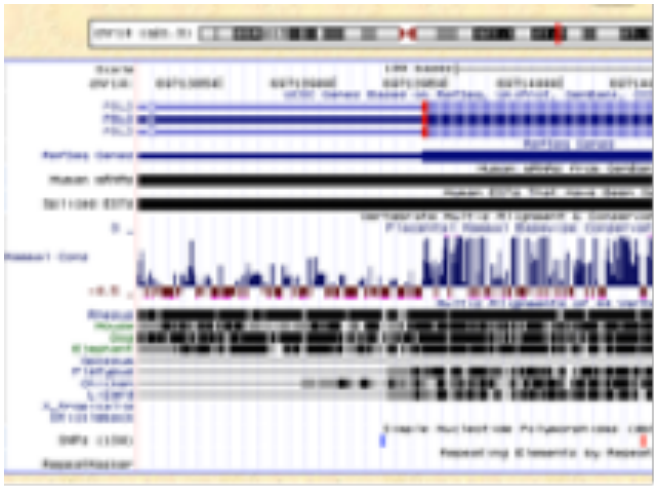
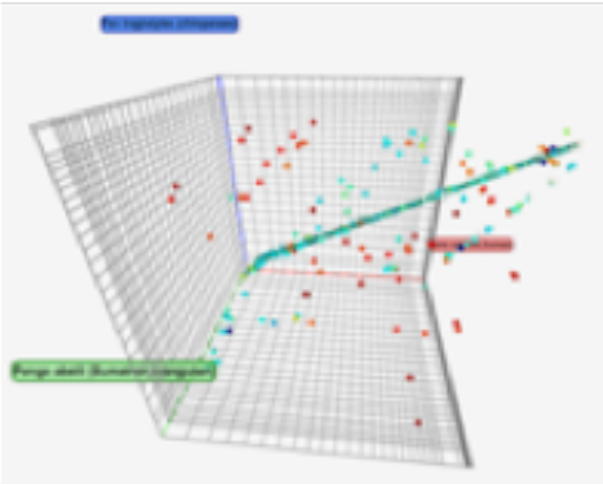
Computational Genomics

