

GigaScience

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

--Manuscript Draft--

Manuscript Number:	GIGA-D-18-00369R2	
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 University (Startup Funds)	Dr. Yan Zhang
	National Cancer Institute (P30CA016058)	Dr. Yan Zhang
Abstract:	<p>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. For example, we find multiple known (e.g., miR-20a-PTEN-PTENP1) and novel (e.g., miR-375-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 co-expression and differential expression with a simple-to-use online tool by bioinformaticians and oncologists alike.</p>	
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:	
Response to Reviewers:	<p>Response Letter</p> <p>Response to editorial issues:</p> <p>1) Please include a citation to your new GigaDB repository to the bibliography, and cite it from the "data availability" section (the same way as you have done it for the GitHub site).</p> <p>The citation is:</p> <p>[xx] Johnson T; Li S; Franz E; Huang Z; Li SD; Campbell MJ; Huang K; Zhang Y Supporting data for "PseudoFuN: Deriving functional potentials of pseudogenes from integrative relationships with genes and miRNAs across 32 cancers" GigaScience Database 2019. https://urldefense.proofpoint.com/v2/url?u=http-3A__dx.doi.org_10.5524_100577&d=DwlGaQ&c=k9MF1d71ITtkuJx-PdWme51dKbmfPEvxwt8SFEkBfs4&r=ekJXXuaVVvw1UTqb5FGVAqKm1D8j8S7Zred9hAkY7k&m=-JC_daUTsqZc2BjQTOY7Gk63PpDti3noA1ESRxQKWG4&s=SyHXIQfP5-vHlv7kVDFTMIWU8RPnS3jS-u9bzUGKalg&e=</p> <p>In the data availability section, please write, for example, "Supporting data and an archival copy of the code are available via the GigaScience database, GigaDB [xx]", or similar wording.</p> <p>Authors' answer: Done.</p> <p>2) Please include a section "Availability of supporting source code and requirements", formatted as described in our Instructions for authors.</p> <p>List the following:</p> <p>Project name: e.g. My bioinformatics project Project home page: e.g. https://urldefense.proofpoint.com/v2/url?u=https-3A__github.com_ISA-2Dtools&d=DwlGaQ&c=k9MF1d71ITtkuJx-PdWme51dKbmfPEvxwt8SFEkBfs4&r=ekJXXuaVVvw1UTqb5FGVAqKm1D8j8S7Zred9hAkY7k&m=-JC_daUTsqZc2BjQTOY7Gk63PpDti3noA1ESRxQKWG4&s=NtadplgDdzem-mFJrMEk2A_FtIP9KrlLUbNh9_7DiCc&e= Operating system(s): e.g. Platform independent Programming language: e.g. Java Other requirements: e.g. Java 1.3.1 or higher, Tomcat 4.0 or higher License: e.g. GNU GPL, FreeBSD etc. RRID: if applicable, e.g. RRID: SCR_014986</p> <p>==</p> <p>Please register any new software application in the https://urldefense.proofpoint.com/v2/url?u=http-3A__SciCrunch.org&d=DwlGaQ&c=k9MF1d71ITtkuJx-PdWme51dKbmfPEvxwt8SFEkBfs4&r=ekJXXuaVVvw1UTqb5FGVAqKm1D8j8S7Zred9hAkY7k&m=-JC_daUTsqZc2BjQTOY7Gk63PpDti3noA1ESRxQKWG4&s=ctsWxoeEpsMQfIAvLUT27I_ssqxLZZ_7mE5J68Ke8hA&e= database to receive a RRID (Research Resource Identification Initiative ID) number, and include this in your manuscript. This will facilitate tracking, reproducibility and re-use of your tool.</p>

	<p>=</p> <p>Related to software availability, you write "we provide another interactive web application 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." - please add details in the manuscript how this web application can be accessed.</p> <p>Authors' answer: All done. We have added a section and made a Research Resource Identification Initiative ID, SCR_017095.</p> <p>3) Please remove any highlighting in red that was made for the purpose of peer review.</p> <p>Authors' answer: Done.</p> <p>4) Please also note the minor comment of the reviewer (below)</p> <p>Authors' answer: Yes. We have added the missing highlighting color back into Table 2.</p> <p>5) Please submit your revised version of the text in Word /LaTex (whatever you used to prepare the document) - we only seem to have a PDF version at the moment.</p> <p>Authors' answer: We will submit Word text with this revision.</p> <p>PS: We added one affiliation for the corresponding author, and modified the Acknowledgements in the main text.</p> <p>Response to reviewer reports: Reviewer #2: I appreciate the effort that the authors put in to addressing my comments and added the additional benchmarking table 2. One minor comment concerning table 2 is that the highlighting described in the figure could not be seen in the table, otherwise I have no further amendments.</p> <p>Authors' answer: We appreciate the great suggestions from the reviewers. We have added the missing highlighting color back into Table 2.</p>
Additional Information:	
Question	Response
Are you submitting this manuscript to a special series or article collection?	No
<p>Experimental design and statistics</p> <p>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.</p> <p>Have you included all the information requested in your manuscript?</p>	Yes
Resources	Yes

<p>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.</p> <p>Have you included the information requested as detailed in our Minimum Standards Reporting Checklist?</p>	
<p>Availability of data and materials</p> <p>All datasets and code on which the conclusions of the paper rely must be either included in your submission or 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.</p> <p>Have you have met the above requirement as detailed in our Minimum Standards Reporting Checklist?</p>	<p>Yes</p>

[Click here to view linked References](#)

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60
61
62
63
64
65

1 **PseudoFuN: Deriving functional potentials of pseudogenes from integrative** 2 **relationships with genes and miRNAs across 32 cancers**

3 Travis S Johnson¹, Sihong Li¹, Eric Franz², Zhi Huang^{3,4}, Shuyu Dan Li⁵, Moray J Campbell⁶, Kun Huang^{4,7}, Yan
4 Zhang^{1,8*}

5
6 ¹ Department of Biomedical Informatics, College of Medicine, The Ohio State University, Columbus, OH
7 43210, USA

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

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

10 ⁴ School of Medicine, Indiana University, Indianapolis, IN 46202, USA

11 ⁵ Department of Genetics and Genomic Sciences, Icahn School of Medicine at Mount Sinai, New York, NY
12 10029, USA

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

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

16 ⁸ The Ohio State University Comprehensive Cancer Center (OSUCCC – James), Columbus, OH 43210, USA

17 * Correspondence: yan.zhang@osumc.edu

18 **Abstract**

19 **Background:** Long thought “relics” of evolution, not until recently have pseudogenes been of
20 medical interest regarding regulation in cancer. Often, these regulatory roles are a direct
21 byproduct of their close sequence homology to protein coding genes. Novel pseudogene-gene
22 functional associations can be identified through the integration of biomedical data, such as
23 sequence homology, functional pathways, gene expression, pseudogene expression, and
24 miRNA expression. However, not all of the information has been integrated, and the vast
25

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60
61
62
63
64
65

1 majority of previous pseudogene studies relied on 1:1 pseudogene-parent gene relationships
2 without leveraging other homologous genes/pseudogenes. **Results:** We produce pseudogene-
3 gene (PGG) families that expand beyond the current 1:1 paradigm. Firstly, we construct
4 expansive PGG databases by i) CUDAlign GPU accelerated local alignment of all pseudogenes
5 to gene families (totaling 1.6 billion individual local alignments and more than 40,000 GPU
6 hours) and ii) BLAST-based assignment of pseudogenes to gene families. Secondly, we create
7 an open-source web application (PseudoFuN) to search for integrative functional relationships
8 of sequence homology, miRNA expression, gene expression, pseudogene expression, and
9 gene ontology. We produce four “flavors” of databases (>462,000,000 pseudogene-gene
10 pairwise alignments and 133,770 PGG families) that can be queried and downloaded using
11 PseudoFuN. These databases are consistent with previous 1:1 pseudogene-gene annotation
12 and also are much more powerful including millions of *de novo* pseudogene-gene associations.
13 For example, we find multiple known (e.g., miR-20a-PTEN-PTENP1) and novel (e.g., miR-375-
14 SOX15- PPP4R1L) miRNA-gene-pseudogene associations in prostate cancer. PseudoFuN
15 provides a “one stop shop” for identifying and visualizing thousands of potential regulatory
16 relationships related to pseudogenes in TCGA cancers. **Conclusions:** Thousands of new
17 pseudogene-gene associations can be explored in the context of miRNA-gene-pseudogene co-
18 expression and differential expression with a simple-to-use online tool by bioinformaticians and
19 oncologists alike.

20
21 **Keywords:** Pseudogenes, database, functional prediction, gene regulation, network analysis,
22 high performance computing, graphics processing unit, competing endogenous RNA

23

1 **Background**

2 Pseudogenes were previously considered unimportant relics of evolution that played an unclear
3 role in biological processes¹. However, more pseudogenes have been discovered to be involved
4 in gene regulation²⁻⁴. These regulatory relationships between pseudogenes and genes have
5 increasingly been explored, such as the transcriptional regulation of PTEN by pseudogene
6 PTENP1 in several cancer conditions⁵. PTEN acts as a tumor suppressor gene, which is
7 underexpressed in gastric cancer. However by overexpressing PTENP1 in gastric cancer, both
8 PTEN underexpression and cell proliferation are mitigated via the regulatory relationship
9 between PTEN and PTENP1⁶. Relationships between these pseudogenes and their parent
10 genes have been found to play critical roles indicating functional potentials of these
11 pseudogenes^{7,8}. This point can most clearly be seen in the importance of sequence homology
12 between pseudogenes and coding genes plays in competing endogenous RNA (ceRNA)
13 networks^{9,10}. In ceRNA networks the pseudogenes act as decoy targets for the miRNAs
14 targeting a protein-coding gene. In short, researchers have made huge strides in understanding
15 pseudogenes from genomic variation to functional potentials^{11,12}, and from “deciphering” the
16 mechanism of ceRNA networks¹³ to experimental validation¹⁴.

