Gene expression databases for kidney epithelial cells

Jennifer C. Huling, Trairak Pisitkun, Jae H. Song, Ming-Jiu Yu, Jason D. Hoffert, and Mark A. Knepper
Epithelial Systems Biology Laboratory, National Heart, Lung, and Blood Institute, National Institutes of Health, Bethesda, Maryland

Submitted 11 August 2011; accepted in final form 18 November 2011

Huling JC, Pisitkun T, Song JH, Yu MJ, Hoffert JD, Knepper MA. Gene expression databases for kidney epithelial cells. Am J Physiol Renal Physiol 302: F401–F407, 2012. First published November 23, 2011; doi:10.1152/ajprenal.00457.2011.—The 21st century has seen an explosion of new high-throughput data from transcriptomic and proteomic studies. These data are highly relevant to the design and interpretation of modern physiological studies but are not always readily accessible to potential users in user-friendly, searchable formats. Data from our own studies involving transcriptomic and proteomic profiling of renal tubule epithelia have been made available on a variety of online databases. Here, we provide a roadmap to these databases and illustrate how they may be useful in the design and interpretation of physiological studies. The databases can be accessed through http://helixweb.nih.gov/ESBL/Database.

proximal tubule; thick ascending limb of Henle; inner medullary collecting duct; mpkCCD

MODERN PHYSIOLOGICAL RESEARCH depends on integration of information at the gene or protein level with functional information obtained from classic technical approaches. For example, the development and exploitation of transgenic and knockout mice benefit from knowledge of gene expression patterns in the tissue of interest. Before initiation of a knockout mouse project, it is useful to know, for example, whether the gene to be knocked out is expressed in the target tissue, whether the target tissue expresses similar genes that may compensate for the deleted gene, and whether the gene of interest is expressed in different isoforms (e.g., splicing variants or with different posttranslational modifications). Similarly, a common approach in modern kidney physiology is to carry out probing genetic manipulations in cultured cell models. These manipulations can involve gene knockdowns, overexpression of dominant negatives, and mutational analysis to explore structure-function relationships. Again, knowledge of the gene expression profile in a cell culture model can be helpful in the design of informative studies. In addition, antibodies have become important tools in physiological research and can readily be designed to localize or quantify virtually every protein expressed in a given tissue. However, the choice of immunogen for antibody production and the interpretation of results from such studies can benefit from prior knowledge of the repertoire of genes expressed in the tissue being targeted.

The turn of the century saw the advent of genome-sequencing projects for a variety of organisms of physiological interest. Of the ~21,000 protein-coding genes in mammalian genomes, ~8,000–10,000 appear to be expressed in specific renal epithelial cell types (22, 25). Transcriptomic and proteomic techniques have been employed to map gene expression lists to specific cell types in the kidney and in cell culture models. We have placed the data acquired in our laboratory on publically accessible WWW sites for use in the design and interpretation of physiological experiments. The object of this short review is to give the reader a roadmap to these sites and to provide examples of how useful information can be extracted from the sites.

Transcriptomic Databases

Using Affymetrix expression arrays, we have reported transcriptomic profiles for three renal tubule segments in rat (proximal tubule (25), medullary thick ascending limb (25), and inner medullary collecting duct (22)) and in a series of clones of the mpkCCD cell line (25). The WWW URLs for the corresponding databases are shown in Table 1. (Tested in Firefox [recommended], Safari, and Internet Explorer). For the renal tubule transcriptomic data, we have built an entry portal (Fig. 1; http://helixweb.nih.gov/ESBL/Database/Transcriptomic/index.html) which offers users two ways of accessing the data. Users can either go directly to one of the renal tubule segment-specific databases by clicking on the appropriate segment in the nephron diagram (Fig. 1A), or they can search all three databases simultaneously to determine in which of the three segments the gene of interest is expressed and the relative expression levels (Fig. 1B). The search is done by entering the amino acid sequence of the coded gene in FASTA protein format, which can be obtained at http://www.ncbi.nlm.nih.gov/protein.1 The search, using the BLAST2 algorithm, then finds

1 FASTA, pronounced “fast-A,” is a sequence alignment algorithm similar in many ways to BLAST (see below). It is not used very frequently now, but it introduced a particular format for representing sequence information and metadata that is now the standard representation format for most sequence alignment tasks, to wit “FASTA format.”

