A peer-reviewed version of this preprint was published in PeerJ on 1 February 2019.

View the peer-reviewed version (peerj.com/articles/6374), which is the preferred citable publication unless you specifically need to cite this preprint.

Gilbert DG. 2019. Genes of the pig, *Sus scrofa*, reconstructed with EvidentialGene. PeerJ 7:e6374 https://doi.org/10.7717/peerj.6374

Genes of the Pig, *Sus scrofa*, reconstructed with EvidentialGene

Donald G Gilbert Corresp. 1

1 Indiana University, Bloomington, IN, United States

Corresponding Author: Donald G Gilbert
Email address: gilbertd@indiana.edu

The pig is a well studied model animal of biomedical and agricultural importance. Genes of this species, *Sus scrofa*, are known from experiments and predictions, and collected at the NCBI Reference Sequence database section. Gene reconstruction from transcribed gene evidence of RNA-seq now can accurately and completely reproduce the biological gene sets of animals and plants. Such a gene set for the pig is reported here, including human orthologs missing from RefSeq and other improvements to the current NCBI pig gene set. Methodology for accurate and complete gene set reconstruction from RNA is used: the automated SRA2Genes pipeline of EvidentialGene project.

Genes of the Pig, *Sus scrofa*, reconstructed with EvidentialGene

Author: Donald G. Gilbert

Affiliation: Indiana University, Bloomington, IN, USA

Email address: gilbertd@indiana.edu or gilbert.bionet@gmail.com

Abstract

The pig is a well studied model animal of biomedical and agricultural importance. Genes of this species, *Sus scrofa,* are known from experiments and predictions, and collected at the NCBI Reference Sequence database section. Gene reconstruction from transcribed gene evidence of RNA-seq now can accurately and completely reproduce the biological gene sets of animals and plants. Such a gene set for the pig is reported here, including human orthologs missing from RefSeq and other improvements to the current NCBI RefSeq pig gene set. Methodology for accurate and complete gene set reconstruction from RNA is used: the automated SRA2Genes pipeline of EvidentialGene project.

Introduction

Precision genomics is essential in medicine, environmental health, sustainable agriculture, and research in biological sciences (eg. Goldfeder et al. 2016). Yet the popular genome informatics methods lag behind the high levels of accuracy and completeness in gene construction that are attainable with today's accurate RNA-seq data.

To demonstrate the accuracy and completeness of gene set reconstruction from expressed gene pieces (RNA-seq) alone, excluding chromosome DNA or other species genes, the pig is a good choice. The pig has a well constructed, partly curated NCBI RefSeq gene set, one of 7 RefSeq top-level model organisms, and is based on extensive expressed sequences dating from the 1990s. Pig has a well-assembled chromosome set from BAC-clone Sanger + Illumina hybrid sequencing in the 2000s (Groenen et al. 2012), and contributions of experimental gene evidence from many projects of agricultural and biomedical focus. As well, published RNA-seq of the pig from over 2,000 studies weighs in the top 10 of model animals and plants. Yet there is just one pubic transcript assembly from these many pig studies, from blood samples only.

If successful, this demonstration can be used to improve RefSeq genes for this species. It will demonstrate to others how to produce reliably accurate gene sets. Unreliability is a continuing problem in gene set reconstructions, whether from RNA assembly or chromosome-based gene modeling. Reasons for this failure are many; the EvidentialGene project aims to provide a solution that others can use, simple in concept, always obtaining an accurate gene set picture from a puzzle box full of gene pieces.

Gene sets reconstructed by the author are more accurate by objective measures of homology and expression recovery, than those of the same species produced by popular methods, including NCBI Eukaryotic Genome Annotation Pipeline (EGAP, Thibaud-Nissen et al. 2013), Ensembl gene annotation pipeline, MAKER genome-gene modeling (Holt & Yandell 2001), Trinity RNA assembly (Grabherr et al 2001), Pacific Biosciences long RNA assemblies, and others. These improved gene sets include popular and model animals and plants such as *Arabidopsis*, *Zea mays* corn, *Th. cacao* chocolate tree, *Danio* zebrafish, *Fundulus* killifish, *Aedes* and *Anopheles* mosquitos, *Nasonia* jewel wasp, *Daphnia* water fleas (Gilbert 2012, 2013, 2016, 2017).

Gene sets reconstructed by others using EvidentialGene methods are also more accurate (Nakasugi et al. 2014, Mamrot et al. 2017), in independent assessments. However some investigators do not apply necessary details of EvidentialGene methodology, or modify portions in ways that reduce accuracy. One impetus for this work is engineering a full, automated pipeline that others can more readily use, for these validated methods.

Methods

EvidentialGene uses several gene modeling and assembly methods, annotates their results with evidence, then classifies and reduces this over-assembly to a set of loci that best recovers the gene evidence. Each modeling method has qualities that others lacks, and produces models with better gene evidence recovery.