17
18 With this progress, there has been renewed interest in pseudogenes, especially in relation to
19 cancer¹⁵. This interest has even uncovered biomarkers in human cancer including but not
20 limited to SUMO1P3 upregulation as a diagnostic biomarker in gastric cancer and OCT4-pg4
21 expression as a prognostic biomarker in hepatocellular carcinoma (HCC)¹⁶⁻¹⁸. Pseudogene
22 expression has been used to stratify tumor subtypes in 7 distinct cancer types¹⁹. However, due
23 to the close sequence homology between pseudogenes and their parent genes, identifying the
24 expression profile unique to a pseudogene or highly homologous gene can be challenging.
25 Efforts have been made to address these technical challenges in estimating pseudogene
26 expression using modified alignment and quantification techniques²⁰. Perhaps more intriguingly

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60
61
62
63
64
65

1 is that pseudogenes can be somatically acquired in cancer development effectively
2 “representing a new class of mutations” that can be either activating or inactivating mutations
3 which function as an “on/off switch”^{21,22}. Specific pseudogenes have been implicated in specific
4 cancers. For example, FTH1 regulates tumorigenesis in prostate cancer²³, TP73-AS1 regulates
5 proliferation in esophageal squamous cell carcinoma²⁴, and pseudogenes NKAPP1, MSTO2P
6 and RPLP0P2 are associated with poor prognosis in lung adenocarcinoma²⁵.

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

13
14 The conventional idea of single parent genes may not be comprehensive enough to model the
15 complex phylogenetic relationships involving multiple genes and pseudogenes in a homolog
16 family. While pseudogenes diverged from their parent genes distantly in the past, only the
17 daughter protein-coding genes other than the original parent gene may now exist. The result is
18 that aligning to the true phylogenetic parent gene itself may not be possible. For this reason, we
19 advocate the use of homologous gene families rather than single parent genes to compare
20 against pseudogenes. By viewing the homologies as a weighted network instead of a single
21 scalar value we believe that new relationships can be uncovered.

22
23 We build the pseudogene-gene (PGG) family databases using two methods: i) CUDAlign²⁹
24 based-local alignment of all pseudogenes to gene families (totaling 1.6 billion individual local
25 alignments and more than 40,000 GPU hours). By aligning all pseudogenes to all gene families
26 (CUDAlign), we can study underlying sequence homology and more easily set cutoffs to assign

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60
61
62
63
64
65

1 pseudogenes to gene families. ii) BLAST³⁰-based assignment of pseudogenes to gene families.
2 This provides a fast heuristic search option. BLAST derivative methods have been commonly
3 used to find parent genes in previous pseudogene studies^{31,32}. Using these two methods we
4 show that these pseudogenes are usually assigned to the gene family of their parent genes but
5 are often not exclusively so. Besides, most pseudogenes can be categorized into processed
6 pseudogenes and unprocessed pseudogenes depending on whether they came from
7 retrotranscription of mRNAs^{11,33,34}. We take these differences into account using both of our
8 methods (CUDAlign and BLAST).

9
10 Furthermore, we make these data publicly downloadable from GitHub³⁵. We also create an R
11 Shiny web application called PseudoFuN³⁶ that supports querying the PGG databases,
12 interactive visualization and functional analysis of the PGG networks, and visualization of
13 pseudogene-gene co-expression and miRNA binding (including binding prediction with
14 Miranda³⁷, PicTar³⁸, and TargetScan³⁹) using The Cancer Genome Atlas and GTEx (Genotype-
15 Tissue Expression) project derived public data^{20,40,41}. Besides, we provide another interactive
16 web application hosted by the Ohio Supercomputer Center⁴² (OSC), which supports querying
17 novel sequences against any of our PGG databases and visualization of the resulting PGG
18 networks.

19
20 The PGG databases can be used to study pseudogene-gene-miRNA co-expression indicative of
21 ceRNA networks across the entire Cancer Genome Atlas. With these diverse tools provided by
22 PseudoFuN, it is possible to generate hypotheses regarding i) the regulatory roles of
23 pseudogenes across tumor and normal tissue, ii) pseudogene-gene relationships through *de*
24 *novo* reassignment of pseudogenes to gene families and iii) functional annotation of
25 pseudogenes. We expect these databases and tools to have more use in cancer studies.

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60
61
62
63
64
65

1 **Methods**

2 ***Construction of Pseudogene-Gene (PGG) Database***

3 To generate these gene families, we use two methods: i) CUDAlign-based local alignment of
4 pseudogenes against consensus sequences representing gene families, and ii) BLAST-based
5 search of pseudogene sequences against all gene sequences (Figure 1). These two
6 approaches can be thought of as heuristic but different processes. The local sequence
7 alignment approach is heuristic in that only two gene sequences are used from each gene
8 family to reduce the search space. These sequences are the most similar and representative
9 sequences to all the other gene sequences in the family. The BLAST-based approach is
10 heuristic in that not all sequences are fully aligned during the process due to the seed-and-
11 extend steps of BLAST⁴³. The result is that not every relationship between pseudogene and
12 gene family is recorded which is an advantage in runtime but a disadvantage in studying
13 underlying sequence homology.

14
15 *i) CUDAlign-based local alignment of gene families*

16 Gene homolog families were generated using the Ensembl biomart gene homolog database^{44,45}.
17 The pairs of homologous genes were separated into connected components using python
18 networkx package⁴⁶. These connected component sub-graphs are considered gene families in
19 this study. To reduce the number of alignments that needed to be performed, we selected
20 consensus genes from each family that would be used to represent the entire family.
21
22 The consensus sequences were selected by aligning every member of the gene family to every
23 other member using local alignment with CUDAlign²⁹. The two members of the family with the
24 largest sum alignment scores across all other family members were selected as the consensus
25 sequences to increase the number of candidate sequences. If only one member existed in the
26 family, then that member was the consensus sequence. Using the list of these consensus

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60
61
62
63
64
65

1 sequences we then aligned every consensus sequence to every pseudogene in the human
2 genome GRCh38 annotated by GENCODE Release 25 (GENCODEv25)⁴⁷.

3
4 Specifically the pseudogenes are split up into processed, unprocessed and other (unclear
5 whether processed or unprocessed), based on their mechanisms of formation⁴⁸. We performed
6 different alignment procedures for processed and unprocessed pseudogenes respectively. The
7 processed pseudogenes were aligned to all consensus gene transcripts with the highest local
8 alignment score recorded. The unprocessed pseudogenes were aligned to the full genomic
9 sequences of each of the consensus genes with the highest local alignment score recorded.

10 Theoretically unprocessed pseudogenes can align to both exonic and intronic regions of DNA,
11 while processed pseudogene can only align to exonic regions. In our previous database we did
12 not perform this two-procedure strategy in part to reduce the runtime of the problem⁴⁹. These
13 changes make the database much more complete and biologically relevant. The other
14 pseudogenes were aligned to both the transcripts and the genomic sequence recording the
15 highest score.

16
17 These scores, one for each combination of pseudogene to gene family, were stored for further
18 analysis. Pseudogenes were assigned to families using a cutoff score (i.e., percentiles of the
19 alignment scores per PGG alignment matrix) and a maximum number of assignments (i.e., the
20 top four alignments above a cutoff). If greater than top four alignments were used, the PGG
21 families were too large to calculate the pairwise alignment matrix. The resulting sets of
22 pseudogenes and genes are called pseudogene-gene (PGG) families. This method was used to
23 allow a pseudogene to be assigned multiple families as well as prevent pseudogenes from
24 being assigned families if their alignment score was low. We used the 99th percentile cutoff
25 (corresponding alignment score 54), 99.9th percentile cutoff (135), and the 99.99th percentile
26 cutoff (198) to generate three resultant databases named CUDAlign54, CUDAlign135, and

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60
61
62
63
64
65

1 CUDAlign198 respectively. A fourth database that is less stringent, CUDAlign18, is also
2 included in the web applications using a 97.5th percentile cutoff (18). All these flavors of
3 databases are available for search in our web applications.

4
5 *ii) BLAST-based generation of PGG families*

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

16
17 ***Comparison between PGG families and pseudogene-parent gene pairs***

18 We also conduct a comparison to the Pseudogene.org resource⁵⁰. In this comparison, we
19 consider pseudogenes and parent gene pairs from pseudogene.org psiDr³¹ database (old)⁵¹ and
20 on GENCODE Release 10 from pseudogene.org psiCube¹¹ database (new)⁵². From our
21 databases, we consider every combination of pseudogene to gene within a PGG family as a
22 pair (for example, a family with 3 genes and 2 pseudogenes would have $C_2^3 = 6$ pairs). Since we
23 have multiple flavors of PGG databases including the BLAST-based version and the CUDAlign-
24 based versions, we compare the intersections between two Pseudogene.org versions and our

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60
61
62
63
64
65

1 BLAST/CUDAlign-based versions. We show the intersections of pseudogene-gene pairs in
2 Venn Diagrams.

3
4 ***Development of PseudoFuN web applications***

5 Aside from generating different flavors of the PGG databases, we assemble them into an online
6 R Shiny application called PseudoFuN³⁶ which supports gene and pseudogene symbol queries
7 against our PGG databases, generates dynamic networks, produces Gene Ontology⁵³ (GO)
8 tables and additional functional analysis features (Table 1). The functionalities, such as
9 calculating the gene co-expression for any resultant PGG network in any of the TCGA⁵⁴ cancers
10 types, are important for ceRNA network hypothesis generation in human cancers. For more
11 information, please visit the PseudoFuN website and follow the README and tutorial.

12
13 Additionally we create another web application hosted by the Ohio Supercomputer Center
14 (OSC) OnDemand⁵⁵ platform. This application has multiple functionalities including the query of
15 Ensembl gene ID or a novel sequence against one selected flavor of our databases. For each of
16 these features we provide a simple-to-use interface that allows users to select which database
17 to query, allows download of the query hits, and allows users to interactively explore the PGG
18 family networks including GO information.

19
20 ***Use cases in multiple cancers***

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

1 prostate cancer cell lines and also differentially expressed in our PGG networks in TCGA
2 prostate cancer samples.

3 4 **Results**

5 *Local alignment of gene families*

6 We performed 1.6 billion local alignments between all pseudogenes and all gene family
7 consensus sequences. The process required over 40,000 GPU hours on the Oakley cluster at
8 the OSC. The highest scores for each gene family and pseudogene were stored in a
9 17,273x26,754 matrix of pseudogene-to-gene-family alignment scores (~462 million elements).
10 From this matrix, we are able to explore global pseudogene-gene family homology relationships
11 and assign pseudogenes to one or more gene families with high sequence homology.