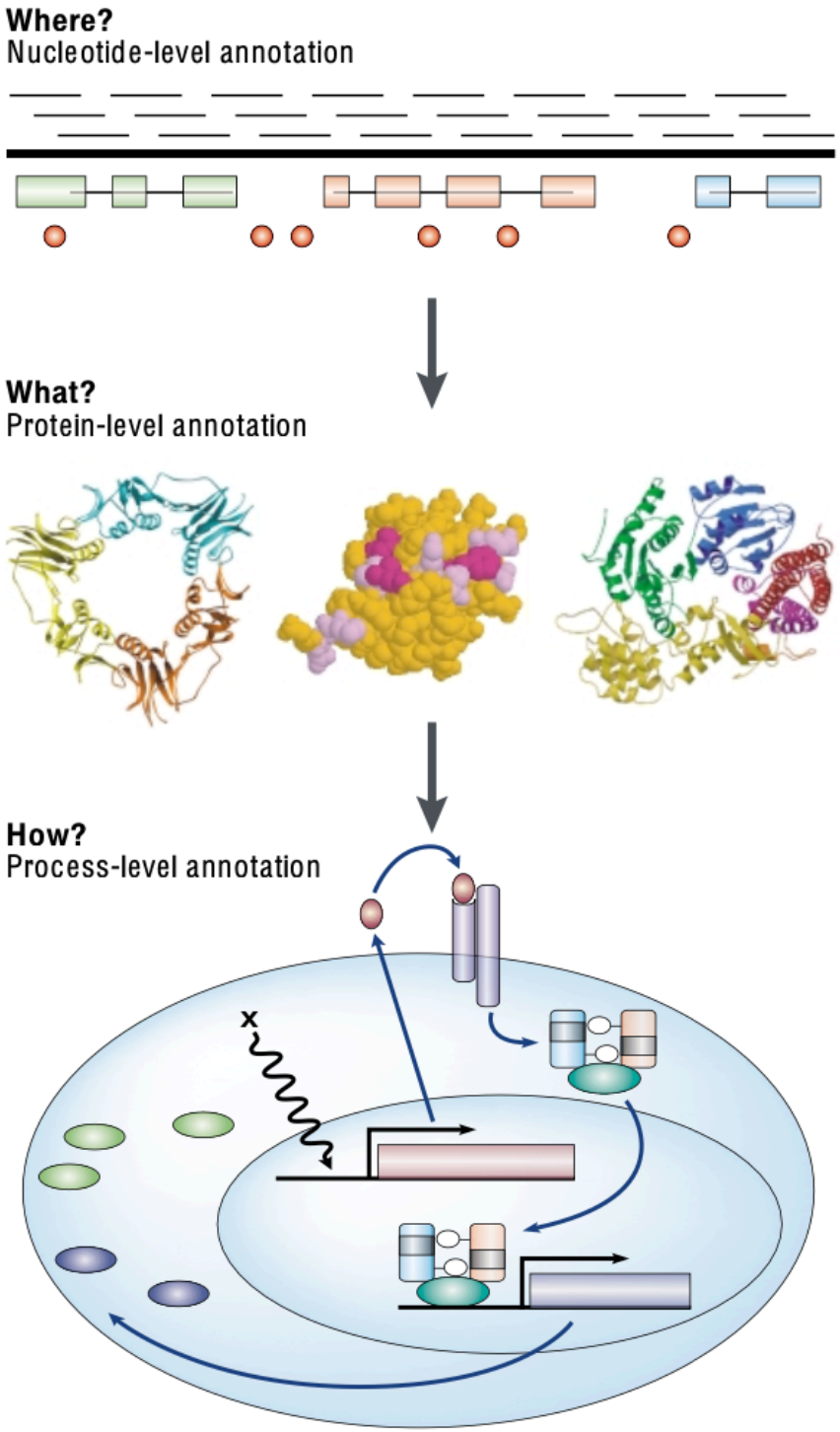
Introduction To Genome Annotation

From Sequence To Biology





The Three Layers of Genome Annotation



https://pubmed.ncbi.nlm.nih.gov/11433356/

The Three Layers of Genome Annotation

Nucleotide-level annotation

- Mapping
- Finding genomic landmarks
- Finding Genes
- Finding non-coding RNAs and regulatory regions
- Identifying repetitive elements
- Mapping segmental duplications
- Mapping variations

Protein-level annotation

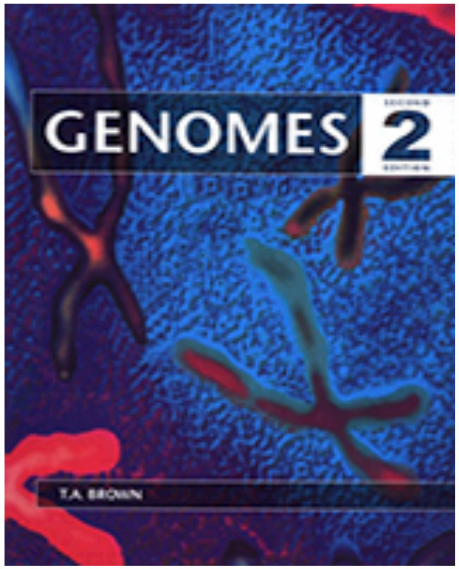
- Generating a “Taxonomy” of proteins. This is organizing them into classes, like DNA Polymerases, Kinases, etc.
- Organizing proteins into protein complexes

Process-level annotation

- How do different proteins interact with each other (same protein complexes) and/or belong to the same metabolic pathway(s)?

The sociology of genome annotation

- Organizing genome annotation efforts
- Publishing and sharing annotations
- Bringing annotation into the mainstream



https://www.ncbi.nlm.nih.gov/books/NBK21134/

Training materials

The Three Layers of Genome Annotation

Locating the Genes in a Genome Sequence

Once a DNA sequence has been obtained, whether it is the sequence of a single cloned fragment or of an entire chromosome, then various methods can be employed to locate the genes that are present. These methods can be divided into those that involve simply inspecting the sequence, by eye or more frequently by computer, to look for the special sequence features associated with genes, and those methods that locate genes by experimental analysis of the DNA sequence. The computer methods form part of the methodology called bioinformatics.

Gene location by sequence inspection. Sequence inspection can be used to locate genes because genes are not random series of nucleotides but instead have distinctive features. These features determine whether a sequence is a gene or not, and so by definition are not possessed by non-coding DNA. At present we do not fully understand the nature of these specific features, and sequence inspection is not a foolproof way of locating genes, but it is still a powerful tool and is usually the first method that is applied to analysis of a new genome sequence.

The coding regions of genes are open reading frames. Genes that code for proteins comprise open reading frames (ORFs) consisting of a series of codons that specify the amino acid sequence of the protein that the gene codes for (see Figure). The ORF begins with an initiation codon - usually (but not always) ATG - and ends with a termination codon: TAA, TAG or TGA. Searching a DNA sequence for ORFs that begin with an ATG and end with a termination triplet is therefore one way of looking for genes. The analysis is complicated by the fact that each DNA sequence has six reading frames, three in one direction and three in the reverse direction on the complementary strand, but computers are quite capable of scanning all six reading frames for ORFs. How effective is this as a means of gene location?