Gene reconstruction steps are (Gilbert 2012): 1. produce several predictions and transcript assembly sets with quality models. 2. annotate models with available gene evidence (transcript introns, exons, protein homology, transposon and other). 3. score models with weighted sum of evidence. 4. remove models below minimum evidence score. 5. select from overlapped models at each locus the highest score, and alternate isoforms, including fusion metrics (longest is not always best). 6. evaluate resulting best gene set (i.e. compare to other sets, examine un-recovered gene evidence). 7. re-iterate the above steps with alternate scoring to refine. Evidence criteria for genes are, in part, protein homology, coding/non-coding ratio, RNA read coverage, RNA intron recovery, and transcript assembly equivalence.

For RNA-only assembly, this paradigm is refined at step 2-4 to introduce a coding-sequence classifier (Gilbert 2013), which reduces large over-assembly sets (e.g., 10 million models of 100,000 biological transcripts) efficiently, using only the self-referential evidence of coding sequence metrics (protein length and completeness, UTR excess).

CDS overlap by self-alignment identifies putative gene loci and their alternate transcripts, similarly to how CDS overlap by alignment to chromosomal DNA is used in traditional genome-gene modeling to classify loci. This CDS classifier, in tr2aacds.pl pipeline script, uses the observed high correlation between protein completeness and homology completeness, making a computationally efficient classifier that will reduce the large over-assembly set to one small enough that the additional evidence classifications are feasible to refine this rough gene set to a finished one, using evidence of protein homology, expression validity, chromosomal alignment, and others.

A fully automated pipeline that includes the above, SRA2Genes, is introduced for this pig gene reconstruction. It includes RNA-seq data fetching from NCBI SRA, over-assembly of these data by several methods and parameters, transcript assembly reduction with coding-sequence classifier, protein homology measurement, sequencing vector and contamination screening, gene annotation to publication quality sequences, and preparation for submission to transcript shotgun assembly archive (TSA).

For pig gene set reconstruction, four RNA source projects were selected, from public RNA in SRA, based on tissue sampling, methodology (all paired-end reads of recent Illumina sequencers), and other factors. See “Data and Software Citations” section for details. Each project RNA set was assembled using SRA2Genes, to the step of non-redundant gene set with alternate isoforms. Then a superset of the best of these projects is produced, using these four reduced assemblies as input over-assembly to SRA2Genes, stepping from assembly reduction to

gene set annotation and publication. Assemblies were done in stages, assessing completeness then adding source data to improve that. Many more RNA source projects are available to improve this set. There are data choice problems for this and other large vertebrates, where most RNA samples are for specific tissues, often for mutant strains, with limited sample documentation. Collecting from public RNA samples to include all expressible genes can be difficult, with some tissue, stage or stress-specific genes missing or weakly expressed.

RNA data source projects are *pig1a:* PRJNA416432 (China Agricultural University), *pig2b:* PRJNA353772 (Iowa State University, USDA-ARS), *pig3c:* PRJEB8784 (Univ. Illinois), and *pig4e*: PRJNA255281 (Jiangxi Agricultural University, Nanchang, China). These comprising 26 read sets of 1,157,824,292 read pairs, or 106,654 megabases. All these are paired-end reads, from Illumina, ranging from 75bp to 150bp read length. PRJEB8784 includes adult female and male tissues of muscle, liver, spleen, heart, lung, kidney. PRJNA416432 includes adult female tissues of two sample types. PRJNA353772 includes tissue samples of brain, liver, pituitary, intestine, and others. PRJNA255281 provides embryonic tissue RNA. Notably missing were head sensory organs, one result being that some eye, ear, nose and taste receptor genes are under-represented or fragmented in this reconstruction.

Assembler software used includes Velvet/Oases (Schulz et al. 2012), idba_trans (Peng et al. 2013), SOAPDenovoTrans (Xie et al. 2013), Trinity (Grabherr et al. 2011), and rnaSPAdes (Bankevich et al. 2012). K-mer shred sizes were selected to span the read sizes, and as observed in many RNA assemblies, 1/2 read-size produces the single most complete set, however most k-mer sizes produce some better gene assemblies, due to wide variation in expression levels and other factors (strongly expressed, long genes tend to assemble well with large k-mer). Both non-normalized and digitally normalized RNA sets were used; each way produces a different set of accurate genes.

Additional assemblies with rnaSPAdes were targeted to unfinished genes, after reference homology measurements identified gene models that were incomplete. Prior work with several methods of assembly extension have proved unreliable, including assemblers Oases, SOAP and idba. These typically work to extend fragments by sequence overlap alone, but rarely produce longer coding sequences, instead indel errors and fused genes are frequent artifacts. rnaSPAdes, unlike the others, uses a graph of paired reads to extend partial transcripts, and may prove more successful.

