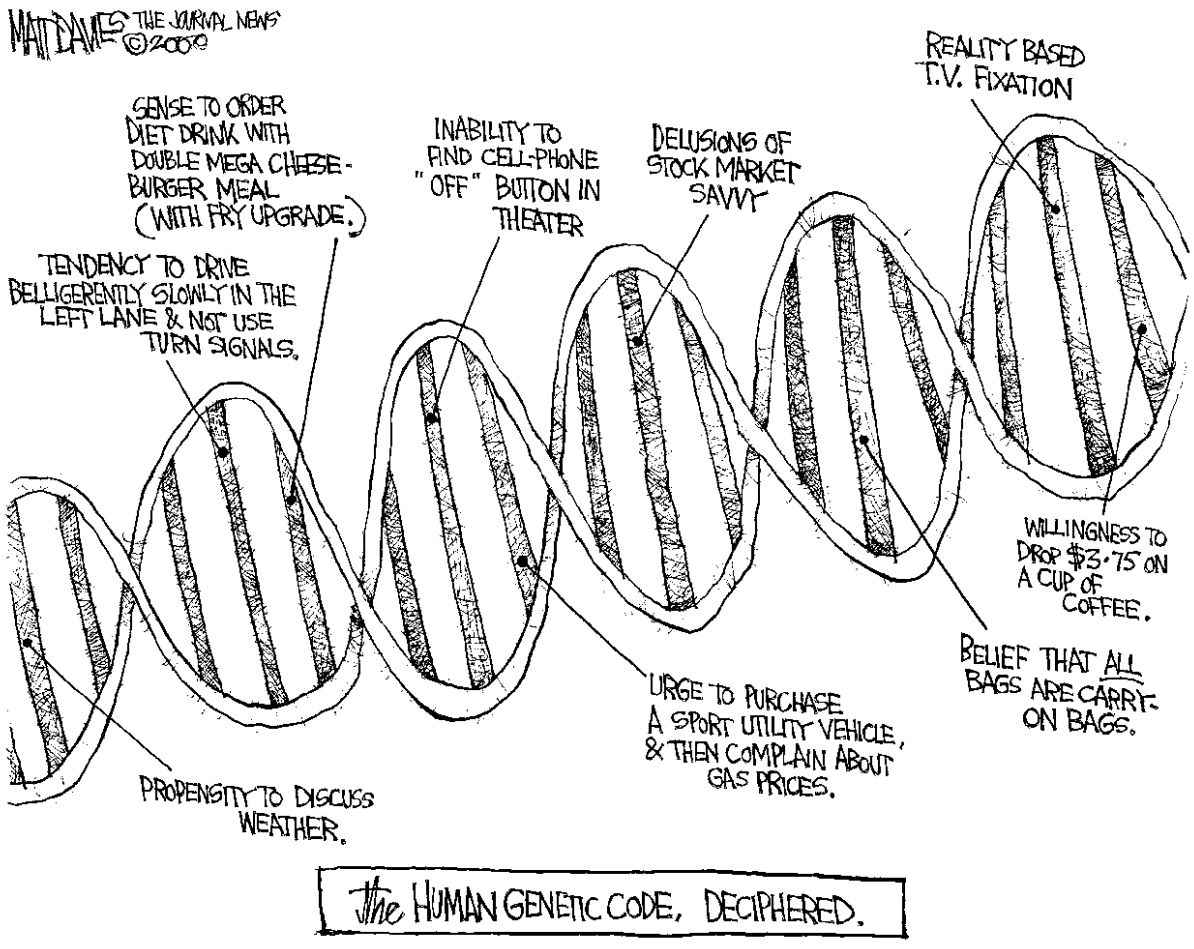
Bacterial Genome Annotation



Bacterial Genome Annotation

For an annotation you want to predict from the sequence, all of ...

- protein-coding genes
  - their stop-start
  - the resulting protein
  - the function
  - the control elements
  - variants
- structural RNAs
- tRNAs
- small RNAs
- pseudogenes,
- control regions
- direct/inverted repeats
- insertion sequences
- transposons,
- other mobile elements

and it must be done quickly without a huge use of resources.

Bacterial Genome Annotation

Fortunately for the protein coding sequences of bacteria there are no introns (well mostly none). But this does not mean that you can simply identify all open reading frames and call the job done.

For example, how small can a protein be?
What do you do if open reading frames overlap?
What about read through?
What about start codons?

Bacterial Genome Annotation

You have mostly been taught that a bacterial coding sequence begins with ATG. True most do!

But you can also have ATG, GTG, TTG, ATT, CTG or ATC.

Similarly stop codons TGA or TAG may encode selenocysteine and pyrrolysine.

Do it fast! Another genome is coming off the sequencer.

Do it fast! Another genome is coming off the sequencer.

Do it without my attention! I have other things to do.

