XScript: GENOA and Hollywoodb .. Dirk Holste 12/XX/04 .. last change 12/XX/2004

```
******************************************************************************
*                         GENOA - Genome Annotation                          *
*             D.Holste, R-F.Yeh, L.P.Lim, G.Yeo and C.Burge                  *
*                         http://genes.mit.edu/genoa                         *
******************************************************************************
```

The Genome Annotation (GENOA_ program maps cDNA and EST sequences to genomic DNA and reports the genomic location of successfully aligned cDNA and EST sequences for gene loci. Genomic DNA input data can be taken as transcriptional units of Ensembl-annotated gene loci, but generally any segment of genomic DNA of size smaller than 1Mb can be annotated. After masking cDNAs for interspersed repeats (rm-cDNAs), GENOA finds locations of significant rm-cDNA:genomic (BLAST) hits and pursues a spliced alignment of cDNAs, using the program mRNAcsGen. For each successful alignment, a GenBank file is created, and cDNA sequence alignments are annotated in GenBank format. This is denoted as 'anchoring'. After anchoring, GENOA finds locations of significant EST:rm-cDNA (BLAST) hits and pursues a spliced alignment of ESTs, using the program sim4. For each successful alignment, the corresponding GenBank file is updated, and cDNA and EST sequence alignments are annotated in GenBank format. Note that cDNA alignments annotate as mRNA and CDS sequences, if given, and EST sequences are annotated with associated cDNA library information (tissue category). In its current setup, EST-only alignments are not supported, but can be achived by mimicking prior cDNA alignments.

This helpfile discusses the following topics:

***

1 Basic INPUT

Input format is either 'FASTA format' for genomic, repetitive and EST sequenes, or 'Genbank format' for cDNAs.

1.1 Genomic sequence data (LIB_genomic)

Multi GenBank file:

```
LOCUS       chr21.G187175.Ctg1  133724 bp    DNA          XXX      00-XXX-0000
DEFINITION  .
ACCESSION   chr21.ENSG00000187175:1..133724
KEYWORDS     ENSG00000187175.Ctg1 chr21:1..133724 133724:133724 ENS:...
SOURCE       EnsEMBL_121503
  ORGANISM   Homo_sapiens
COMMENT
            no comment.
FEATURES             Location/Qualifiers
BASE COUNT
ORIGIN
        1 GAAGGGCTTT GGTCAAACAT CTCAAGCAGA GGCCTTCCTC CCCCGCCTGC CTGCACGTGG
       61 C...
    10081 T...
    10141 AAAAATGATA ACTCTATAAT AGAGAAGCAT GGCAAACACT ACCTTA
//
```

1.2 cDNA sequence data (LIB_mRNA)

