GigaScience

PseudoFuN: Deriving functional potentials of pseudogenes from integrative relationships with genes and miRNAs across 32 cancers

--Manuscript Draft--

Manuscript Number:	GIGA-D-18-00369
Full Title:	PseudoFuN: Deriving functional potentials of pseudogenes from integrative relationships with genes and miRNAs across 32 cancers
Article Type:	Technical Note
Funding Information:	U.S. National Library of Medicine (US) (4T15LM011270-05) — Mr. Travis Johnson Ohio State Univerisy (Startup Funds) — Dr. Yan Zhang
Abstract:	Background: Long thought “relics” of evolution, not until recently have pseudogenes been of medical interest regarding regulation in cancer. Often, these regulatory roles are a direct byproduct of their close sequence homology to protein coding genes. Novel pseudogene-gene functional associations can be identified through the integration of biomedical data, such as sequence homology, functional pathways, gene expression, pseudogene expression, and miRNA expression. However, not all of the information has been integrated, and the vast majority of previous pseudogene studies relied on 1:1 pseudogene-parent gene relationships without leveraging other homologous genes/pseudogenes. Results: We produce pseudogene-gene (PGG) families that expand beyond the current 1:1 paradigm. Firstly, we construct expansive PGG databases by i) CUDAlign GPU accelerated local alignment of all pseudogenes to gene families (totaling 1.6 billion individual local alignments and more than 40,000 GPU hours) and ii) BLAST-based assignment of pseudogenes to gene families. Secondly, we create an open-source web application (PseudoFuN) to search for integrative functional relationships of sequence homology, miRNA expression, gene expression, pseudogene expression, and gene ontology. We produce four “flavors” of databases (>462,000,000 pseudogene-gene pairwise alignments and 133,770 PGG families) that can be queried and downloaded using PseudoFuN. These databases are consistent with previous 1:1 pseudogene-gene annotation and also are much more powerful including millions of de novo pseudogene-gene associations. We find multiple known (e.g., miR20a-PTEN-PTENP1) and novel (e.g., miR375-SOX15- PPP4R1L) miRNA-gene-pseudogene associations in prostate cancer. PseudoFuN provides a “one stop shop” for identifying and visualizing thousands of potential regulatory relationships related to pseudogenes in TCGA cancers. Conclusions: Thousands of new pseudogene-gene associations can be explored in the context of miRNA-gene-pseudogene coexpression and differential expression with a simple-to-use online tool by bioinformaticians and oncologists alike.
Corresponding Author:	Yan Zhang Ohio State University Columbus, Ohio UNITED STATES
Corresponding Author Secondary Information:	
Corresponding Author's Institution:	Ohio State University
Corresponding Author's Secondary Institution:	
First Author:	Travis Johnson
First Author Secondary Information:	
Order of Authors:	Travis Johnson
	Sihong Li
	Eric Franz
	Zhi Huang

	Shuyu Dan Li
	Moray J Campbell
	Kun Huang
	Yan Zhang
Order of Authors Secondary Information:	
Additional Information:	
Question	**Response**
Are you submitting this manuscript to a special series or article collection?	No
Experimental design and statistics Full details of the experimental design and statistical methods used should be given in the Methods section, as detailed in our Minimum Standards Reporting Checklist. Information essential to interpreting the data presented should be made available in the figure legends. Have you included all the information requested in your manuscript?	Yes
Resources A description of all resources used, including antibodies, cell lines, animals and software tools, with enough information to allow them to be uniquely identified, should be included in the Methods section. Authors are strongly encouraged to cite Research Resource Identifiers (RRIDs) for antibodies, model organisms and tools, where possible. Have you included the information requested as detailed in our Minimum Standards Reporting Checklist?	Yes
Availability of data and materials All datasets and code on which the conclusions of the paper rely must be either included in your submission or	Yes

deposited in publicly available repositories (where available and ethically appropriate), referencing such data using a unique identifier in the references and in the “Availability of Data and Materials” section of your manuscript. Have you have met the above requirement as detailed in our Minimum Standards Reporting Checklist?	

PseudoFuN: Deriving functional potentials of pseudogenes from integrative relationships with genes and miRNAs across 32 cancers

Travis S Johnson[1], Sihong Li[1], Eric Franz[2], Zhi Huang[3,4], Shuyu Dan Li[5], Moray J Campbell[6], Kun Huang[4,7], Yan Zhang[1]*

[1] Department of Biomedical Informatics, College of Medicine, The Ohio State University, Columbus, OH 43210, USA

[2] Ohio Supercomputer Center (OSC), Columbus, OH 43212, USA

[3] Department of Electrical and Computer Engineering, Purdue University, West Lafayette, IN 47907, USA

[4] School of Medicine, Indiana University, Indianapolis, IN 46202, USA

[5] Department of Genetics and Genomic Sciences, Icahn School of Medicine at Mount Sinai, New York, NY 10029, USA

[6] Division of Pharmaceutics and Pharmaceutical Chemistry, College of Pharmacy, The Ohio State University, Columbus, OH 43210, USA

[7] School of Informatics and Computing, Indiana University, Indianapolis, IN 46262, USA

* Correspondence: yan.zhang@osumc.edu

Abstract

Background: Long thought "relics" of evolution, not until recently have pseudogenes been of medical interest regarding regulation in cancer. Often, these regulatory roles are a direct byproduct of their close sequence homology to protein coding genes. Novel pseudogene-gene functional associations can be identified through the integration of biomedical data, such as sequence homology, functional pathways, gene expression, pseudogene expression, and miRNA expression. However, not all of the information has been integrated, and the vast majority of previous pseudogene studies relied on 1:1 pseudogene-parent gene relationships without leveraging other homologous genes/pseudogenes. **Results:** We produce pseudogene-gene (PGG) families that expand beyond the current 1:1 paradigm. Firstly, we construct expansive PGG databases by i) CUDAlign GPU accelerated local alignment of all pseudogenes to gene families (totaling 1.6 billion individual local alignments and more than 40,000 GPU hours) and ii) BLAST-based assignment of pseudogenes to gene families. Secondly, we create an open-source web application (PseudoFuN) to search for integrative functional relationships of sequence homology, miRNA expression, gene expression, pseudogene expression, and gene ontology. We produce four “flavors” of databases (>462,000,000 pseudogene-gene pairwise alignments and 133,770 PGG families) that can be queried and downloaded using PseudoFuN. These databases are consistent with previous 1:1 pseudogene-gene annotation and also are much more powerful including millions of *de novo* pseudogene-gene associations. We find multiple known (e.g., miR20a-PTEN-PTENP1) and novel (e.g., miR375-SOX15-PPP4R1L) miRNA-gene-pseudogene associations in prostate cancer. PseudoFuN provides a “one stop shop” for identifying and visualizing thousands of potential regulatory relationships related to pseudogenes in TCGA cancers. **Conclusions:** Thousands of new pseudogene-gene associations can be explored in the context of miRNA-gene-pseudogene coexpression and differential expression with a simple-to-use online tool by bioinformaticians and oncologists alike.

Keywords: Pseudogenes, database, functional prediction, gene regulation, network analysis, high performance computing, graphics processing unit, competing endogenous RNA

Background

Pseudogenes were previously considered unimportant relics of evolution that played an unclear role in biological processes[1]. However, more pseudogenes have been discovered to be involved in gene regulation[2-4]. These regulatory relationships between pseudogenes and genes have increasingly been explored, such as the transcriptional regulation of PTEN by pseudogene PTENP1 in several cancer conditions[5]. PTEN acts as a tumor suppressor gene, which is underexpressed in gastric cancer. However by overexpressing PTENP1 in gastric cancer, both PTEN underexpression and cell proliferation are mitigated via the regulatory relationship between PTEN and PTENP1[6]. Relationships between these pseudogenes and their parent genes have been found to play critical roles indicating functional potentials of these pseudogenes[7,8]. This point can most clearly be seen in the importance of sequence homology between pseudogenes and coding genes plays in competing endogenous RNA (ceRNA) networks[9,10]. In ceRNA networks the pseudogenes act as decoy targets for the miRNAs targeting a protein-coding gene. In short, researchers have made huge strides in understanding pseudogenes from genomic variation to functional potentials[11,12], and from “deciphering” the mechanism of ceRNA networks[13] to experimental validation[14].

With this progress, there has been renewed interest in pseudogenes, especially in relation to cancer[15]. This interest has even uncovered biomarkers in human cancer including but not limited to SUMO1P3 upregulation as a diagnostic biomarker in gastric cancer and OCT4-pg4 expression as a prognostic biomarker in hepatocellular carcinoma (HCC)[16-18]. Pseudogene expression has been used to stratify tumor subtypes in 7 distinct cancer types[19]. However, due to the close sequence homology between pseudogenes and their parent genes, identifying the expression profile unique to a pseudogene or highly homologous gene can be challenging. Efforts have been made to address these technical challenges in estimating pseudogene expression using modified alignment and quantification techniques[20]. Perhaps more intriguingly is that pseudogenes can be somatically acquired in cancer development effectively “representing a new class of mutations” that can be either activating or inactivating mutations which function as an “on/off switch”[21,22]. Specific pseudogenes have been implicated in specific cancers. For example, FTH1 regulates tumorgenesis in prostate cancer[23], TP73-AS1 regulates proliferation in esophageal squamous cell carcinoma[24], and NKAPP1/MSTO2P/RPLP0P2 is associated with poor prognosis in lung adenocarcinoma[25].