Bacterial Genome Annotation

5S, 16S, and 23S rRNA are found using BLASTn to Refseq

ncRNAs are found using BLASTn to Rfam families

Both 5S rRNA and ncRNA use `cmsearch`[†] to refine hits.

tRNAs are found using `tRNAscan-SE`[*].

[†]covariance model search - a dynamic programming approach to include known sequence **and** 2D structural information to find a match in sequences alone.

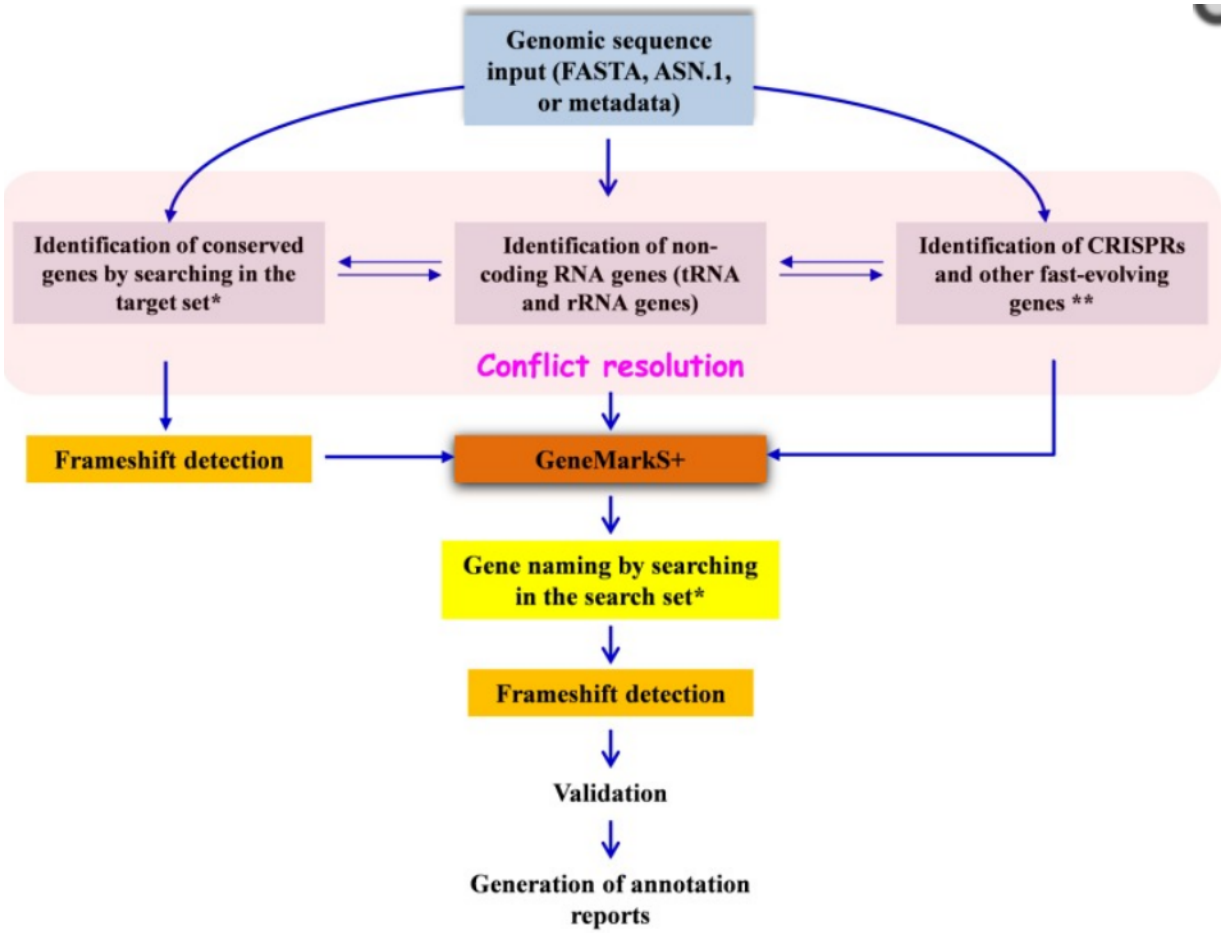
[*]tRNAscan-SE - less than one false positive per 15 gigabases. A multi-pass algorithm that in the final pass uses a probabilistic secondary structure profile.

Bacterial Genome Annotation

For proteins they use the pan-genome, and target sets of proteins expected to be found (e.g. ribosomal proteins) and compare these using a frameshift/error aware aligner `ProSplign`. Non-perfect alignments are passed to `GeneMarkS+` which is a HMM gene caller (NCBI also hosts `GLIMMER` for secondary gene identification). If the hit is poor, it might be searched against a wider set of proteins and then passed through `ProSplign/GeneMarkS+` again.

Rapidly evolving regions such as phage related proteins and CRISPRs are treated separately.