Multi GenBank file:

```
LOCUS       AB000095                2399 bp    mRNA    linear   PRI 04-MAR-1998
DEFINITION  Homo sapiens mRNA for hepatocyte growth factor activator inhibitor,
            complete cds.
ACCESSION   AB000095
VERSION     AB000095.1  GI:2924600
KEYWORDS    hepatocyte growth factor activator inhibitor.
SOURCE      Homo sapiens (human)
  ORGANISM  Homo sapiens
            Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;
            Mammalia; Eutheria; Primates; Catarrhini; Hominidae; Homo.
REFERENCE   1  (sites)
  AUTHORS   Shimomura,T., Denda,K., Kitamura,A., Kawaguchi,T., Kito,M.,
            Kondo,J., Kagaya,S., Qin,L., Takata,H., Miyazawa,K. and Kitamura,N.
  TITLE     Hepatocyte growth factor activator inhibitor, a novel Kunitz-type
            serine protease inhibitor
  JOURNAL   J. Biol. Chem. 272 (10), 6370-6376 (1997)
  MEDLINE   97197808
   PUBMED   9045658
REFERENCE   2  (bases 1 to 2399)
  AUTHORS   Denda,K.
  TITLE     Direct Submission
  JOURNAL   Submitted (24-DEC-1996) Kimitoshi Denda, Tokyo Institute of
            Technology, Department of Life Science; 4259 Nagatsuta, Midori-ku,
            Yokohama, Kanagawa 227, Japan (E-mail:kdenda@bio.titech.ac.jp,
            Tel:45-924-5702, Fax:45-924-5771)
FEATURES             Location/Qualifiers
     source          1..2399
                     /organism="Homo sapiens"
                     /mol_type="mRNA"
                     /db_xref="taxon:9606"
     CDS             176..1717
                     /codon_start=1
                     /product="hepatocyte growth factor activator inhibitor"
                     /protein_id="BAA25014.1"
                     /db_xref="GI:2924601"
                     /translation="MAPARTMARARLAPAGIPAVALWLLCTLGLQGTQAGPPPAPPGL
                     PAGADCLNSFTAGVPGFVLDTNASVSNGATFLESPTVRRGWDCVRACCTTQNCNLALV
                     ELQPDRGEDAIAACFLINCLYEQNFVCKFAPREGFINYLTREVYRSYRQLRTQGFGGS
                     GIPKAWAGIDLKVQPQEPLVLKDVENTDWRLLRGDTDVRVERKDPNQVELWGLKEGTY
                     LFQLTVTSSDHPEDTANVTVTVLSTKQTEDYCLASNKVGRCRGSFPRWYYDPTEQICK
                     SFVYGGCLGNKNNYLREEECILACRGVQGPSMERRHPVCSGTCQPTQFRCSNGCCIDS
                     FLECDDTPNCPDASDEAACEKYTSGFDELQRIHFPSDKGHCVDLPDTGLCKESIPRWY
                     YNPFSEHCARFTYGGCYGNKNNFEEEQQCLESCRGISKKDVFGLRREIPIPSTGSVEM
                     AVAVFLVICIVVVVAILGYCFFKNQRKDFHGHHHHPPPTPASSTVSTTEDTEHLVYNH
                     TTRPL"
     polyA_signal    2379..2384
ORIGIN
```

```
        1 cggccgagcc cagctctccg agcaccgggt cggaagccgc gacccgagcc gcgcaggaag
       61 c...
     2281 g...
     2341 aggttctcca acatcacagc ccagcccacc cactgggtaa taaaagtggt ttgtggaaa
//
```

1.3 Expressed sequence tag data (LIB_EST)

Multi FASTA file:

```
>dbEST|Acc|M61958|id|5|tissueESTHs.1|hippocampus|Homo|sapiens
TGCACAACCAAGTTTTGTGACTACGGGAAGGCTCCCGGGGCAGAGGAGTACGCTCAACAA
GATGTGTTAAAGAAATCTTACTCCAAGGCCTTCACGCTGACCATCTCTGCCCTCTTTGTG
ACACCCAAGACGACTGGGGCCCNGGTGGAGTTAAGCGAGCAGCAACTNCAGTTGTNGCCG
AGTGATGTGGACAAGCTGTCACCCACTGACA
```

In addition, GENOA expects to find three BLAST−formated file of the EST file LIB_EST:

```
        (1) LIB_EST.nhr
        (2) LIB_EST.seq
        (3) LIB_EST.nin.
```

1.4 Repetitive sequence data (LIB_repeat)

Multi FASTA file

```
>ALU
ggccgggcgcggtggctcacgcctgtaatcccagcactttgggaggccgaggcgggaggattgcttgagcc
caggagttcgagaccagcctgggcaacatagcgagaccccgtctctacaaaaaatacaaaaattagccggg
cgtggtggcgcgcgcctgtagtcccagctactcgggaggctgaggcaggaggatcgcttgagcccaggagt
tcgaggctgcagtgagctatgatcgcgccactgcactccagcctgggcgacagagcgagaccctgtctca
```

In addition, GENOA expects to find three BLAST−formated file of the repeat file LIB_repeat:

```
        (1) LIB_repeat.nhr
        (2) LIB_repeat.seq
        (3) LIB_repeat.nin.
```

Obtain repeats files from http://www.girinst.org/index.html, will require Username and Password (obtained after registration).

2. Basic OPTIONS

The general usage for a Genome Annotation (GENOA) run can be obtained from %runGENOA.pl −h. For instance, runGENOA.pl can be started via command line with the following usage: %runGENOA.pl [−options], where options is set to ’genus Homo’, ’species sapines’, ’chr chr21’, ’mrna DIR/LIB_mRNA’, ’genomic DIR/LIB_genomic.chr21’, ’genomesize 150000’, ’ngenes 1000’ and ’clean’.

The option ’clean’ will delete numerous tmp files at the end of each GENOA run, including the basic input files LIB_EST.*, LIB_repeat.* as well as LIB_genomic.* files. These options can individually be modified at the end of the runGENOA.pl script.