For these reasons, having a complete understanding of these pseudogene-gene relationships is important. While studying these relationships, a common conception is to only consider the pseudogenes in relation to their parent genes with highest homology[7-9,26]. There have also been pioneer studies probing pseudogene functions through aligning them to parent proteins (corresponding to the parent genes) and then to parent protein domains[7,27,28].

The conventional idea of single parent genes may not be comprehensive enough to model the complex phylogenetic relationships involving multiple genes and pseudogenes in a homolog family. While pseudogenes diverged from their parent genes distantly in the past, only the daughter protein-coding genes other than the original parent gene may now exist. The result is that aligning to the true phylogenetic parent gene itself may not be possible. For this reason, we advocate the use of homologous gene families rather than single parent genes to compare against pseudogenes. By viewing the homologies as a weighted network instead of a single scalar value we believe that new relationships can be uncovered.

We build the pseudogene-gene (PGG) family databases using two methods: i) CUDAlign[29] based-local alignment of all pseudogenes to gene families (totaling 1.6 billion individual local alignments and more than 40,000 GPU hours). By aligning all pseudogenes to all gene families (CUDAlign), we can study underlying sequence homology and more easily set cutoffs to assign pseudogenes to gene families. ii) BLAST [30]-based assignment of pseudogenes to gene families. This provides a fast heuristic search option. BLAST derivative methods have been commonly used to find parent genes in previous pseudogene studies[31,32]. Using these two methods we show that these pseudogenes are usually assigned to the gene family of their parent genes but are often not exclusively so. Besides, most pseudogenes can be categorized into processed pseudogenes and unprocessed pseudogenes depending on whether they came from retrotranscription of mRNAs[11,33,34]. We take these differences into account using both of our methods (CUDAlign and BLAST).

Furthermore, we make these data publically downloadable from GitHub[35]. We also create an R Shiny web application called PseudoFuN[36] that supports querying the PGG databases, interactive visualization and functional analysis of the PGG networks, and visualization of pseudogene-gene co-expression and miRNA binding using The Cancer Genome Atlas and GTEx (Genotype-Tissue Expression) project derived public data[20,37,38]. Besides, we provide another interactive web app hosted by the Ohio Supercomputer Center[39] (OSC), which supports querying novel sequences against any of our PGG databases and visualization of the resulting PGG networks.

The PGG databases can be used to study pseudogene-gene-miRNA co-expression indicative of ceRNA networks across the entire Cancer Genome Atlas. With these diverse tools provided by PseudoFuN, it is possible to generate hypotheses regarding i) the regulatory roles of pseudogenes across tumor and normal tissue, ii) pseudogene-gene relationships through *de novo* reassignment of pseudogenes to gene families and iii) functional annotation of pseudogenes. We expect these databases and tools to have more use in cancer studies.

Methods

Construction of Pseudogene-Gene (PGG) Database

To generate these gene families, we use two methods: i) CUDAlign-based local alignment of pseudogenes against consensus sequences representing gene families, and ii) BLAST-based search of pseudogene sequences against all gene sequences (Figure 1). These two approaches can be thought of as heuristic but different processes. The local sequence alignment approach is heuristic in that only two gene sequences are used from each gene family to reduce the search space. These sequences are the most similar and representative sequences to all the other gene sequences in the family. The BLAST-based approach is heuristic in that not all sequences are fully aligned during the process due to the seed-and-extend steps of BLAST[40]. The result is that not every relationship between pseudogene and gene family is recorded which is an advantage in runtime but a disadvantage in studying underlying sequence homology.

i) CUDAlign-based local alignment of gene families

Gene homolog families were generated using the Ensembl biomart gene homolog database[41,42]. The pairs of homologous genes were separated into connected components using python networkx package[43]. These connected component sub-graphs are considered gene families in this study. To reduce the number of alignments that needed to be performed, we selected consensus genes from each family that would be used to represent the entire family.

The consensus sequences were selected by aligning every member of the gene family to every other member using local alignment with CUDAlign[29]. The two members of the family with the largest sum alignment scores across all other family members were selected as the consensus sequences to increase the number of candidate sequences. If only one member existed in the family, then that member was the consensus sequence. Using the list of these consensus sequences we then aligned every consensus sequences to every pseudogene in the human genome GRCh38 annotated by GENCODE Release 25[44].

Specifically the pseudogenes are split up into processed, unprocessed and other (unclear whether processed or unprocessed), based on their mechanisms of formation[45]. We performed different alignment procedures for processed and unprocessed pseudogenes respectively. The processed pseudogenes were aligned to all of the consensus gene transcripts with the highest local alignment score recorded. The unprocessed pseudogenes were aligned to the full genomic sequences of each of the consensus genes with the highest local alignment score recorded. Theoretically unprocessed pseudogenes can align to both exonic and intronic regions of DNA, while processed pseudogene can only align to exonic regions. In our previous database we did not perform this two-procedure strategy in part to reduce the runtime of the problem[46]. These changes make the database much more complete and biologically relevant. The other pseudogenes were aligned to both the transcripts and the genomic sequence recording the highest score.

These scores, one for each combination of pseudogene to gene family, were stored for further analysis. Pseudogenes were assigned to families using a cutoff score (i.e., percentiles of the alignment scores per PGG alignment matrix) and a maximum number of assignments (i.e., the top four alignments above a cutoff). If greater than top four alignments were used, the PGG families were too large to calculate the pairwise alignment matrix. The resulting sets of pseudogenes and genes are called pseudogene-gene (PGG) families. This method was used to allow a pseudogene to be assigned multiple families as well as prevent pseudogenes from being assigned families if their alignment score was low. We used the 99th percentile cutoff (corresponding alignment score 54), 99.9th percentile cutoff (135), and the 99.99th percentile cutoff (198) to generate three resultant databases named CUDAlign54, CUDAlign135, and

CUDAlign198 respectively. All these flavors of databases are available for search in our web apps.

ii) BLAST-based generation of PGG families

In contrast to the local alignment of every combination of pseudogene to gene family, PGG families were also created by assigning the pseudogenes to the family containing its closest BLAST search match. This approach was used to contrast with the CUDAlign method, which uses up to the top 4 matches. The pseudogenes were separated into processed, unprocessed and other. Then, all genes in the GENCODE Release 25 annotation were used to generate genomic, transcript, and combined BLAST databases (blastdb). The processed pseudogenes would be blasted against transcript blastdb, unprocessed against the genomic sequence blastdb, and the rest pseudogenes were blasted against the combined genomic/transcript blastdb. The pseudogene was assigned to the gene family containing the best match from the BLAST search.

Comparison between PGG families and pseudogene-parent gene pairs

We also conduct a comparison to the Pseudogene.org resource[47]. In this comparison, we consider pseudogenes and parent gene pairs from pseudogene.org psiDr[31] database (old)[48] and on GENCODE Release 10 from pseudogene.org psiCube[11] database (new)[49]. From our databases, we consider every combination of pseudogene to gene within a PGG family as a pair (for example, a family with 3 genes and 2 pseudogenes would have $C_2^3 = 6$ pairs). Since we have multiple flavors of PGG databases including the BLAST-based version and the CUDAlign-based versions, we compare the intersections between two Pseudogene.org versions and our BLAST/CUDAlign-based versions. We show the intersections of pseudogene-gene pairs in Venn Diagrams.

Development of PseudoFuN web applications

Aside from generating different flavors of the PGG databases, we assemble them into an online R Shiny application called PseudoFuN[36] which supports gene and pseudogene symbol queries against out PGG databases, generates dynamic networks, produces Gene Ontology[50] (GO) tables and additional functional analysis features (Table 1). The functionalities, such as calculating the gene co-expression for any resultant PGG network in any of the TCGA[51] cancers types, are important for ceRNA network hypothesis generation in human cancers. For more information, please visit the PseudoFuN website and follow the README and tutorial.

Additionally we create another web app hosted by the Ohio Supercomputer Center (OSC) OnDemand[52] platform. This application has multiple functionalities including the query of Ensembl gene ID or a novel sequence against one selected flavor of our databases. For each of these features we provide a simple-to-use interface that allows users to select which database to query, allows download of the query hits, and allows users to interactively explore the PGG family networks including GO information.

Use cases in multiple cancers

Furthermore three use cases are provided to show the potential utility of PseudoFuN to researchers and oncologists looking for functional relationships between pseudogenes, genes, and miRNAs. Use Case I validates known pseudogene-gene functional relationships. Use Case II identifies high confidence novel miRNA-pseudogene-gene relationships. Use Case III is primarily focused on agreement with a validation study. We focused on pseudogenes/genes that were differentially expressed in low RARG/low TACC1/high miR-96 compared to the reverse in prostate cancer cell lines and also differentially expressed in our PGG networks in TCGA prostate cancer samples.