Bacterial Genome Annotation



"NCBIs Prokaryotic Genome Annotation Pipeline combines a computational gene prediction algorithm with a similarity-based gene detection approach. The pipeline currently predicts protein-coding genes, structural RNAs (5S, 16S, 23S), tRNAs, and small non-coding RNAs. The flowchart describes the major components of the pipeline."

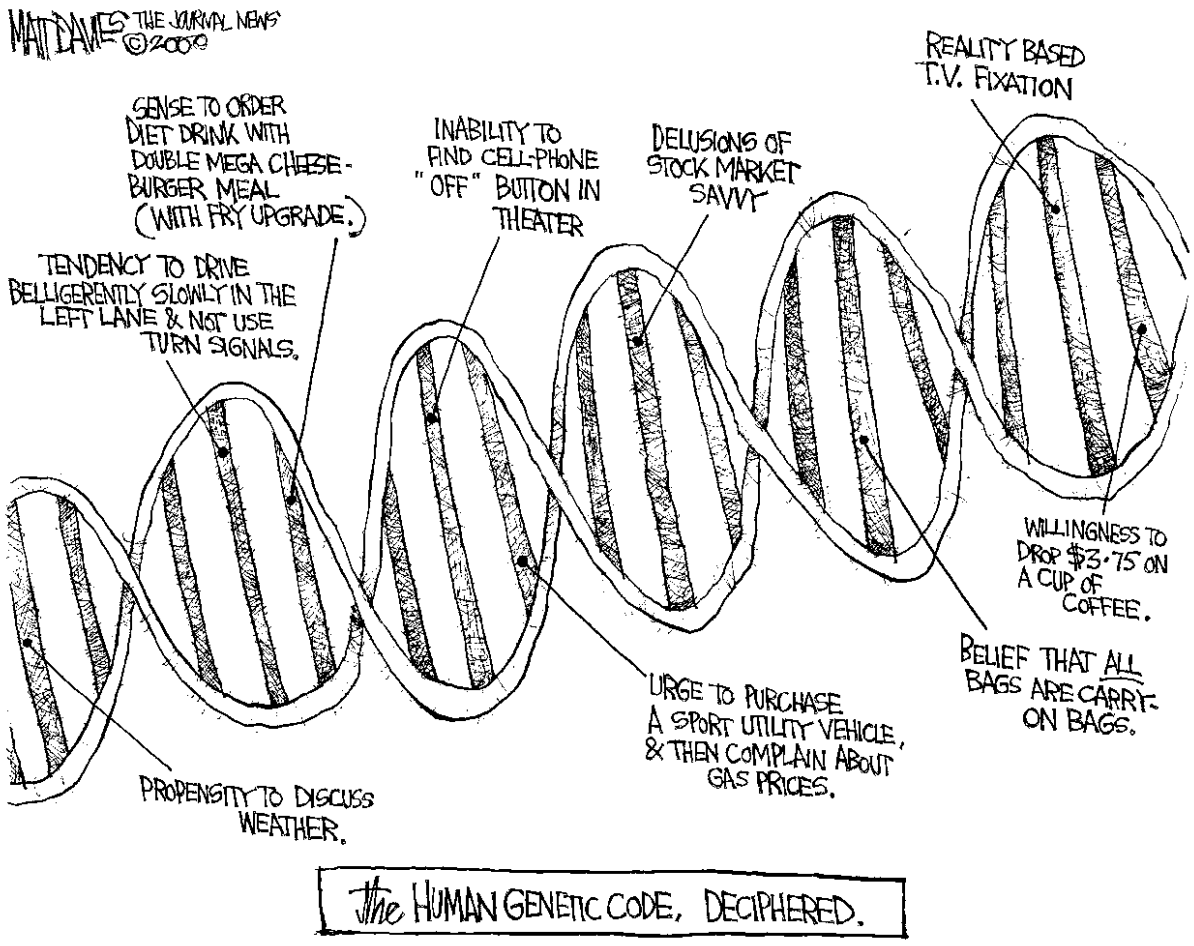
From http://www.ncbi.nlm.nih.gov/books/NBK174280/

Bacterial Genome Annotation

In a Genbank file you will often find a meta-data entry about the annotation. For example for *E. coli* strain JJ1897, you will find,

```
##Genome-Assembly-Data-START##
Assembly Date           :: SEP-2015
Assembly Method         :: HGAP 3 (PacBio) v. SEPT-2015
Genome Representation   :: Full
Expected Final Version :: Yes
Genome Coverage         :: 100.0x
Sequencing Technology   :: PacBio
##Genome-Assembly-Data-END##

##Genome-Annotation-Data-START##
Annotation Provider          :: NCBI
Annotation Date              :: 01/05/2016 15:37:32
Annotation Pipeline          :: NCBI Prokaryotic Genome Annotation Pipeline
Annotation Method            :: Best-placed reference protein set; GeneMarkS+
Annotation Software revision :: 3.0
Features Annotated           :: Gene; CDS; rRNA; tRNA; ncRNA; repeat_region
Genes                        :: 5,336
CDS                          :: 4,695
Pseudo Genes                 :: 532
rRNAs                        :: 8, 7, 7 (5S, 16S, 23S)
complete rRNAs               :: 8, 7, 7 (5S, 16S, 23S)
tRNAs                        :: 87
ncRNAs                       :: 0
##Genome-Annotation-Data-END##
```

