EvidentialGene Gene Set Reconstruction Software
Don Gilbert, gilbertd At indiana edu, 2018
http://eugenes.org/EvidentialGene/

=item ABOUT

About Evigene-R : genes assembled from RNA pieces
About evigene/scripts/evgpipe_sra2genes.pl

  evgpipe_sra2genes is an omnibus pipeline to reconstruct genes, using
  several EvidentialGene methods, starting at Public SRA database of RNA-Seq
  (but can start w/ own RNA), and finishing with publication quality,
  annotated gene sequences. The basic steps are outlined below.
  See SRA2Genes Test Drive below.

About evigene/scripts/prot/tr2aacds.pl
  See also http://eugenes.org/EvidentialGene/about/EvidentialGene_trassembly_pipe.html

  Too many transcript assemblies are much better than too few. It allows one
  then to apply biological criteria to pick out the best ones. Don't be
  misled by the "right number" of transcripts that one or other transcript
  assembler may produce. It is the "right sequence" you want, and now the
  only way to get it is to produce too many assemblies (an "over-assembly")
  from RNA data, with several methods and several parameter settings.

  EvidentialGene tr2aacds.pl is a pipeline script for
  processing large piles of transcript assemblies, from several methods
  such as Velvet/Oases, Trinity, Soap, etc, into the most biologically useful
  "best" set of mRNA, classified into primary and alternate transcripts.

  It takes as input the transcript fasta produced by any/all of the
  transcript assemblers. These are classified (not clustered) into valid,
  non-redundant coding sequence transcripts ("okay"), and and redundant,
  fragment or non-coding transcripts ("drop"). The okay set is close to a
  biologically real set regardless of how many millions of input assemblies
  you start with.

  It solves major problems in gene set reconstruction found in other methods:

  1. Others do not not classify alternate transcripts of same locus properly,
  Alternates may differ in protein quite a bit, but share identical exon parts.

  2. Others remove paralogs with high identity in protein sequence, which
  is a problem for some very interesting gene families.

  3. Others may select artifacts with insertion errors by selecting longest transcripts
  Evigene works first with coding sequences.

  Quality assessment of this Transcript Assembly Software is
  described in http://eugenes.org/EvidentialGene/about/EvidentialGene_quality.html

About Evigene-G : traditional genes modeled on genome
About evigene/scripts/overbestgene2.pl

  This works on gene locations on chromosome assembly, in GFF v3 format tables.

  Gene models with overlapping CDS exons are "the same locus", each model has
  some form of evidence score, and the method picks out those models with
  highest evidence score.   The trick or trouble is mainly in applying
  various evidence scores, and their combination, to return the best models
  that a human expert would pick.

  See also evigene/docs/evg_overbestgenes.help.txt for details

About Evigene-N : non-coding gene reconstruction
  in progress
  See also evigene/docs/evigene_goals2015b.txt

About Evigene-H : gene reconstruction with hybrid of methods
  in progress
  See also evigene/docs/evigene_goals2015b.txt

=item HOW TO GET SOFTWARE

EvidentialGene software packaged as tar files are what you want, from here
  ftp://arthropods.eugenes.org/evigene.tar
  http://arthropods.eugenes.org/EvidentialGene/other/evigene_old/
  https://sourceforge.net/projects/evidentialgene/files/

EvidentialGene software in unpackaged form (lots of files) is here
  http://eugenes.org/EvidentialGene/evigene/
  e.g. http://eugenes.org/EvidentialGene/evigene/docs/
  is same as evigene/docs/ in your copy of this package.

=item WHO USES IT?

  See evigene/docs/evigene-cites.txt

=item IS IT ANY GOOD?

  See evigene/docs/evigene_plantsanimals_2017sum.txt

=item HOW TO INSTALL

Extract the tar archive this way, into current folder, preserving run permission.
  tar -xf evigene.tar

Run the Perl ".pl" scripts from extracted evigene folder, as they are a package.
  export evigene=`pwd`/evigene; # Unix bash shell, or
  set evigene=`pwd`/evigene;   # Unix csh/tcsh shell