12
13 As one might expect, the number of pseudogenes with high alignments (defined as above a
14 percentile threshold) to many gene families is relatively low. It can be seen that the majority of
15 pseudogenes will align to one gene family in the CUDAlign databases (Figure 2). We evaluate
16 alignment of pseudogenes to genes using the Smith-Watermann local pairwise alignment
17 score⁵⁶ between a pseudogene and a gene. These scores indicate the highest score possible
18 for two sequences based on their specific dynamic programming matrix which is solved by the
19 Smith-Watermann algorithm. The cutoffs we use, 18, 54, 135, and 198, indicate the 97.50th,
20 99.0th, 99.90th and 99.99th percentiles of alignment scores in our alignment matrix between all
21 pseudogenes and consensus sequences. Another feature of note is that there are some
22 pseudogenes that align to many gene families (e.g., 9 pseudogenes, UBE2Q2P1, RP11-
23 313J2.1, TPTEP1, BMS1P1, CTD-2245F17.3, SCAND2P, GTF2IP7, WHAMMP3, IGLV3-2,
24 have alignment scores above 54 in 15,000 gene families and 571 pseudogenes, see
25 Supplementary Table 2, have alignment scores above 54 in 1,000 gene families).

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60
61
62
63
64
65

1 In contrast to previous belief in single gene-pseudogene homology, some pseudogenes are
2 related to many genes. It is worth considering that these high homology pseudogenes (e.g.,
3 FTLP10 with 3,006 gene family pairwise alignments over a 54 threshold) may have a role in
4 regulating major biological processes⁵⁷ and disease⁵⁸. Of the 9 highest homology pseudogenes
5 (Supplementary Table 2), one, RP11-313J2.1, is a zinc finger pseudogene and two, CTD-
6 2245F17.3 and SCAND2P, are located in the promoters of zinc finger genes. Four
7 pseudogenes in the 9 highest homology pseudogenes (RP11-313J2.1, CTD-2245F17.3,
8 SCAND2P, and WHAMMP3) also have 92-96% sequence identity with zinc finger genes
9 (ZNF72P, ZNF518A, ZNF37A and ZNF788P/ZNF20 respectively) when BLAST searched
10 against the human genome. Of the 571 highest homology pseudogenes (Supplementary Table
11 2), we found 27 zinc finger pseudogenes. Using EnrichR⁵⁹ we identified enrichment in GO
12 Molecular Function GO:0004430 1-phosphatidylinositol 4-kinase activity (Fisher's exact test p-
13 value = 0.001), and enrichment for GO Biological Process GO:0070475 rRNA base methylation
14 (Fisher's exact test p-value = 0.003). In the ARCHS4 database⁶⁰ 324 transcription factors were
15 significantly co-expressed (Benjamini-Hochberg adjusted Fisher's exact test p-value < 0.05)
16 with members of the 571 highest homology pseudogenes. Of those 324 transcription factors,
17 228 were zinc finger genes. These findings show that the highest homology pseudogenes, like
18 zinc finger genes, likely contain repetitive elements that align to many genomic loci.

19

20 *BLAST generation of PGG families*

21 The BLAST generated database was larger than the CUDAlign generated databases with
22 68,578 total connections. This database was also much simpler to compute with since it was not
23 an exhaustive search. These conclusions make it a simple method to quickly estimate the
24 pseudogene-to-gene relationships.

25

26 *Direct comparison to pseudogene parents*

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60
61
62
63
64
65

1 We compare our databases to the previous pseudogene-parent gene databases retrieved from
2 Pseudogene.org resources (Figure 3). It shows that our methods reconstruct most of the
3 pseudogene-parent-gene relationships identified by Pseudogene.org. The overall consistency of
4 our databases (BLAST and CUDAlign) with both Pseudogenes.org databases (new and old)
5 was 75% (i.e., all our databases combined). Individually, the BLAST-based database contained
6 61% of the Pseudogene.org relationships (both new and old) and the CUDAlign 54 cutoff
7 contained 60% of the Pseudogene.org relationships (both new and old). Our databases also
8 generate a larger pool of possible interactions. It is worth noting that 391 pseudogenes and 152
9 genes in the new Pseudogene.org (GENCODE Release 10) are not present in the GENCODE
10 Release 25 annotation used in our analysis. These genes and pseudogenes together account
11 for 1030 edges that were used in our comparison. Accounting for these differences in the
12 annotation, we are able to reconstruct 85% of the pseudogene-gene relationships in the new
13 Pseudogene.org database. Since these associations were generated without prior pseudogene-
14 gene relationship information and the annotations have changed slightly since
15 Pseudogenes.org, our methods prove to independently identify known and unknown
16 pseudogene-gene relationships at a high rate.

17
18 *Development of a pseudogene query tool*

19 The R Shiny application is a comprehensive hypothesis generating tool that is freely available
20 on the internet³⁶. This tool provides a wide array of functionality that a researcher can access
21 quickly and download results as the raw data for more in-depth analysis. These features are
22 outlined in detail in Table 1.

23
24 *Use Cases: Assisting functional study of ceRNA networks in cancer*

25 To illustrate the utility of our databases and tools we present three use cases.
26

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60
61
62
63
64
65

1 Use Case I: To validate known pseudogene-gene relationships, we first identified 31 benchmark
2 pseudogene-gene relationships from three studies^{15,16,23} and query our databases. These
3 studies represent prominent regulatory pseudogenes in cancers by established laboratories. We
4 query a gene/pseudogene name one at a time and PseudoFuN will return the top PGG
5 network(s) that contain the query (Table 2). In general, we found that our databases together
6 were able to identify 87% of the benchmarking cases (Table 2) and the CUDAlign versions were
7 able to identify 65% of the benchmarking cases. Perhaps most importantly, three of the cases
8 identified by CUDAlign (ATP8A2, CXADR, PERP) were not identified by the more traditional
9 BLAST approach (Table 2) showing that consensus sequence alignment can identify some
10 overlooked relationships. Next, individual benchmark cases were evaluated in more detail
11 (Supplementary Figure 2).

12
13 PTENP1 is a processed pseudogene homologous to PTEN, a tumor suppressor gene. PTENP1
14 is selectively lost in cancer and may regulate PTEN expression as a miRNA decoy target^{5,6}. We
15 have observed differential co-expression patterns of PGG families in tumor vs. normal for
16 PTENP1 network in multiple cancers including breast cancer (Supplementary Figure 3B,C). We
17 identified known miRNAs (has-miR-93 targets PTEN in breast cancer⁶¹) targeting PTEN PGG
18 network nodes providing insights into ceRNA regulation (Supplementary Figure 3D). These
19 insights are important since some pseudogenes competitively bind to miRNAs thus regulate
20 gene expression. We also identify hsa-miR-103a-3p, known to regulate PTEN in endometrial⁶²
21 and colorectal cancers⁶³, in breast cancer (Supplementary Figure 3D). The miRNA hsa-miR-
22 20a, known to regulate PTEN by the ceRNA mechanism in prostate cancer⁶⁴, was also
23 identified in breast cancer. The ceRNA network regulatory relationship is governed by effect
24 modulation of miRNA on gene expression by pseudogene expression (Supplementary Figure
25 1A,C,E). This leads to a correlation between pseudogene (miRNA decoy targets) and gene
26 (miRNA targets) expression (Supplementary Figure 1D). That means both these pseudogenes

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60
61
62
63
64
65

1 and homologous genes competitively bind to miRNAs. KRAS-KRAS1 regulatory network was
2 also identified by our database (Supplementary Figure 2). KRAS and KRASP1 are known to be
3 involved in ceRNA network regulation^{5,10,64}. PseudoFuN query of KRAS identified co-expression
4 patterns in prostate cancer consistent with ceRNA network regulation by hsa-miR-145, a known
5 modulator of KRAS in prostate cancer⁶⁵. The FTH1 query also resulted in the identification of
6 pseudogenes (FTH1P2, FTH1P8, FTH1P11, FTH1P16) that regulate FTH1 in prostate cancer²³
7 as well novel miRNAs that may be involved in ceRNA network regulation of FTH1 in prostate
8 cancer. GBP1 is an IFN- α induced transcript that is involved in immune response in prostate
9 cancer⁶⁶. The GBP1 involved PGG network also contained the pseudogene GBP1P1 which may
10 have a ceRNA regulatory role in breast cancer⁶⁷ and in some neurodegenerative diseases⁶⁸.

11
12 Use Case II: We wanted to identify possible gene-miRNA relationships of interest within our
13 database. We chose to study these relationships with respect to miR-96, a known cancer
14 regulator microRNA in prostate cancer⁶⁹. Through differential expression analysis between
15 tumors in the TCGA-PRAD cohort with lower expression of RARG and TACC1 (also a miR-96
16 target) and high expression miR-96 (low RARG/low TACC1/high miR-96), compared to the
17 reverse, we previously identified altered SOX15 gene expression is significantly associated with
18 worse disease-free survival. We visualized expression patterns of SOX15 PGG families, and
19 corresponding miRNA associations. miR-96 is included as a validation.