Eukaryotic Genome Annotation



Eukaryotic Genome Annotation

Core components of the pipeline are the alignment programs `Splign` (a splicing aware alignment algorithm) and `ProSplign` (a protein to genomic sequence alignment method) and `Gnomon`, a gene prediction program combining information from alignments of experimental evidence and from models produced *ab initio* with an HMM-based algorithm.

from http://www.ncbi.nlm.nih.gov/books/NBK169439/

Eukaryotic Genome Annotation

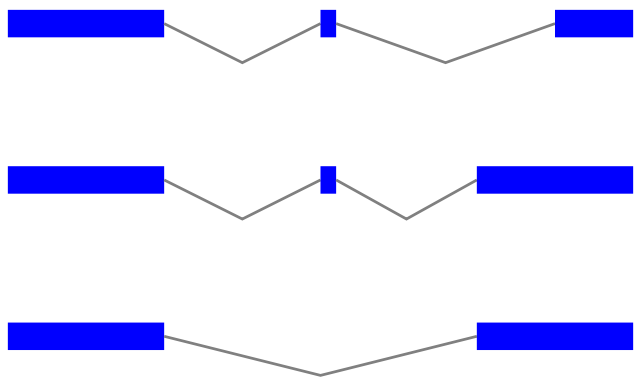
The annotation pipeline produces comprehensive sets of genes, transcripts, and proteins derived from multiple sources, depending on the data available. In order of preference, the following sources are used:

1. RefSeq curated annotated genomic sequences
2. Known RefSeq transcripts
3. `Gnomon`-predicted models - generates gene models using mRNA, EST and protein alignments as evidence, supplemented by *ab initio* models.

Eukaryotic Genome Annotation

The problem with eukaryotes is splicing.

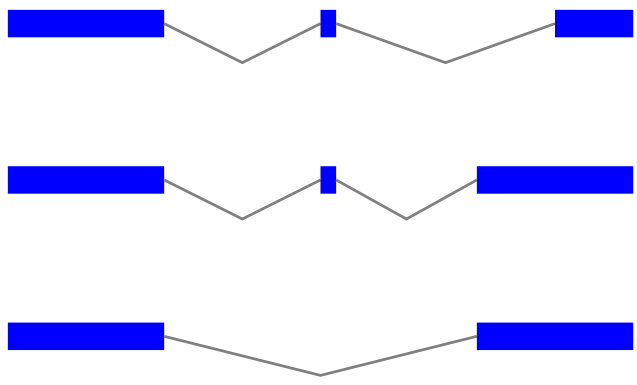
Which is the correct annotation?



Eukaryotic Genome Annotation

The problem with eukaryotes is splicing.

Which is the correct annotation?



But the correct answer could be YES!

Eukaryotic Genome Annotation

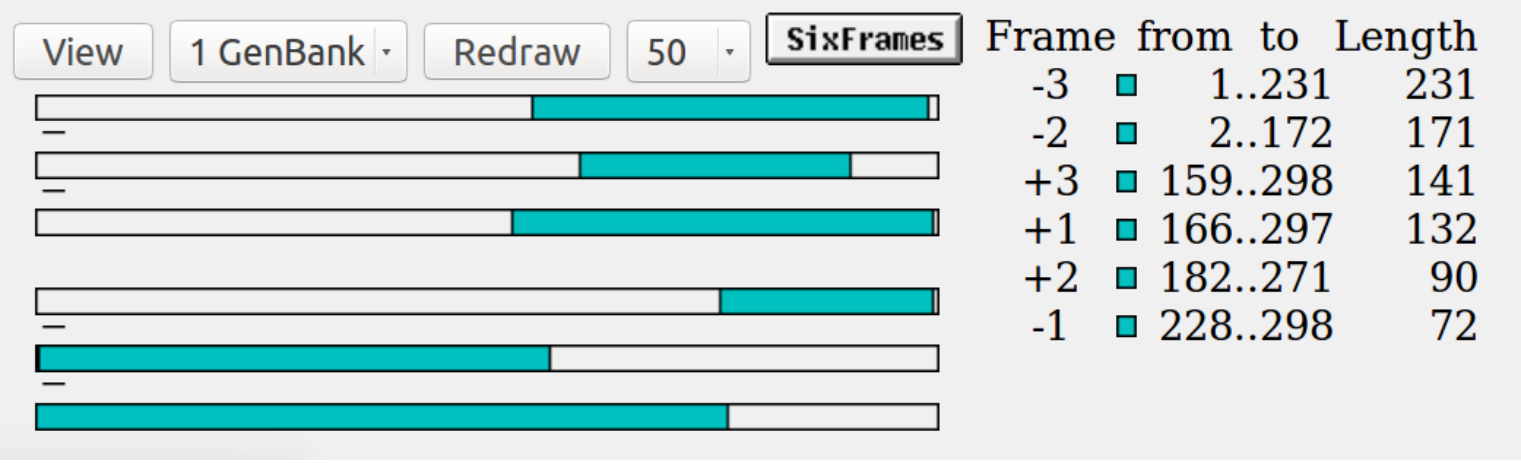
Splices sites make "coding" bits potentially very small. (below is a segment of human chr 4).

```
cgcctcccct ggccccgtgc acacacacgc ccaccgcggc
tcgggctggc tgagcgcggg cgagtgtgag cgcgagtgtg
cgcacgccgc gggagcctct ctgccctctc ctcgcaccct
gctcagggca tctgaagagc ctggaaacgt gaacaggctt
gaagtatggc atgttgcaaa gatggtttct gccaagaagg
tacccgcgat cgctctgtcc gccggggtca gtttcgccct
cctgcgcttc ctgtgcctgg cggtttggtg agttctctgg
ggtgcgagga ggtgggcaa
```

