to ask biological questions through a user-friendly interface. For example, the user can determine in which

samples their gene-of-interest is expressed or which genes are differentially expressed between the 2 samples without having to handle the raw data directly.

One of the key added-value aggregators of transcriptomic information is the Gene Expression Atlas (<http://www.ebi.ac.uk/gxa>) [Kapushesky et al., 2009], which provides information about gene expression in different cell types and organs in addition to different developmental stages, disease states, and biological/experimental conditions. This atlas contains expression data from a large number of species including all common model organisms. The user can query individual genes, looking for differential gene expression by gene names or by searching for genes correlated with an attribute such as cell types. For example, *Sox9*/SOX9 expression can be queried in a general (to pull up all available species) or in a specific species (Fig. 1C, D). In mouse, expression data are available for a variety of embryonic and adult stages from the FANTOM 49 Consortium project and other individual projects. Examining expression in the testis reveals expression during embryonic development, specifically at E16 or 16.5 dpc in the testis, pancreas, kidney, and other organs (purple in Figure 1C). In human, expression data are available for a number of Consortia, including FANTOM 68, which examines expression in adult tissues. SOX9 is expressed in the testis, skin, and a number of distinct brain regions (purple in Figure 1D).

Pathway Analysis: GO Term Enrichment Analysis

Pathway analysis is one of the most biologist-accessible ways to look for patterns in candidate gene lists. Many similar publicly available analysis software and tools that were developed in the early 2000s can be used to function-

Fig. 1. Overview and examples of gene entries in the Eurexpress, Gene Expression Atlas, and OMIM databases. **A** In situ hybridization images of 14.5 dpc embryos are compiled in the Eurexpress Transcriptome Atlas Database for Mouse Embryo. Expression observed in an image is annotated and encoded as text that can be queried. The more general anatomical terms branch into more specific terms. For example, the nested terms that describe the testis are displayed. **B** The Eurexpress viewer scrolls through sections in a ‘virtual embryo environment’ and zooms into regions similar to a digital microscope. For the gene *Star*, the testis can be visualized in the 6th section. The testis region is demarcated by a dashed line. a, adrenal; k, kidney; m, mesonephros. **C, D** The Gene Expression Atlas houses easy to access transcriptomic data for a wide variety of species from all major consortia and sequencing efforts. For the gene *Sox9*, detailed information is available on expression in the mouse (**C**) and human (**D**). Data is presented in a tabular

format that can be sorted using a number of parameters (graphics shown here are part of the full table). Selecting data from a project in the table, such as 16.5 dpc 49 FANTOM (**C**) or 68 FANTOM (**D**), highlights the organs (in our case, the testis) that are included in that dataset in purple on the interactive body map. **E** The OMIM database catalogues human disease variants. Entries can be either by phenotype or gene. In the example of a search for the gene *NR5A1* (also known as *SF1*), there are many alternative gene symbols that have been used historically. However, the disease information is catalogued under the HUGO approved symbol (*NR5A1*). Information on the location of the gene and all phenotypes caused by mutations in *NR5A1* are listed in a table with ID numbers for each phenotype and additional information on inheritance. The entry number and mapping key encode additional information about the disorder.

ally analyze large genes-of-interest lists [Huang et al., 2009b; Khatri et al., 2012]. However, many of these databases are no longer fully updated and maintained. For this reason, out of the freely available tools we recommend DAVID (Database for Annotation, Visualization and Integrated Discover; <https://david.ncifcrf.gov/>) (Table 1) [Huang et al., 2007, 2009b] and GSEA (Gene Set Enrichment Analysis; <http://software.broadinstitute.org/gsea/index.jsp>) (Table 1) [Subramanian et al., 2005]. DAVID v6.8 is now available with an updated knowledgebase. v6.7 will be available for continued use until January 15, 2017 (data here is analyzed in v6.8Beta). GSEA and DAVID use the same core approach of searching a genes-of-interest list and then systematically map the list against a bank of GO terms in order to identify the most overrepresented or enriched terms out of all the linked terms that associate with the genes on the list. This kind of enrichment analysis strategy allows investigators to identify biological patterns and processes that may be relevant to their area of study that would never be discovered by looking at the list with the naked eye. Knowledgebases, such as DAVID and GSEA, draw on different repositories (including NCBI, <https://www.ncbi.nlm.nih.gov/>) and UniProt (<http://www.uniprot.org/>), therefore hosting multiple GO terms for a single gene in an attempt to increase the comprehensiveness of the query output. A single gene can be mapped to many different and redundant terms, just as a single term maps to many genes. DAVID deals with this GO term redundancy by clustering and classifying the redundant terms into themed ‘annotation clusters’ that can be searched by the user.

When submitting a query to a pathway analysis tool, the lists of genes can be in a number of formats. DAVID’s flexible input allows for a broad range of identifiers to be used as initial search criteria. However, it is recommended that unique universal gene identifiers are used as inputs so that redundant gene names are not confused. GSEA also has a series of curated gene lists that cover a curated range of gene sets drawn from chemical and genetic perturbation experiments from PubMed, in addition to genes that share conserved *cis*-regulatory motifs [Xie et al., 2005] and transcription factor targets (using TRANSFAC, BioBase licensed through Qiagen; <http://www.gene-regulation.com/pub/databases.html>) [Wingender et al., 2000; Matys et al., 2003]. In both GSEA and DAVID the output or results of the search can be viewed and exported in a variety of ways. Be aware that these tools will restart to the main page after a period of inactivity, regardless of which step the analysis is paused at (you can set the timeout in GSEA, but for DAVID it is auto-

matically set at 20 min inactivity). Both tools have features that link pathways and genes to disease associations, in addition to more protein-based tools that highlight protein functional domains and motifs.