20
21 Interestingly we identified the pseudogene PPP4R1L as a potential member of a SOX15 ceRNA
22 network (Figure 4A). PPP4R1L and SOX15 are both significantly differentially expressed
23 between tumor and normal controls (Bonferroni corrected p-value = 3.42×10^{-7} , 2.01×10^{-14}
24 respectively, Figure 4E). PPP4R1L and SOX15 are significantly co-expressed (Pearson
25 correlation coefficient (PCC)=0.51, p-value< 2.2×10^{-16}) in tumor tissue but much less correlated
26 in normal controls in prostate cancer (PCC=0.24, p-value=0.09, Figure 4B,C). Positively

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60
61
62
63
64
65

1 correlated expression is an assumption when determining ceRNA network relationships⁷⁰
2 (Supplementary Figure 1). Both SOX15 and PPP4R1L are likely regulated by hsa-miR-375
3 based on the TCGA prostate cancer dataset. hsa-miR-375 is associated with docetaxel
4 resistance in prostate cancer^{71,72} and PPP4R1L knock-down in HeLa cells induces taxol
5 resistance⁷³. These findings are intriguing since taxol and docetaxel are closely related chemical
6 compounds. PPP4R1L is also located in a region associated with high mutation rates in cancer
7 cell lines⁷³ which could be indicative of mutational “on/off switches” in pseudogene regulation.
8
9 Use Case III: We were most interested in the differentially expressed (DE) genes (and related
10 pseudogenes) that both appeared in our PGG database and were contained in networks with
11 genes differentially expressed in low RARG/low TACC1/high miR-96 compared to vice versa.
12 We searched the DE genes in our PGG database, and identified the top networks with enriched
13 number of DE genes. As a result, parent genes HTR7, CNN2, MSN and TAGLN2 are
14 differentially expressed; they generate pseudogenes, which are specifically expressed in
15 prostate cancer samples¹⁶. These four parent genes are also detected in our 5 top PGG families
16 involving miR-96 regulated (direct or indirect) DE genes. We identified HTR7P1 pseudogene in
17 the same PGG family as HTR7 gene, which is potentially regulated by hsa-miR-607 and has-
18 miR-3654 in the TCGA prostate cancer dataset (Supplementary Figure 4). 11 CNN2
19 pseudogenes (CNN2P1-CCN2P4, CNN2P6-CNN2P12) were identified in the CNN2 PGG family
20 along with TAGLN2 and TAGLN2P1. TAGLN2P1 is differentially expressed between the tumor
21 and normal samples in the prostate dataset (Supplementary Figure 5, Bonferroni corrected p-
22 value = 6.23x10-4). MSN and MSNP1 were in the same PGG family and hsa-miR-96 potentially
23 regulates MSN in the TCGA prostate cancer dataset (Supplementary Figure 5). In addition,
24 although our DE genes were detected from prostate cancer, we further compared them with DE
25 pseudogenes identified in four other cancer types and we observed interesting results (see
26 Supplementary Materials - *Potential regulatory roles in cancer*).

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60
61
62
63
64
65

1

2 **Discussion**

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

14

15 From these networks, we can also identify possible relationships of differentially expressed
16 pseudogenes in various cancers. For instance, both PPP4R1L pseudogene and SOX15 are
17 differentially expressed in prostate cancer and associated with hsa-miR-375. These types of
18 relationships should be further evaluated along with more complex regulation with multiple
19 miRNAs, pseudogenes, and genes. It is experimentally shown that SOX15 is regulated by hsa-
20 miR-96⁶⁹. It may be important to include hsa-miR-96 in the hsa-miR-375-SOX15-PPP4R1L
21 potential ceRNA network. Aside from PGG family specific differential pseudogene expression,
22 the PseudoFuN application allows for comprehensive differential pseudogene expression
23 (DPgE) analysis in any of the TCGA cancer datasets.

24

25 The use of this database also has utility in integrative analysis where the databases can be
26 used as a mask for other data modalities. Some examples would be using the nodes (genes

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60
61
62
63
64
65

1 and pseudogenes) in each of the PGG families as groups in gene expression experiments.
2 Similarly, these groups could be used for feature reduction when visualizing data. We hope
3 researchers can use these relationships we have identified to reduce large numbers of
4 candidate associations down to numbers that can be easily validated and generate new
5 candidates when querying novel sequences. For instance, miRNA-gene pairs filtered through
6 the sets of PGG families would identify high priority ceRNA candidates.

7

8 **Conclusions**

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

22

23 **Availability of Supporting Data**

24 We have made the PGG family data publicly downloadable from GitHub³⁵. We also created an
25 R Shiny web application called PseudoFuN³⁶ that supports querying the PGG databases,

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60
61
62
63
64
65

1 interactive visualization and functional analysis of the PGG networks, and visualization of
2 pseudogene-gene co-expression and miRNA binding. Apache License 2.0 is associated with
3 PseudoFuN (R Shiny web application). These data are also available on GigaDB⁷⁴. Besides, we
4 provide another interactive web application hosted on Ohio Supercomputer Center (OSC)
5 OnDemand, which supports querying novel sequences against any of our PGG databases and
6 visualization of the resulting PGG networks.

8 **Availability of supporting source code and requirements**

- 9 Project name: PseudoFuN
- 10 Project home page: https://github.com/yanzhanglab/PseudoFuN_app,
- 11 <https://github.com/OSC/pseudofun>, https://integrativeomics.shinyapps.io/pseudofun_app/
- 12 Operating system: Platform independent
- 13 Programming language: R, Python, Java Script
- 14 Other requirements: NA
- 15 License: CC, MIT
- 16 RRID: SCR_017095
- 17 OSC OnDemand application access: Please contact yan.zhang@osumc.edu.

19 **Additional Files**

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

22 **Supplementary Figure 1. Example of ceRNA network regulation of gene expression.** A) A
23 graphical view of how pseudogene expression can regulate gene expression. B) A cellular view
24 of ceRNA network regulation. C) Equations used to model the correlation between gene and
25 pseudogene expression in a ceRNA network. D) The distribution of the gene-pseudogene

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60
61
62
63
64
65

1 correlations based on the models in C. E) The effect that pseudogene expression has on the
2 miRNA induced change in gene expression.

3 **Supplementary Figure 2. PseudoFuN online output for PTEN PGG family.** A) Interactive
4 graph visualization of the PTEN PGG network. B) TCGA prostate co-expression matrix for
5 PTEN PGG family genes and pseudogenes across normal samples. C) TCGA prostate co-
6 expression matrix for PTEN PGG family genes and pseudogenes across tumor samples. D)
7 Negatively correlated miRNAs for all members of the PTEN PGG family. E) Differential gene
8 and pseudogene expression for tumor and normal samples for each member of the PTEN PGG
9 family in the prostate cancer TCGA dataset.

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

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

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60
61
62
63
64
65

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

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

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

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

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

21 Abbreviations

- 22 PseudoFuN: Pseudogene Functional Networks
- 23 PGG: Pseudogene-Gene (i.e., PGG families)
- 24 TCGA: The Cancer Genome Atlas

- 1
- 2
- 3
- 4 1 ceRNA: Competing Endogenous RiboNucleic Acid
- 5
- 6 2 HCC: HepatoCellular Carcinoma
- 7
- 8 3 BLAST: Basic Local Alignment and Search Tool
- 9
- 10 4 OSC: Ohio Supercomputer Center
- 11
- 12 5 GO: Gene Ontology
- 13
- 14 6 DE: Differential Expression
- 15
- 16 7 DGE: Differential Gene Expression
- 17
- 18 8 DPgE: Differential Pseudogene Expression
- 19
- 20 9
- 21
- 22
- 23

24 **10 Acknowledgments**

27 11 This work is partially supported by NIH-NLM MIDAS Training Fellowship (4T15LM011270-05)
28
29 12 awarded to Travis Johnson, and The Ohio State University Startup Funds and OSU
30
31 13 Comprehensive Cancer Center Support Grant (P30CA016058) to Yan Zhang. The authors also
32
33 14 thank the Ohio Supercomputer Center (OSC) for providing computing resources.
34
35

36 15

37

38 **16 Author contributions**

39
40 17 TSJ, SL, ZH and YZ performed data analyses. TSJ, EF and ZH developed the web applications.
41
42 18 YZ and TSJ conceived and initiated this project. YZ and KH supervised the project. MJC
43
44 19 provided experimental data. All authors contributed to biological interpretation. TSJ, YZ, MJC
45
46 20 and SDL wrote the manuscript. All authors read and approved the manuscript.
47
48
49

50 21

51

52 **22 Ethics, consent and permissions**

53
54 23 Not applicable.
55
56 24

57

58

59 **25 Consent to publish**

1
2
3
4 1 Not applicable.
5
6 2
7
8
9 3 **Competing interests**
10
11 4 The authors declare that they have no competing interests.
12
13 5
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60
61
62
63
64
65

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60
61
62
63
64
65

1 **Figure Captions**

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

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

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

15 **Figure 4. PseudoFuN online output for SOX15 PGG family.** A) Interactive graph visualization
16 of the SOX15 PGG network. B) TCGA prostate co-expression matrix for SOX15 PGG family
17 genes and pseudogenes across normal samples. C) TCGA prostate co-expression matrix for
18 SOX15 PGG family genes and pseudogenes across tumor samples. D) Negatively correlated
19 miRNAs for all members of the SOX15 PGG family. E) Differential gene and pseudogene
20 expression for tumor and normal samples for each member of the SOX15 PGG family in the
21 prostate cancer TCGA dataset.

22 **Tables**

23 **Table 1 Summary of features that are freely available at the PseudoFuN website.**

PseudoFuN features	Additional description
--------------------	------------------------

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60
61
62
63
64
65

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 ⁷⁵ 2. Kolmogorov-Smirnov (KS) Classic ⁷⁶ 3. Kolmogorov-Smirnov (KS) Elim ⁷⁶
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 and Ensembl websites ^{44,77} 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 information²⁰. The online tool allows for manipulation and download of the table.

Table 2. Benchmarking analysis of PseudoFuN databases. Genes indicate the gene with which the pseudogenes are associated in the literature. BLAST and CUDAlign columns indicate the specific databases. PMID indicates the literature from which the gene-pseudogene relationship was derived. Benchmark totals are included at the bottom of the table. Yellow highlighting indicates gene-pseudogene relationships found using BLAST but not CUDAlign. Green highlighting indicate indicates gene-pseudogene relationships found by CUDAlign but not by BLAST. Orange highlighting indicate where neither type of database identified the benchmark gene-pseudogene relationship. Benchmark totals are included at the bottom of the table.