Eukaryotic Genome Annotation

Splices sites make "coding" bits potentially very small. (below is a segment of human chr 4).

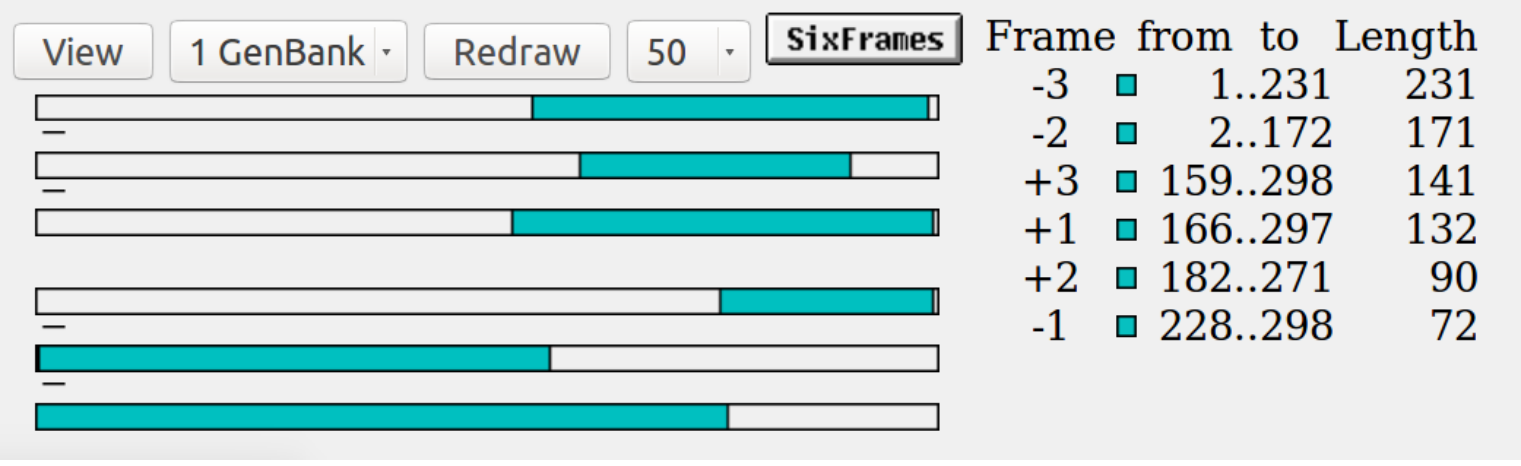
Open reading frames for this region are ...



Eukaryotic Genome Annotation

Splices sites make "coding" bits potentially very small. (below is a segment of human chr 4).

Open reading frames for this region are ...



They are all wrong!