Results

Data and software result public access: An open access, persistent repository of this annotated pig gene data set is at https://scholarworks.iu.edu/ with DOI 10.5967/K8DZ06G3. Transcriptome Shotgun Assembly accession is DQIR01000000 at DDBJ/EMBL/GenBank, BioProject PRJNA480168, for these annotated transcript sequences. Preliminary gene set is at http://eugenes.org/EvidentialGene/vertebrates/pig/pig18evigene/ . EvidentialGene software package is available at http://eugenes.org/EvidentialGene/ and at http://sourceforge.net/projects/evidentialgene/ .

The results of gene assembly for each of 4 data sources are summarized as *pig1a* 11,691,549 assemblies reduced to 595,497 non-redundant coding sequences (5%), *pig2b* 3,984,284 assemblies reduced to 404,908 (10%), *pig3c* 8,251,720 assemblies reduced to 564,523 (7%), and *pig4e*, a smaller embryo-only RNA set, of 1,955,018 assemblies to 134,156 (7%). These 4 reduced assemblies are then used in secondary runs of SRA2Genes, stepping from assembly inputs. Several secondary runs were performed, with reference homology assessment, to ensure all valid homologs are captured. Additional assemblies with rnaSPAdes of incomplete genes improved some of the fragment models (16,168 or 5% of final transcripts, including 1571 loci with best homology).

The final gene set is summarized in Table 1 by categories of gene qualities and evidences. The number of retained loci include all with measurable homology to 4 related vertebrate species gene sets, and a set of non-homolog, but expressed with introns in gene structure, two forms of gene evidence that provide a reliable criterion. The number with homology is similar to that of RefSeq genes for pig. The expressed, multi-exon genes add 15,000 loci, which may be biologically informative in further studies. The pig RefSeq gene set has 63,586 coding-sequence transcripts at 20,610 loci, of which 5,177 have exceptions to chromosome location (indels, gaps and RNA/DNA mismatch).

The extended gene data set includes culled transcript sequences, which do not meet criteria for homology or unique expression, but which pass other criteria for unique transcripts: 92627 culled loci, and 175,793 culled alternate transcripts. Further evidence may indicate some of these are valid. The published gene data set includes mRNA, coding and protein sequences in FastA format for the public set (pig18evigene_m4wf.public mrna, cds and aa), and the culled set (pig18evigene_m4wf.xcull_ mrna, cds and aa). There are two sequence object-annotation tables, pig18evigene_m4wf.pubids (gene locus and alternate public ids, object ids, class, protein and homology attributes), and pig18evigene_m4wf.mainalt.tab (locus main/alternate linkage for original object ids). A gene annotation table pig18evigene_m4wf.ann.txt contains public ids, name, protein, homology, database cross references, and chromosome location annotations. Chromosome assembly locations to RefSeq pig genome are given in pig18evigene_m4wf.mrna.gmap.gff in GFF version 3 format.

[insert **Table 1**. *Sus scrofa* (pig) gene set numbers, version Susscr4EVm]

The table 2a scores are measured against vertebrate conserved BUSCO subset of OrthoDB v9, and are counts relative to 2586 total conserved genes, but for the Align average in aminos. Full is the count of pig genes completely aligned to conserved proteins. Table 2b has scores for human gene alignments, percentages relative to all reference genes found in either pig set (n= 37,883), calculated from table of "blastp -query human.proteins -db two_pigsets.proteins -evalue 1e-5". These proteins include 19122 of 20191 (95%) of human gene loci. NCBI has 25% of best match, Evigene 20% of best, and 55% of comparisons are equal for the human proteins that are found in either pig gene set. Scores are Align = alignment to reference proteins, as percent of human gene, Frag = percent with fragment alignment, size < 50% of reference, Short = percent with size < 95% of reference, Miss = percent with no alignment, Best = percent or count of greater alignments in pairwise match to each reference gene. Supplemental tables 1 and 2 have the pair-wise pig gene alignment scores of summary tables 2a and 2b.

[insert **Table 2**. *Sus scrofa* gene sets compared for gene evidence recovery]

Average homology scores are nearly same for both of these gene sets, but they differ for individual loci. The "Best" columns in Table 2 indicate a subset of Evigene that can usefully improve the RefSeq gene set: 4,521 proteins have improved human gene homology to greater or lesser extent. 283 of Evigene improvements have no pig RefSeq equivalent, including the 9 vertebrate conserved BUSCO genes missing from the NCBI set. 121 of the improved coding genes are modeled as non-coding in RefSeq ([NX]R_ IDs), that can be better modeled as coding genes with exceptions in chromosome mapping. 548 have a RefSeq mRNA that is co-located with an Evigene model, but notably deficient in human gene alignment (i.e. a fragment or divergent model), while a majority of 1048 improvements have small, exon-sized differences, as alternate transcripts to existing RefSeq loci.