https://www.ncbi.nlm.nih.gov/books/NBK21134/

The Three Layers of Genome Annotation

Locating the Genes in a Genome Sequence

- The key to the success of ORF scanning is the frequency with which termination codons appear in the DNA sequence.
- If the DNA has a random sequence and a GC content of 50% then each of the three termination codons - TAA, TAG and TGA - will appear, on average, once every $4^3 = 64$ bp.
- If the GC content is > 50% then the termination codons, being AT-rich, will occur less frequently but one will still be expected every 100–200 bp.
- This means that random DNA should not show many ORFs longer than 50 codons in length, especially if the presence of a starting ATG is used as part of the definition of an 'ORF'.
- Most genes, on the other hand, are longer than 50 codons: the average lengths are 317 codons for *Escherichia coli*, 483 codons for *Saccharomyces cerevisiae*, and approximately 450 codons for *Homo sapiens*.
- ORF scanning, in its simplest form, therefore takes a figure of, say, 100 codons as the shortest length of a putative gene and records positive hits for all ORFs longer than this.
- How well does this strategy work in practice? With bacterial genomes, simple ORF scanning is an effective way of locating most of the genes in a DNA sequence.

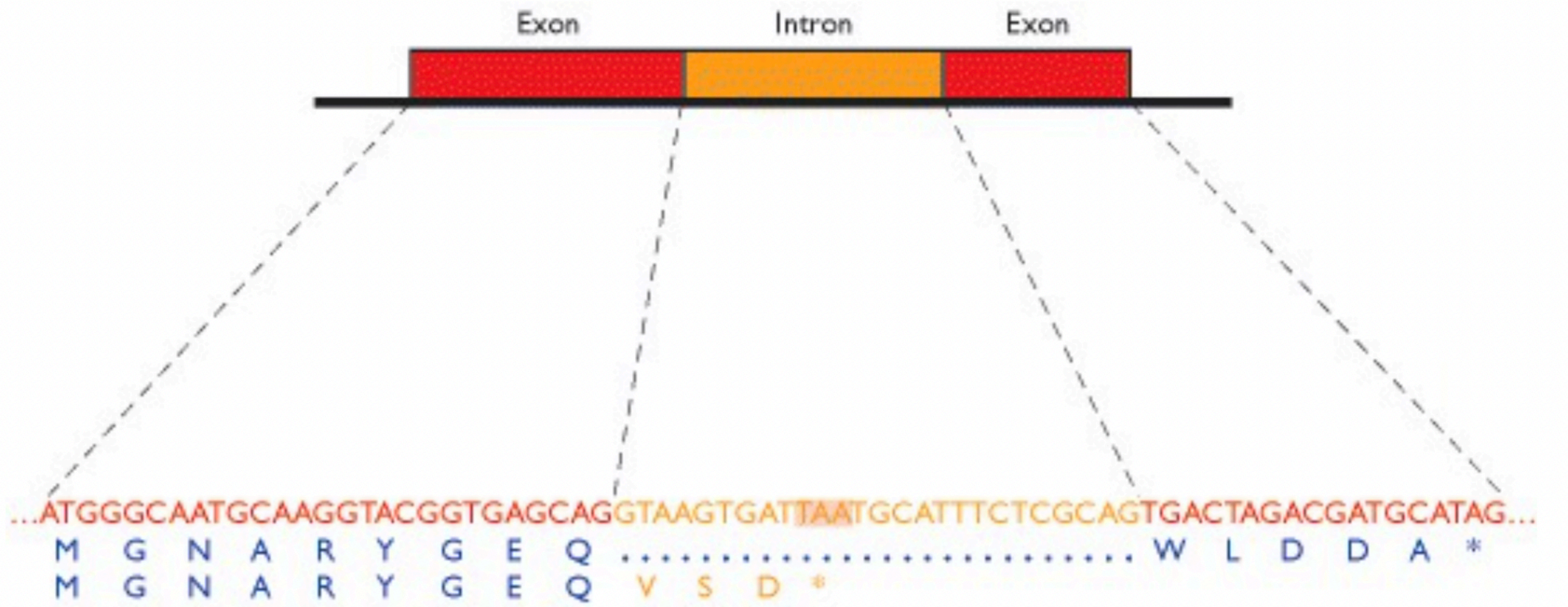
Playing with Open Reading Frame Finder

https://www.ncbi.nlm.nih.gov/books/NBK21134/

The Three Layers of Genome Annotation

Simple ORF scans are less effective with higher eukaryotic DNA

- ORF scans work well for bacterial genomes, but they are less effective for locating genes in DNA sequences from higher eukaryotes. This is partly because there is substantially more space between the real genes in a eukaryotic genome
- The main problem with the human genome and the genomes of higher eukaryotes in general is that their genes are often split by introns, and so do not appear as continuous ORFs in the DNA sequence.
- Many exons are shorter than 100 codons, some fewer than 50 codons, and continuing the reading frame into an intron usually leads to a termination sequence that appears to close the ORF. In other words, the genes of a higher eukaryote do not appear in the genome sequence as long ORFs, and simple ORF scanning cannot locate them.



https://www.ncbi.nlm.nih.gov/books/NBK21134/

Solving The Intron Challenge In Gene Annotation

Codon Bias

- Codon bias is taken into account. ‘Codon bias’ refers to the fact that not all codons are used equally frequently in the genes of a particular organism. For example, leucine is specified by six codons in the genetic code (TTA, TTG, CTT, CTC, CTA and CTG), but in human genes leucine is most frequently coded by CTG and is only rarely specified by TTA or CTA. Similarly, of the four valine codons, human genes use GTG four times more frequently than GTA. The biological reason for codon bias is not understood, but all organisms have a bias, which is different in different species. Real exons are expected to display the codon bias whereas chance series of triplets do not. The codon bias of the organism being studied is therefore written into the ORF scanning software.