For enrichment analysis, it is important to remember that size matters: a larger gene list generally results in higher statistical confidence and more significant *p* values in lowly enriched terms and more specific ontological terms at the ends of the ontological branches. Conversely, the broader and more general ontological terms are less enriched. The effect of list size on the absolute enrichment *p* values means that it is not recommended that the users directly compare the absolute enrichment *p* values across gene lists [Huang et al., 2009b]. In cases where well under 100 genes are in the genes-of-interest list, tools like DAVID can still be used. But using the statistical *p* values as metrics of significance must be used cautiously, as the statistical power behind the enrichment analysis is limited by the small number of genes [Huang et al., 2009a]. Searching such a small list will produce a very focused list of ontologies and annotations that can be thoroughly explored by the user; however, the statistics produced by the software are largely meaningless.

The basis of enrichment analysis is that there are differences between the biological processes in an abnormal or perturbed state (or in the case of developmental biology often a different cell type or time-point or genotype). The assumption is that the co-functioning genes (or related genes, which can be determined by looking at GO terms) should be enriched together and that these terms will therefore be selected as a relevant or significant group [Huang et al., 2007, 2009a, b; Huang and Yao, 2010]. The degree of enrichment depends on the background of gene expression or the noise. The background is a key factor that can influence the conclusions drawn from the data and the certainty with which we can say genes are enriched [Huang et al., 2009a]. The background must be set up to perform the comparison in tools such as DAVID and should be carefully considered. The background gene set for analysis should only include the genes that have a chance of being selected. For this reason, choosing the whole genome when the whole genome is not represented on the Affymetrix Chip can skew the data. Similarly, when analyzing RNA-seq data, including all genes in the genome instead of discarding those genes for which no counts were recorded will skew the data [Huang et al., 2009a]. Carefully selecting a background should be a priority for each study. Customized chip background lists are available for all commonly used microarray platforms, and individualized background

lists can be easily imported to meet the user's individual needs.

Mouse Knockout Databases

In order to build a comprehensive functional catalogue of the mammalian genome, a collaboration was launched to create a comprehensive library of knockout/conditional allele mouse models for researchers to utilize. In mice, this project has been spearheaded by the International Knockout Mouse Consortium (IKMC; <https://www.mousephenotype.org/>) [Bradley et al., 2012]). The aim of IKMC is to generate targeted embryonic stem (ES) cells of all known protein-coding genes in mice and companion Cre driver lines [Rosen et al., 2015]. The current design used for targeting vectors allows for the production of reporter, conditional, and knockout alleles and provides researchers with flexibility in the design of their experiments and the ability to complement CRISPR/Cas9 strategies [Rosen et al., 2015]. The International Mouse Phenotyping Consortium (IMPC; www.knockoutmouse.org) [Ring et al., 2015] builds on the work of the IKMC to generate the mouse strains and perform standardized phenotyping. The production of mouse strains from these ES cells is tracked, and this information is freely available to the research community. The IKMC/IMPC web portals were merged to create a central hub (<http://www.mousephenotype.org>) [Rosen et al., 2015] that has detailed information about the available resources including a catalogue of the targeting vectors, targeted alleles, ES cell clones, and mutant mouse strains generated and links out to other repositories. All gene trap alleles are housed at Jax (<http://www.informatics.jax.org/allele>). In addition, the International Mouse Strain Resource (IMSR; www.findmice.org) [Eppig, and Strivens, 1999; Eppig et al., 2015] has a searchable catalogue of over 2,000 Cre strains produced by the scientific community that links to the MGD and the repository holding the material.

At least 30% of the targeted knockouts generated in mouse by programs such as the IKMC and IMPC result in embryonic or perinatal death [Adams et al., 2013]. This led to the inception of the Wellcome Trust-funded research program Deciphering the Mechanisms of Developmental Disorders (DMDD; <https://dmdd.org.uk>) [Mohun et al., 2013; Adams et al., 2013; Wilson et al., 2016] that aims to characterize these lethal mutants further. DMDD focuses on phenotyping embryonic lethal mutants to shed light on the genetic regulation of tissue differentiation, organ formation, and embryo morphogenesis [Adams et al., 2013]. This resource is designed for developmental biologists and clinicians. It also comple-

ments existing United Kingdom clinical programs focused on better understanding low frequency and rare genetic changes leading to human disease, such as the Deciphering Developmental Disorders (<https://www.ddduk.org>) and UK10K (<http://www.uk10k.org>) projects.

Human Disease Variant Databases

Online Mendelian Inheritance in Man (OMIM; <http://www.omim.org/>) [Hamosh et al., 2005; Amberger et al., 2009, 2011] is a resource, published since the 1960s, aimed at cataloguing known human disease variants and improving disease classification with a focus on diseases that have a significant genetic basis. One of the goals of OMIM is to develop a standard nomenclature for features of a disorder (traits) through the Human Phenotype Ontology [Amberger et al., 2011]. OMIM is a curated resource based exclusively on the biomedical literature. The database links to genomic databases and model organism information, as well as other clinical resources. OMIM serves both molecular biologists and healthcare providers by classifying disorders and biological variation reported in the literature [Amberger et al., 2011]. Entries in OMIM can be classified under the phenotype or the gene. For example, the gene *NR5A1* (also known as *SF1*), is located on chromosome 9q33.3 (Fig. 1E). Mutations in this gene result in 4 known phenotypes: 46,XY sex reversal, premature ovarian failure, spermatogenic failure, and adrenocortical insufficiency (Fig. 1E). The allelic variants associated with each phenotype listed for a gene are briefly described under the gene entry. More detailed information about the disorder is provided in a separate descriptive entry corresponding to the phenotype (not a unique locus), and this entry has a separate identifier. The entry number and mapping key encode additional information about the disorder. More information about the assignment of different MIM numbers can be found at <http://www.omim.org/help/faq>.