Many of the 15,000 putative genes that lack homology to human, cow, mouse or fish RefSeq genes do have homology by other measures. With non-redundant NCBI protein database, 11% of these have a significant match, to uncharacterized genes in other mammals or vertebrates, or endonuclease/reverse transcriptase transposon-like proteins, or as fragment alignments to characterized proteins. Coding alignment of these putative genes to the cow (*Bos taurus*) chromosome set, and calculation of synonymous/non-synonymous substitutions (Ka/Ks), identifies from 13% to 28% have coding sequence conservation, the majority not identified as having protein homology in the other tests. These putative genes may include recently duplicated and modified coding genes, ambiguous non-coding/coding genes, as well as fragments of other genes, putative transposon residue, and untranslated but expressed genome regions.

[insert **Table 3**. Assembler method effects on Human reference gene recovery in Pig gene sets] The table 3a scores are for alignments to human gene with blastp, subset by assembler method for data of Bioproject PRJNA416432. Table 3b scores for a second pig project are also subset by methods for alignment to human genes . This second project collected both Illumina RNA-seq (75bp paired reads) and PacBio (<1-2kb, 2-3kb, 3-5kb, >5kb single reads from Pacific Biosciences instrument) from the same set of tissue samples. This PacBio assembly, which includes improvement using the Illumina RNA with Proovread, was done by the project authors, and published in SRA, under Bioproject ID PRJNA351265, while the Illumina RNA is under Bioproject PRJNA353772.

The major option used for these various assemblies is k-mer size, the sub-sequence length for placing reads in the assembly graph structure. Different genes are best assembled with different k-mer sizes, depending on expression level, gene complexity, and other factors, that indicates why many assemblies of the same data but different options result in a larger set of accurate gene reconstructions. For Table 3a sample, with read size of 150 bp, k-mers from 25 to 125 were used. k-mer of 105 returned the most accurate genes, for both velvet and idba methods. The range k70..k125 produced 5/10 of best models, range k40..k65 produced 4/10, and range k25..k35 the remaining 1/10 of best models. The popular Trinity method underperforms all others, due part to its limited low k-mer option.

Sample 2 (Table 3b) demonstrates the value of assembling accurate gene pieces (Illumina, 80% of reads have highest quality score in SRA), over inaccurate but longer sequences (PacBio, 15%

of reads have highest quality score in SRA). This project sequenced pig RNA with both technologies, and PacBio assembly software plus Illumina RNA to improve PacBio sequence quality, to produce a gene set that is less accurate than that produced from the Illumina-only RNA, assembled with a competent short-read assembler.

Discussion

The main result of this demonstration compared with the NCBI RefSeq pig gene set is, on average, they are equally valid by homology measures, but differ at many gene loci, with Evigene adding many alternate transcripts. The Evigene set also retains more putative loci, lacking measured homology but with other evidence, that further study will clarify their value. Improvements to the pig gene set are numerous enough to warrant updating RefSeq with those from this work. These include 1,500 missing or poorly modeled genes with homology to human, and improved vertebrate conserved genes. Between RefSeq and Evigene sets, all highly conserved vertebrate genes of the BUSCO set exist in pig. The another 3,000 improvements are mostly alternate transcripts with greater alignment to other species, by changes in an exon or two.

This Evigene set has demonstrated objectively accurate gene assemblies that improve the reference gene set of the pig model organism. It has been submitted for that purpose to NCBI as a third party annotation/assembly (TPA) of a transcriptome shotgun assembly (TSA), which are International Nucleotide Sequence Database Collaboration (INSDC) classifications. There are policy reasons to limit inferential or computational TPA entries, and there are also policy reasons to accept these. On one hand, objectively accurate gene and chromosome assemblies of experimental RNA and DNA fragments are the desired contents of public sequence databases. On the other hand, having many assemblies of the same RNA or DNA fragments is confusing and could overwhelm databases devoted to experimentally derived genome sequences. This pig gene set adheres to the described policy of TPA in that (a) it is assembled from primary data already represented in the INSDC databases (SRA sequence read section); (b) it is indirectly experimentally supported by reference gene homology measures; (c) it is published in a peer-reviewed scientific journal. Additionally this gene set provides thousands of improvements to the reference gene set. The author produced no wet-lab experimental evidence, but has assembled gene sequence evidence from several sources into a gene set that substantially improves upon NCBI EGAP and Ensembl gene sets. Review of this data set, by NCBI and independent peers, weighs the above dilemma: improve public genome sequences or limit independent computational assemblies.

