The human (PEDB) and mouse (mPEDB) Prostate Expression Databases

Peter S. Nelson¹,²*, Colin Pritchard¹, Denise Abbott¹ and Nigel Clegg¹

¹Division of Human Biology and ²Division of Clinical Research, Fred Hutchinson Cancer Research Center, 1100 Fairview Avenue North, Seattle, WA 98109-1024, USA

Received September 17, 2001; Accepted September 18, 2001

ABSTRACT

The Prostate Expression Databases (PEDB and mPEDB) are online resources designed to allow researchers to access and analyze gene expression information derived from the human and murine prostate, respectively. Human PEDB archives more than 84 000 Expressed Sequence Tags (ESTs) from 38 prostate cDNA libraries in a curated relational database that provides detailed library information including tissue source, library construction methods, sequence diversity and sequence abundance. The differential expression of each EST species can be viewed across all libraries using a Virtual Expression Analysis Tool (VEAT), a graphical user interface written in Java for intra- and inter-library sequence comparisons. Recent enhancements to PEDB include (i) the development of a murine prostate expression database, mPEDB, that complements the human gene expression information in PEDB, (ii) the assembly of a non-redundant sequence set or ‘prostate unigene’ that represents the diversity of gene expression in the prostate, and (iii) an expanded search tool that supports both text-based and BLAST queries. PEDB and mPEDB are accessible via the World Wide Web at http://www.pedb.org and http://www.mpedb.org.

INTRODUCTION

Diseases of the prostate are among the most common pathologies to afflict aging men. Prostate carcinoma is the most frequently diagnosed non-cutaneous malignancy in the US with more than 180 000 new cases estimated for 2001 (1). In order to characterize molecular alterations that accompany prostate disease processes and provide resources for virtual and physical analyses, we have developed the Prostate Expression Database (PEDB) (2). PEDB serves as a centralized collection of gene expression information derived from the human prostate that is organized in a fashion suitable for sequence-based queries, assessment of gene expression diversity, and comparative expression analyses. Expressed Sequence Tags (ESTs) and full-length cDNA sequences derived from 38 human prostate cDNA libraries are archived and represent gene expression profiles reflecting a wide spectrum of normal, benign and malignant prostate disease states. Detailed library information including tissue source, library construction methods, sequence diversity and sequence abundance are maintained in a relational database management system (RDBMS). Prostate ESTs are assembled into distinct species groups using the sequence assembly program Phrap, and annotated with information from the GenBank, dbEST and Unigene public sequence databases.

In recognition of the emerging uses of the mouse as a model system for the study of normal and pathological prostate development, we have developed a database complementary to PEDB that serves to archive and analyze murine prostate gene expression information. The mouse Prostate Expression Database (mPEDB) currently comprises >6000 ESTs from five mouse prostate cDNA libraries constructed from distinct developmental stages and anatomical locations. A detailed description of the database development, data inventory and utilities is available online: www.pedb.org/OVERVIEW/ (Table 1).

*To whom correspondence should be addressed at: Division of Human Biology, Fred Hutchinson Cancer Research Center, 1100 Fairview Avenue North, PO Box 19024, Seattle, WA 98109-1024, USA. Tel: +1 206 667 3377; Fax: +1 206 667 2917; Email: pnelson@fhcrc.org

Table 1. Table of contents for PEDB overview (http://www.pedb.org/OVERVIEW)

1. Introduction to PEDB
2. PEDB Information
   2.a. Construction
   2.b. Dataflow
   2.c. Build Process
   2.d. Current PEDB Build
3. PEDB Utilities
   3.a. BLAST Queries
   3.b. Search Engine
   3.c. Prostate Unigene
   3.d. Virtual Analysis Expression Tool (VEAT)
   3.e. Prostate Transcriptome
   3.f. Prostate Proteome
4. mPEDB
5. PEDB References and Resources
PEDB DATA AND ANALYSIS TOOLS

PEDB consists of archives of ESTs derived from 38 human prostate cDNA libraries. These ESTs are obtained from public sequence repositories such as GenBank (3), the database of ESTs (dbEST) (4), the Cancer Genome Anatomy Project (CGAP) (5), The Institute for Genome Research (TIGR) or from in-house EST sequencing projects. Sequence processing and curation involves a pipeline of sequence submission, sequence masking, sequence assembly and assembly annotation that now incorporates quality-based assemblies using Phred and Phrap base-calling and sequence assembly algorithms (6,7) (www.pedb.org/OVERVIEW/). Assembled consensus sequences are used for BLAST queries against the Unigene, GenBank and dbEST databases to provide cluster annotation and to further facilitate the assembly process.