2 BLAST (or Basic Local Alignment Search Tool) is an computer algorithm that is used to compare a biological sequence (here, an amino acid sequence) with all sequences in a library of sequences to determine what members of the target library are most similar to the test sequence. The output is dependent on choice of internal parameters and the target library. Here, the purpose of the BLAST search is to find out whether a particular protein, represented by the test sequence, is present in a given cell-specific database. Accordingly, for this purpose, we use a library that we have constructed from a list of proteins or transcripts that have been detected in the renal cell type of interest.

http://www.ajprenal.org
the best protein matches from the entries in the three renal tubule databases. Note that the BLAST algorithm used on our website searches databases containing only entries for proteins found in the three renal tubule segments, in contrast to BLAST searches performed at the NCBI website (www.ncbi.nlm.nih.gov/BLAST/). Figure 2 shows a typical output for such a search using the amino acid sequence for “integrin-β6 precursor” (RefSeq identifier NP_001004263) as input.3 As can be seen, the search not only identified integrin-β6 as expressed in all three segments but also identified other integrins, two of

Table 1. Database URLs

<table>
<thead>
<tr>
<th>Transcriptome Databases</th>
<th>Proteome Databases</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rat proximal tubule transcriptome</td>
<td>IMCD proteome database</td>
</tr>
<tr>
<td>Rat TAL transcriptome</td>
<td>IMCD membrane protein Database</td>
</tr>
<tr>
<td>Rat IMCD transcriptome</td>
<td>Mouse mpkCCD clone 11 proteome database</td>
</tr>
<tr>
<td>Mouse mpkCCD clone 11 vs. clone 2 transcriptome</td>
<td>Mouse mpkCCD clone 11 proteome database</td>
</tr>
<tr>
<td></td>
<td>IMCD membrane protein Database</td>
</tr>
<tr>
<td></td>
<td>IMCD phosphoproteome database</td>
</tr>
<tr>
<td></td>
<td>Mouse mpkCCD phosphoproteome database</td>
</tr>
<tr>
<td></td>
<td>Rat medullary TAL phosphoproteome database</td>
</tr>
<tr>
<td></td>
<td>Renal cortical membrane phosphoproteome database</td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
</tbody>
</table>

3 This sequence can conveniently be found by clicking on “How do I find this information?” near the BLAST input box.
which are selectively expressed in only one renal tubule segment.

The output of each search includes the Affymetrix Probe Set ID, RefSeq number and Swiss-Prot number for each identified gene. The latter two numbers also serve as links to the corresponding protein records. The output also shows the BLAST similarity score and the corresponding expect value (E-value) using the selected substitution matrix (1) to help the user gauge the degree of similarity.4 The amino acid alignments for each match are provided at the bottom of the page (not shown in Fig. 2). The search uses protein sequences corresponding to transcripts and presents amino acid alignments here because of the physiological context of the tool, recognizing that proteins are the macromolecules that determine most molecular functions.

The three columns highlighted in Fig. 2 correspond to each of the 3 segment-specific databases being searched. A green icon in a column means that the protein does appear in the corresponding database, and a red icon means the protein does not (determined if the Probe Set ID had a signal strength <0.2). The numbers beside each icon are the signal strengths. B: when the user holds the cursor over the word “view” a visual summary of the previous three columns is shown. The signal strengths are displayed over the correct nephron region. Each region is also shaded in blue, with darker blue shading corresponding to a higher signal level. IMCD, inner medullary collecting duct.