Combining and selecting by evidence criteria the assemblies of several methods improves gene reconstruction to a higher level of accuracy. The individual methods return from 77% (Velvet) down to 50% (Trinity) of the best gene models, and a hybrid PacBio+Illumina assembly is intermediate at 66%. K-mer sizes are an important parameter, as noted by others: "smaller values of k collapse more repeats together, making the assembly graph more tangled. Larger values of k may fail to detect overlaps between reads, particularly in low coverage regions, making the graph more fragmented." (SPAdes, Bankevich et al 2012). Alternate isoforms of each gene, which share exons and differ in expression levels, are more accurately distinguished

from other genes at large k-mer sizes (idba_tran, Peng et al. 2013). These results are consistent with multi-method reconstructions for arabidopsis, corn, zebrafish, mosquitos, and water fleas.

The main flaw in this Evigene pig set is incomplete reconstruction of many genes, especially longer ones. While this is not always a problem with RNA-only assemblies, it is a common one. Importantly, there does not appear to be a reliable method for improving gene assemblies identified as fragmentary, using de-novo RNA assembly. While there are several methods that attempt to address this, those tested by the author are unreliable. A trial of rnaSPAdes to extend fragments did improve some genes, but not as many as the RNA data warrants.

A second flaw in EvidentialGene's method of classifying loci from self-referential alignment of coding sequences, is that some paralogs are confused as alternate transcripts of the same locus. With high sequence identity, paralogs align to each other similarly to transcripts of one locus (a class termed "altpar" or "paralt"), though with mismatches that chromosome alignment can resolve. This has been measured at a rate of about 5% for reference gene sets of mouse and zebrafish, and 3% for arabidopsis; a smaller 0.5% portion of alternates at one locus are misclassified as paralog loci. Several de-novo gene assembly methods that classify loci have similar altpar confusion, as RNA-seq reads are often shared among paralogs as well as alternate transcripts. These altpar transcripts have not been resolved for this pig gene set, though it is an improvement in development.

This demonstration excluded the use of chromosomes and other species genes to assemble or extend assemblies. Both methods can be employed to advantage to reconstruct genes, where there are few errors in these additional evidences. An important reason to limit initial gene reconstruction to RNA-only assembly is to avoid compounding errors from several sources. This limited-palette reconstruction is validated with independent evidence from genomic DNA and other species sources; genes identified as mis-assembled, or missing, in such RNA-only sets can be improved with these other methods. Many discrepancies between RNA-only reconstruction and the other evidences are from flaws in chromosome assemblies or other species genes that can be identified with careful evaluations.

Gene transcripts from any source, such as EST and PacBio, may be added into SRA2Genes pipeline. Excluded from this reconstruction are the extensive public set of pig ESTs, and the PacBio+Illumina assembly of sample *pig2b*. These contribute a small number of improved transcripts not in this EvidentialGene set (8 missed human orthologs in ESTs, 12 EST and 24 PacBio with significant improvements), and are used in the RefSeq set. However as these are already in the public databases, this demonstration reconstruction adds no value to them.

While these gene data and paper were in review at repositories, Zhao P, X Zheng, Y Yu et al (2018) pre-published a reconstruction of pig genes, with newly sampled proteomic and transcriptomic sequences. The authors provide public access to these under BioProject PRJNA392949 for SRA RNA-Seq, and a bioRxiv preprint with sequences of 3,703 novel protein isoforms. The experimental design of this work is well suited to gene set reconstruction, as it sampled 34 tissues of adult male, female and juvenile pigs. Unlike the samples winnowed from prior SRA entries by this author, each from a pig portion, this new work is comprehensive in collecting expressed and translated genes.

Zhao and colleagues compare their results to the same RefSeq gene set and chromosome assembly as this paper. In brief, of the 3,700 novel proteins, most align to other gene sets and chromosome assembly: 74% are contained in this paper's transcripts, 65% are contained in the RefSeq transcript set, and 61% are contained in the pig chromosome set, at 75% or greater alignment (protein to RNA/DNA aligned with tBLASTn). None the less, most of these novel proteins do not have a protein equivalent in the gene sets: about 800 novel proteins align to Evigene proteins, and about 600 to Refseq proteins. A main difference here lies in measures from RNA to protein, including new alternate transcripts and discrepancies in RNA to protein reconstruction, rather than in newly identified gene loci, and is beyond scope of this note to resolve. A rough draft with SRA2Genes of this recent RNA-Seq, assembling only well-expressed genes, contains about the same 74% of novel proteins as for this paper's set. An application suited to SRA2Genes is to update with these completely sampled pig genes, including depositing an improved version to Transcriptome Shotgun Assembly public database for further uses.