Exon-intron boundaries

- ***Exon-intron boundaries can be searched for as these have distinctive sequence features, although unfortunately the distinctiveness of these sequences is not so great as to make their location a trivial task. The sequence of the upstream, exon-intron boundary is usually described as:***

  5’-AG↓GTAAGT-3’

  the arrow indicating the precise boundary point. However, only the ‘GT’ immediately after the arrow is invariable; elsewhere in the sequence nucleotides other than the ones shown are quite often found. In other words, the sequence shown is a consensus - the average of a range of variabilities. The downstream intron-exon boundary is even less well defined:

  5’-PyPyPyPyPyPyNCAG↓-3’

  where ‘Py’ means one of the pyrimidine nucleotides (T or C) and ‘N’ is any nucleotide. Simply searching for the consensus sequences will not locate more than a few exon-intron boundaries because most have sequences other than the ones shown. Writing software that takes account of the known variabilities has proven difficult, and at present locating exon-intron boundaries by sequence analysis is a hit-and-miss affair.

Solving The Intron Challenge In Gene Annotation

Upstream regulatory sequences

- **Upstream regulatory sequences can be used to locate the regions where genes begin. This is because these regulatory sequences, like exon-intron boundaries, have distinctive sequence features that they possess in order to carry out their role as recognition signals for the DNA-binding proteins involved in gene expression. Unfortunately, as with exon-intron boundaries, the regulatory sequences are variable, more so in eukaryotes than in prokaryotes, and in eukaryotes not all genes have the same collection of regulatory sequences. Using these to locate genes is therefore problematic.**

Final Considerations

- ***These three extensions of simple ORF scanning are generally applicable to all higher eukaryotic genomes. Additional strategies are also possible with individual organisms, based on the special features of their genomes.***
- ***For example, vertebrate genomes contain CpG islands upstream of many genes, these being sequences of approximately 1 kb in which the GC content is greater than the average for the genome as a whole.***
- ***Some 40–50% of human genes are associated with an upstream CpG island. These sequences are distinctive and when one is located in vertebrate DNA, a strong assumption can be made that a gene begins in the region immediately downstream.***

Gene annotation in Ensembl

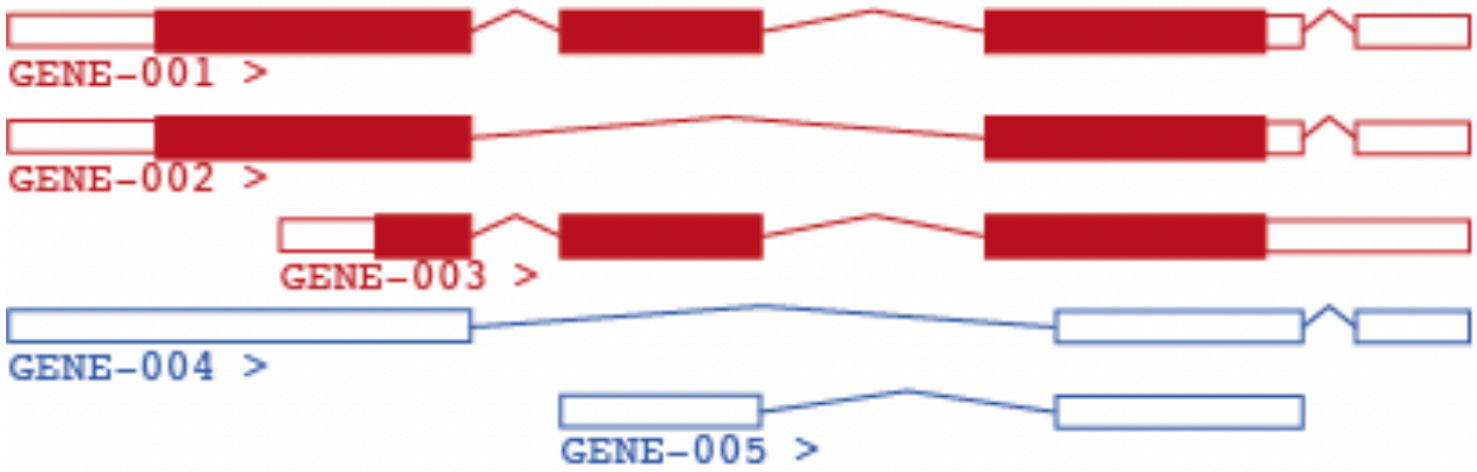
Gene annotation is the plotting of genes onto genome assemblies, and indexing their genomic coordinates.