The options ’genomesize’ (in Mb) and ’ngenes’ for the number of expected gene loci, are not crucial in the sense that they will limit the number of genomic DNA or genes to be annotated, but are approximate memory pre−allocation parameters and ought to be set according to available prior information or guess.

2.1 cDNA Repeat masking

GENOA incorporates masking of repetitive sequences in cDNAs by the following

usage: %runGENOA.pl [−options], where the additional options are ’mask LIB_repeat’ and ’MBLAST’. The options ’mask’ and ’MBLAST’ incorporate the masking of repetitive sequences detected in cDNAs, and the usage of the program MaskBLAST. Alternatively, the program RepeatMasker can be used (’RM’).

2.2 EST sequences

GENOA incorporates the masking of repetitive sequences in cDNAs by the following usage: %runGENOA.pl [−options], where the additional option is ’est DIR/LIB_EST’ used.

2.3 Extraction of cDNA sequences from GenBank

GENOA incorporates the extraction of GenBank report files for an individual genus and species by the following usage: %runGENOA.pl [−options], where the additional options are ’db LIB_GenBank’ and ’noimmune’.

In order to use runGENOA.pl with the above options, obtain all corresponding GenBank files, e.g., the flat files gbprim.seq, gbhtc.seq or gbrod.seq. Then, modify runGENOA.pl in the line following line:

```
"# ##################################################################
 # TOSET: extract LIB_species, LIB_mRBA and LIB_genomics (and die)
 $DIEAFTERGB = $FALSE;
"
```

and set the parameter $DIEAFTERGB to the value $TRUE. GENOA will extract genus and species as set in options, collect all GenBank records in a multi GenBank file and terminate after extraction. Use the option ’noimmune’ to screen for immunoglobulin genes and to not include those Ig’s into the multi GenBank file.

3 Advanced options

Advanced option can be set and modified for BLAST searches and parsing of sim4 output within the scripts runGENOA.pl and sim4Gb.pl, respectively.

3.1 BLAST parameters

For each BLAST search, (1) cDNA vs repetitive sequences and RepeatMasking, (2) cDNA vs genomic sequences, and (3) ESTs vs cDNA sequences, the word size, the the E values and the number of BLAST hits per search can be set individually.

3.2 EST spliced alignments (sim4 output)

A set of parameters affecting both sim4 output and parsing can be modified in sim42Gb.pl, the dependence of which is outlined for each parameter, starting in the following line:

```
"# ################################################################## #
 # TOSET:
 $PARTIALCHECK
"
```

In particular, for each alignment the EST:genomic sequence similarity, the EST alignment size, the first and the last EST fragment similarities and minimum sizes can be controlled.

4 Basic OUTPUT

GENOA stored any output in separated sub directories and files, created during the run in the current working directory. The subdirectories are as follows:

/genoa, /genoa−nested, /genoa−blastout and /genoa−monitor. The last directory contains files for monitoring and reprocessing the alignment, and comprises the following subdirectories: /genomicseq−cds, /genomicseq−err, /genomicseq−est, and files: sim4Gb.err.fl and logGENOA.day.time.

4.1 The subdirectory /genoa and how to read results

GENOA stores the final output of cDNA and EST sequence annotated genomic files in the subdirectory /genoa. The output files are formated in GenBank format, and can be used for further downstream processing.

Example output: (one gene)

```
LOCUS       chr21.G154721.Ctg1.G1−5  85647 bp    DNA       XXX      00−XXX−0000
DEFINITION  .
ACCESSION   chr21.ENSG00000154721:1..85647:1..85647
KEYWORDS     ENSG00000154721.Ctg1 chr21:1..85647 85647:85647 ENS:...
SOURCE       EnsEMBL_121503
  ORGANISM   Homo_sapiens
COMMENT
            no comment.
FEATURES             Location/Qualifiers
     mRNA            join(49607..49672,55590..59632,
                     64401..64603)
                     /certainty=111
                     /introncertainty=111
                     /match="AF255910:39..1245"
     mRNA            join(49607..49672,55590..55697,
                     64401..67993)
                     /certainty=111
                     /introncertainty=111
                     /match="AY077698:1..1087"
     mRNA            join(49607..49672,59480..59632,
                     64401..64603)
                     /certainty=111
                     /introncertainty=111
                     /match="AY358361:1..1295"
BASE COUNT    25594 a  16965 c  17140 g  25948 t
ORIGIN
        1 tgaattcaga attagaatgg gtggaaagaa ttaaaaatgg taagctgtcc caaaacacca
       61 a...
    85561 t...
    85621 tatgtgaagt tacaaagttg ttccatg
```

4.2 The log file (logGENOA.*)

The log file (logGENOA.day.time) monitors all major steps of the alignment, and includes all BLAST outout (hitlists), the cDNA repeat−masking, the spliced cDNA alignments, the EST alignments. For each process step, several statistics are monitored and provide overiew about GENOA’s task and performance, with time stamps where appropriate.