Gene	BlastDB	CUDAlign18	CUDAlign54	CUDAlign135	CUDAlign198	PMID
PTEN	Yes	No	No	No	No	26442270
TUSC	No	No	No	No	No	26442270
INTS6	Yes	No	No	No	No	26442270
OCT4	Yes	Yes	Yes	Yes	Yes	26442270
HMGA1	Yes	Yes	Yes	Yes	Yes	26442270
CYP4Z1	No	No	No	No	No	26442270
BRAF	Yes	No	No	No	No	26442270
KLK4	No	No	No	No	No	22726445
ATP8A2	No	Yes	Yes	No	No	22726445
CXADR	No	Yes	Yes	Yes	Yes	22726445
CALM2	Yes	Yes	Yes	Yes	Yes	22726445
TOMM40	Yes	Yes	Yes	Yes	Yes	22726445
NONO	Yes	Yes	Yes	Yes	Yes	22726445
PERP	No	Yes	Yes	Yes	Yes	22726445
DUSP8	Yes	Yes	No	No	No	22726445
YES1	Yes	Yes	No	No	No	22726445
GJA1	Yes	No	No	No	No	22726445
AURKA	Yes	Yes	Yes	Yes	Yes	22726445
RHOB	No	No	No	No	No	22726445
HMGB1	Yes	Yes	Yes	Yes	Yes	22726445
EIF4A1	Yes	Yes	No	No	No	22726445
EIF4H	Yes	Yes	Yes	Yes	Yes	22726445
SNRP6	Yes	Yes	Yes	Yes	Yes	22726445
RAB1	Yes	No	No	No	No	22726445
VDAC1	Yes	Yes	No	No	No	22726445
RCC2	Yes	No	No	No	No	22726445
PTMA	Yes	Yes	Yes	Yes	Yes	22726445
NDUFA9	Yes	Yes	Yes	Yes	Yes	22726445
CES7	Yes	No	No	No	No	22726445
EPCAM	Yes	Yes	Yes	Yes	Yes	22726445
FTH1	Yes	Yes	Yes	Yes	Yes	29240947
Hits	24/31	20/31	16/31	15/31	15/31	
Total hits	27/31					

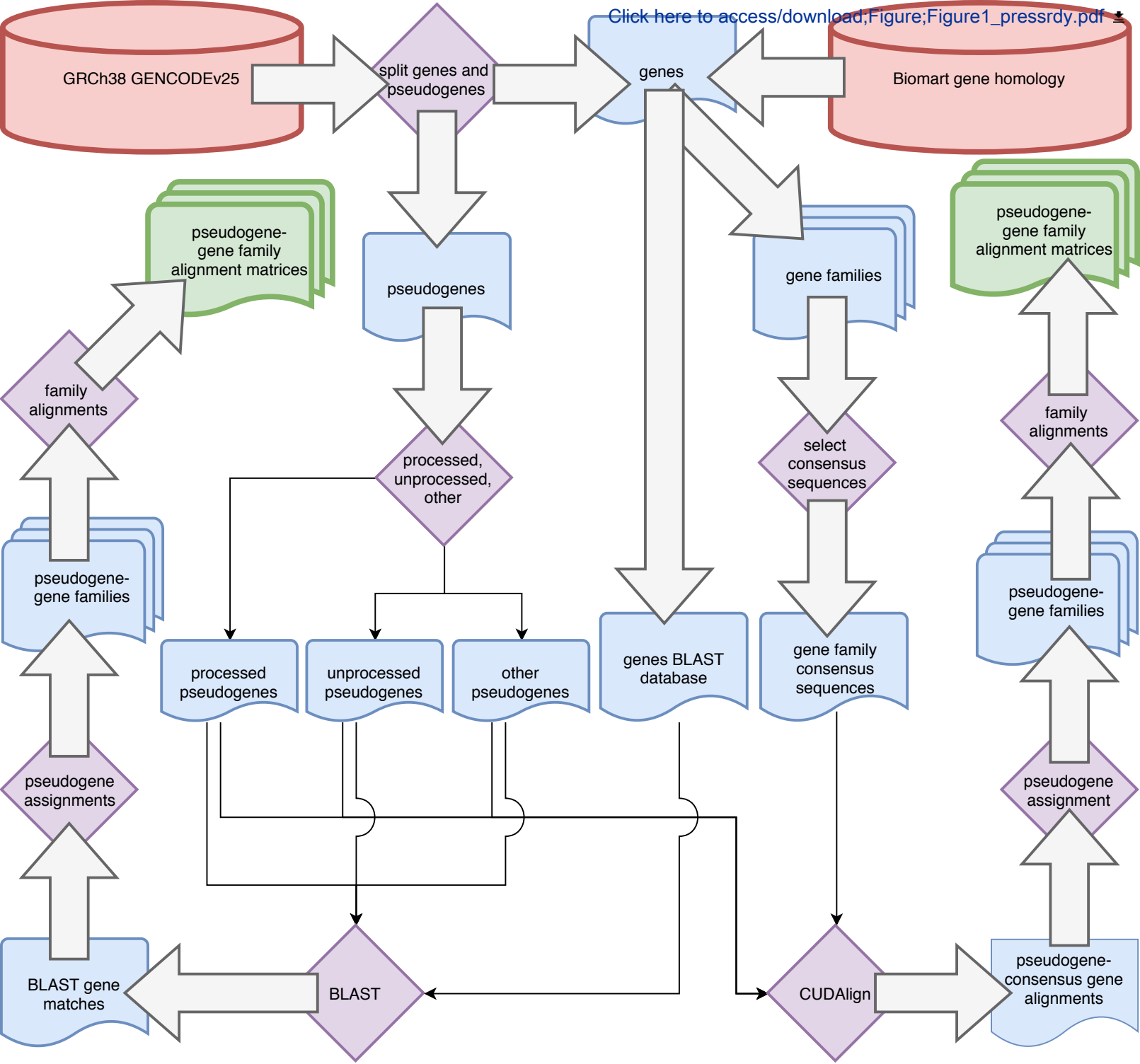
1 References

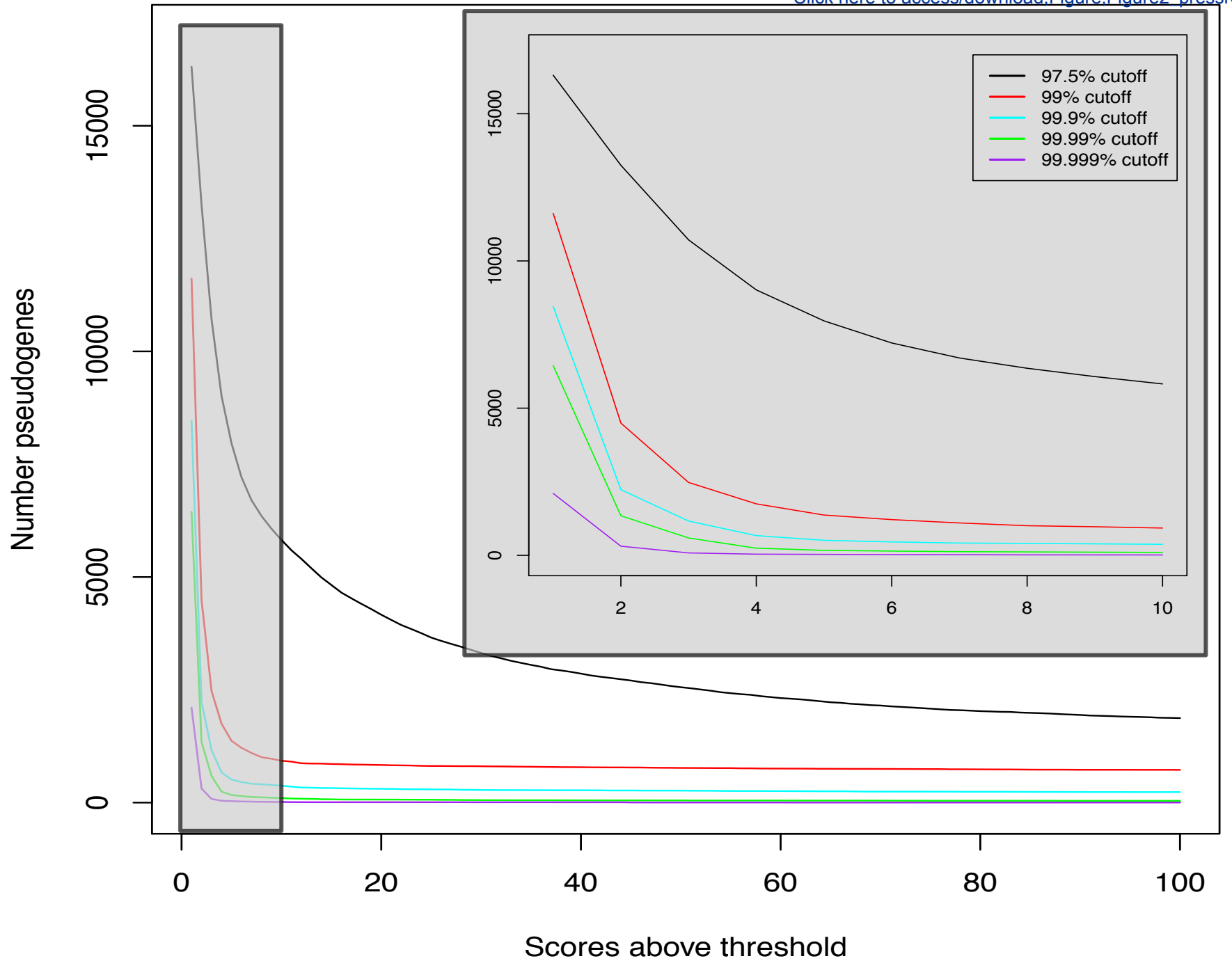
- 2 1. Vanin EF: Processed pseudogenes: characteristics and evolution. *Annu Rev Genet* 19:253-72, 1985
- 3 2. Mighell AJ, Smith NR, Robinson PA, et al: Vertebrate pseudogenes. *FEBS Lett* 468:109-14, 2000
- 4 3. Pink RC, Wicks K, Caley DP, et al: Pseudogenes: pseudo-functional or key regulators in health and disease? *RNA* 17:792-8, 2011
- 5 4. Chan JJ, Tay Y: Noncoding RNA:RNA Regulatory Networks in Cancer. *Int J Mol Sci* 19, 2018
- 6 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
- 7 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
- 8 7. Lam HY, Khurana E, Fang G, et al: Pseudofam: the pseudogene families database. *Nucleic Acids Res* 37:D738-43, 2009
- 9 8. Zheng D, Gerstein MB: A computational approach for identifying pseudogenes in the ENCODE regions. *Genome Biol* 7 Suppl 1:S13 1-10, 2006
- 10 9. An Y, Furber KL, Ji S: Pseudogenes regulate parental gene expression via ceRNA network. *J Cell Mol Med* 21:185-192, 2017
- 11 10. Poliseno L, Pandolfi PP: PTEN ceRNA networks in human cancer. *Methods* 77-78:41-50, 2015
- 12 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
- 13 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
- 14 13. Cesana M, Daley GQ: Deciphering the rules of ceRNA networks. *Proc Natl Acad Sci U S A* 110:7112-3, 2013
- 15 14. Chiu HS, Martinez MR, Bansal M, et al: High-throughput validation of ceRNA regulatory networks. *BMC Genomics* 18:418, 2017
- 16 15. Poliseno L, Marranci A, Pandolfi PP: Pseudogenes in Human Cancer. *Front Med (Lausanne)* 2:68, 2015
- 17 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
- 18 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
- 19 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
- 20 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
- 21 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