Gene annotation provided by Ensembl includes automatic annotation, ie genome-wide determination of transcripts. For selected species (ie human, mouse, zebrafish, rat), gene annotation may also include manual curation, ie reviewed determination of transcripts on a case-by-case basis. Furthermore, Ensembl imports annotation from FlyBase, WormBase and SGD.

Ensembl transcripts displayed on our website are products of the Ensembl automatic gene annotation system (a collection of gene annotation pipelines), termed the Ensembl annotation process. All Ensembl transcripts are based on experimental evidence and thus the automated pipeline relies on the mRNAs and protein sequences deposited into public databases from the scientific community. Manually-curated transcripts are produced by the HAVANA group.

An Ensembl gene (with a unique ENSG... ID) includes any spliced transcripts (ENST...) with overlapping coding sequence, with the exception of manually annotated readthrough genes which are annotated as a separate locus. Transcripts from the Ensembl annotation process, the Havana/Vega set and the Consensus Coding Sequence (CCDS project) set may all be clustered into the same gene. Transcripts that belong to the same gene ID may differ in transcription start and end sites, splice events and exons, and can give rise to very different proteins. Transcript clusters with no overlapping coding sequence are annotated as separate genes. Two transcripts may overlap in non-coding sequence (ie intronic sequence or UnTranslated Region (UTR), and be classified under two separate genes. After the Ensembl gene and transcript sequences are defined, the gene and transcript names are assigned.

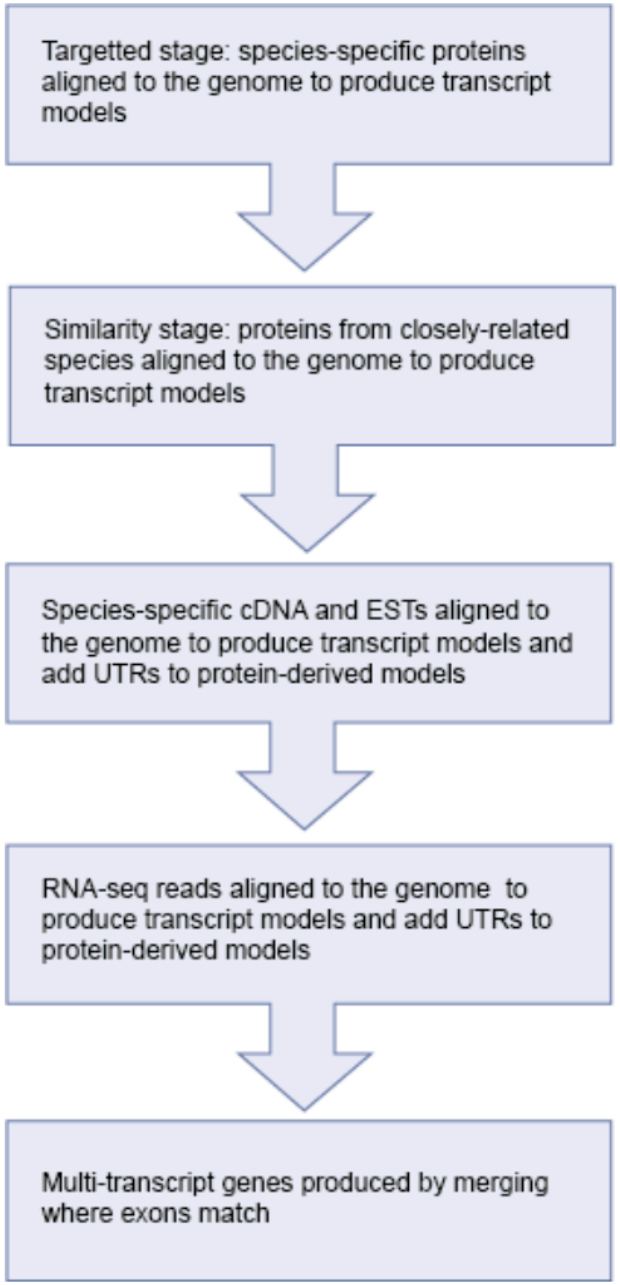
The image below shows a cartoon of a gene ("GENE") with five transcripts, some coding (red) and non-coding (blue).