OMIM now facilitates a series of more advanced search options, such as retrieval of similar concepts, clinical or anatomical features. Currently, the user is still restricted to querying a single gene at a time through the OMIM interface. However, OMIM does actively encourage the large-scale mining of its repository. API (Application Program Interface) access to the entire OMIM repository (updated nightly) is freely accessible for individual research use with a reasonable fair-use license signed upon download (<https://omim.org/api/>). Once downloaded, tools such as 'R' can be used to query batches of genes using gene names. For example, users can query all the genes on their genes-of-interest list against the OMIM

(For legend see next page.)

database and pull out features such as the gene name, MIM number, and OMIM description into a searchable Excel file. The OMIM descriptions can then be searched as a batch for pathologies and disorders of interest.

The Human Gene Mutation Database (HGMD) represents the other comprehensive collection of mutations that underlie or are associated with human inherited diseases. From its inception the HGMD (which is run out of Cardiff University, Cardiff, UK) was part of a commercial agreement, meaning it runs 2 versions. The online HGMD version that is 'free' for academics (<http://www.hgmd.cf.ac.uk/ac/index.php>) [Cooper et al., 2006; Stenson et al., 2009] is available via the Cardiff University, and a licensed version of the database, 'HGMD Professional' (Table 1) is available through Qiagen. Newly added mutational information is available to paid users for 2.5 years from the date of initial inclusion in the database before it can be accessed in the free academic resource [Stenson et al., 2009]. Therefore, we recommend using OMIM if you do not have access to the HGMD Professional version of the database.

Using Databases as a Discovery Tool to Construct a Genes-Of-Interest List for Testis Development

In addition to interrogate genes-of-interest lists, databases can also be used as a discovery tool to identify new candidate genes. By querying annotated testis expressed genes generated by the Eurexpress expression screen, we identified a pseudo-candidate list with genes expressed in the testis at 14.5 dpc. This list contains 298 entries with 289 protein-coding genes, 6 unannotated transcripts, and 3 microRNAs (online suppl. Table 1; for all online suppl. material, see www.karger.com/doi/10.1159/000455113). Among the 289 annotated protein-encoding genes, 34 of them are still listed under Rik ID numbers (numbered genes ending with 'Rik' are annotated genes without a canonical name yet) although the gene has subsequently been renamed. To ensure the remaining 255 genes are listed under the current approved gene names, we ran them through the HGNC (HUGO Gene Nomenclature

Committee) Multi-Symbol Checker (http://www.gene-names.org/cgi-bin/symbol_checker). This tool checks all the names in the submitted list against HUGO verified names and their known synonyms but does not cross-reference Rik IDs. We therefore cross-referenced the 34 genes under Rik IDs in the MGI database to locate the current gene identifier (online suppl. Table 1). As many of the databases began cataloguing entries over 15 years ago, not all genes are listed under the current HUGO approved name. This means that the investigator may overlook entries that remain curated under defunct synonyms of IDs. Once we confirmed/identified the gene names for all 295 genes (289 protein-encoding genes and 6 unannotated transcripts), we obtained the in situ hybridization images and scored them on an arbitrary metric of 1–4 (1 = very high publication quality; 2 = high publication quality; 3 = moderate quality (requires confirmation); 4 = low quality (not publication quality)). For every gene that scored a 1 or 2 (142 genes; online suppl. Table 2), we downloaded a representative full embryo image containing the testis and cropped a 600 × 600 pixel section containing the testis in Photoshop. In most cases the first testis of the pair appeared in slice 3–7 of the slide deck (before the kidney), whereas the second testis appeared between slice 11–21 (after the kidney).

We first identified previously described marker genes and validated their expression. We compared the list of 289 protein-encoding genes and 6 unannotated transcripts (a total of 295) from the Eurexpress database to published microarray data for the genes enriched in 5 key testicular cell types (supporting or Sertoli cells, germ cells, interstitial cells, Leydig cells, and endothelial cells; online suppl. Table 3) at 13.5 dpc [Jameson et al., 2012]. For the supporting or Sertoli cell-enriched genes, among 295 testis-expressing genes from the Eurexpress database and 491 genes from the Sertoli cell-enriched genes from the Jameson et al. [2012] microarray, 35 genes were found in both datasets (Fig. 2A), including 5 known Sertoli cell genes *Amh*, *Ptgds*, *Etv5*, *Tyro3*, and *Col9a3* (Fig. 2B). Sixteen of the 35 genes were of publication quality (Fig. 2C). For the male germ cell genes, 25 genes were shared by both the Eurexpress database and the Jameson et al.

Fig. 2. Sertoli cell-expressed genes identified from testis-annotated entries in the Eurexpress database. **A** The list of 295 entries annotated as testis-expressed in the Eurexpress database were compared to 491 Sertoli cell-expressed genes from a microarray of sorted cells at 13.5 dpc [Jameson et al., 2012]. A total of 35 genes were represented in both sources (online suppl. Table 2). **B** Out of these 35 genes, 5 known Sertoli cell marker genes were identified: *Amh*,

Col9a3, *Etv5*, *Ptgds*, and *Tyro3*. **C** Sixteen novel Sertoli cell candidate genes with a quality ranking of 1 or 2 were annotated in the Eurexpress database: *Adamts16*, *Atp8b2*, *Ctnna2*, *Cxadr*, *Dsp*, *Gjb2*, *Gstm7*, *Lsr*, *Npr1*, *Pgpep1*, *Slc20a1*, *Slc38a1*, *Tesc*, *Tmcc3*, *Tpd52*, and *Trank1*. The testis region is demarcated by a dashed line. a, adrenal; k, kidney; m, mesonephros. Scale bar, 0.5 mm.

(For legend see next page.)

[2012] microarray (Fig. 3A), including *Sox2*, *Dazl*, and *Dppa3* (Fig. 3B). Among the 25 genes, 10 were of publication quality in situ hybridization images from the Eurexpress database (Fig. 3C).