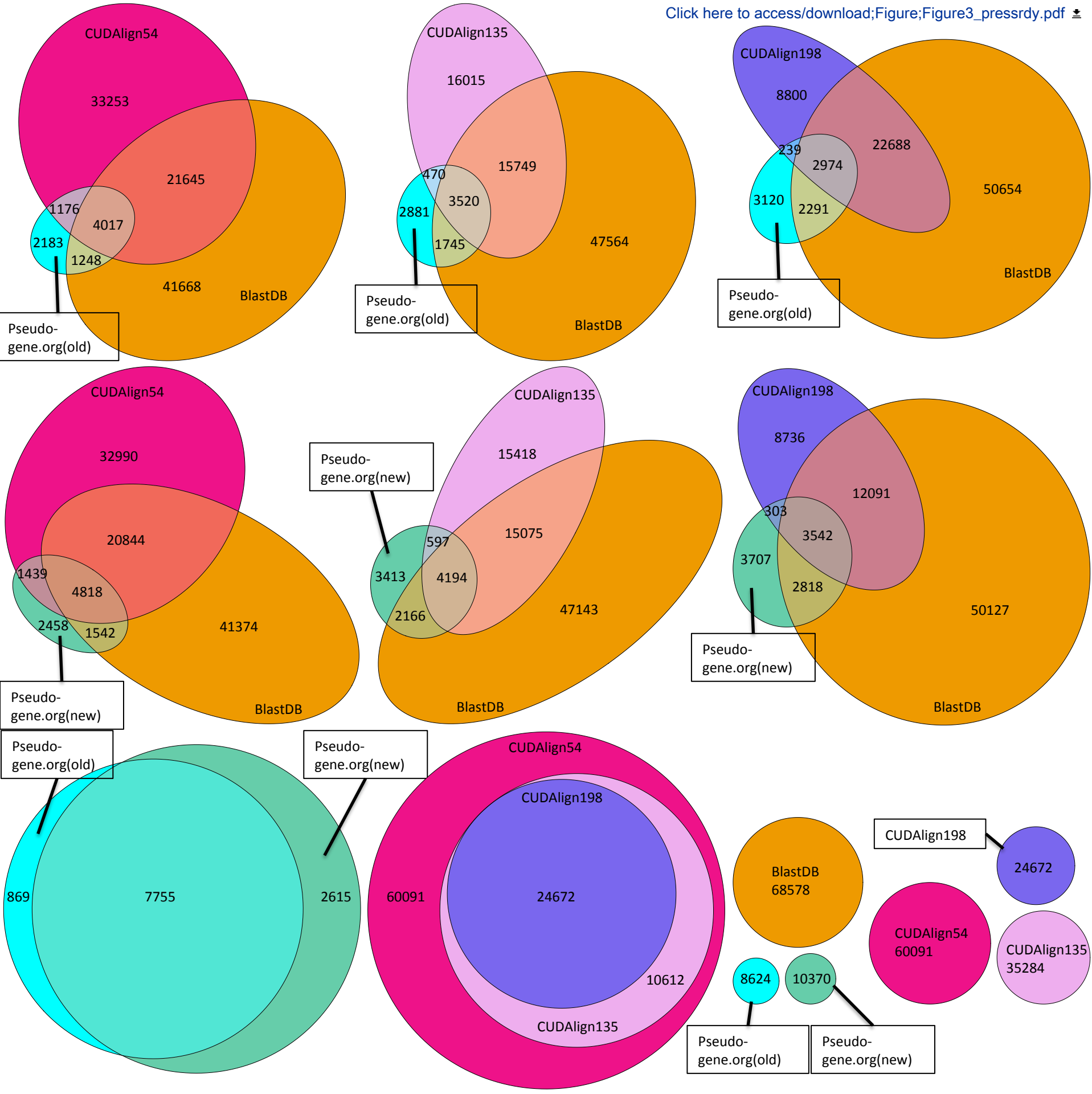
- 1 21. Cooke SL, Shlien A, Marshall J, et al: Processed pseudogenes acquired
2 somatically during cancer development. *Nat Commun* 5:3644, 2014
- 3 22. Shukla R, Upton KR, Munoz-Lopez M, et al: Endogenous retrotransposition
4 activates oncogenic pathways in hepatocellular carcinoma. *Cell* 153:101-11, 2013
- 5 23. Chan JJ, Kwok ZH, Chew XH, et al: A FTH1 gene:pseudogene:microRNA
6 network regulates tumorigenesis in prostate cancer. *Nucleic Acids Res* 46:1998-2011, 2018
- 7 24. Zang W, Wang T, Wang Y, et al: Knockdown of long non-coding RNA TP73-
8 AS1 inhibits cell proliferation and induces apoptosis in esophageal squamous cell
9 carcinoma. *Oncotarget* 7:19960-74, 2016
- 10 25. Wei Y, Chang Z, Wu C, et al: Identification of potential cancer-related
11 pseudogenes in lung adenocarcinoma based on ceRNA hypothesis. *Oncotarget* 8:59036-
12 59047, 2017
- 13 26. Milligan MJ, Lipovich L: Pseudogene-derived lncRNAs: emerging regulators of
14 gene expression. *Front Genet* 5:476, 2014
- 15 27. Bateman A, Birney E, Durbin R, et al: The Pfam protein families database.
16 *Nucleic Acids Res* 28:263-6, 2000
- 17 28. Finn RD, Mistry J, Schuster-Bockler B, et al: Pfam: clans, web tools and
18 services. *Nucleic Acids Res* 34:D247-51, 2006
- 19 29. Chirag Jain SK: Fine-grained GPU parallelization of pairwise local sequence
20 alignment. Presented at the 21st International Conference on High Performance Computing
21 (HiPC, 2014)
- 22 30. Soroceanu L, Matlaf L, Khan S, et al: Cytomegalovirus Immediate-Early
23 Proteins Promote Stemness Properties in Glioblastoma. *Cancer Res* 75:3065-76, 2015
- 24 31. Pei B, Sisu C, Frankish A, et al: The GENCODE pseudogene resource. *Genome*
25 *Biology* 13:R51, 2012
- 26 32. Zhang Z, Carriero N, Zheng D, et al: PseudoPipe: an automated pseudogene
27 identification pipeline. *Bioinformatics* 22:1437-1439, 2006
- 28 33. Lynch M, Conery JS: The evolutionary fate and consequences of duplicate
29 genes. *Science* 290:1151-5, 2000
- 30 34. Baertsch R, Diekhans M, Kent WJ, et al: Retrocopy contributions to the
31 evolution of the human genome. *BMC Genomics* 9:466, 2008
- 32 35. Zhang Y: PseudoFuN GitHub.
33 https://github.com/yanzhanglab/PseudoFuN_app, 2018
- 34 36. Johnson TS, Li S, Franz E, et al: PseudoFuN.
35 https://integrativeomics.shinyapps.io/pseudofun_app/, 2018
- 36 37. Miranda KC, Huynh T, Tay Y, et al: A pattern-based method for the
37 identification of MicroRNA binding sites and their corresponding heteroduplexes. *Cell*
38 126:1203-17, 2006
- 39 38. Krek A, Grun D, Poy MN, et al: Combinatorial microRNA target predictions.
40 *Nat Genet* 37:495-500, 2005
- 41 39. Agarwal V, Bell GW, Nam JW, et al: Predicting effective microRNA target sites
42 in mammalian mRNAs. *Elife* 4, 2015
- 43 40. Grossman RL, Heath AP, Ferretti V, et al: Toward a Shared Vision for Cancer
44 Genomic Data. *N Engl J Med* 375:1109-12, 2016
- 45 41. Carithers LJ, Moore HM: The Genotype-Tissue Expression (GTEx) Project.
46 *Biopreserv Biobank* 13:307-8, 2015