https://www.ensembl.org/info/genome/genebuild/index

Automatic annotation of coding genes

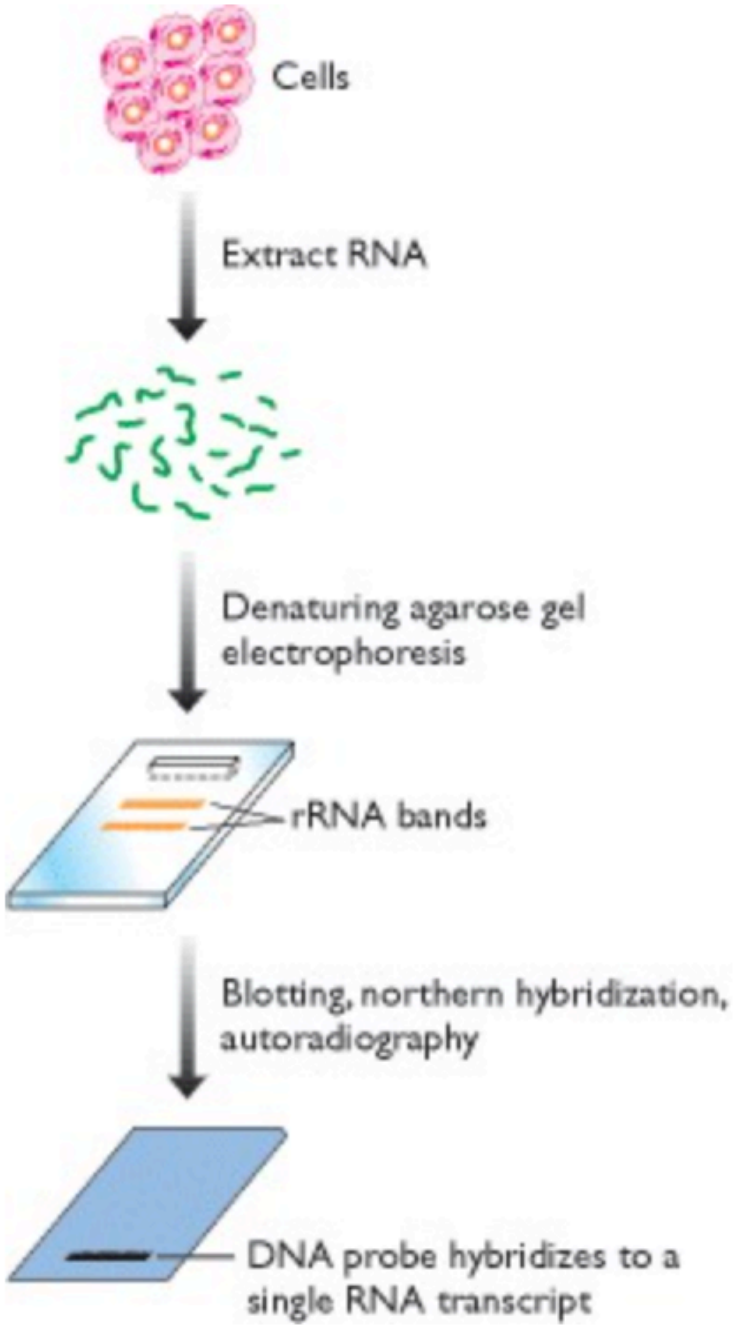
There are several steps in the annotation of coding genes onto repeat-masked genomes. These steps are summarized in the following diagram and described in detail below:



- **The first stage of the Ensembl annotation process is known as the Targeted stage. Here, species-specific proteins are aligned to the genome and Genewise and/or Exonerate is used to build a transcript structure for the protein on the genome.**
- **The Targeted stage is followed by the Similarity stage in which proteins from closely related species are used to build transcript structure in regions were a Targeted transcript structure is absent. For those species having a lot of experimentally-generated protein sequences, the Targeted stage tends to contribute most of the gene structures in the Ensembl annotation process. However, for species with fewer species-specific protein sequences the Similarity stage plays a much more important role in predicting gene structures.**
- **The next stage in the Ensembl annotation process is to align species-specific cDNA and EST sequences to the genome. Where cDNA alignments overlap transcripts predicted in the preceding stages, any non-translated region from the cDNA is spliced onto the transcript prediction as UTR. EST alignments are displayed on our website but are usually not used as supporting evidence in the Ensembl annotation process.**
- **Where available, we also align RNA-seq reads to the genome and build RNA-seq-based transcript models from these data. These models may also be used to add new genes or transcript isoforms into the gene set, and to add UTR to protein-coding models.**
- **The final set of transcript predictions is obtained by merging identical transcripts built from different proteins sequences to produce multi-transcript gene predictions, each with a non-redundant set of transcripts models. For every transcript model, the protein and mRNA sequences used to predict the model is viewable in the browser as 'supporting evidence'.**
- **Where transcripts have identical exons, these are combined to form a gene.**