The testis interstitium houses heterogeneous populations of cells including vasculature, steroidogenic fetal Leydig cells, and non-steroidogenic interstitial cells. In the Jameson et al. [2012] microarray, a *Maflb*-eGFP line was used to isolate both steroidogenic and non-steroidogenic interstitial cells and an *Flk*-mCherry line was used to isolate endothelial cells. However, expression of steroidogenic Leydig cell genes was found in cells isolated from both these lines. As a result, a mixed interstitial cell list without steroidogenic genes and a steroidogenic Leydig cell list were generated by the authors. Of the 130 mixed interstitial cell genes identified in the Jameson et al. [2012] microarray, 9 genes were also represented in the Eurexpress database (Fig. 4A). In addition to 4 publication quality in situ hybridization images for known interstitial marker genes (Fig. 4B), we identified images for 5 of the 9 overlapping interstitial cell genes (Fig. 4C). RNA-seq data on sorted nonsteroidogenic interstitial cells and fetal Leydig cells at 12.5 dpc suggests that the expression of 3 of these genes (*Clca1*, *Itga9*, and *Nrg1*) may be in fetal Leydig cells [McClelland et al., 2015]. The Jameson et al. [2012] microarray produced a list of Leydig cell-specific genes, 11 of which were also represented in the Eurexpress database (Fig. 4D), and 8 of them are of publication quality (Fig. 4E).

In addition to verifying microarray and other transcriptomic data, the data from expression atlases can be used to identify novel putative markers of the different cell lineages. The resolution of the Eurexpress images is sufficient for characterizing the subdivision of organs or mapping regional differences within structurally complex organs [Diez-Roux et al., 2011; Yang and Chen, 2014; McClelland et al., 2015]. We subsequently examined the images with a quality score of a 1 or 2 that were not represented in the Jameson et al. [2012] microarray and categorized them based on testis structures. Inside the testis cords, we found 28 genes with putative Sertoli cell expres-

sion and 24 genes with putative expression in germ cells. Twenty-six genes were found in the interstitium, 9 genes were expressed in the entire testis, and 3 genes were expressed putatively in the vasculature. For the genes expressed in the testis cords, we could not determine with certainty whether the expression was localized to the Sertoli cells or the germ cells. Likewise, interstitium-expressed genes could be expressed in the fetal Leydig cells or in nonsteroidogenic interstitial cells.

This list of candidate genes (295 genes) can be considered as a small randomized data set similar to an experiment examining the expression patterns of genes in a 14.5 dpc testis. As a first pass for the following pathway analysis, we used the 'Tissue Expression' feature in DAVID to get a quick look at a list of reported expression patterns in different tissues for each gene [Huang et al., 2009a]. Using this tool we were able to interrogate expression patterns of 162 genes, and we determined that 55 genes had expression reported in the fetal or adult testis (online suppl. Table 4). We then queried the complete list of genes in DAVID (V6.8Beta) to search for enriched associations among these genes. As expected, because this gene list was randomly assembled, not a complete representation of gene expression at 14.5 dpc and not directly testing a hypothesis, there were few strongly enriched processes or pathways. Subsequently, as an example, we performed pathway enrichment analysis on the list of 289 protein-encoding genes to look for enriched clusters, and identified an overrepresentation of genes involved in glutathione metabolism (6 genes) and transcriptional regulation (39 genes) in the gene list (online suppl. Table 5). When looking at a newly generated gene list, an initial submission to DAVID after validating expression of a few genes on the list can provide directions for specific niche searches. Once a pathway or process is identified as being enriched using DAVID, expression of other pathway components can be validated using expression databases and transcriptomic resources. These pathways can then be queried in OMIM to look for an association with human diseases. For example, from RNA-seq data profiling the interstitial cells of the fetal testis, 'neuroactive ligand

Fig. 3. Germ cell-expressed genes identified from testis-annotated entries in the Eurexpress database. **A** The list of 295 entries annotated as testis-expressed in the Eurexpress database were compared to 433 germ cell-expressed genes from a microarray of sorted cells at 13.5 dpc [Jameson et al., 2012]. A total of 25 genes were represented in both sources (online suppl. Table 2). **B** These genes included the known germ cell markers *Sox2* and *Dppa3*. The germ

cell marker *Dazl* was also included as a positive control. **C** In the Eurexpress database, there were 10 novel germ cell genes that were also represented in the microarray: *Ap3b2*, *Bub1*, *Cdk5rap1*, *Irf1*, *Liph*, *Mybbp1a*, *Rcc2*, *Rcl1*, *Trip13*, and *Uchl1*. The testis region is demarcated by a dashed line. a, adrenal; k, kidney; m, mesonephros. Scale bar, 0.5 mm.

interaction' is overrepresented by analysis using DAVID. By identifying the receptor/ligand pairs in different testicular cell populations, a putative model could be constructed. Several components of this model had a known association with DSD [Diez-Roux et al., 2011; McClelland et al., 2015]. By reconstructing the pathways in silico, expression of the gene *Frem2/FREM2* (Fras1 related extracellular matrix protein 2) and its family member *Fras1* (Fraser syndrome 1 homolog), known DSD genes (OMIM:219000), were identified in the developing gonad [Jadeja et al., 2005; McClelland et al., 2015].

Tools like DAVID rely on querying what is known about the function of a gene. As a result, although powerful in uncovering functionality, this approach cannot make inferences or uncover unknown functionality. This limitation is illustrated by the example of *Pdgfa* (platelet-derived growth factor- α), which encodes the PDGF-A and is annotated in GO for the molecular function of platelet-derived growth factor receptor binding. However, although it is annotated for the biological processes of lung alveolar development (GO:0048286), salivary gland morphogenesis (GO:0060683), and bone development (GO:0060348; among others), it is not annotated for the biological process of male gonad development (GO:0008584). To a researcher in the field of sex determination and germ cell development, it may be obvious that *Pdgfa* should be involved in male gonad development, as we know this pathway is important for gonadogenesis [Brennan et al., 2003; Schmahl et al., 2008; Cool et al., 2011]. However without appropriate annotation for this biological process, the program cannot computationally retrieve *Pdgfa* as associated with testicular development. The tool is only as complete and up-to-date as the GO terms it uses.