Eukaryotic Genome Annotation

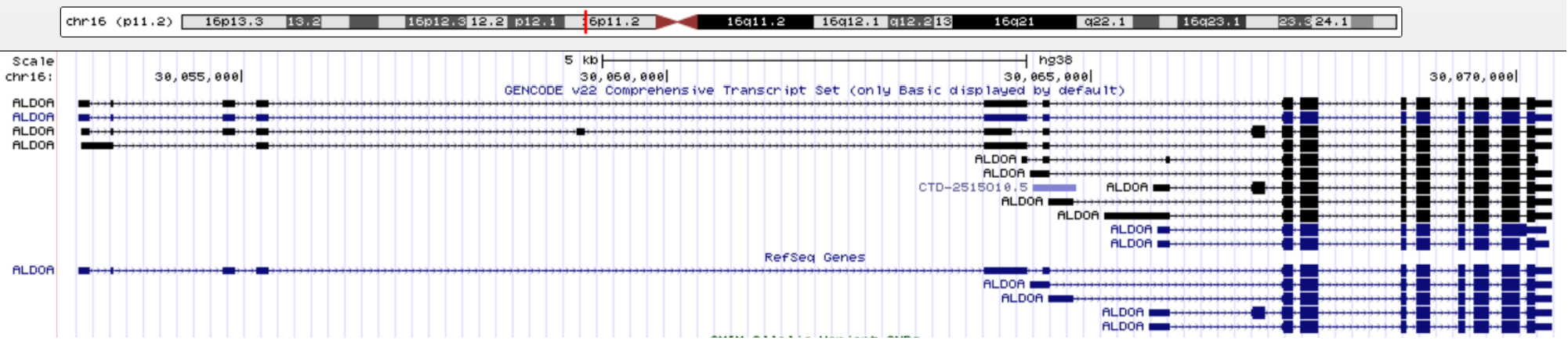
Splices sites make "coding" bits potentially very small. (below is a segment of human chr 4).

```
cgcctcccct ggccccgtgc acacacacgc ccaccgcggc
tcgggctggc tgagcgcggg cgagtgtgag cgcgagtgtg
cgcacgccgc gggagcctct CTGCCCTCTC CTCGCACCCT
GCTCAGGGCA TCTGAAGAGC CTGGAAACGT GAACAGGCTT
GAAGTATGGC ATGTTGCAAA GATGGTTTCT GCCAAGAAGg
tacccgcgat cgctctgtcc gccggggtca gtttcgccct
cctgcgcttc ctgtgcctgg cggtttggtg agttctctgg
ggtgcgagga ggtgggcaa
```

It codes for just 9 amino acids `MLQRWFLPR` and is the first exon of a GABA receptor isoform.

Eukaryotic Genome Annotation

Nor do exons have to be coding. These are splice variants of the human aldolase A gene. For just the first four variants there are 14, 14, 16 and 12 exons of which the first 6, 6, 7 and 4 are non-coding.



How do you find these?

Eukaryotic Genome Annotation

Currently the pipeline incorporates RNA-Seq data where available (as of 2013).

Eukaryotic Genome Annotation

"Both Splign and ProSplign are global alignment tools that enable alignment of transcripts and proteins with high resolution of splice sites. The computational cost of these algorithms requires that approximate placements of the query sequences (transcripts or proteins) on the target (genome) be first identified with a local alignment tool, such as BLAST.

Since a query often aligns at multiple locations, the BLAST hits are analyzed by the Compart algorithm to identify compartments prior to running Splign or ProSplign."

http://www.ncbi.nlm.nih.gov/books/NBK169439/

Eukaryotic Genome Annotation

The `Compart` algorithm sorts through the BLAST hits to find a compatible set that make sense relative to query sequences and to the genome. For example, they cannot be in the order ABC in the query and then in order CAB in the genome.

The resulting compartments of hits are then given to the global alignment tools `Splign` and `ProSplign`.

Eukaryotic Genome Annotation

The global alignment tools `Splign` and `ProSplign` use a modified Needleman-Wunsch algorithm with a score modified to include a difference between alignment (or evolutionary) gaps and intronic gaps.

Thus the score is constructed from the number of matches, number of mismatches, and normal gap penalties, plus gap penalities for intronic gaps.

The intron gap opening cost is different between the most frequent consensus splices (GT/AG), less frequent consensus splices (GC/AG, AT/AC), and non-consensus splice sites.

These alignments are then given to a `chainer` algorithm.

Eukaryotic Genome Annotation

The `chainer` combines RNA-Seq alignments together in order to reduce the complexity of the problem.

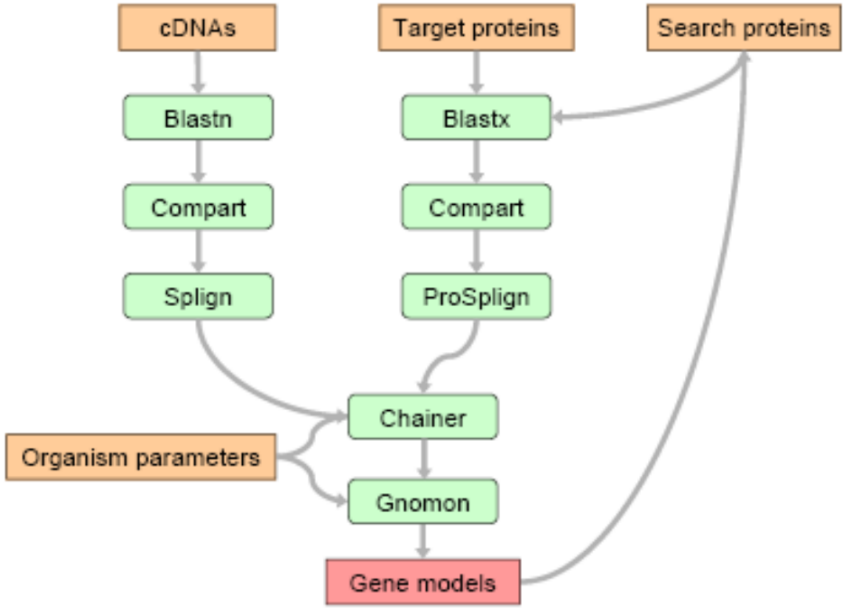
The `chainer` also takes into account intron structure and coding potential.