The most recent build of PEDB ESTs was assembled starting with 84 832 prostate ESTs. Portions of EST sequences with homology to cloning vectors, *Escherichia coli* genomic DNA and human repetitive DNA sequences were masked. Sequences annotating to the mitochondrial genome were removed and the remaining ESTs with >300 bp of high quality sequence were admitted to the assembly process. A total of 68 426 high-quality ESTs were assembled using Phrap and Phrap base-calling and sequence assembly algorithms (6,7) (www.pedb.org/OVERVIEW/). Assembled consensus sequences are used for BLAST queries against the Unigene, GenBank and dbEST databases to provide cluster annotation and to further facilitate the assembly process.

The most recent build of PEDB ESTs was assembled starting with 84 832 prostate ESTs. Portions of EST sequences with homology to cloning vectors, *Escherichia coli* genomic DNA and human repetitive DNA sequences were masked. Sequences annotating to the mitochondrial genome were removed and the remaining ESTs with >300 bp of high quality sequence were admitted to the assembly process. A total of 68 426 high-quality ESTs were assembled using Phrap and Phrap base-calling and sequence assembly algorithms (6,7) (www.pedb.org/OVERVIEW/). Assembled consensus sequences are used for BLAST queries against the Unigene, GenBank and dbEST databases to provide cluster annotation and to further facilitate the assembly process.

The most recent build of PEDB ESTs was assembled starting with 84 832 prostate ESTs. Portions of EST sequences with homology to cloning vectors, *Escherichia coli* genomic DNA and human repetitive DNA sequences were masked. Sequences annotating to the mitochondrial genome were removed and the remaining ESTs with >300 bp of high quality sequence were admitted to the assembly process. A total of 68 426 high-quality ESTs were assembled using Phrap and Phrap base-calling and sequence assembly algorithms (6,7) (www.pedb.org/OVERVIEW/). Assembled consensus sequences are used for BLAST queries against the Unigene, GenBank and dbEST databases to provide cluster annotation and to further facilitate the assembly process.

The most recent build of PEDB ESTs was assembled starting with 84 832 prostate ESTs. Portions of EST sequences with homology to cloning vectors, *Escherichia coli* genomic DNA and human repetitive DNA sequences were masked. Sequences annotating to the mitochondrial genome were removed and the remaining ESTs with >300 bp of high quality sequence were admitted to the assembly process. A total of 68 426 high-quality ESTs were assembled using Phrap and Phrap base-calling and sequence assembly algorithms (6,7) (www.pedb.org/OVERVIEW/). Assembled consensus sequences are used for BLAST queries against the Unigene, GenBank and dbEST databases to provide cluster annotation and to further facilitate the assembly process.

The most recent build of PEDB ESTs was assembled starting with 84 832 prostate ESTs. Portions of EST sequences with homology to cloning vectors, *Escherichia coli* genomic DNA and human repetitive DNA sequences were masked. Sequences annotating to the mitochondrial genome were removed and the remaining ESTs with >300 bp of high quality sequence were admitted to the assembly process. A total of 68 426 high-quality ESTs were assembled using Phrap and Phrap base-calling and sequence assembly algorithms (6,7) (www.pedb.org/OVERVIEW/). Assembled consensus sequences are used for BLAST queries against the Unigene, GenBank and dbEST databases to provide cluster annotation and to further facilitate the assembly process.
SUMMARY AND FUTURE DEVELOPMENTS

The human and mouse Prostate Expression Databases serve as centralized archives of gene expression information derived from the human and murine prostate that can be utilized by investigators studying normal and neoplastic prostate development. The assembled human prostate transcriptome currently comprises 20,187 distinct transcripts. Ongoing work involves the characterization of additional cDNA libraries representing specific prostate cell types and early developmental stages, the virtual comparative analyses of human and mouse prostate gene expression, and a database extension for archiving and analyzing cDNA microarray data derived from PEDB and mPEDB sequence resources. PEDB is accessible via the World Wide Web at http://www.pedb.org. mPEDB is accessible at http://www.mpedb.org or through a link from PEDB.

ACKNOWLEDGEMENTS

We thank Alec Mooreman, Stacy Boness and Josh Epstein for sequencing assistance. We thank Burak Eroglu for database administration. This work was supported by the CaPCURE Foundation and grants to P.S.N. from the NCI (K08 CA75173-01A1), NIDDK (RO1 DK59125-02) and the USAMRMC Program in Prostate Cancer Research (DAMD17-98-1-8499).

REFERENCES