The Power of Annotation: The Case of the Missing Ovary

If 298 entries are annotated as expressed in the testis in the Eurexpress database, we would assume that the same discovery capacity should be found in querying the ovary.

Fig. 4. Interstitial and fetal Leydig cell-expressed genes identified from testis-annotated entries in the Eurexpress database. **A** The list of 295 entries annotated as testis-expressed in the Eurexpress database were compared to 130 interstitial cell-expressed genes from a microarray of sorted cells at 13.5 dpc [Jameson et al., 2012]. A total of 9 genes were represented in both sources (online suppl. Table 2). **B** Among these genes, 4 known interstitial/Leydig cell marker genes were identified: *Cyp11a1*, *Prlr*, *Star*, and *Tacr3*.

When queried under the anatomy search for 'ovary', only 28 entries were retrieved. These numbers of genes are far less than the numbers of genes catalogued from the microarray data [Nef et al., 2005; Liu et al., 2010; Jameson et al., 2012]. We have discussed that many genes are annotated in several biological processes, yet we know that some biological processes are better studied than others. This should be a consideration when interpreting query results relying on GO terms, as some processes and gene families are more thoroughly annotated than the others largely due to historical reasons or the size of the research field. Similarly, the expression of genes annotated in some organs, for example the ovary, depends on semi-experts being able to accurately identify an organ in a tissue section [Diez-Roux et al., 2011]. In the 298 entries annotated for the testis, 11 (3.7%) genes were annotated incorrectly, where expression was actually in the neighboring pancreas or kidney. In an organ like the ovary, which does not develop apparent distinguishing features like testis cords, it is a lack of annotation in databases that hampers researchers, rather than a lack of expression of ovarian genes.

In some cases, a lack of annotation can become a serious problem that skews the results. All GO terms tend to be weighted equally in enrichment tools, whereas in reality certain ontologies are better described than others [Huang et al., 2009b]. This is often more of a problem in emerging or smaller fields. A better described ontology, based on gene families or specific processes where there is more available data, tends to have a higher chance of being associated with any list. The search is only as good as the ontology and the annotation behind it.

Conclusion

We are surrounded by transcriptomic data telling us what genes go up and what genes go down. Thanks to the affordability of transcriptomics, we have more data than we could have ever imaged. Each microarray or RNA-seq experiment produces an extensive list of candidate genes. What we have not had is a way to prioritize the genes on

C High-quality in situ hybridization images were available for 5 of the 9 overlapping interstitial genes: *Clca1*, *Cxcl12*, *Itga9*, *Nrg1*, and *Sostdc1*. **D** Eleven genes were represented in both the Leydig cell microarray and the Eurexpress database (online suppl. Table 2). **E** High-quality in situ hybridization images were available for 8 genes: *Fads1*, *Gramd1b*, *Hsd17b7*, *Prkar2b*, *Ren1*, *Scarb1*, *Sct*, and *Stc1*. The testis region is demarcated by a dashed line. a, adrenal; k, kidney; m, mesonephros. Scale bar, 0.5 mm.

these lists for functional analysis. In this review, we have provided strategies to narrow down the genes-of-interest list and identify which genes in a candidate list to pursue for further functional analyses. The key frameworks used by databases revolve around standardization of terms, both GO terms and anatomical descriptions. Understanding how GO term descriptors work and are used by different databases is a critical first step for the researchers. Coupled with an understanding of GO terms, we provide a strategy using 4 types of databases to help researchers prioritize candidate genes: (1) expression databases can be used to verify the expression of genes of interest in the different organs throughout development; (2) after confirming expression of the genes of interest, investigators can use databases like DAVID and GSEA to uncover interesting processes and pathways present in the candidate gene list. Focused searches that expand out from these pathways can identify candidate genes that may be involved in central processes in the system of interest; (3) resources such as that EMBL-EBI Expression Atlas that annotate reanalyzed transcriptomic data can provide important information about gene expression patterns across developmental time points, during disease states, and in different species including human; and (4) the importance of a gene in a human disease can then be investigated by querying knowledgebases such as OMIM.

Using this approach, a candidate gene list can be whittled down to a selection of promising candidates that fit the investigator's criteria for a strong candidate for functional analysis.

This prioritizing approach requires the investigator to move between multiple databases. Luckily, there are new efforts underway to centralize the data housed by different Consortia and to better store and annotate the kind of visual data that developmental biologists generate. The best example of this evolving database and framework is the Edinburgh Mouse Atlas Project (EMAP), which aims to build a 3D anatomical atlas of mouse embryo development using a flexible anatomical ontology. The ontology uses a thesaurus of alternative anatomical terminologies that was built upon *The Atlas of Mouse Development* [Kaufman, 1994] and allows flexibility within queries. This framework, by enabling visual tracing of progenitors and derivatives for each structure or organ, aims to provide the kind of fate tracing map that are missing in other databases. The structures in the digital model are anchored by key anatomical terms, and the data are then associated with that tissue 'lineage' within the spatial framework. This kind of framework provides information about lineages belong-

ing to structures, like an organ, over time. The EMAP database currently links out to and draws data from Eurexpress, and both EMAP and Eurexpress link out to interface with resources such as MGI.