Results

Local alignment of gene families

We performed 1.6 billion local alignments between all pseudogenes and all gene family consensus sequences. The process required over 40,000 GPU hours on the Oakley cluster at the OSC. The highest scores for each gene family and pseudogene were stored in a 17,273x26,754 matrix of pseudogene-to-gene-family alignment scores (~462 million elements). From this matrix, we are able to explore global pseudogene-gene family homology relationships and assign pseudogenes to one or more gene families with high sequence homology.

As one might expect, the number of pseudogenes with high alignments (defined as above a percentile threshold) to many gene families is relatively low. It can be seen that the majority of pseudogenes will align to one gene family in the CUDAlign databases (Figure 2). Another feature of note is that there are some pseudogenes that align to many gene families (e.g., 9 pseudogenes have alignment scores above 54 in 15,000 gene families and 571 pseudogenes have alignment scores above 54 in 1,000 gene families). In contrast to previous belief in single gene-pseudogene homology, some pseudogenes are related to many genes. It is worth considering that these high homology pseudogenes (e.g., FTLP10 with 3,006 gene family pairwise alignments over a 54 threshold) may have a role in regulating major biological processes[53] and disease[54].

BLAST generation of PGG families

The BLAST generated database was larger than the CUDAlign generated databases with 68,578 total connections. This database was also much simpler to compute with since it was not an exhaustive search. These conclusions make it a simple method to quickly estimate the pseudogene-to-gene relationships.

Direct comparison to pseudogene parents

We compare our databases to the previous pseudogene-parent gene databases retrieved from Pseudogene.org resources (Figure 3). It shows that our methods reconstruct most of the pseudogene-parent-gene relationships identified by Pseudogene.org. The overall consistency of our databases (BLAST and CUDAlign) with both Pseudogenes.org databases (new and old) was 75% (i.e., all our databases combined). Individually, the BLAST-based database contained 61% of the Pseudogene.org relationships (both new and old) and the CUDAlign 54 cutoff contained 60% of the Pseudogene.org relationships (both new and old). Our databases also generate a larger pool of possible interactions.

Development of a pseudogene query tool

The R Shiny application is a comprehensive hypothesis generating tool that is freely available on the internet[36]. This tool provides a wide array of functionality that a researcher can access quickly and download results as the raw data for more in-depth analysis. These features are outlined in detail in Table 1.

Use Cases: Assisting functional study of ceRNA networks in cancer

To illustrate the utility of our databases and tools we present three use cases.

Use Case I: To validate known pseudogene-gene relationships we query pseudogenes or genes of interest individually, e.g., PTENP1, or KRASP1, FTH1P1, GBP1P1. We query a gene/pseudogene name one at a time, PseudoFuN will return the top PGG network(s) that contain the query (Figure 4). PTENP1 is a processed pseudogene homologous to PTEN, a tumor suppressor gene. PTENP1 is selectively lost in cancer and may regulate PTEN expression as a miRNA decoy target[5,6]. We have observed differential co-expression patterns of

PGG families in tumor vs. normal for PTENP1 network in multiple cancers including prostate cancer (Supplementary Figure 2B,C). We identified known miRNAs (hsa-miR-20a in prostate cancer[55]) targeting PTEN PGG network nodes providing insights into ceRNA regulation (Supplementary Figure 2D). These insights are important since some pseudogenes competitively bind to miRNAs thus regulate gene expression. We also identify hsa-miR103a-3p as potentially targeting both PTEN and PTENP1 (Supplementary Figure 2D). The ceRNA network regulatory relationship is governed by effect modulation of miRNA on gene expression by pseudogene expression (Supplementary Figure 1A,C,E). This leads to a correlation between pseudogene (miRNA decoy targets) and gene (miRNA targets) expression (Supplementary Figure 1D). That means both these pseudogenes and homologous genes competitively bind to miRNAs. KRAS-KRASP1 regulatory network was also identified by our database (Figure 4). KRAS and KRASP1 are known to be involved in ceRNA network regualtion[5,10,55]. PseudoFuN query of KRAS identified co-expression patterns in prostate cancer consistent with ceRNA network regulation by hsa-miR-145, a known modulator of KRAS in prostate cancer[56]. The FTH1 query also resulted in the identification of pseudogenes (FTH1P2, FTH1P8, FTH1P11, FTH1P16) that regulate FTH1 in prostate cancer[23] as well novel miRNAs that may be involved in ceRNA network regulation of FTH1 in prostate cancer. GBP1 is an IFN-α induced transcript that is involved in immune response in prostate cancer[57]. The GBP1 involved PGG network also contained the pseudogene GBP1P1 which may have a ceRNA regulatory role in breast cancer[58] and in some neurodegenerative diseases[59].

Use Case II: We wanted to identify possible gene-miRNA relationships of interest within our database. We chose to study these relationships with respect to miR-96, a known cancer regulator microRNA in prostate cancer[60]. Through differential expression analysis between tumors in the TCGA-PRAD cohort with lower expression of RARG and TACC1 (also a miR-96 target) and high expression miR-96 (low RARG/low TACC1/high miR-96), compared to the reverse, we previously identified altered SOX15 gene expression is significantly associated with worse disease free survival. We visualized expression patterns of SOX15 PGG families, and corresponding miRNA associations. miR-96 is included as a validation.

Interestingly we identified the pseudogene PPP4R1L as a potential member of a SOX15 ceRNA network (Figure 5A). PPP4R1L and SOX15 are both significantly differentially expressed between tumor and normal controls (Bonferroni corrected p-value = 3.42×10^{-7}, 2.01×10^{-14} respectively, Figure 5E). PPP4R1L and SOX15 are significantly co-expressed (Pearson correlation coefficient (PCC)=0.51, p-value<2.2×10^{-16}) in tumor tissue but much less correlated in normal controls in prostate cancer (PCC=0.24, p-value=0.09, Figure 5B,C). Positively correlated expression is an assumption when determining ceRNA network relationships[61] (Supplementary Figure 1). Both SOX15 and PPP4R1L are likely regulated by hsa-miR-375 based on the TCGA prostate cancer dataset. hsa-miR-375 is associated with docetaxel resistance in prostate cancer[62,63] and PPP4R1L knock-down in HeLa cells induces taxol resistance[64]. These findings are intriguing since taxol and docetaxel are closely related chemical compounds. PPP4R1L is also located in a region associated with high mutation rates in cancer cell lines[64] which could be indicative of mutational “on/off switches” in pseudogene regulation.

Use Case III: We were most interested in the deferentially expressed (DE) genes (and related pseudogenes) that both appeared in our PGG database and were contained in networks with genes differentially expressed in low RARG/low TACC1/high miR-96 compared to vice versa. We searched the DE genes in our PGG database, and identified the top networks with enriched number of DE genes. As a result, parent genes HTR7, CNN2, MSN and TAGLN2 are differentially expressed; they generate pseudogenes, which are specifically expressed in prostate cancer samples[16]. These four parent genes are also detected in our 5 top PGG families involving miR-96 regulated (direct or indirect) DE genes. We identified HTR7P1 pseudogene in the same PGG family as HTR7 gene, which is potentially regulated by hsa-miR-607 and has-miR-3654 in the TCGA prostate cancer dataset (Supplementary Figure 3). 11 CNN2 pseudogenes (CNN2P1-CCN2P4, CNN2P6-CNN2P12) were identified in the CNN2 PGG family along with TAGLN2 and TAGLN2P1. TAGLN2P1 is differentially expressed between the tumor and normal samples in the prostate dataset (Supplementary Figure 4, Bonferroni corrected p-value = 6.23×10-4). MSN and MSNP1 were in the same PGG family and hsa-miR-96 potentially regulates MSN in the TCGA prostate cancer dataset (Supplementary Figure 4). In addition, although our DE genes were detected from prostate cancer, we further compared them with DE pseudogenes identified in four other cancer types and we observed interesting results (see Supplementary Materials - *Potential regulatory roles in cancer*).

Discussion

We identify 133,770 PGG families that have significant potential to reveal important information about regulatory pseudogene-gene relationships in health and disease. Within these families we identify both new and existing regulatory networks that contain pseudogenes such as PTENP1, KRAS1P, FTH1P8/11/16, and GBP1P1 (Figure 4). Since all genes and all pseudogenes are included in our database there are thousands of opportunities to identify new regulatory relationships. These thousands of opportunities can be easily stratified using gene name, pseudogene name and cancer type. Our web application makes it a simple and intuitive process to query pseudogenes (or genes) to identify which gene families they may be regulating as well as the functions that are attributed to the members of the network. We also have an application hosted by the OSC that allows the querying of novel sequences against our database.

From these networks, we can also identify possible relationships of differentially expressed pseudogenes in various cancers. For instance, both PPP4R1L pseudogene and SOX15 are differentially expressed in prostate cancer and associated with hsa-miR-375. These types of relationships should be further evaluated along with more complex regulation with multiple miRNAs, pseudogenes, and genes. It is experimentally shown that SOX15 is regulated by hsa-miR-96[60]. It may be important to include hsa-miR-96 in the hsa-miR-375-SOX15-PPP4R1L potential ceRNA network. Aside from PGG family specific differential pseudogene expression, the PseudoFuN app allows for comprehensive differential pseudogene expression (DPgE) analysis in any of the TCGA cancer datasets.