Conclusions

The SRA2Genes pipeline is demonstrated, for the pig model organism, as a reliable gene reconstruction method, useful to other projects and for improving public reference gene sets. The resulting complete transcriptome assembly of pig fills a void at public repositories. Reconstruction from RNA only provides independent gene evidence, free of errors and biases from chromosome assemblies and other species gene sets. Not only are the easy, well known ortholog genes reconstructed well, but harder gene problems of alternate transcripts, paralogs, and complex structured genes are usually more complete with EvidentialGene methods.

Acknowledgements

XSEDE/TeraGrid shared computational resources, for a decade of development and implementation, Award# MCB100147, to Genome Informatics for Animals and Plants, D.G. Gilbert. IUScholarWorks staff, including Richard Higgins, for providing a permanent open-access repository of EvidentialGene animal and plant gene sets. NCBI GenBank submissions staff for reviewing effort to deposit TPA/TSA gene data sets.

Supplemental Information

Supplemental Table 1. Conserved vertebrate genes recovered in Pig Evigene vs NCBI gene sets, as computed with vertebrate conserved genes of OrthoDB v9, BUSCO and hmmer software. Columns include gene ids of BUSCO_ID, Evigene_ID, and NCBI RefSeq ID. Other columns: Cmp, the qualitative comparison (evgain, same, evloss) of alignment difference; Diff, numeric difference in alignment score to conserved protein; dEvg-Ncb, the two alignment scores; BC, the BUSCO complete/fragment/missing quality score; and Product_Name, the vertebrate protein product. File name: pig18evg_ncbi_busco.xlsx

Supplemental Table 2. Human genes recovered in Pig Evigene vs NCBI gene sets, as computed with human and pig RefSeq and Evigene proteins and NCBI BLASTP software. This includes only unique alignments of isoforms of both pig gene sets to isoforms of human genes. Columns include gene ids for Human RefSeq ID, Evigene_pig_ID, NCBI_pig_ID; AAsize, human protein size; EvAlign, NcAlign, alignment scores to Evigene and NCBI proteins; DiffA, difference in alignments; and Human_Gene_Name. File name: pig18evg_ncbi_human.xlsx

Data and Software Citations

NCBI pig gene set used in comparison, from ftp://ftp.ncbi.nlm.nih.gov/refseq/S_scrofa/mRNA_Prot/pig.1.rna.gbff.gz, accessed on 27 Apr 2018.

Ensembl pig gene set used in comparison, from ftp://ftp.ensembl.org/pub/current_fasta/sus_scrofa/pep/Sus_scrofa.Sscrofa11.1.pep.all.fa.gz, accessed on 28 Jul 2018.

NCBI RefSeq pig chromosome assembly Sscrofa11.1, accession: GCF_000003025.6, dated 2017-2-7, is used for chromosome mapping.

NCBI RefSeq gene sets used as reference genes are H_sapiens, M_musculus, B_taurus, and D_rerio, accessed at same location and date as pig genes.

RNA data sources with NCBI BioProject ID are
SRA data *pig1a:* PRJNA416432 (China Agricultural University),
SRA data *pig2b:* PRJNA353772 (Iowa State University, USDA-ARS),
SRA data *pig3c:* PRJEB8784 (Univ. Illinois),
SRA data *pig4e*: PRJNA255281 (Jiangxi Agricultural University, Nanchang, China).

The SRA read table of these data sets is the starting point for SRA2Genes, and provided at http://eugenes.org/EvidentialGene/vertebrates/pig/pig18evigene/

Expressed sequences of the pig from dbEST, by Sanger and 454 sequencing (max length 900 bases), from projects reported in PubMedID:14681463, dbEST n=304,418, and PubMedID: 17407547, dbEST n=716,260.
Vertebrate conserved single-copy genes, of OrthoDB v9 (http://www.orthodb.org), BUSCO.py software, with hmmer (v3.1, http://hmmer.org/).

Software components of EvidentialGene SRA2Genes:
fastq-dump, of sratoolkit281, https://www.ncbi.nlm.nih.gov/sra/docs/toolkitsoft/
blastn, blastp of https://blast.ncbi.nlm.nih.gov/
vecscreen, tbl2asn of http://ncbi.nlm.nih.gov/tools/vecscreen/, /genbank/tbl2asn2/
fastanrdb, of exonerate, https://www.ebi.ac.uk/about/vertebrate-genomics/software/exonerate
cd-hit, cd-hit-est, of https://github.com/weizhongli/cdhit/
normalize-by-median.py, of khmer, https://github.com/ged-lab/khmer
velvet, oases of velvet1210 assembler, https://www.ebi.ac.uk/~zerbino/oases/
idba_tran, of idba assembler, http://hku-idba.googlecode.com/files/idba-1.1.1.tar.gz
SOAPdenovo-Trans, http://soap.genomics.org.cn/SOAPdenovo-Trans.html
Trinity, of trinityrnaseq assembler, https://github.com/trinityrnaseq/trinityrnaseq
rnaSPAdes, of SPAdes assembler, http://cab.spbu.ru/software/spades/