Ultimately, the community needs these resources to become more centralized. To this end, the MGI Gene Expression Database (GXD) project and the Edinburgh Mouse Atlas (EMA) project are working together to develop a comprehensive resource called the Mouse Gene Expression Information Resource (MGEIR) (<http://www.emouseatlas.org/emap/about/mgeir.html>). MGEIR aims to provide a central resource that will house text- and spatial-based gene expression data which will be acquired from the literature and by direct data submission from external labs/consortia. This resource merges 'spatial' in situ hybridization data into the EMAP's 3D format and flexible anatomical framework with the 'non-spatial' expression data (i.e., RT-PCR) collected in the GXD [Hill et al., 2004; Christiansen et al., 2006; de Boer et al., 2009; Geffers et al., 2012]. The development of MGEIR and the creation of centralized databanks will provide the investigators with an amazing array of information at the click of a mouse.

The next big challenge for expression databases will be the onslaught of single cell RNA-seq data. This new technology offers an unprecedented amount of detail about the transcriptome of any tissue. It's reasonable to expect that in the next few years, high quality single cell datasets will be available for many organs, including the developing mouse and human gonads. How these datasets are integrated into expression databases will be a challenge for the use of ontologies. The flexible anatomical design, used by databases like EMAP, may provide enough flexibility to trace progenitor cell populations and population derivatives for each tissue or organ (for a detailed overview of current eMouseAtlas capabilities see Armit et al. [2015]). How tissue 'lineages' are managed within the spatial framework of the expression atlas will provide a challenge for database custodians.

The next time when an investigator is facing a 'gene prioritization problem', we encourage him/her to step away from the bench and use the tools and databases described here to move the projects and hypotheses forward. Publically available data can be used to efficiently narrow down a candidate gene list to produce strong candidates for functional analysis. Using databases, genes can be selected that have known expression in the organ of interest and are involved in important signaling pathways or biological processes. The genes of interest can be represented in other transcriptomic studies or associated

with human disease. So, before you decide to make a genetically modified mouse, sit down at your desk and put the other type of mouse to work.

Acknowledgements

We thank members of the Yao Lab for helpful discussions on the concepts in this paper. The concept for this review was born out of work done by K.S.M. during her time in the Koopman Lab, University of Queensland, Australia. All work contributing to this review was completed in the Yao Lab, NIEHS/NIH, USA. K.S.M.

would like to thank Peter Koopman and Josephine Bowles for their encouragement and support. This research was supported by the Intramural Research Program (ES102965 to H.H.-C.Y.) of the NIH, National Institute of Environmental Health Sciences. The *in situ* hybridization data in this review is drawn exclusively from the Euxpress Transcriptome Atlas Database in May 2016.

Disclosure Statement

The authors have no conflicts of interest to declare.

References

- Adams D, Baldock R, Bhattacharya S, Copp AJ, Dickinson M, et al: Bloomsbury report on mouse embryo phenotyping: recommendations from the IMPC workshop on embryonic lethal screening. *Dis Model Mech* 6:571–579 (2013).
- Albrecht KH, Eicher EM: Evidence that Sry is expressed in pre-Sertoli cells and Sertoli and granulosa cells have a common precursor. *Dev Biol* 240:92–107 (2001).
- Amberger J, Bocchini CA, Scott AF, Hamosh A: McKusick's Online Mendelian Inheritance in Man (OMIM). *Nucleic Acids Res* 37:D793–796 (2009).
- Amberger J, Bocchini CA, Hamosh A: A new face and new challenges for Online Mendelian Inheritance in Man (OMIM®). *Hum Mutat* 32:564–567 (2011).
- Armit C, Richardson L, Hill B, Yang Y, Baldock RA: eMouseAtlas informatics: embryo atlas and gene expression database. *Mamm Genome* 26:431–440 (2015).
- Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, et al: Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet* 25:25–29 (2000).