The use of this database also has utility in integrative analysis where the databases can be used as a mask for other data modalities. Some examples would be using the nodes (genes and pseudogenes) in each of the PGG families as groups in gene expression experiments. Similarly, these groups could be used for feature reduction when visualizing data. We hope researchers can use these relationships we have identified to reduce large numbers of candidate associations down to numbers that can be easily validated and generate new candidates when querying novel sequences. For instance, miRNA-gene pairs filtered through the sets of PGG families would identify high priority ceRNA candidates.

Conclusions

We generate multiple large databases of pseudogene gene family relationships and the tools to study them for use by biomedical researchers. These databases are more comprehensive than previous pseudogene-gene databases by including many more homology relationships in PGG families, thus more powerful for experiment validation and knowledge discovery. These databases are useful in identifying pseudogene-gene regulatory relationships in 32 cancer types and show high similarity with known pseudogene-gene relationships. Aside from the known relationships we identify many unknown relationships. Furthermore, these databases and associated analyses can be easily accessed online or through the OSC OnDemand platform, allowing for novel hypotheses to be assessed quickly by biomedical researchers. We find evidence of both known regulatory pseudogene-gene relationships and novel hypothesized relationships that we plan to validate. PseudoFuN is a comprehensive, dynamic tool that allows any bioinformatician or oncologist to find novel regulatory pseudogenes within their cancer or gene of interest.

Availability of Supporting Data

We have made the PGG family data publically downloadable from GitHub[35]. We also created an R Shiny web application called PseudoFuN[36] that supports querying the PGG databases, interactive visualization and functional analysis of the PGG networks, and visualization of pseudogene-gene co-expression and miRNA binding. Besides, we provide another interactive web app hosted on Ohio Supercomputer Center (OSC) OnDemand, which supports querying novel sequences against any of our PGG databases and visualization of the resulting PGG networks.

Additional Files

There is an additional Supplementary Materials file containing additional information on the data and additional analyses. It includes the following figures and tables:

Supplementary Figure 1. Example of ceRNA network regulation of gene expression. A) A graphical view of how pseudogene expression can regulate gene expression. B) A cellular view of ceRNA network regulation. C) Equations used to model the correlation between gene and pseudogene expression in a ceRNA network. D) The distribution of the gene-pseudogene correlations based on the models in C. E) The effect that pseudogene expression has on the miRNA induced change in gene expression.

Supplementary Figure 2. PseudoFuN online output for PTEN PGG family. A) Interactive graph visualization of the PTEN PGG network. B) TCGA prostate co-expression matrix for

PTEN PGG family genes and pseudogenes across normal samples. C) TCGA prostate co-expression matrix for PTEN PGG family genes and pseudogenes across tumor samples. D) Negatively correlated miRNAs for all members of the PTEN PGG family. E) Differential gene and pseudogene expression for tumor and normal samples for each member of the PTEN PGG family in the prostate cancer TCGA dataset.

Supplementary Figure 3. PseudoFuN online output for HTR7 PGG family. A) Interactive graph visualization of the HTR7 PGG network. B) TCGA prostate co-expression matrix for HTR7 PGG family genes and pseudogenes across normal samples. C) TCGA prostate co-expression matrix for HTR7 PGG family genes and pseudogenes across tumor samples. D) Negatively correlated miRNAs for all members of the HTR7 PGG family. E) Differential gene and pseudogene expression for tumor and normal samples for each member of the HTR7 PGG family in the prostate cancer TCGA dataset.

Supplementary Figure 4. PseudoFuN online output for CNN2/TAGLN2 PGG family. A) Interactive graph visualization of the CNN2/TAGLN2 PGG network. B) TCGA prostate co-expression matrix for CNN2/TAGLN2 PGG family genes and pseudogenes across normal samples. C) TCGA prostate co-expression matrix for CNN2/TAGLN2 PGG family genes and pseudogenes across tumor samples. D) Negatively correlated miRNAs for all members of the CNN2/TAGLN2 PGG family. E) Differential gene and pseudogene expression for tumor and normal samples for each member of the CNN2/TAGLN2 PGG family in the prostate cancer TCGA dataset.

Supplementary Figure 5. PseudoFuN online output for MSN PGG family. A) Interactive graph visualization of the MSN PGG network. B) TCGA prostate co-expression matrix for MSN PGG family genes and pseudogenes across normal samples. C) TCGA prostate co-expression matrix for MSN PGG family genes and pseudogenes across tumor samples. D) Negatively correlated miRNAs for all members of the MSN PGG family. E) Differential gene and pseudogene expression for tumor and normal samples for each member of the MSN PGG family in the prostate cancer TCGA dataset.

Supplementary Figure 6. The PGG families in our network with the most DE genes after mir-96 treatment. The line weights indicate the sequence homology between members of the PGG family. Red nodes indicate mir96 targets. Yellow nodes with names indicate other genes contained in the PGG family. Yellow nodes without names are pseudogenes contained within the network.

Supplementary Figure 7. The user interface of the OSC OnDemand web application. A) is the main query page where a user can search either sequences or ensemble gene IDs. B) is a representative output of one of the gene searches. This includes an interactive network and the GO information.

Supplementary Figure 8. GBP1P1 DE in TCGA prostate cancer (information retrieved from Han et al.)**.**

Supplementary Table 1. DE parent gene/pseudogenes potentially regulated by miRr-96 in prostate cancer vs. TCGA derived DE pseudogenes.

Abbreviations

PseudoFuN: Pseudogene Functional Networks

PGG: Pseudogene-Gene (i.e., PGG families)

TCGA: The Cancer Genome Atlas

ceRNA: Competing Endogenous RiboNucleic Acid

HCC: HepatoCellular Carcinoma

BLAST: Basic Local Alignment and Search Tool

OSC: Ohio Supercomputer Center

GO: Gene Ontology

DE: Differential Expression

DGE: Differential Gene Expression

DPgE: Differential Pseudogene Expression

Acknowledgments

This work is partially supported by NIH-NLM MIDAS Training Fellowship (4T15LM011270-05) awarded to Travis Johnson and The Ohio State University Startup Funds to Yan Zhang. The authors also thank the Ohio Supercomputer Center (OSC) for providing computing resources.

Author contributions

TSJ, SL, ZH and YZ performed data analyses. TSJ, EF and ZH developed the web applications. YZ and TSJ conceived and initiated this project. YZ and KH supervised the project. MJC provided experimental data. All authors contributed to biological interpretation. TSJ, YZ, MJC and SDL wrote the manuscript. All authors read and approved the manuscript.

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Figure Captions

Figure 1. Workflow for both CUDAlign and BLAST databases. Left side PGG families are produced using the BLAST matches. Right side PGG families are produced using the pseudogene-gene-family alignment matrix with percentile cutoffs using CUDAlign.

Figure 2. The number pseudogenes that align to gene families. The x-axis is the number of gene families which have an alignment score above a specified cutoff (the different colored lines). The y-axis is the number of pseudogenes with an alignment score higher than the cutoff to the number of gene families on the x-axis. The inset grey box is a closer view of the low range gene family numbers (1-10) to show more granular patterns.

Figure 3. Comparison of database members. The top 6 plots are comparisons between the CUDAlign databases using different cutoffs, the BLAST database, and the Pseudogene.org parent genes. The bottom row shows intra-database comparisons, left: Pseudogene.org, middle: CUDAlign databased of different alignment score cutoffs, right: relative size of all databases.

Figure 4. Representative examples of our OSC OnDemand pseudogene query tool. Displayed are the network relationships from our databases for three common ceRNA network examples (queries: FTH1, KRAS, PTEN), and a relationship of interest (GBP1-GBP1P1).

Figure 5. PseudoFuN online output for SOX15 PGG family. A) Interactive graph visualization of the SOX15 PGG network. B) TCGA prostate co-expression matrix for SOX15 PGG family genes and pseudogenes across normal samples. C) TCGA prostate co-expression matrix for SOX15 PGG family genes and pseudogenes across tumor samples. D) Negatively correlated miRNAs for all members of the SOX15 PGG family. E) Differential gene and pseudogene expression for tumor and normal samples for each member of the SOX15 PGG family in the prostate cancer TCGA dataset.

Tables

Table 2. Summary of PseudoFuN features that are freely available at the PseudoFuN website.

PseudoFuN features	Additional description
Interactive visualization of PGG family networks including the query pseudogene/gene	Users can query any single gene or pseudogene symbol, e.g., PTENP1. Nodes are colored by sub-clusters within the network.
Functional enrichment analysis of PGG family	Functional enrichment can be conducted on the genes within the PGG family on Biological Process, Molecular Function or Cellular Components annotations. The GO functional enrichment is calculated with: 1. Fisher's exact test[65] 2. Kolmogorov-Smirnov (KS) Classic[66] 3. Kolmogorov-Smirnov (KS) Elim[66]
Genomic loci mapping of PGG family	The genes in the PGG family can be mapped back to the genome using a circus plot to identify potential loci of interest.
Data download for all of the figures	Users can also download results including: 1. the differential pseudogene expression (DPgE) table for all pseudogenes in the selected cancer 2. the gene and pseudogene expression 3. miRNA correlation table
Links to other gene databases for more information	By directly clicking the node in the network, users can open the GeneCards website[67] for detailed gene information.
Gene/pseudogene co-expression analysis across the entire TCGA	Once a PGG family has been identified the gene/pseudogene co-expression matrix is calculated across one of the 32 available TCGA cancer types.
Tumor vs. normal differential expression of genes/pseudogenes across all TCGA cancer types	The gene/pseudogene differential expression is calculated for all members of the selected PGG family. There is also an option to run differential expression on a specified cancer for all pseudogenes which can be viewed or downloaded as a table.
Predicted miRNA targets involved in the PGG families across all TCGA cancer types	The miRNA targets involved in the selected cancer and PGG family are displayed to show which miRNAs could regulate the PGG family members. This is by using the miRNA correlation tables from

	the TCGA.
Differential Pseudogene Expression (DPgE) Analysis	Differential pseudogene expression is calculated for each of the pseudogenes in TCGA cancers using dreamBase expression information[20]. The online tool allows for manipulation and download of the table.