  $evigene/scripts/prot/tr2aacds.pl [options] ..;

```
  $evigene/scripts/evgpipe_sra2genes.pl [options]  .. ;
  $evigene/scripts/evgmrna2tsa.pl [options]  .. ;

Required additional software
  You need these additional software for tr2aacds, installed in Unix PATH or via
  run scripts.

  * fastanrdb of exonerate package, quickly reduces perfect duplicate sequences
    https://www.ebi.ac.uk/~guy/exonerate/  OR
    https://www.ebi.ac.uk/about/vertebrate-genomics/software/exonerate
  * cd-hit, cd-hit-est, clusters protein or nucleotide sequences.
    http://cd-hit.org/ OR https://github.com/weizhongli/cdhit/
  * blastn and makeblastdb of NCBI BLAST, https://blast.ncbi.nlm.nih.gov/
    Basic Local Alignment Search Tool, finds regions of local similarity between sequen

Most of the shell ".sh" scripts require editing for your cluster; consider
them examples. Perl scripts have brief -help, but most of their
documentation is perl POD. This is a complex package, including my working
scripts for several genome projects, some are obsolete now. One needs a
cheat-sheet to make sense of what is good and I am working on such.


=item TEST DRIVE

SRA2Genes Test Drive, 2019-May
  http://eugenes.org/EvidentialGene/other/sra2genes_testdrive/

SRA2Genes is a full gene-set reconstruction pipeline, and includes
the well-used tr2aacds, and the other parts of Evigene needed for
complete gene set assembly, annotation and publication.

tr2aacds reduces a large, over-assembly of transcripts by using
only self-referential coding-gene metrics.  That is very useful
but also fairly limited and rough, in that it uses only the
gene evidence from that transcript assembly.

The more complete gene reconstruction pipeline of SRA2Genes
brings in external gene evidence, notably the wealth of
conserved gene information.  Genome biologists should consider
using this SRA2Genes pipeline in place of tr2aacds.

As configured now, this test starts at transcript assembly input,
step7, thru public gene set production, step10, and gene set
submission to TSA, step11. Later test drives will start from RNA-seq
inputs. See also evigene/docs/evgpipe_sra2genes.help.txt

When this test drive works for you, the sample data sets can be
replaced with your own transcript and reference gene sets for useful
results.  This basic SRA2Genes Test Drive will run for up to a few
hours on a minimal computer, e.g. on my Mac laptop in a Linux virtual
machine with 2 CPU.
```

```
                    *---------------------*
TR2AACDS only:
Try this test case with small input data (TAIR10 mRNA) and tr2aacds outputs,
  http://eugenes.org/EvidentialGene/plants/arabidopsis/evigene_tr2aacds_test/

It is worth your time to set up and run this with same input data to see
that you get same results.

RUN-LOCAL:
  env trset=arath_TAIR10_20101214up.cdna.gz datad=`pwd` ./tr2aacds_test.sh
RUN-CLUSTER:
  env trset=arath_TAIR10_20101214up.cdna.gz datad=`pwd` qsub -q normal tr2aacds_test.sh

You should be able to get same result from same Arabidopsis transcripts
input data file, and where problems appear, please consult a local computer
expert familiar with your cluster computer to resolve.  After you get that
test set working ok, running on your data set should be simpler.


=item BASIC USAGE of Evigene-R

  See steps in evigene/scripts/evgpipe_sra2genes.pl

Options:
  myspecies=Genus_Species_v1
  trset=$myspecies.cdna   # 1 fasta input file with many transcript sequences, assemble
  evigene=/your/path/to/evigene #  where you un-tarred evigene.tar
  ncpu=1 # or 2 or 8      # 8 cpu probably enough, each uses 2+ GB memory
  maxmem=32000   # in megabytes, expect 2+GB per cpu, maybe more for complex large over

STEP 1. get RNA-Seq data

STEP 4. run assemblers of RNA-seq, with kmer size options, other opts
    4a. velvet/oases, ~10 kmer steps
    4b. idba_tran, ~10 kmer steps
    4c. soap_trans, ~10 kmer steps
    4d. trinity / other / user choices
  ...

STEP 5. trformat.pl, post process assembly sets

  subd=veloset   # velvet run directory with several velvet kmer subfolders
  $evigene/scripts/rnaseq/trformat.pl -pre $myspecies -out trsets/$subd.tr -log -in $su

  subd=idbaset   # idba run directory, several transcripts-kmer.fa outputs
  $evigene/scripts/rnaseq/trformat.pl -pre $myspecies -out trsets/$subd.tr -log -in $su

  Catenate all transcript sets to one file:
  cat trsets/*.tr > $myspecies.cdna

STEP 7. tr2aacds, reduce over-assembly to draft gene set
```

```
  $evigene/prot/tr2aacds.pl -tidy -NCPU $ncpu -MAXMEM $maxmem -log -cdna $myspecies.cdn

STEP 10. evgmrna2tsa,  produce public gene sequences
  $myspecies.trclass is a result from STEP 7, tr2aacds

  $evigene/scripts/evgmrna2tsa2.pl -onlypubset -idprefix $myspeciesEVm -class $myspecie


=item TR2AACDS PIPELINE ALGORITHM

Prerequisite is that you create transcript assemblies (with any/all
useful methods). This software reads all the transcripts.fasta you have,
chews on them and puts them into good and bad piles (with extras).

  0. collect input transcripts.tr,
  You supply input transcript sequences in .fasta, an over-assembly with
  redundant and variable quality transcripts, as one file.

  1. perfect redundant removal: exonerate/fastanrdb input.cds > input_nr.cds,
  and protein qualities are used for choosing among cds identicals.

  2. perfect fragment removal: cd-hit-est -c 1.0 -l $MINCDS ..
     Cluster  *identical coding sequences*, short and long, keep the longest

  3. blastn, basic local align hi-ident subsequences for alternate tr.,
  with -perc_identity CDSBLAST_IDENT (98%), to find high-identity
  exon-sized alignments.

  4. classify main/alternate cds, okay & drop subsets, using
  evigene/rnaseq/asmrna_dupfilter2.pl
  merges alignment table, protein-quality and identity, to score
  okay-main, ok-alt, and drop sets.

  5. final output files from outclass: okay-main, okay-alts, drops
  okayset is for public consumption.  The drop set of redundant, fragment,
  non-coding sequences, may contain valid coding sequences (more details).


=item OTHER EVIDENTIALGENE COMPONENTS

evigene/scripts/rnaseq/trformat.pl
    See STEP 5  of evgpipe_sra2genes.pl
    Use BEFORE tr2aacds to regularize IDs in fasta of
    Velvet,Soap,Trinity, ensure unique IDs, add prefixes for parameter sets.

evigene/scripts/prot/namegenes.pl
    See STEPs 8-9 of evgpipe_sra2genes.pl
    Use AFTER tr2aacds on okayset, add gene function names from
    UniProt-Ref and CDD blastp.

    blastp -db refprots -query okay_all.aa  -outfmt 7  -out $name.blastp
    namegenes.pl -refnames $refdb.names -blast $name.blastp
```