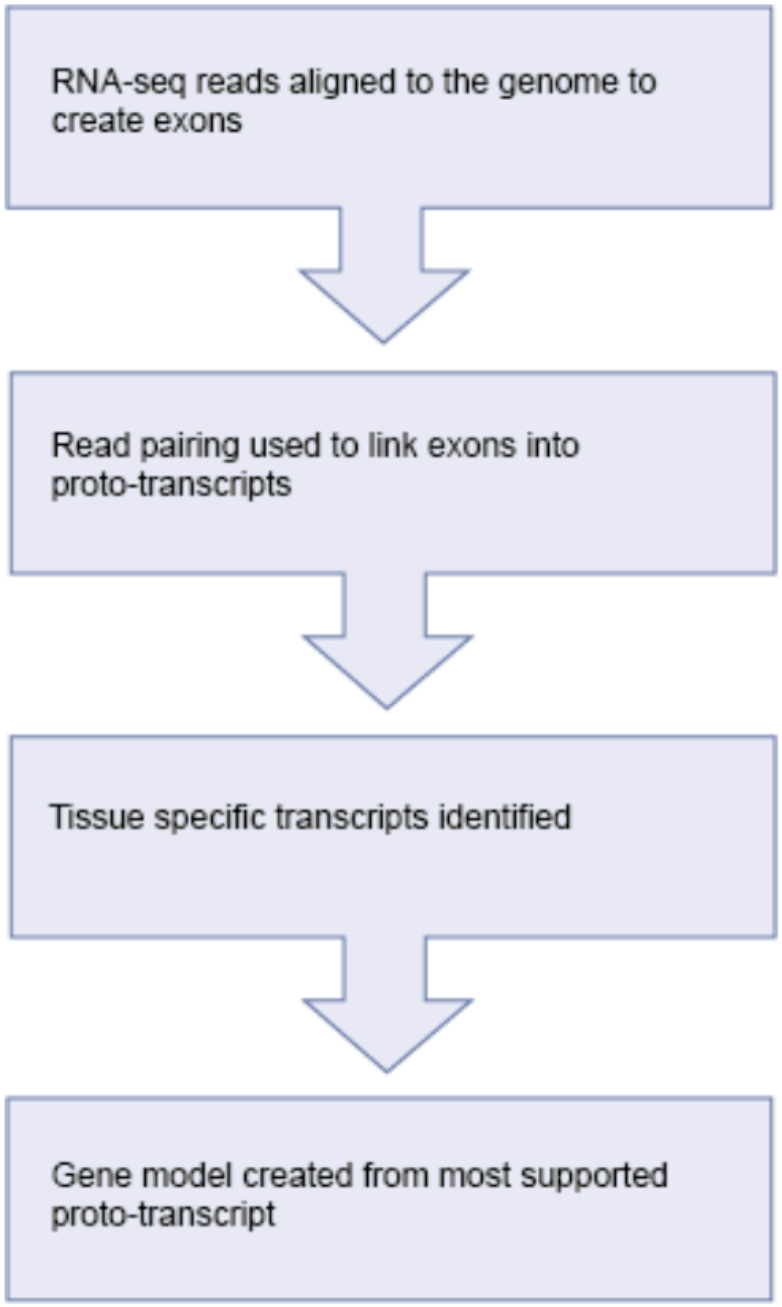
https://www.ensembl.org/info/genome/genebuild/index

Experimental techniques for gene location



https://www.ncbi.nlm.nih.gov/books/NBK21134/

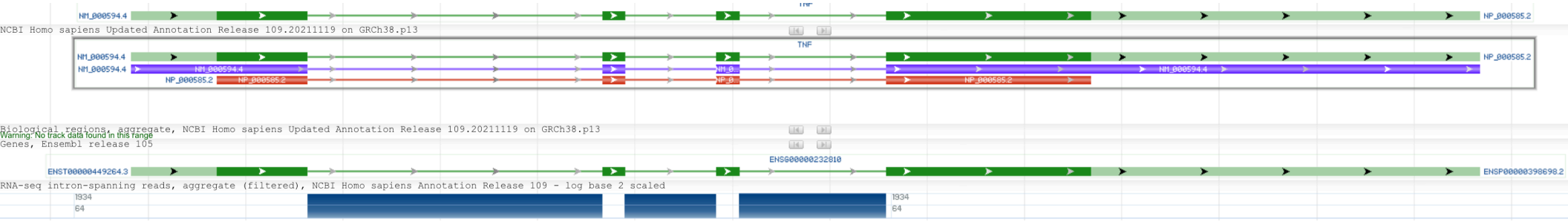
Automatic annotation using RNA-seq data



https://www.ensembl.org/info/genome/genebuild/index.html

Experimental techniques for gene location

Automatic annotation using RNA-seq data



https://www.ncbi.nlm.nih.gov/genome/gdv/browser/gene/?

Automatic annotation of non-coding genes

Non-coding RNAs (ncRNAs) are involved in many biological processes and are increasingly seen as important. As is the case with proteins, it is the overall structure of the molecule which imparts function. However, while similar protein structures are often reflected in a conserved amino acid sequence, sequences underlying RNA secondary structure are very variable; this makes ncRNAs difficult to detect using sequence alone.