References

1. Vanin EF: Processed pseudogenes: characteristics and evolution. Annu Rev Genet 19:253-72, 1985
2. Mighell AJ, Smith NR, Robinson PA, et al: Vertebrate pseudogenes. FEBS Lett 468:109-14, 2000
3. Pink RC, Wicks K, Caley DP, et al: Pseudogenes: pseudo-functional or key regulators in health and disease? RNA 17:792-8, 2011
4. Chan JJ, Tay Y: Noncoding RNA:RNA Regulatory Networks in Cancer. Int J Mol Sci 19, 2018
5. Poliseno L, Salmena L, Zhang J, et al: A coding-independent function of gene and pseudogene mRNAs regulates tumour biology. Nature 465:1033-8, 2010
6. Zhang R, Guo Y, Ma Z, et al: Long non-coding RNA PTENP1 functions as a ceRNA to modulate PTEN level by decoying miR-106b and miR-93 in gastric cancer. Oncotarget 8:26079-26089, 2017
7. Lam HY, Khurana E, Fang G, et al: Pseudofam: the pseudogene families database. Nucleic Acids Res 37:D738-43, 2009
8. Zheng D, Gerstein MB: A computational approach for identifying pseudogenes in the ENCODE regions. Genome Biol 7 Suppl 1:S13 1-10, 2006
9. An Y, Furber KL, Ji S: Pseudogenes regulate parental gene expression via ceRNA network. J Cell Mol Med 21:185-192, 2017
10. Poliseno L, Pandolfi PP: PTEN ceRNA networks in human cancer. Methods 77-78:41-50, 2015
11. Sisu C, Pei B, Leng J, et al: Comparative analysis of pseudogenes across three phyla. Proc Natl Acad Sci U S A 111:13361-6, 2014
12. Zhang Y, Li S, Abyzov A, et al: Landscape and variation of novel retroduplications in 26 human populations. PLoS Comput Biol 13:e1005567, 2017
13. Cesana M, Daley GQ: Deciphering the rules of ceRNA networks. Proc Natl Acad Sci U S A 110:7112-3, 2013
14. Chiu HS, Martinez MR, Bansal M, et al: High-throughput validation of ceRNA regulatory networks. BMC Genomics 18:418, 2017
15. Poliseno L, Marranci A, Pandolfi PP: Pseudogenes in Human Cancer. Front Med (Lausanne) 2:68, 2015
16. Kalyana-Sundaram S, Kumar-Sinha C, Shankar S, et al: Expressed pseudogenes in the transcriptional landscape of human cancers. Cell 149:1622-34, 2012
17. Mei D, Song H, Wang K, et al: Up-regulation of SUMO1 pseudogene 3 (SUMO1P3) in gastric cancer and its clinical association. Med Oncol 30:709, 2013
18. Wang L, Guo ZY, Zhang R, et al: Pseudogene OCT4-pg4 functions as a natural micro RNA sponge to regulate OCT4 expression by competing for miR-145 in hepatocellular carcinoma. Carcinogenesis 34:1773-81, 2013
19. Han L, Yuan Y, Zheng S, et al: The Pan-Cancer analysis of pseudogene expression reveals biologically and clinically relevant tumour subtypes. Nat Commun 5:3963, 2014
20. Zheng LL, Zhou KR, Liu S, et al: dreamBase: DNA modification, RNA regulation and protein binding of expressed pseudogenes in human health and disease. Nucleic Acids Res 46:D85-D91, 2018

21. Cooke SL, Shlien A, Marshall J, et al: Processed pseudogenes acquired somatically during cancer development. Nat Commun 5:3644, 2014

22. Shukla R, Upton KR, Munoz-Lopez M, et al: Endogenous retrotransposition activates oncogenic pathways in hepatocellular carcinoma. Cell 153:101-11, 2013

23. Chan JJ, Kwok ZH, Chew XH, et al: A FTH1 gene:pseudogene:microRNA network regulates tumorigenesis in prostate cancer. Nucleic Acids Res 46:1998-2011, 2018

24. Zang W, Wang T, Wang Y, et al: Knockdown of long non-coding RNA TP73-AS1 inhibits cell proliferation and induces apoptosis in esophageal squamous cell carcinoma. Oncotarget 7:19960-74, 2016

25. Wei Y, Chang Z, Wu C, et al: Identification of potential cancer-related pseudogenes in lung adenocarcinoma based on ceRNA hypothesis. Oncotarget 8:59036-59047, 2017

26. Milligan MJ, Lipovich L: Pseudogene-derived lncRNAs: emerging regulators of gene expression. Front Genet 5:476, 2014

27. Bateman A, Birney E, Durbin R, et al: The Pfam protein families database. Nucleic Acids Res 28:263-6, 2000

28. Finn RD, Mistry J, Schuster-Bockler B, et al: Pfam: clans, web tools and services. Nucleic Acids Res 34:D247-51, 2006

29. Chirag Jain SK: Fine-grained GPU parallelization of pairwise local sequence alignment. Presented at the 21st International Conference on High Performance Computing (HiPC, 2014

30. Soroceanu L, Matlaf L, Khan S, et al: Cytomegalovirus Immediate-Early Proteins Promote Stemness Properties in Glioblastoma. Cancer Res 75:3065-76, 2015

31. Pei B, Sisu C, Frankish A, et al: The GENCODE pseudogene resource. Genome Biology 13:R51, 2012

32. Zhang Z, Carriero N, Zheng D, et al: PseudoPipe: an automated pseudogene identification pipeline. Bioinformatics 22:1437-1439, 2006

33. Lynch M, Conery JS: The evolutionary fate and consequences of duplicate genes. Science 290:1151-5, 2000

34. Baertsch R, Diekhans M, Kent WJ, et al: Retrocopy contributions to the evolution of the human genome. BMC Genomics 9:466, 2008

35. Zhang Y: PseudoFuN GitHub. https://github.com/yanzhanglab/PseudoFuN_app, 2018

36. Johnson TS, Li S, Franz E, et al: PseudoFuN. https://integrativeomics.shinyapps.io/pseudofun_app/, 2018

37. Grossman RL, Heath AP, Ferretti V, et al: Toward a Shared Vision for Cancer Genomic Data. N Engl J Med 375:1109-12, 2016

38. Carithers LJ, Moore HM: The Genotype-Tissue Expression (GTEx) Project. Biopreserv Biobank 13:307-8, 2015

39. Center OS: Ohio Supercomputer Center. Columbus OH, Ohio Supercomputer Center, 1987

40. Altschul SF, Gish W, Miller W, et al: Basic local alignment search tool. J Mol Biol 215:403-10, 1990

41. Zerbino DR, Achuthan P, Akanni W, et al: Ensembl 2018. Nucleic Acids Res 46:D754-D761, 2018

42. Ensembl: Ensembl Biomart. ensembl.org/biomart/martview, 2018

43. Hagberg A, Swart P, S Chult D: Exploring network structure, dynamics, and function using NetworkX, Los Alamos National Lab.(LANL), Los Alamos, NM (United States), 2008

44. Harrow J, Frankish A, Gonzalez JM, et al: GENCODE: the reference human genome annotation for The ENCODE Project. Genome Res 22:1760-74, 2012

45. Echols N, Harrison P, Balasubramanian S, et al: Comprehensive analysis of amino acid and nucleotide composition in eukaryotic genomes, comparing genes and pseudogenes. Nucleic Acids Res 30:2515-23, 2002

46. Johnson TS, Li S, Kho JR, et al: Network analysis of pseudogene-gene relationships: from pseudogene evolution to their functional potentials. Pac Symp Biocomput 23:536-547, 2018

47. Karro JE, Yan Y, Zheng D, et al: Pseudogene.org: a comprehensive database and comparison platform for pseudogene annotation. Nucleic Acids Res 35:D55-60, 2007

48. pseudogenes.org: psiDr. pseudogenes.org/psidr/similarity.dat

49. pseudogenes.org: psiCube. http://pseudogene.org/psicube/

50. Ashburner M, Ball CA, Blake JA, et al: Gene Ontology: tool for the unification of biology. Nature genetics 25:25, 2000

51. Cancer Genome Atlas Research N, Weinstein JN, Collisson EA, et al: The Cancer Genome Atlas Pan-Cancer analysis project. Nat Genet 45:1113-20, 2013