Fig. 2. Sample output of a BLAST search. The Results page gives a brief explanation, a diagram of the nephron for reference, and a summary table displaying the key results of the search. The table lists matches based on amino acid sequence and for each protein displays general information about the protein and about the protein’s location. A: 3 columns highlighted correspond to each of the 3 segment-specific databases being searched. A green icon in a column means that the protein does appear in the corresponding database, and a red icon means the protein does not (determined if the Probe Set ID had a signal strength <0.2). The numbers beside each icon are the signal strengths. B: when the user holds the cursor over the word “view” a visual summary of the previous three columns is shown. The signal strengths are displayed over the correct nephron region. Each region is also shaded in blue, with darker blue shading corresponding to a higher signal level. IMCD, inner medullary collecting duct.

4 “E-value” is a value generated by BLAST that summarizes the chances of finding the observed degree of overlap between the input sequence and the target sequence purely by chance. The lower the E-value, the more likely the match indicates that the selected target protein is biologically related to the input sequence. If the specific protein of interest is present in the database, the E-value is usual 0.0, indicating a complete match. However, if the target sequence contains genetic variant sequences or sequencing errors, a very small non-zero value may be obtained.

<table>
<thead>
<tr>
<th>Probe Set ID</th>
<th>RefSeq Number</th>
<th>Protein Name</th>
<th>Swiss-Prot Score</th>
<th>E-Value</th>
<th>PT Signal Strength</th>
<th>TAL Signal Strength</th>
<th>IMCD Signal Strength</th>
<th>Display (not in IE)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1392419.1d</td>
<td>NM_001098243</td>
<td>integrin beta-6 precursor</td>
<td>Q647F4</td>
<td></td>
<td>1.53</td>
<td>2.03</td>
<td>9.52</td>
<td>view</td>
</tr>
<tr>
<td>1398358.1a</td>
<td>NM_001098243</td>
<td>integrin, beta 5</td>
<td>...</td>
<td>766</td>
<td>1.77</td>
<td>7.65</td>
<td>4.74</td>
<td>view</td>
</tr>
<tr>
<td>1360519.1d</td>
<td>NM_001098243</td>
<td>integrin beta-1</td>
<td>P49138</td>
<td>1113.2</td>
<td>4.51</td>
<td>13.7</td>
<td>17.35</td>
<td>view</td>
</tr>
<tr>
<td>1368612.1d</td>
<td>NM_001098243</td>
<td>integrin beta-4 precursor</td>
<td>Q64612</td>
<td>404</td>
<td>0.05</td>
<td>0.00</td>
<td>0.00</td>
<td>view</td>
</tr>
<tr>
<td>1383912.1a</td>
<td>NM_001098243</td>
<td>laminin, alpha 5</td>
<td>41.1</td>
<td>5e-04</td>
<td>8.23</td>
<td>1.01</td>
<td>5.65</td>
<td>view</td>
</tr>
<tr>
<td>1389321.1d</td>
<td>NM_001098243</td>
<td>laminin, alpha 5</td>
<td>41.1</td>
<td>5e-04</td>
<td>8.23</td>
<td>1.01</td>
<td>5.65</td>
<td>view</td>
</tr>
<tr>
<td>1391041.1d</td>
<td>NM_001098243</td>
<td>laminin, beta 3</td>
<td>42.4</td>
<td>7e-04</td>
<td>0.0</td>
<td>0.00</td>
<td>4.42</td>
<td>view</td>
</tr>
</tbody>
</table>
base expansion is expected to involve the development and exploitation of transgenic mice that target selectable fluorescent proteins to specific cell types to allow sorting on a cellular or organellar level. Details of these methods are beyond the scope of the current review. Beyond this, it will be useful to carry out transcriptomic profiling of epithelial cell lines other than the mpkCCD cells discussed above.