- 1
2
3
4 1 42. Center OS: Ohio Supercomputer Center. Columbus OH, Ohio Supercomputer
5 2 Center, 1987
6 3
7 3 43. Altschul SF, Gish W, Miller W, et al: Basic local alignment search tool. J Mol
8 4 Biol 215:403-10, 1990
9 5 44. Zerbino DR, Achuthan P, Akanni W, et al: Ensembl 2018. Nucleic Acids Res
10 6 46:D754-D761, 2018
11 7 45. Ensembl: Ensembl Biomart. ensembl.org/biomart/martview, 2018
12 8 46. Hagberg A, Swart P, Schult D: Exploring network structure, dynamics, and
13 9 function using NetworkX, Los Alamos National Lab.(LANL), Los Alamos, NM (United
14 10 States), 2008
15 11 47. Harrow J, Frankish A, Gonzalez JM, et al: GENCODE: the reference human
16 12 genome annotation for The ENCODE Project. Genome Res 22:1760-74, 2012
17 13 48. Echols N, Harrison P, Balasubramanian S, et al: Comprehensive analysis of
18 14 amino acid and nucleotide composition in eukaryotic genomes, comparing genes and
19 15 pseudogenes. Nucleic Acids Res 30:2515-23, 2002
20 16 49. Johnson TS, Li S, Kho JR, et al: Network analysis of pseudogene-gene
21 17 relationships: from pseudogene evolution to their functional potentials. Pac Symp
22 18 Biocomput 23:536-547, 2018
23 19 50. Karro JE, Yan Y, Zheng D, et al: Pseudogene.org: a comprehensive database
24 20 and comparison platform for pseudogene annotation. Nucleic Acids Res 35:D55-60, 2007
25 21 51. [pseudogenes.org: psiDr. pseudogenes.org/psiDr/similarity.dat](http://pseudogenes.org/psiDr)
26 22 52. [pseudogenes.org: psiCube. http://pseudogene.org/psicube/](http://pseudogenes.org/psiCube/)
27 23 53. Ashburner M, Ball CA, Blake JA, et al: Gene Ontology: tool for the unification
28 24 of biology. Nature genetics 25:25, 2000
29 25 54. Cancer Genome Atlas Research N, Weinstein JN, Collisson EA, et al: The
30 26 Cancer Genome Atlas Pan-Cancer analysis project. Nat Genet 45:1113-20, 2013
31 27 55. Hudak D, Johnson D, Chalker A, et al: Open OnDemand: A web-based client
32 28 portal for HPC centers.
33 29 56. Smith TF, Waterman MS: Identification of common molecular subsequences. J
34 30 Mol Biol 147:195-7, 1981
35 31 57. Carmona U, Li L, Zhang L, et al: Ferritin light-chain subunits: key elements for
36 32 the electron transfer across the protein cage. Chem Commun (Camb) 50:15358-61, 2014
37 33 58. Wu T, Li Y, Liu B, et al: Expression of Ferritin Light Chain (FTL) Is Elevated in
38 34 Glioblastoma, and FTL Silencing Inhibits Glioblastoma Cell Proliferation via the
39 35 GADD45/JNK Pathway. PLoS ONE 11:e0149361, 2016
40 36 59. Kuleshov MV, Jones MR, Rouillard AD, et al: Enrichr: a comprehensive gene
41 37 set enrichment analysis web server 2016 update. Nucleic Acids Res 44:W90-7, 2016
42 38 60. Lachmann A, Torre D, Keenan AB, et al: Massive mining of publicly available
43 39 RNA-seq data from human and mouse. Nat Commun 9:1366, 2018
44 40 61. Li N, Miao Y, Shan Y, et al: MiR-106b and miR-93 regulate cell progression by
45 41 suppression of PTEN via PI3K/Akt pathway in breast cancer. Cell Death Dis 8:e2796, 2017
46 42 62. Guo C, Song WQ, Sun P, et al: LncRNA-GAS5 induces PTEN expression
47 43 through inhibiting miR-103 in endometrial cancer cells. J Biomed Sci 22:100, 2015
48 44 63. Geng L, Sun B, Gao B, et al: MicroRNA-103 promotes colorectal cancer by
49 45 targeting tumor suppressor DICER and PTEN. Int J Mol Sci 15:8458-72, 2014
50
51
52
53
54
55
56
57
58
59
60
61
62
63
64
65

- 1
2
3
4 1 64. Yang C, Wu D, Gao L, et al: Competing endogenous RNA networks in human
5 2 cancer: hypothesis, validation, and perspectives. *Oncotarget* 7:13479-90, 2016
6 3
7 3 65. Cui SY, Wang R, Chen LB: MicroRNA-145: a potent tumour suppressor that
8 4 regulates multiple cellular pathways. *J Cell Mol Med* 18:1913-26, 2014
9 5
10 5 66. Persano L, Moserle L, Esposito G, et al: Interferon-alpha counteracts the
11 6 angiogenic switch and reduces tumor cell proliferation in a spontaneous model of prostatic
12 7 cancer. *Carcinogenesis* 30:851-60, 2009
13 8
14 8 67. Welch JD, Baran-Gale J, Perou CM, et al: Pseudogenes transcribed in breast
15 9 invasive carcinoma show subtype-specific expression and ceRNA potential. *BMC Genomics*
16 10 16:113, 2015
17 11
18 11 68. Costa V, Esposito R, Aprile M, et al: Non-coding RNA and pseudogenes in
19 12 neurodegenerative diseases: "The (un)Usual Suspects". *Front Genet* 3:231, 2012
20 13
21 13 69. Long MD, Singh PK, Russell JR, et al: The miR-96 and RARgamma signaling
22 14 axis governs androgen signaling and prostate cancer progression. *Oncogene*, 2018
23 15
24 15 70. Xu J, Feng L, Han Z, et al: Extensive ceRNA-ceRNA interaction networks
25 16 mediated by miRNAs regulate development in multiple rhesus tissues. *Nucleic Acids Res*
26 17 44:9438-9451, 2016
27 18
28 18 71. Costa-Pinheiro P, Ramalho-Carvalho J, Vieira FQ, et al: MicroRNA-375 plays a
29 19 dual role in prostate carcinogenesis. *Clin Epigenetics* 7:42, 2015
30 20
31 20 72. Wang Y, Lieberman R, Pan J, et al: miR-375 induces docetaxel resistance in
32 21 prostate cancer by targeting SEC23A and YAP1. *Mol Cancer* 15:70, 2016
33 22
34 22 73. MacKeigan JP, Murphy LO, Blenis J: Sensitized RNAi screen of human kinases
35 23 and phosphatases identifies new regulators of apoptosis and chemoresistance. *Nat Cell Biol*
36 24 7:591-600, 2005
37 25
38 25 74. Johnson TS, Li S, Franz E, et al: Supporting data for "PseudoFuN: Deriving
39 26 functional potentials of pseudogenes from integrative relationships with genes and
40 27 miRNAs across 32 cancers". *Gigascience Database* 2019. <http://dx.doi.org/10.5524/100577>
41 28
42 28 75. F.R.S. RAF: Tests of significance in harmonic analysis. *Proceedings of the*
43 29 *Royal Society of London. Series A* 125:54, 1929
44 30
45 30 76. Alexa A RJ: Gene set enrichment analysis with topGO.
46 31 <http://www.bioconductor.org>, Bioconductor, 2009
47 32
48 32 77. Stelzer G, Rosen N, Plaschkes I, et al: The GeneCards Suite: From Gene Data
49 33 Mining to Disease Genome Sequence Analyses. *Curr Protoc Bioinformatics* 54:1 30 1-1 30
50 34 33, 2016
51 35
52
53
54
55
56
57
58
59
60
61
62
63
64
65



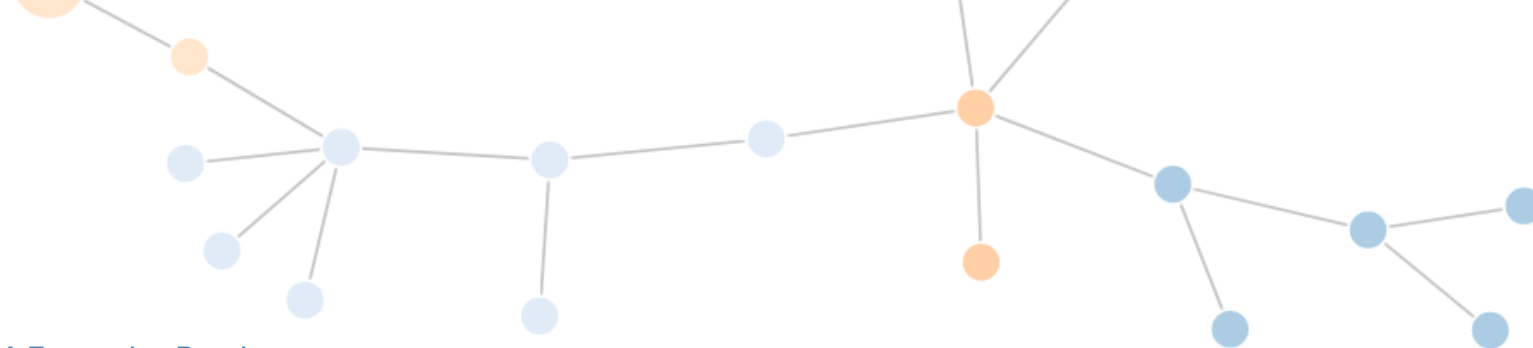




A

Gene: SOX15: ENSG00000129194

Pseudogene: PPP4R1L: ENST00000422302.2

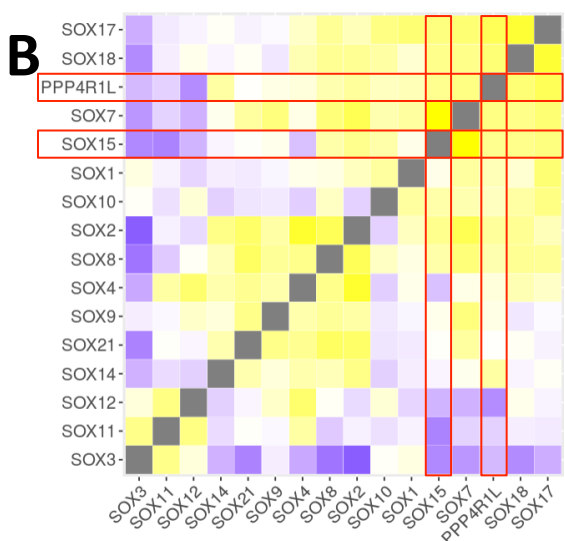


TCGA Expression Panel

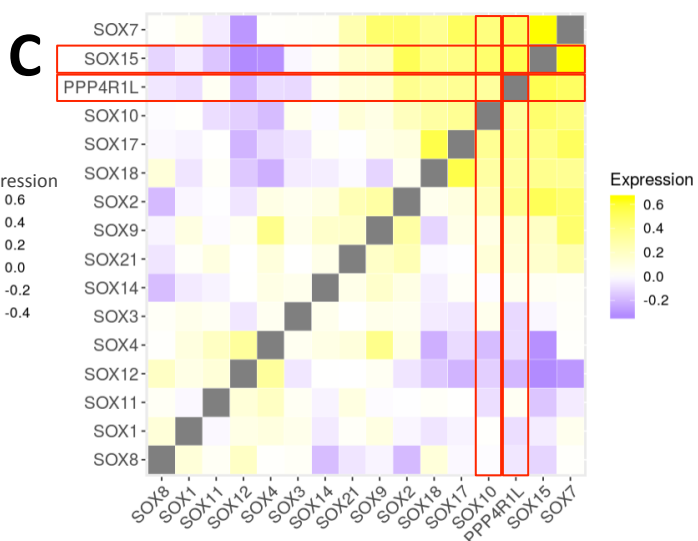
Gene: sox15; Database: CUAlign18; Cancer: PRAD; Network: 1.