52. Hudak D, Johnson D, Chalker A, et al: Open OnDemand: A web-based client portal for HPC centers.

53. Carmona U, Li L, Zhang L, et al: Ferritin light-chain subunits: key elements for the electron transfer across the protein cage. Chem Commun (Camb) 50:15358-61, 2014

54. Wu T, Li Y, Liu B, et al: Expression of Ferritin Light Chain (FTL) Is Elevated in Glioblastoma, and FTL Silencing Inhibits Glioblastoma Cell Proliferation via the GADD45/JNK Pathway. PLoS ONE 11:e0149361, 2016

55. Yang C, Wu D, Gao L, et al: Competing endogenous RNA networks in human cancer: hypothesis, validation, and perspectives. Oncotarget 7:13479-90, 2016

56. Cui SY, Wang R, Chen LB: MicroRNA-145: a potent tumour suppressor that regulates multiple cellular pathways. J Cell Mol Med 18:1913-26, 2014

57. Persano L, Moserle L, Esposito G, et al: Interferon-alpha counteracts the angiogenic switch and reduces tumor cell proliferation in a spontaneous model of prostatic cancer. Carcinogenesis 30:851-60, 2009

58. Welch JD, Baran-Gale J, Perou CM, et al: Pseudogenes transcribed in breast invasive carcinoma show subtype-specific expression and ceRNA potential. BMC Genomics 16:113, 2015

59. Costa V, Esposito R, Aprile M, et al: Non-coding RNA and pseudogenes in neurodegenerative diseases: "The (un)Usual Suspects". Front Genet 3:231, 2012

60. Long MD, Singh PK, Russell JR, et al: The miR-96 and RARgamma signaling axis governs androgen signaling and prostate cancer progression. Oncogene, 2018

61. Xu J, Feng L, Han Z, et al: Extensive ceRNA-ceRNA interaction networks mediated by miRNAs regulate development in multiple rhesus tissues. Nucleic Acids Res 44:9438-9451, 2016

62. Costa-Pinheiro P, Ramalho-Carvalho J, Vieira FQ, et al: MicroRNA-375 plays a dual role in prostate carcinogenesis. Clin Epigenetics 7:42, 2015

63. Wang Y, Lieberman R, Pan J, et al: miR-375 induces docetaxel resistance in prostate cancer by targeting SEC23A and YAP1. Mol Cancer 15:70, 2016

64. MacKeigan JP, Murphy LO, Blenis J: Sensitized RNAi screen of human kinases and phosphatases identifies new regulators of apoptosis and chemoresistance. Nat Cell Biol 7:591-600, 2005

65. F.R.S. RAF: Tests of significance in harmonic analysis. Proceedings of the Royal Society of London. Series A 125:54, 1929

66. Alexa A RJ: Gene set enrichment analysis with topGO. http://www.bioconductor.org, Bioconductor, 2009

67. Stelzer G, Rosen N, Plaschkes I, et al: The GeneCards Suite: From Gene Data Mining to Disease Genome Sequence Analyses. Curr Protoc Bioinformatics 54:1 30 1-1 30 33, 2016

Figure 1

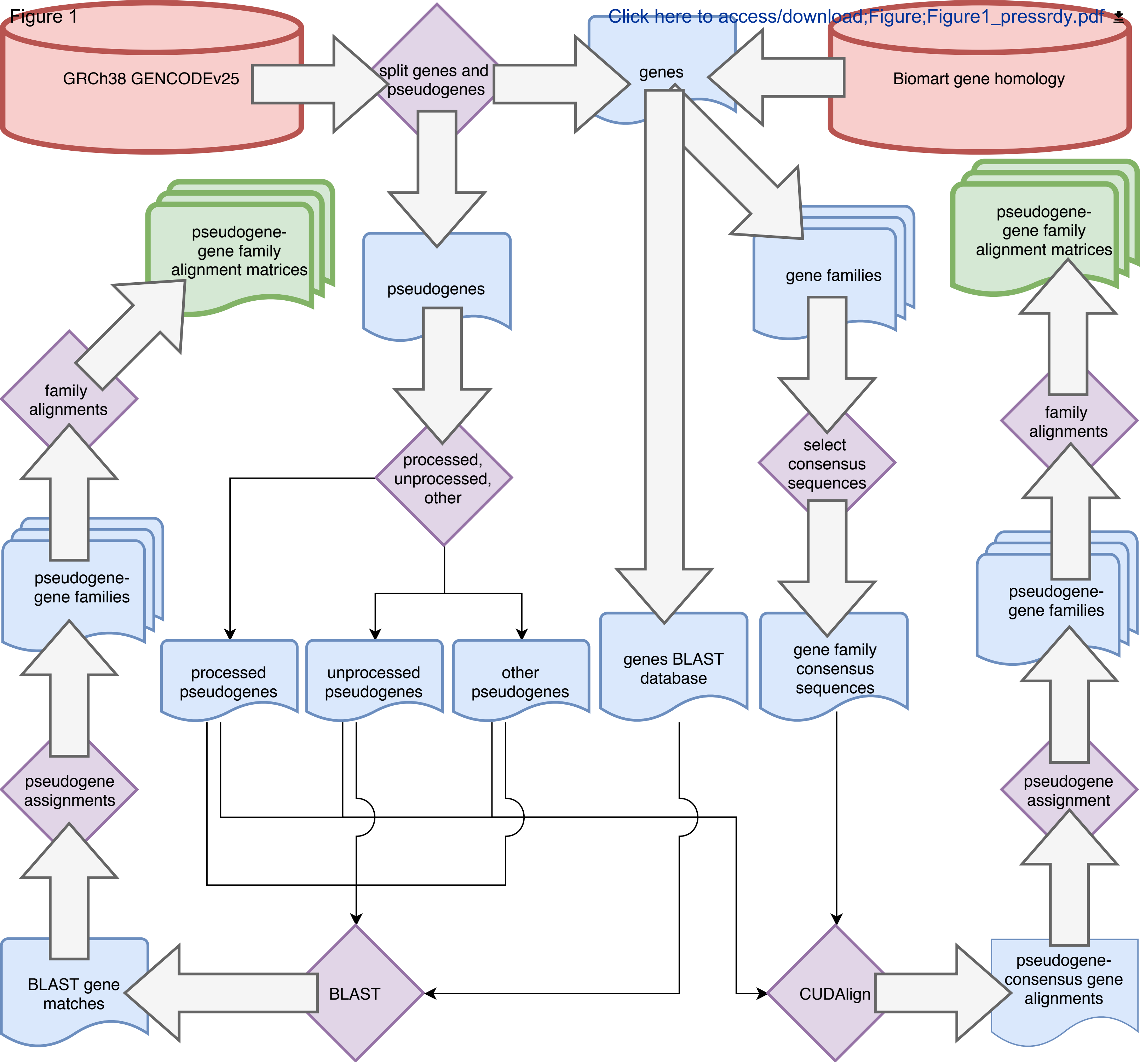

GRCh38 GENCODEv25
split genes and pseudogenes
genes
Biomart gene homology
pseudogene-gene family alignment matrices
pseudogenes
gene families
pseudogene-gene family alignment matrices
family alignments
processed, unprocessed, other
select consensus sequences
family alignments
pseudogene-gene families
pseudogene-gene families
processed pseudogenes
unprocessed pseudogenes
other pseudogenes
genes BLAST database
gene family consensus sequences
pseudogene assignments
pseudogene assignment
BLAST gene matches
BLAST
CUDAlign
pseudogene-consensus gene alignments