- Ball CA, Dolinski K, Dwight SS, Harris MA, Issel-Tarver L, et al: Integrating functional genomic information into the Saccharomyces genome database. *Nucleic Acids Res* 28:77–80 (2000).
- Beverdam A, Koopman P: Expression profiling of purified mouse gonadal somatic cells during the critical time window of sex determination reveals novel candidate genes for human sexual dysgenesis syndromes. *Hum Mol Genet* 15:417–431 (2006).
- Blake JA, Eppig JT, Richardson JE, Davisson MT: The Mouse Genome Database (MGD): expanding genetic and genomic resources for the laboratory mouse. The Mouse Genome Database Group. *Nucleic Acids Res* 28:108–111 (2000).
- Bouma GJ, Hart GT, Washburn LL, Recknagel AK, Eicher EM: Using real time RT-PCR analysis to determine multiple gene expression patterns during XX and XY mouse fetal gonad development. *Gene Expr Patterns* 5:141–149 (2004).
- Bouma GJ, Affourtit JP, Bult CJ, Eicher EM: Transcriptional profile of mouse pre-granulosa and Sertoli cells isolated from early-differentiated fetal gonads. *Gene Expr Patterns* 7:113–123 (2007).
- Bouma GJ, Hudson QJ, Washburn LL, Eicher EM: New candidate genes identified for controlling mouse gonadal sex determination and the early stages of granulosa and Sertoli cell differentiation. *Biol Reprod* 82:380–389 (2010).
- Bowles J, Bullejos M, Koopman P: A subtractive gene expression screen suggests a role for *vanin-1* in testis development in mice. *Genesis* 27:124–135 (2000).
- Bradley A, Anastassiadis K, Ayadi A, Battey JF, Bell C, et al: The mammalian gene function resource: the International Knockout Mouse Consortium. *Mamm Genome* 23:580–586 (2012).
- Brennan J, Tilmann C, Capel B: *Pdgfr-α* mediates testis cord organization and fetal Leydig cell development in the XY gonad. *Genes Dev* 17:800–810 (2003).
- Brunskill EW, Aronow BJ, Georgas K, Rumballe B, Valerius MT, et al: Atlas of gene expression in the developing kidney at microanatomic resolution. *Dev Cell* 15:781–791 (2008).
- Christiansen JH, Yang Y, Venkataraman S, Richardson L, Stevenson P, et al: EMAGE: a spatial database of gene expression patterns during mouse embryo development. *Nucleic Acids Res* 34:D637–41 (2006).
- Cool J, DeFalco TJ, Capel B: Vascular-mesenchymal cross-talk through *Vegf* and *Pdgf* drives organ patterning. *Proc Natl Acad Sci USA* 108:167–172 (2011).
- Cooper DN, Stenson PD, Chuzhanova NA: The Human Gene Mutation Database (HGMD) and its exploitation in the study of mutational mechanisms. *Curr Protoc Bioinformatics* chap 1:unit 1.13 (2006).
- Coveney D, Ross AJ, Slone JD, Capel B: A microarray analysis of the XX *Wnt4* mutant gonad targeted at the identification of genes involved in testis vascular differentiation. *Gene Expr Patterns* 8:529–537 (2008).
- de Bakker BS, de Jong KH, Hagoort J, de Bree K, Besselink CT, et al: An interactive three-dimensional digital atlas and quantitative database of human development. *Science* 354:aag0053 (2016).
- de Boer BA, Ruijter JM, Voorbraak FPJM, Moorman AFM: More than a decade of developmental gene expression atlases: where are we now? *Nucleic Acids Res* 37:7349–7359 (2009).
- Diez-Roux G, Banfi S, Sultan M, Geffers L, Anand S, et al: A high-resolution anatomical atlas of the transcriptome in the mouse embryo. *PLoS Biol* 9:e1000582 (2011).
- Eggers S, Sadedin S, van den Bergen JA, Robevska G, Ohnesorg T, et al: Disorders of sex development: insights from targeted gene sequencing of a large international patient cohort. *Genome Biol* 17:243 (2016).
- Eppig JT, Strivens M: Finding a mouse: the International Mouse Strain Resource (IMSR). *Trends Genet* 15:81–82 (1999).
- Eppig JT, Motenko H, Richardson JE, Richards-Smith B, Smith CL: The International Mouse Strain Resource (IMSR): cataloging worldwide mouse and ES cell line resources. *Mamm Genome* 26:448–455 (2015).
- FlyBase Consortium: The FlyBase database of the Drosophila Genome Projects and community literature. *Nucleic Acids Res* 27:85–88 (1999).
- Geffers L, Herrmann B, Eichele G: Web-based digital gene expression atlases for the mouse. *Mamm Genome* 23:525–538 (2012).
- Gitton Y, Dahmane N, Baik S, Altaba ARI, Neidhardt L, et al: A gene expression map of human chromosome 21 orthologues in the mouse. *Nature* 420:586–590 (2002).
- Gong S, Zheng C, Doughty ML, Losos K, Didkovsky N, et al: A gene expression atlas of the central nervous system based on bacterial artificial chromosomes. *Nature* 425:917–925 (2003).