Additional transcriptomic data are available from other sources for renal tubule segments and renal epithelial cell culture models. For example, transcriptomic data have been reported from application of Serial Analysis of Gene Expression (SAGE) for mpkCCD cells (17) and for various microdissected renal tubule segments (4–6, 15, 20, 24, 28). In addition, microarray data are available from a cultured renal proximal tubule cell line (8).

Proteomic Databases

Using protein mass spectrometry, we have identified the proteomes of both the renal inner medullary collecting duct (2, 3, 10–13, 18, 19, 21, 23, 26, 27) and cultured mpkCCD cells (clone 11) (16). The URLs for the corresponding databases are shown in Table 1. The IMCD Proteome Database (http://helixweb.nih.gov/ESBL/Database/IMCD_Proteome/index.html) is organized in a manner similar to that of the transcriptomic databases discussed above. The “Tech view” feature allows the user to find the correct reference with the appropriate description of the techniques used to generate the data. One additional feature of the database is that each individual protein name links to the “Protein Viewer” feature, which shows a linear map of the protein from N terminus to C terminus with various features plotted as a function of amino acid number, including Kyte-Doolittle hydrophathy, Chou-Fasman secondary structure predictions, and relative immunogenicity (Fig. 4). The Protein Viewer feature has been adapted from another program, NHLBI-AbDesigner (14), for design of peptide-directed antibodies. Note that Java Runtime Environment (JRE) is needed to show Java applets within Protein Viewer. JRE is preloaded on most desktop and laptop computers. If not, the user can download JRE free at http://www.java.com/en/.

Phosphoproteomic Databases

Protein phosphorylation is an important posttranslational modification that is part of virtually every signaling pathway in eukaryotic organisms. Consequently, we have carried out extensive studies in which we have identified and quantified phosphorylation sites in proteins using LC-MS/MS techniques in various renal tubule epithelia, viz., inner medullary collecting duct (2, 10), medullary thick ascending limb (9), proximal tubule and other renal cortical segments (7), and cultured mpkCCD (clone 11) cells (16). The URLs for the corresponding databases are shown in Table 1. Figure 5 shows a screen shot of the proximal tubule phosphoproteomics site. This and other phosphoproteomic databases show major features of the phosphopeptides detected in a tabular format listing the RefSeq Accession Number of the protein, the Official Gene Symbol, the protein name, the sequence of the identified phosphopeptide, the amino acid(s) phosphorylated, and the specificity of the site assignment (i.e., whether it is assigned with certainty or with lower confidence).
Many of the phosphorylation sites detected in these studies are novel, i.e., not previously reported. Consequently, investigators may discover something new about a given protein just by looking it up on each of the phosphoproteomic databases.

**DISCUSSION**

Here, we have presented a series of transcriptomic and proteomic databases that provide a resource for modern renal physiological research. These databases are not comprehensive in the sense that not all renal cell types are covered and data from laboratories other than ours have not been incorporated. We propose that further work toward a comprehensive set of databases be carried out collaboratively among members of the renal community. Meanwhile, the data included in the databases presented here have been extensively employed in our own studies and hopefully provide information useful throughout the renal physiology community.

**ACKNOWLEDGMENTS**

Present address for M.-J. Yu: Institute of Biochemistry and Molecular Biology, National Taiwan University College of Medicine, Taipei, Taiwan.
Fig. 5. Phosphoproteomic database for proximal tubule and other renal cortical membrane proteins. The page provides a sortable table of phosphorylation sites detected along with the accession number, official gene symbol, protein name, phosphopeptide sequence detected, and site assignment information. In addition the user is able to download all the data as a spreadsheet or a PDF file.

National Heart, Lung, and Blood Institute (NHLBI; project Z01-HL001285, M. A. Knepper).

DISCLOSURES
No conflicts of interest, financial or otherwise, are declared by the authors.

AUTHOR CONTRIBUTIONS

REFERENCES

AJP-Renal Physiol • doi:10.1152/ajprenal.00457.2011 • www.ajprenal.org