Figure 2

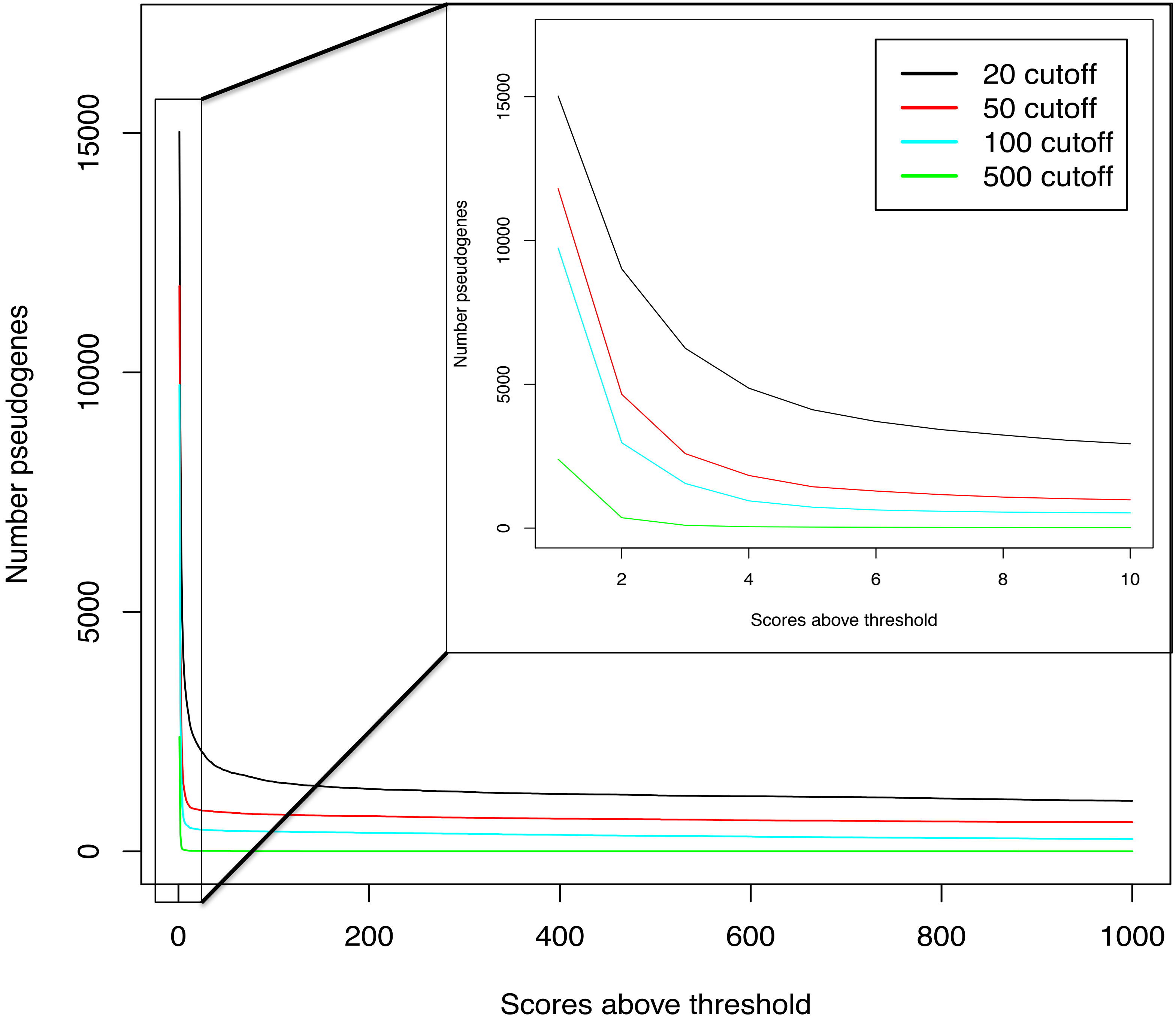

20 cutoff
50 cutoff
100 cutoff
500 cutoff
Number pseudogenes
0
5000
10000
15000
2
4
6
8
10
Scores above threshold
Number pseudogenes
0
5000
10000
15000
0
200
400
600
800
1000
Scores above threshold

Figure 3

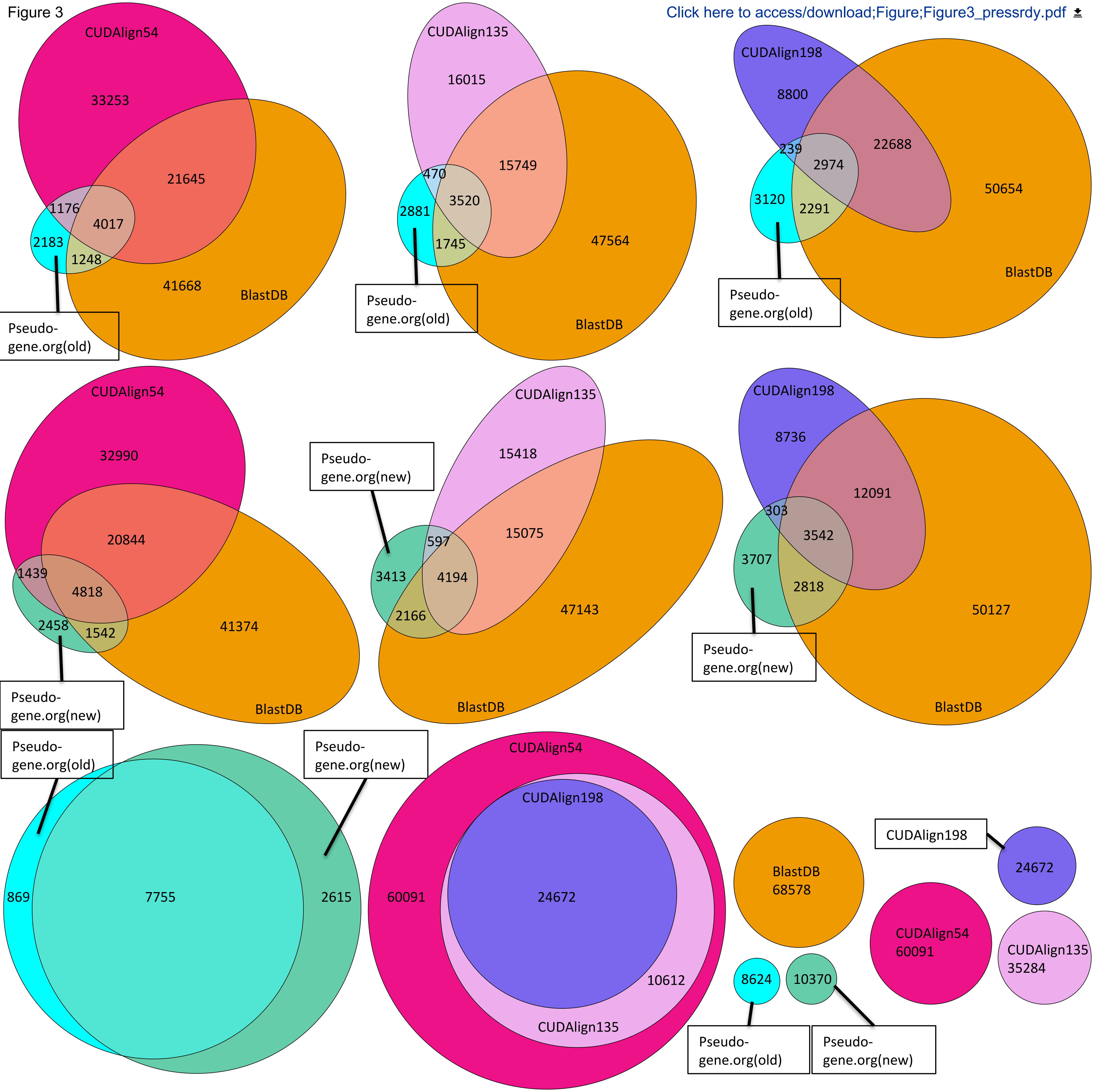
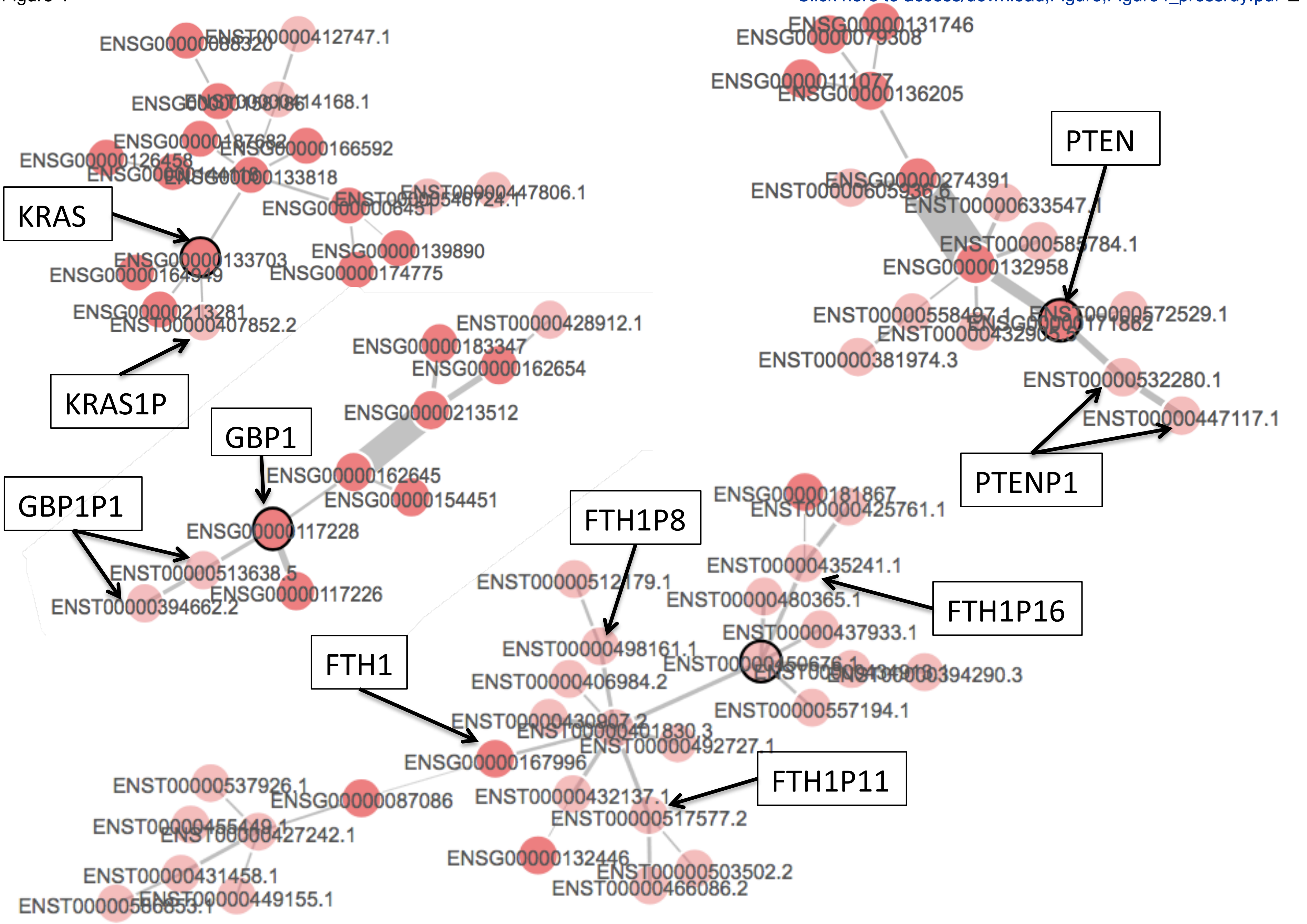
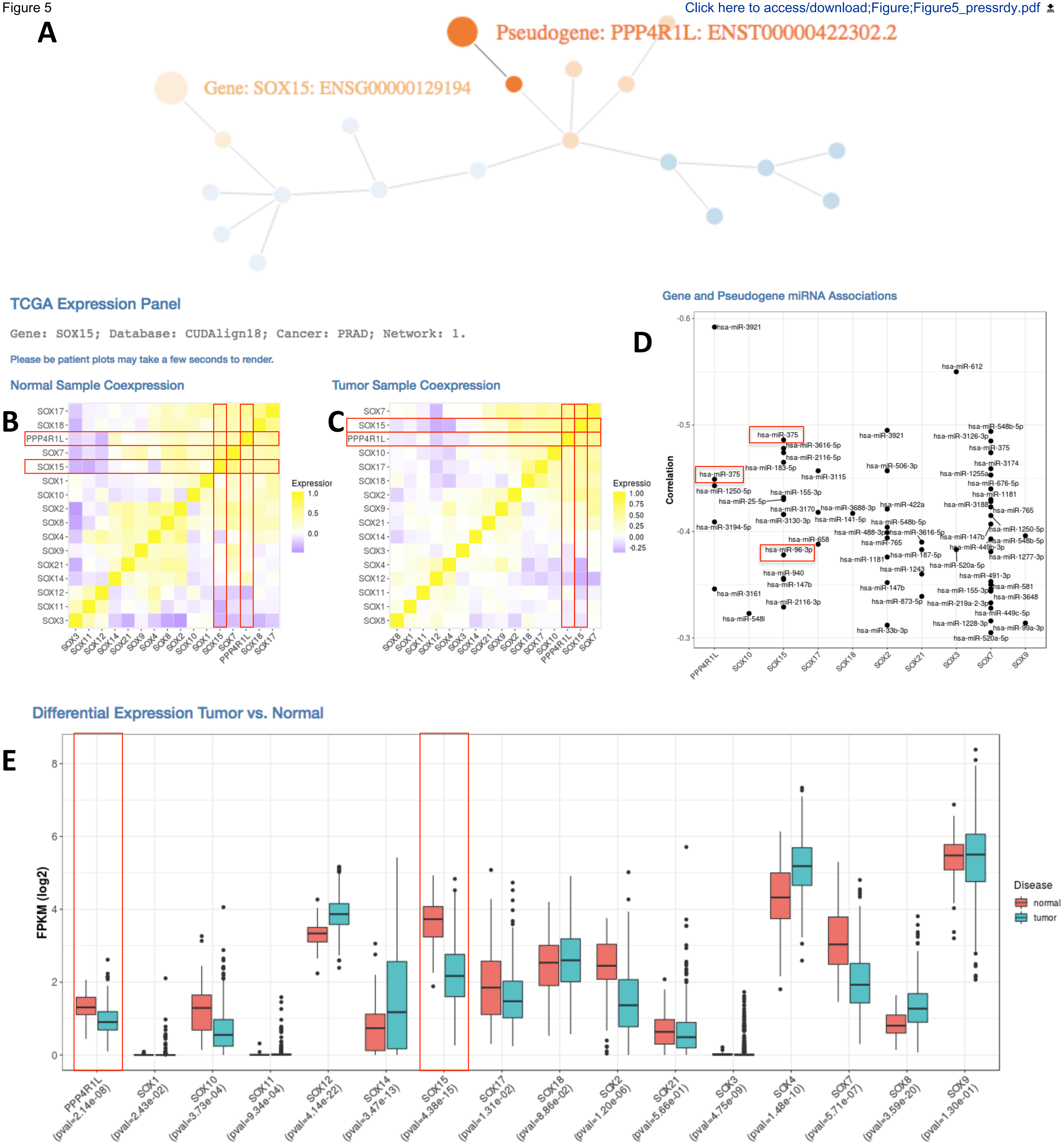