- Grimmond S, Van Hateren N, Siggers P, Arkell R, Larder R, et al: Sexually dimorphic expression of protease nexin-1 and vanin-1 in the developing mouse gonad prior to overt differentiation suggests a role in mammalian sexual development. *Hum Mol Genet* 9:1553–1560 (2000).
- Hamosh A, Scott AF, Amberger JS, Bocchini CA, McKusick VA: Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. *Nucleic Acids Res* 33:D514–517 (2005).
- Hayamizu TF, Baldock RA, Ringwald M: Mouse anatomy ontologies: enhancements and tools for exploring and integrating biomedical data. *Mamm Genome* 26:422–430 (2015).
- Hill DP, Begley DA, Finger JH, Hayamizu TF, McCright IJ, et al: The mouse Gene Expression Database (GXD): updates and enhancements. *Nucleic Acids Res* 32:D568–71 (2004).
- Huang CC, Yao HH: Diverse functions of Hedgehog signaling in formation and physiology of steroidogenic organs. *Mol Reprod Dev* 77:489–496 (2010).
- Huang DW, Sherman BT, Tan Q, Kir J, Liu D, et al: DAVID bioinformatics resources: expanded annotation database and novel algorithms to better extract biology from large gene lists. *Nucleic Acids Res* 35:W169–175 (2007).
- Huang DW, Sherman BT, Lempicki RA: Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat Protoc* 4:44–57 (2009a).
- Huang DW, Sherman BT, Lempicki RA: Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. *Nucleic Acids Res* 37:1–13 (2009b).
- Inoue M, Shima Y, Miyabayashi K, Tokunaga K, Sato T, et al: Isolation and characterization of fetal Leydig progenitor cells of male mice. *Endocrinology* 157:1222–1233 (2016).
- Jadeja S, Smyth I, Pitera JE, Taylor MS, van Haelst M, et al: Identification of a new gene mutated in Fraser syndrome and mouse myelencephalic blebs. *Nat Genet* 37:520–525 (2005).
- Jameson SA, Natarajan A, Cool J, Defalco T, Maatouk DM, et al: Temporal transcriptional profiling of somatic and germ cells reveals biased lineage priming of sexual fate in the fetal mouse gonad. *PLoS Genet* 8:e1002575 (2012).
- Jones AR, Overly CC, Sunkin SM: The Allen Brain Atlas: 5 years and beyond. *Nat Rev Neurosci* 10:821–828 (2009).
- Kapuszesky M, Emam I, Holloway E, Kurnosov P, Zorin A, et al: Gene expression atlas at the European bioinformatics institute. *Nucleic Acids Res* 38:D690–D698 (2009).
- Kaufman MH: The Atlas of Mouse Development (Elsevier, Amsterdam 1994).
- Khatri P, Sirota M, Butte AJ: Ten years of pathway analysis: current approaches and outstanding challenges. *PLoS Comput Biol* 8:e1002375 (2012).
- Lein ES, Hawrylycz MJ, Ao N, Ayres M, Bensinger A, et al: Genome-wide atlas of gene expression in the adult mouse brain. *Nature* 445:168–176 (2007).
- Lindeman RE, Gearhart MD, Minkina A, Krentz AD, Bardwell VJ, Zarkower D: Sexual cell-fate reprogramming in the ovary by DMRT1. *Curr Biol* 25:764–771 (2015).
- Liu CF, Liu C, Yao HH: Building pathways for ovary organogenesis in the mouse embryo. *Curr Top Dev Biol* 90:263–290 (2010).
- Matys V, Fricke E, Geffers R, Gössling E, Hubbard M, et al: TRANSFAC: transcriptional regulation, from patterns to profiles. *Nucleic Acids Res* 31:374–378 (2003).
- McClelland KS, Bell K, Larney C, Harley VR, Sinclair AH, et al: Purification and transcriptomic analysis of mouse fetal Leydig cells reveals candidate genes for specification of gonadal steroidogenic cells. *Biol Reprod* 92:145 (2015).
- McClive PJ, Hurley TM, Sarraj MA, van den Bergen JA, Sinclair AH: Subtractive hybridisation screen identifies sexually dimorphic gene expression in the embryonic mouse gonad. *Genesis* 37:84–90 (2003).
- Menke DB, Page DC: Sexually dimorphic gene expression in the developing mouse gonad. *Gene Expr Patterns* 2:359–367 (2002).
- Mohun T, Adams DJ, Baldock R, Bhattacharya S, Copp AJ, et al: Deciphering the Mechanisms of Developmental Disorders (DMDD): a new programme for phenotyping embryonic lethal mice. *Dis Model Mech* 6:562–566 (2013).
- Munger SC, Aylor DL, Syed HA, Magwene PM, Threadgill DW, Capel B: Elucidation of the transcription network governing mammalian sex determination by exploiting strain-specific susceptibility to sex reversal. *Genes Dev* 23:2521–2536 (2009).
- Munger SC, Natarajan A, Looger LL, Ohler U, Capel B: Fine time course expression analysis identifies cascades of activation and repression and maps a putative regulator of mammalian sex determination. *PLoS Genet* 9:e1003630 (2013).
- Nef S, Schaad O, Stallings NR, Cederroth CR, Pitetti JL, et al: Gene expression during sex determination reveals a robust female genetic program at the onset of ovarian development. *Dev Biol* 287:361–377 (2005).
- Ono M, Harley VR: Disorders of sex development: new genes, new concepts. *Nat Rev Endocrinol* 9:79–91 (2013).
- Ostrer H: Disorders of sex development (DSDs): an update. *J Clin Endocrinol Metab* 99:1503–1509 (2014).
- Reymond A, Marigo V, Yaylaoglu MB, Leoni A, Ucla C, et al: Human chromosome 21 gene expression atlas in the mouse. *Nature* 420:582–586 (2002).
- Ring N, Meehan TF, Blake A, Brown J, Chen C-K, et al: A mouse informatics platform for phenotypic and translational discovery. *Mamm Genome* 26:413–421 (2015).
- Ringwald M, Baldock R, Bard J, Kaufman M, Eppig JT, et al: A database for mouse development. *Science* 265:2033–2034 (1994).
- Ringwald M, Eppig JT, Kadin JA, Richardson JE: GXD: a Gene Expression Database for the laboratory mouse: current status and recent enhancements. *The Gene Expression Database group. Nucleic Acids Res* 28:115–119 (2000).
- Rolland AD, Lehmann KP, Johnson KJ, Gaido KW, Koopman P: Uncovering gene regulatory networks during mouse fetal germ cell development. *Biol Reprod* 84:790–800 (2011).
- Rosen B, Schick J, Wurst W: Beyond knockouts: the International Knockout Mouse Consortium delivers modular and evolving tools for investigating mammalian genes. *Mamm Genome* 26:456–466 (2015).
- Schmahl J, Rizzolo K, Soriano P: The PDGF signaling pathway controls multiple steroid-producing lineages. *Genes Dev* 22:3255–3267 (2008).
- Stenson PD, Ball E, Howells K, Phillips A, Mort M, Cooper DN: Human Gene Mutation Database: towards a comprehensive central mutation database. *J Med Genet* 45:124–126 (2008).
- Stenson PD, Mort M, Ball EV, Howells K, Phillips AD, et al: The Human Gene Mutation Database: 2008 update. *Genome Med* 1:13 (2009).
- Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, et al: Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci USA* 102:15545–15550 (2005).
- Van Otterloo E, Williams T, Artinger KB: The old and new face of craniofacial research: how animal models inform human craniofacial genetic and clinical data. *Dev Biol* 415:171–187 (2016).
- Visel A, Thaller C, Eichele G: GenePaint.org: an atlas of gene expression patterns in the mouse embryo. *Nucleic Acids Res* 32:D552–6 (2004).
- Wertz K, Herrmann BG: Large-scale screen for genes involved in gonad development. *Mech Dev* 98:51–70 (2000).
- Wilson R, McGuire C, Mohun T, DMDD Project: Deciphering the mechanisms of developmental disorders: phenotype analysis of embryos from mutant mouse lines. *Nucleic Acids Res* 44:D855–61 (2016).
- Wingender E, Chen X, Hehl R, Karas H, Liebig I, et al: TRANSFAC: an integrated system for gene expression regulation. *Nucleic Acids Res* 28:316–319 (2000).
- Xie X, Lu J, Kulbokas EJ, Golub TR, Mootha V, et al: Systematic discovery of regulatory motifs in human promoters and 3' UTRs by comparison of several mammals. *Nature* 434:338–345 (2005).
- Yang J, Chen J: Developmental programs of lung epithelial progenitors: a balanced progenitor model. *Wiley Interdiscip Rev Dev Biol* 3:331–347 (2014).