Example output: (one gene)

```
"Mon Nov 15 12:06:12 EST 2004
  _____________________________________________________________
  GENOA Genome Annotation System                          v1.01
  _____________________________________________________________
  massachusetts institute 31 ames str     department of biology
  of technology           68−211          holste@mit.edu
  cambridge, ma 02139     617.253.7039    (c) 2004

  Usage: runGENOA.pl
         −db LIB_GenBank || (−mrna LIB_mRNA && −genomic LIB_genomic)
         −genus genus
         −species species
```

```
        −chr chromosome
        −genomesize genomesize
        −ngenes ngenes

 Options:
 Genus:                         Homo
 Species:                       sapiens
 Chromosome:                    chr21
 Flatfile:                      <>
 Genomic file:                  <LIB_genomic.chr21>
 mRNA file:                     <LIB_mRNA.noimmune>
 Genome size (kb):              150000
 Number of genes:               5000
 BLAST:
       E/ W mRNA vs repeats:    1e−10/ 11
       E/ W mRNA vs genomic:    1e−15/ 25/ 1
       E/ W mRNA vs ESTs:       1e−25/ 25/ 1000
 Repeats:
       Use Repeat file:         <LIB_repeat>
       Use RepeatMasker:        0
       Use RM option −pa:
       Use MaskBLAST:           1
 EST data:
       EST data file:           <LIB_EST>
       EST tissue entry:        5
 Clean up files:                0
 Log file:                      <logGENOA.Wed.Nov.10.153727.EST.2004>
 Program name:                  ./runGENOA.pl
 Arguments given:               −genus homo −species sapiens −chr chr21
                                −mrna LIB_mRNA.noimmune
                                −genomic LIB_genomic.chr21 −genomesize 150000
                                −ngenes 5000
                                −mask LIB_repeat −MBLAST −est LIB_EST
Arguments not set in ./runGENOA.pl:
       "intron gap"    | 30bp   [ "mrnavsgen.c" ]
       "MAXINT"         | 200000        [ "genomicSeq2Newagain..pl " ]
       "$PARTIALCHECK"  | FALSE         [ "sim4Gb.pl" ]
       "$QUALITYCHECK"  | TRUE  [ "sim4Gb.pl" ]
       "$STRANDCHECK"   | TRUE  [ "sim4Gb.pl" ]
       "$QualityCutOff"         | 90%   [ "sim4Gb.pl" ]
       "$GenGapCutOff"  | 30bp  [ "sim4Gb.pl" ]
       "$GenIntCutOff"  | 200000bp      [ "sim4Gb.pl" ]
       "$ESTGapCutOff"  | 1bp   [ "sim4Gb.pl" ]
       "$ESTLenCutOff"  | 90%   [ "sim4Gb.pl" ]
       "$FirstESTLen"   | 30    [ "sim4Gb.pl" ]
       "$FirstESTQual"  | 90%   [ "sim4Gb.pl" ]
       "$LastESTLen"    | 30    [ "sim4Gb.pl" ]
       "$LastESTQual"   | 90%   [ "sim4Gb.pl" ]

# runGENOA:01
#       Files <LIB_mRNA.noimmune> and <LIB_genomic.chr21> present
# runGENOA:02|12
#       Count loci in mRNA files multi GenBank file
#       Number of mRNA files found: 10
#       Number of genomic files found: 264
# runGENOA:02
#       Converting mRNAs to FastA format...
#       Indexing mRNA library files...
# runGENOA:03|12
#       Masking repetitive sequenences...
#       Use MaskBLAST, repeat library data in FastA format & pre−formatdb data
#       BLAST mRNAs vs repeat library...
# runGENOA:04|12
#       Indexing genomics library files...
#       Converting genomics to FastA format and formatDB...
#       BLAST mRNAs vs genomic sequences...
#       Create hitlists...
# runGENOA:05|12
#       Count loci in LIB_genomic...
```

```
 #       Number of overall genomic BLAST hits: 1 | 1 uniq
 #       Number of mRNAs involved in hits: 1 | 1 uniq
 #       For each CTGhit: 1,2,... do
 #
 #       * chr21.G184029.Ctg1
 #       Number of genomic BLAST hits for chr21.G184029.Ctg1: 1
 #       Aligning mRNAs to genomics...
 #       Number of genomic files for chr21.G184029.Ctg1 aligned with mRNAs: 1
 #       Renumber of overall genomic BLAST hits after each CTG: 1 ( cf 1 )
 #       Number of overall genomic files aligned with mRNAs: 1
 #       Done overall aligning mRNAs to genomics
 # runGENOA:06|12
 #       Store re−annotation of mRNAvsgen in LIB_genomic.new
 #       Number of re−annotated genomic files: 1
 # runGENOA:07|12
 #       Trimming transcripts in genomicseq (intron−less, size and certainty)
 #       Move old and error files to genomicseq−old respectively genomicseq−err
 # runGENOA:08|12
 #       Separating single gene regions
 #       Check cDNAs for genes on opposite strands...
 #       For each genon, 1,2,... do
 #
 #       * chr21.G184029.Ctg1
 #         separate strands f 0 | r 1
 #         inverse complement genes r 1 | ic 1
 #         chop genes on ic strand...
 #       Number of cDNAs, and ( f | r | ic ) files: 1, and ( 0 | 1 | 1 )
 # runGENOA:09|12
 #       Moving chopped cDNA transcripts to new directory...
 #       Number of multi−cDNA (single−cDNA) genes 0 (1)
 #       Rename genomicseq to genomicseq−tmp (historical reasons)
 #       Rename genomicseq−new to genomicseq (historical reasons)
 # runGENOA:10|12
 #       Use ESTs in FastA format and use pre−formatDBed EST data
 #       Select mRNAs from genomicseq/chr21 with alignments to BLAST ESTs...
 #
 #       * Number of overall aligned mRNAs to CTG chr21.G184029.Ctg1.ic.G1
 #         found 1 | 1 uniq
 #       BLAST aligned mRNAs from vs ESTs...
 #       Making index file of hits...
 #       Aligning EST2genomics...
 #       Creating new genomic sequence annotation based on ESTs...
 #       found EST alignments ... completed 1 | 1
 #       Number of overall EST hits 1 ( 1 uniq )
 #       Number of overall sim4 alignments 1 ( 1 uniq )
 # runGENOA:11|12
 #       Add EST alignments to mRNA chopped CTGs...
 #       Number of transcript units: 1
 #       Number of transcript units with single   mRNA hits: 1
 #       Number of transcript units with multiple mRNA hits: 0
 # runGENOA:12|12
 #       Classifying mRNA hits:
 #       Alternatively spliced:      0
 #       Nested:                     1
 #       On opposite strands:        0
 #       Overlapping but not nested: 0
"
```