CUDAlign54
33253
21645
1176
4017
2183
1248
41668
BlastDB
Pseudo-
gene.org(old)
CUDAlign135
16015
15749
470
3520
2881
1745
47564
BlastDB
Pseudo-
gene.org(old)
CUDAlign198
8800
22688
239
2974
3120
2291
50654
BlastDB
Pseudo-
gene.org(old)
CUDAlign54
32990
20844
1439
4818
2458
1542
41374
BlastDB
Pseudo-
gene.org(new)
Pseudo-
gene.org(new)
CUDAlign135
15418
15075
597
3413
4194
2166
47143
BlastDB
CUDAlign198
8736
12091
303
3542
3707
2818
50127
Pseudo-
gene.org(new)
BlastDB
Pseudo-
gene.org(old)
Pseudo-
gene.org(new)
869
7755
2615
CUDAlign54
CUDAlign198
60091
24672
10612
CUDAlign135
BlastDB
68578
CUDAlign198
24672
CUDAlign54
60091
CUDAlign135
35284
8624
10370
Pseudo-
gene.org(old)
Pseudo-
gene.org(new)

Figure 4


KRAS
KRAS1P
GBP1
GBP1P1
PTEN
PTENP1
FTH1P8
FTH1P16
FTH1
FTH1P11
ENSG00000088320
ENST00000412747.1
ENST00000614168.1
ENSG00000166592
ENSG00000126458
ENSG00000133818
ENST00000447806.1
ENSG00000006451
ENSG00000133703
ENSG00000139890
ENSG00000174775
ENSG00000213281
ENST00000407852.2
ENST00000428912.1
ENSG00000183347
ENSG00000162654
ENSG00000213512
ENSG00000162645
ENSG00000154451
ENSG00000117228
ENST00000513638.5
ENSG00000117226
ENST00000394662.2
ENSG00000131746
ENSG00000111077
ENSG00000136205
ENSG00000274391
ENST00000633547.1
ENST00000585784.1
ENSG00000132958
ENST00000572529.1
ENST00000381974.3
ENST00000532280.1
ENST00000447117.1
ENSG00000181867
ENST00000425761.1
ENST00000435241.1
ENST00000480365.1
ENST00000437933.1
ENST00000498161.1
ENST00000406984.2
ENST00000557194.1
ENST00000492727.1
ENSG00000167996
ENST00000537926.1
ENSG00000087086
ENST00000432137.1
ENST00000427242.1
ENST00000517577.2
ENSG00000132446
ENST00000431458.1
ENST00000503502.2
ENST00000449155.1
ENST00000466086.2



TCGA Expression Panel

Gene: SOX15; Database: CUDAlign18; Cancer: PRAD; Network: 1.

Please be patient plots may take a few seconds to render.

Supplementary Material

Sept 22, 2018

Dear Colleagues,

We are excited to present our new resource PseudoFuN (https://integrativeomics.shinyapps.io/pseudofun_app/) for consideration of publication in GigaScience. Here we submit the manuscript entitled "PseudoFuN: Deriving functional potentials of pseudogenes from integrative relationships with genes and miRNAs across 32 cancers".

In the past 1.5 years, we have been working on generating comprehensive pseudogene-gene (PGG) family databases. Unlike previous pseudogene-gene databases which conventionally only considered the 1:1 pseudogene-parent gene pairs, we considered all the homologous genes and pseudogenes as a PGG family. We believe PGG families are more comprehensive in modeling evolutionary relationship and functional relationships of pseudogenes and genes.

These PGG families can be used as input to study gene-pseudogene-miRNA co-expression indicative of ceRNA networks (e.g., across the entire Cancer Genome Atlas), individually downloaded with pairwise sequence homology, mapped to functional annotation, and mapped back to the genomic location. With these databases and tools provided by PseudoFuN, it is possible to generate hypotheses regarding i) the regulatory roles of pseudogenes across tumor and normal tissue, ii) pseudogene gene relationships through our de novo reassignment of pseudogenes to gene families and iii) functional annotation of pseudogenes. We expect our databases and tools to have more applications in cancer studies.

Best,
Yan

--
Yan Zhang, Ph.D.
Assistant Professor
Department of Biomedical Informatics
College of Medicine
The Ohio State University
310-B Lincoln Tower, 1800 Cannon Drive, Columbus, OH 43210
Phone: (614) 688-9643 | Email: Yan.Zhang@osumc.edu
https://medicine.osu.edu/bmi/people/yan_zhang/Pages/index.aspx