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

Normal Sample Coexpression

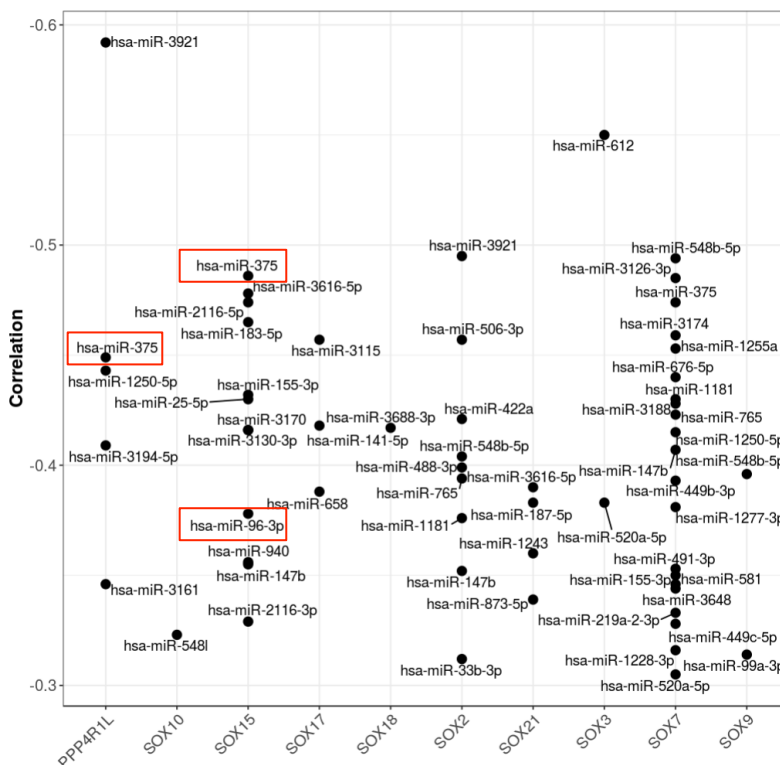


Tumor Sample Coexpression



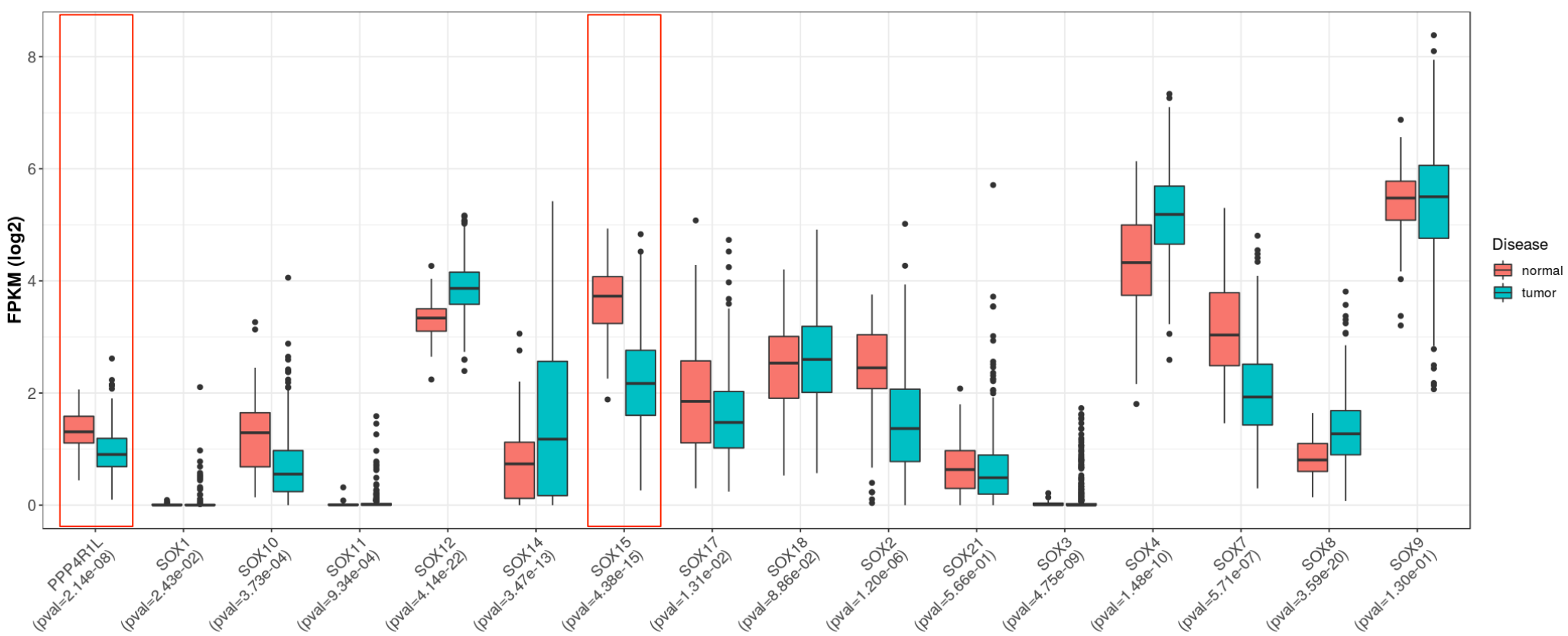
Gene and Pseudogene miRNA Associations

D



Differential Expression Tumor vs. Normal

E





[Click here to access/download](#)

Supplementary Material

PseudoFuN_suppl_Giga_20190328.docx





THE OHIO STATE UNIVERSITY

March 28, 2019

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

Dear Editor,

Here attached is our manuscript “PseudoFuN: Deriving functional potentials of pseudogenes from integrative relationships with genes and miRNAs across 32 cancers” after minor revision. We have solved all the editorial issues and addressed all the left points raised by the reviewers.

We thank you and the reviewers for all the great suggestions that have helped us make the project better. And we hope that this improved version is ready to be handed over to the journal production team. We look forward to your feedback.

Sincerely,



Yan Zhang, Ph.D.
Assistant Professor
Department of Biomedical Informatics
The Ohio State University

Response Letter

Response to editorial issues:

1) Please include a citation to your new GigaDB repository to the bibliography, and cite it from the "data availability" section (the same way as you have done it for the GitHub site).

The citation is:

[xx] Johnson T; Li S; Franz E; Huang Z; Li SD; Campbell MJ; Huang K; Zhang Y Supporting data for "PseudoFuN: Deriving functional potentials of pseudogenes from integrative relationships with genes and miRNAs across 32 cancers" GigaScience Database 2019. https://urldefense.proofpoint.com/v2/url?u=http-3A_dx.doi.org_10.5524_100577&d=DwIGaQ&c=k9MF1d71ITtkuJx-PdWme51dKbmfPEvxwt8SFEkBs4&r=ekJXXuaVVww1UTqb5FGVAqKm1D8j8S7Zredd9hAkY7k&m=-JC_daUTsqZc2BjqT0Y7Gk63PpDti3noA1ESRxQKWG4&s=SyHXIQfP5-vHlv7kVDFTMIWU8RPnS3jS-u9bzUGKalg&e=

In the data availability section, please write, for example, "Supporting data and an archival copy of the code are available via the GigaScience database, GigaDB [xx]", or similar wording.

Authors' answer: Done.

2) Please include a section "Availability of supporting source code and requirements", formatted as described in our Instructions for authors.

List the following:

Project name: e.g. My bioinformatics project

Project home page: e.g. https://urldefense.proofpoint.com/v2/url?u=https-3A_github.com_ISA-2Dtools&d=DwIGaQ&c=k9MF1d71ITtkuJx-PdWme51dKbmfPEvxwt8SFEkBs4&r=ekJXXuaVVww1UTqb5FGVAqKm1D8j8S7Zredd9hAkY7k&m=-JC_daUTsqZc2BjqT0Y7Gk63PpDti3noA1ESRxQKWG4&s=NtadplgDdzem-mFjrMEk2A FtIP9KrILUbNh9_7DiCc&e=

Operating system(s): e.g. Platform independent

Programming language: e.g. Java

Other requirements: e.g. Java 1.3.1 or higher, Tomcat 4.0 or higher

License: e.g. GNU GPL, FreeBSD etc.

RRID: if applicable, e.g. RRID: SCR_014986

==

Please register any new software application in the https://urldefense.proofpoint.com/v2/url?u=http-3A_SciCrunch.org&d=DwIGaQ&c=k9MF1d71ITtkuJx-PdWme51dKbmfPEvxwt8SFEkBs4&r=ekJXXuaVVww1UTqb5FGVAqKm1D8j8S7Zredd9hAkY7k&m=-

<https://pubmed.ncbi.nlm.nih.gov/35444444/> database to receive a RRID (Research Resource Identification Initiative ID) number, and include this in your manuscript. This will facilitate tracking, reproducibility and re-use of your tool.

=

Related to software availability, you write "we provide another interactive web application 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." - please add details in the manuscript how this web application can be accessed.

Authors' answer: All done. We have added a section and made a Research Resource Identification Initiative ID, SCR_017095.

3) Please remove any highlighting in red that was made for the purpose of peer review.

Authors' answer: Done.

4) Please also note the minor comment of the reviewer (below)

Authors' answer: Yes. We have added the missing highlighting color back into Table 2.

5) Please submit your revised version of the text in Word /LaTeX (whatever you used to prepare the document) - we only seem to have a PDF version at the moment.

Authors' answer: We will submit Word text with this revision.

PS: We added one affiliation for the corresponding author, and modified the Acknowledgements in the main text.

Response to reviewer reports:

Reviewer #2: I appreciate the effort that the authors put in to addressing my comments and added the additional benchmarking table 2.

One minor comment concerning table 2 is that the highlighting described in the figure could not be seen in the table, otherwise I have no further amendments.

Authors' answer: We appreciate the great suggestions from the reviewers. We have added the missing highlighting color back into Table 2.