5 Scripts, binaries and system requirements

The main script supervising the reading in of files and performing the BLAST and alignment steps is runGENOA.pl. The subdirectories /bin and /pl contain the corresponding binaries and perl scripts. GENOA requires preformated BLAST files for cDNAs, repeats and EST sequences. The BLAST v2.2.5 (or higher) has been tested and performed well for large (>2Gb) EST multi FASTA files. Note that in order to process >2Gb files, Perl v5.8.0 (or higher) is required. GENOA has been developed using libraries available under the OpenSource Linux RedHat distribution 7.3

5.1 Preprocessing

Genomic:GenBank flatfiles can be formated by using three the format*.pl scripts available under /pl subdirectory: formatCtg2chop.pl and formatFasta2Gb.pl, and a batch file formatCtgchop+Fasta2Gb.pl. In order, these scripts chop genomic DNA into contigs of size 1Mb (current setting), and create GenBank files for each contig.

6 References

GENOA program.
D.Holste, R−F.Yeh, L.P.Lim, G.Yeo and C.Burge.
http://genes.mit.edu/genoa

Variation in alternative splicing across human tissues.
G. Yeo, D.Holste, G.Kreiman, and C.B.Burge.
Genome Biology 5 (2004)

SNP−based validation of exonic splicing enhancers.
W.G.Faibrother, D.Holste}, C.B.Burge, and P.A.Sharp.
PLoS Biology 2 (2004)