International Nucleotide Sequence Database Collaboration (INSDC) policy documents pertaining to these data:
About TSA, https://www.ncbi.nlm.nih.gov/genbank/TSA
About TPA, https://www.ncbi.nlm.nih.gov/genbank/TPA
TPA FAQ, https://www.ncbi.nlm.nih.gov/genbank/tpafaq
TPA-Inferential, https://www.ncbi.nlm.nih.gov/genbank/TPA-Inf

References

Bankevich A, S Nurk, D Antipov, A A. Gurevich, M Dvorkin, A S. Kulikov, V M. Lesin, S I. Nikolenko, S Pham, A D. Prjibelski, A V. Pyshkin, A V. Sirotkin, M Vyahhi, G Tesler, M A. Alekseyev, and P A. Pevzner. 2012. SPAdes: A New Genome Assembly Algorithm and Its Applications to Single-Cell Sequencing. J. Computational Biology, 19:5 pp. 455–477 DOI 10.1089/cmb.2012.0021

Gilbert D. 2012. Perfect Arthropod Genes constructed with Gigabases of RNA. 6th annual Arthropod Genomics Symposium. Kansas State U. F1000Research (poster) DOI 10.7490/f1000research.1112595.1

Gilbert D. 2013. Gene-omes built from mRNA seq not genome DNA. 7th annual arthropod genomics symposium. Notre Dame. F1000Research (poster) DOI 10.7490/f1000research.1112594.1

Gilbert D. 2016. Accurate & complete gene construction with EvidentialGene. Galaxy Community Conference 2016, Bloomington IN. F1000Research, 5:1567 (slide set). DOI 10.7490/f1000research.1112467.1

Gilbert D. 2017. Animal and Plant gene set reconstructions with EvidentialGene. http://arthropods.eugenes.org/EvidentialGene/about/evigene_plantsanimals_2017sum.html

Goldfeder, et al. 2016. Medical implications of technical accuracy in genome sequencing. Genome Medicine. DOI 10.1186/s13073-016-0269-0

456 Grabherr MG, Haas BJ, Yassour M, Levin JZ, Thompson DA, et al. 2011 Full- length
457 transcriptome assembly from RNA-Seq data without a reference genome. Nat Biotech 29:
458 644–652. DOI 10.1038/nbt.1883
459 Groenen MAM, et al. 2012. Analyses of pig genomes provide insight into porcine demography
460 and evolution. Nature, 491(7424): 393–398. DOI 10.1038/nature11622.
461 Holt C, Yandell M. 2011. MAKER2: an annotation pipeline and genome- database management
462 tool for second- generation genome projects. BMC Bioinformatics, 12:491 DOI
463 10.1186/1471-2105-12-491
464 Mamrot J, R Legaie, SJ. Ellery, T Wilson, T Seemann, DR. Powell, DK. Gardner, DW. Walker,
465 P Temple-Smith, AT. Papenfuss & H Dickinson. 2017. De novo transcriptome assembly for
466 the spiny mouse (*Acomys cahirinus*). Scientific Reports 7, A# 8996. DOI 10.1038/s41598-
467 017-09334-7
468 Nakasugi K, Crowhurst R, Bally J, Waterhouse P. 2014. Combining Transcriptome Assemblies
469 from Multiple De Novo Assemblers in the Allo-Tetraploid Plant *Nicotiana benthamiana*.
470 PLoS ONE 9(3): e91776. DOI 10.1371/journal.pone.0091776
471 Peng Y, Leung HC, Yiu S-M, Lv M-J, Zhu X-G, Chin FY. 2013. IDBA-tran: a more robust de
472 novo de Bruijn graph assembler for transcriptomes with uneven expression levels.
473 Bioinformatics 29:i326–i334; DOI 10.1093/bioinformatics/btt219
474 Schulz MH, Zerbino DR, Vingron M, Birney E. 2012. Oases: Robust de novo RNA-seq
475 assembly across the dynamic range of expression levels. Bioinformatics 28: 1086–1092. DOI
476 10.1093/bioinformatics/bts094
477 Simao FA, Waterhouse RM, Ioannidis P, Kriventseva EV & Zdobnov EM. 2015. BUSCO:
478 assessing genome assembly and annotation completeness with single-copy orthologs.
479 Bioinformatics 31: 3210-3212. DOI 10.1093/bioinformatics/btv351
480 Thibaud-Nissen F, A Souvorov, T Murphy, M DiCuccio, and P Kitts. 2013. NCBI Eukaryotic
481 Genome Annotation Pipeline. The NCBI Handbook [Internet]. 2nd edition.
482 https://www.ncbi.nlm.nih.gov/books/NBK169439/
483 Waterhouse RM, et al. 2013. OrthoDB: a hierarchical catalog of animal, fungal and bacterial
484 orthologs. Nucleic Acids Res. 2013:D358-65, DOI 10.1093/nar/gks1116
485 Xie Y, Wu G, Tang J, Luo R, Patterson J, et al. 2013. SOAPdenovo-Trans: De novo
486 transcriptome assembly with short RNA-Seq reads. Bioinformatics 30: 1660–1666. DOI
487 10.1093/bioinformatics/btu077
488 Zhao Q-Y, Wang Y, Kong Y-M, Luo D, Li X, et al. 2011. Optimizing de novo transcriptome
489 assembly from short-read RNA-Seq data: a comparative study. BMC Bioinformatics 12(Suppl
490 14):S2. DOI 10.1186/1471-2105-12-S14-S2
491 Zhao P, X Zheng, Y Yu, Z Hou, C Diao, H Wang, H Kang, C Ning, J Li, W Feng, W Wang, G E
492 Liu, B Li, J Smith, Y Chamba, J-F Liu. 2018. Mining unknown porcine protein isoforms by
493 tissue-based map of proteome enhances the pig genome annotation. bioRxiv preprint, Aug.
494 14, 2018; DOI: 10.1101/391466.