Eukaryotic Genome Annotation

The `Gnomon` algorithm takes the results from `chainer` and tries to predict coding sequence and possible isoforms.

Possible coding regions are predicted and scored using a "3-periodic fifth-order Hidden Markov Model (HMM) for coding propensity and Weight Matrix Method (WMM) models for splice signals and translation initiation and termination signals".

Eukaryotic Genome Annotation



The availabe cDNAs and proteins are used to build the first predictions. These gene models (or 'candidates') are compared with a broad set of proteins. Good matches continue in the pipeline. `Compart` finds the positions of the target sequences, `Splign` & `ProSplign` build spliced alignments, `Chainer` combines these into longer models, `Gnomon` extends and finalizes the gene model.

This then goes into all of the databases.

Eukaryotic Genome Annotation

- Example entries from GRCh38.p7 chr 21

```
LOCUS       NC_000021           46709983 bp    DNA     linear   CON 06-JUN-2016
DEFINITION  Homo sapiens chromosome 21, GRCh38.p7 Primary Assembly.
ACCESSION   NC_000021 GPC_000001313
VERSION     NC_000021.9  GI:568815577
DBLINK      BioProject: PRJNA168
            Assembly: GCF_000001405.33
KEYWORDS    RefSeq.
SOURCE      Homo sapiens (human)

.......

##Genome-Annotation-Data-START##
Annotation Provider         :: NCBI
Annotation Status           :: Full annotation
Annotation Version          :: Homo sapiens Annotation Release 108
Annotation Pipeline         :: NCBI eukaryotic genome annotation
                               pipeline
Annotation Software Version :: 7.0
Annotation Method           :: Best-placed RefSeq; Gnomon
Features Annotated          :: Gene; mRNA; CDS; ncRNA
##Genome-Annotation-Data-END##

.......
```

Eukaryotic Genome Annotation

- Example entries from GRCh38.p7 chr 21

```
.......

ncRNA  complement(join(5011976..5012370,5012556..5012684))
       /ncRNA_class="lncRNA"
       /gene="LOC105379484"
       /product="uncharacterized LOC105379484"
       /note="Derived by automated computational analysis using
       gene prediction method: Gnomon. Supporting evidence
       includes similarity to: 1 EST, and 100% coverage of the
       annotated genomic feature by RNAseq alignments, including
       37 samples with support for all annotated introns"
       /transcript_id="XR_951076.1"
       /db_xref="GI:768019556"
       /db_xref="GeneID:105379484"

.......
```

Eukaryotic Genome Annotation

- Example entries from GRCh38.p7 chr 21

```
.......

mRNA    join(5022044..5022693,5025009..5025049,5026280..5026630,
        5027935..5028225,5032053..5032217,5033408..5033443,
        5045419..5046678)
        /gene="LOC102723996"
        /product="ICOS ligand, transcript variant X2"
        /note="Derived by automated computational analysis using
        gene prediction method: Gnomon. Supporting evidence
        includes similarity to: 65 ESTs, 5 long SRA reads, 4
        Proteins, and 100% coverage of the annotated genomic
        feature by RNAseq alignments, including 13 samples with
        support for all annotated introns"
        /transcript_id="XM_011546078.2"
        /db_xref="GI:1034627737"
        /db_xref="GeneID:102723996"
mRNA    join(5022044..5022693,5025009..5025049,5026280..5026630,
        5027935..5028225,5032053..5032217,5033408..5033443,
        5034582..5036775)
        /gene="LOC102723996"
        /product="ICOS ligand, transcript variant X4"
        /note="Derived by automated computational analysis using
        gene prediction method: Gnomon. Supporting evidence
        includes similarity to: 8 mRNAs, 78 ESTs, 108 long SRA
        reads, 4 Proteins, and 100% coverage of the annotated
        genomic feature by RNAseq alignments, including 101
        samples with support for all annotated introns"
        /transcript_id="XM_006723900.2"
        /db_xref="GI:768019569"
        /db_xref="GeneID:102723996"

.......
```