evigene/scripts/rnaseq/asmrna_trimvec.pl
    See STEPs 8-9 of evgpipe_sra2genes.pl
    UniVec vector screening and NNN-end trimming, per NCBI or INSDC desires

evigene/scripts/evgmrna2tsa.pl
    See STEPs 10 of evgpipe_sra2genes.pl, See evigene/docs/evgmrna2tsa_help.txt
    make public mRNA gene set, with pubIDs,
    main/alternates, names and annotation, and Genbank TSA format for
    public submission

=item HELP AND METHOD DOCUMENTS

How To get Best mRNA Transcript assemblies
  evigene/docs/perfect-mrna-assembly-2013jan.txt

  Please read this brief How-To document that summarizes my tests on best
  transcript assembly methods. It points out tips for assembly parameters,
  such as using scaffolding and multi-kmer settings (for Velvet, Soap,
  others that allow; not Trinity), that will improve your transcript
  assemblies.

Best Assembler methods
  evigene/docs/evg_geneassembly_bestmethods1603.html
  Best assembly methods compared for mosquito genes
  has recent comparison of gene assembler accuracy,

EvidentialGene tr2aacds mRNA classifier description
  evigene/docs/evigene-tr2aacds-classifier.txt

  Classification is based mainly on CDS-dna local alignment identity.
  Perfect fragment CDS are dropped, those with some CDS base differences
  are kept, with longest CDS as primary transcript. UTR identity is
  ignored (for now) because many of the mis-assemblies are from
  joined/mangled genes in UTR region.

Error of selecting longest transcripts, as with CD-HIT-EST
  evigene/docs/cdhiterr-arabidopsis-example.txt
  Selecting genes by longest transcripts is a mistake.

EvidentialGene ORF/protein computation
  evigene/docs/protein_calcs_compared.txt
  update to evigene/docs/evigene_vs_transdecoder_arabidopsis.txt

  This document compares several ORF methods for recovering
  proteins of well-constructed Human RefSeq and Arabidopsis
  reference genes.

  EvidentialGene computes ORFs (proteins and coding sequences of
  those) in a standard way, and recovers reference proteins properly,
  as does NCBI's ORF calculator. Other methods do not do as well,

mismodeling 10 to 20% of reference proteins, including TransDecoder
and GeneMark methods.

Validating with RNA-Seq map-back
evigene/docs/transrate_err_arabid17.txt

RNA-seq mapping methods are influenced strongly by presence of duplicated
sequence spans, as with alternate transcripts and high identity paralogs.
An accurate statistic of proper paired fragment mapping to a
given transcript should give the same value regardless of whether
alternates to that transcript exist.

TransRate, and some other RNA map-back validation methods, produce
inaccurate statistics, that are influenced by presence/absence of other
biological alternates and paralogs.