Table 1(on next page)

pig18evg_datadesc.pages *Sus scrofa* (pig) gene set numbers, version Susscr4EVm

Table 1. *Sus scrofa* (pig) gene set numbers, version Susscr4EVm

39879 gene loci, all supported by RNA-seq, most also have protein homology evidence
39879 (100%) are protein coding, 0 are non-coding
All genes (100%) are assembled from RNA evidence, 0 are genome-modeled
25383/39879 (64%) have protein homology to other species genes.
316491 alternate transcripts are at 25512 (64%) loci, with 5 median, 12.4 ave, transcripts per locus, with 756 alts maximum, 1079 loci have 50+ alts, 8453 have 10+ alts,
27473 (69%) have complete proteins, 12406 have partial proteins, of 39879 coding genes
37918 (95%) are properly mapped to chromosome assembly (>=80% align),
1144 partial-mapped coverage (10% < align <80%),
817 are ~un-mapped genes (align < 10%),
6746/37918 (18%) are single-exon loci of those mapping >= 50% to genome,
3274 of these have homology to other species genes.

92627 are culled loci, not in public gene set, but with some unique sequences.
99 culls are multi-exon, well aligned; 87515 are single exon, well aligned,
1082 are partially mapped, and 3931 are poorly aligned to chromosomes.
13658 culls have protein homology, 78969 lack it.
175793 are culled alternate transcripts, at both public and culled loci, redundant in splicing patterns to public alternates, or lacking in alignment or evidence.

Gene locus IDs: Susscr4EVm000001t1 .. Susscr4EVm137575t1, Alternate transcripts have ID suffix t2 .. t100. EVm000001 is the longest protein, ID numbers are ordered by protein size, mostly. Culled transcripts are those initially classed as unique coding sequences, but re-classified as redundant, or lacking sufficient evidence, by chromosome alignment and homology evidence. These are separate from the public gene set as low quality, but are available as expressed transcripts, that may be recovered with further evidence.

Table 2(on next page)

pig18evg_datadesc.pages *Sus scrofa* gene sets compared for gene evidence recovery: 2a. Conserved vertebrate genes in pig gene sets (BUSCO), 2b. Human reference genes (Homo_sapiens RefSeq).

Table 2. *Sus scrofa* gene sets compared for gene evidence recovery: 2a. Conserved vertebrate genes in pig gene sets (BUSCO), 2b. Human reference genes (Homo_sapiens RefSeq).

2a. Vertebrate conserved genes

Geneset	Full	Align	Miss	Best
Evigene	2568	447 aa	8	776
NCBI	2567	440 aa	17	80
Ensembl	2552	431 aa	14	na

2b. Human reference genes

Geneset	Align	Miss	Frag	Best
Evigene	96.0%	0.7%	1.7%	20%
NCBI	97.2%	0.7%	0.6%	25%

Table 3(on next page)

pig18evg_datadesc.pages Assembler method effects on Human reference gene recovery in Pig gene sets: 3a. Sample set 1 (PRJNA416432), 3b. Sample set 2 (PRJNA353772)

Table 3. Assembler method effects on Human reference gene recovery in Pig gene sets: 3a. Sample set 1 (PRJNA416432), 3b. Sample set 2 (PRJNA353772).

3a. Sample set 1

Method	Miss	Frag	Short
Velvet	5%	7%	23%
Idba	8%	12%	30%
Soap	12%	16%	36%
Trinity	20%	28%	49%

3b. Sample set 2

Method	Miss	Frag	Short
Illumina_all	4%	6%	20%
Illum_velvet	5%	7%	23%
PacBio+	12%	15%	33%