Eukaryotic Genome Annotation

- Example entries from GRCh38.p7 chr 21

```
.......

gene    5553637..5592004
        /gene="LOC102724219"
        /note="uncharacterized LOC102724219; Derived by automated
        computational analysis using gene prediction method:
        BestRefSeq,Gnomon."
        /db_xref="GeneID:102724219"
mRNA    join(5553637..5553838,5585823..5585987,5590019..5592004)
        /gene="LOC102724219"
        /product="uncharacterized LOC102724219"
        /note="Derived by automated computational analysis using
        gene prediction method: BestRefSeq."
        /transcript_id="NM_001322044.1"
        /db_xref="GI:1015130023"
        /db_xref="GeneID:102724219"

.......
```

Eukaryotic Genome Annotation

- Example entries from GRCh38.p7 chr 21

```
.......

gene    complement(17454789..17454859)
        /gene="TRG-GCC1-5"
        /note="Derived by automated computational analysis using
        gene prediction method: tRNAscan-SE."
        /db_xref="GeneID:100189284"
        /db_xref="HGNC:HGNC:34850"
tRNA    complement(17454789..17454859)
        /gene="TRG-GCC1-5"
        /product="tRNA-Gly"
        /inference="COORDINATES: profile:tRNAscan-SE:1.23"
        /note="transfer RNA-Gly (GCC) 1-5; tRNA features were
        annotated by tRNAscan-SE; Derived by automated
        computational analysis using gene prediction method:
        tRNAscan-SE."
        /anticodon=(pos:complement(17454825..17454827),aa:Gly,
        seq:gcc)
        /db_xref="GeneID:100189284"
        /db_xref="HGNC:HGNC:34850"

.......
```

Eukaryotic Genome Annotation

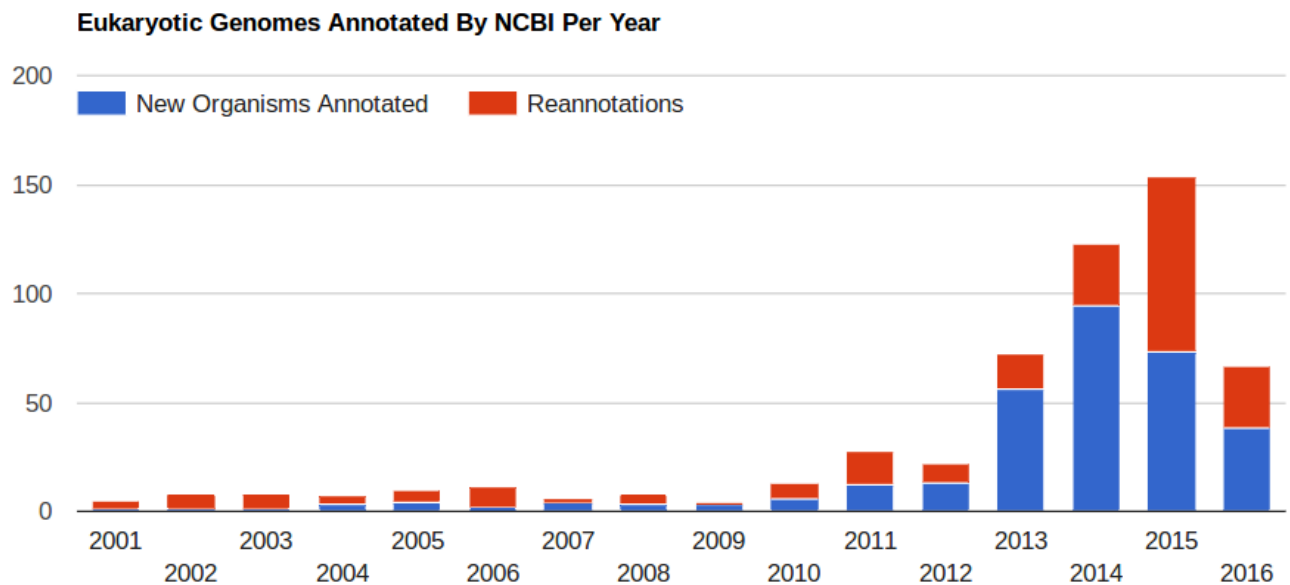
- Example entries from GRCh38.p7 chr 21

```
.......

gene  19620752..19631774
      /gene="NIPA2P3"
      /note="non imprinted in Prader-Willi/Angelman syndrome 2
      pseudogene 3; Derived by automated computational analysis
      using gene prediction method: Curated Genomic."
      /pseudo
      /db_xref="GeneID:100128057"
      /db_xref="HGNC:HGNC:42043"

.......
```

Not the end

Finally, note that this is not the end.

Genomes are constantly being re-annotated as new information is gathered and as algorithms are improved.



... hence record the annotation version if your work uses it.