Types of ncRNA

Abbreviation	Definition
tRNA	transfer RNA
Mt-tRNA	transfer RNA located in the mitochondrial genome
rRNA	ribosomal RNA
scRNA	small cytoplasmic RNA
snRNA	small nuclear RNA
snoRNA	small nucleolar RNA
miRNA	microRNA precursors
misc_RNA	miscellaneous other RNA
lincRNA	Long intergenic non-coding RNAs

Annotation Details

Most ncRNAs are annotated by aligning genomic sequence against RFAM using BLASTN. The BLAST hits are clustered and filtered by E-value and are used to seed Infernal searches of the locus with the corresponding RFAM covariance models. The purpose of this is to reduce the search space required, as to scan the entire genome with all the RFAM covariance models would be extremely CPU-intensive. The resulting BLAST hits are then used as supporting evidence for ncRNA genes.

Some ncRNAs have specialised annotation.

- **miRNAs.** miRNAs are imported from miRBase. All species are used.
- **tRNAs.** tRNAs are annotated as part of the raw compute process using tRNAscan-SE. Because of this, they are not included as genes in the database, but as Simple Features instead.
- **lincRNA.** lincRNA (Long intergenic non-coding RNAs) Ensembl gene annotation, cDNA alignments and chromatin-state map data from the Ensembl regulatory build are used to predict lincRNAs for human and mouse. We do not import the lincRNAs identified by Guttman et al [1], but their publication guided us to our current approach for automatically annotating lincRNAs. First, regions of chromatin methylation (H3K4me3 and H3K36me3) outside known protein-coding loci are identified. Next, cDNAs which overlap with H3K4me3 or H3K36me3 features are identified as candidate lincRNAs. A final evaluation step investigates if each candidate lincRNA has any protein-coding potential. Any candidate lincRNA containing a substantial open reading frame (ORF) covering 35% or more of its length and containing PFAM/tigrfam protein domains will be rejected. Candidate lincRNAs that pass the final evaluation step are included in the human or mouse gene set as lincRNA genes.

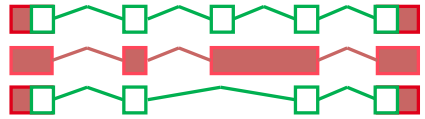
https://www.ensembl.org/info/genome/genebuild/index

1. Ensembl / GENCODE basics

GENCODE / Ensembl human gene annotation

60,651 genes in human v37, 234,485 transcript models

Transcript models



19,951 protein-coding genes

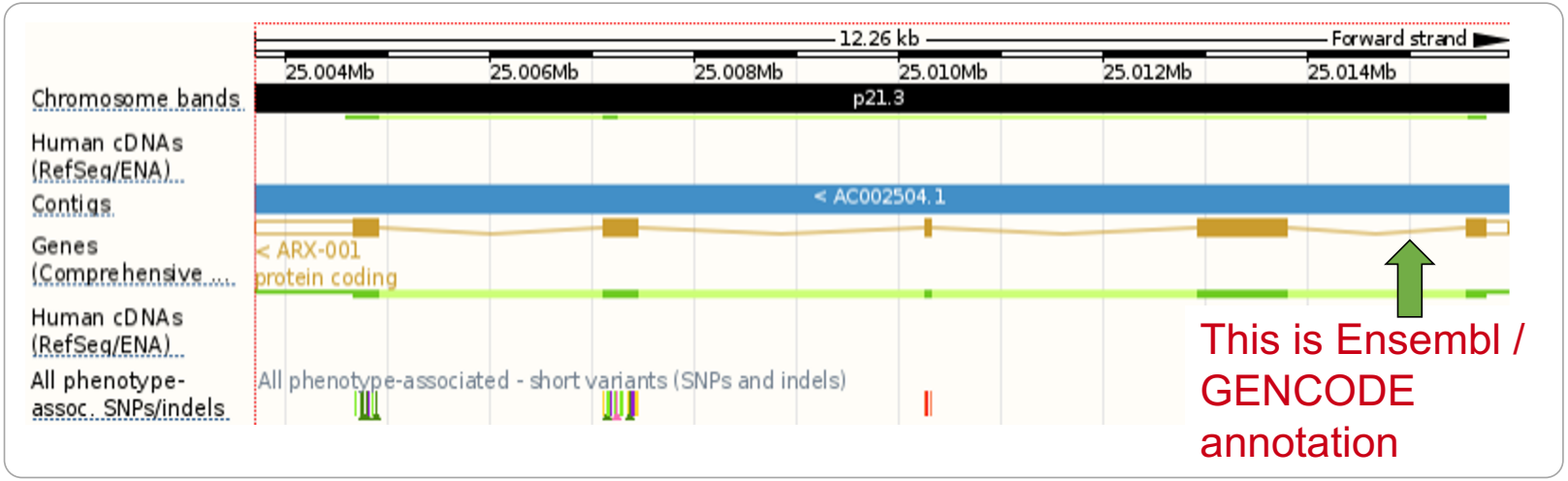
17,948 long non-coding RNA genes

… also pseudogenes and small RNAs

coding sequence
non-coding

The GENCODE <–> Ensembl relationship

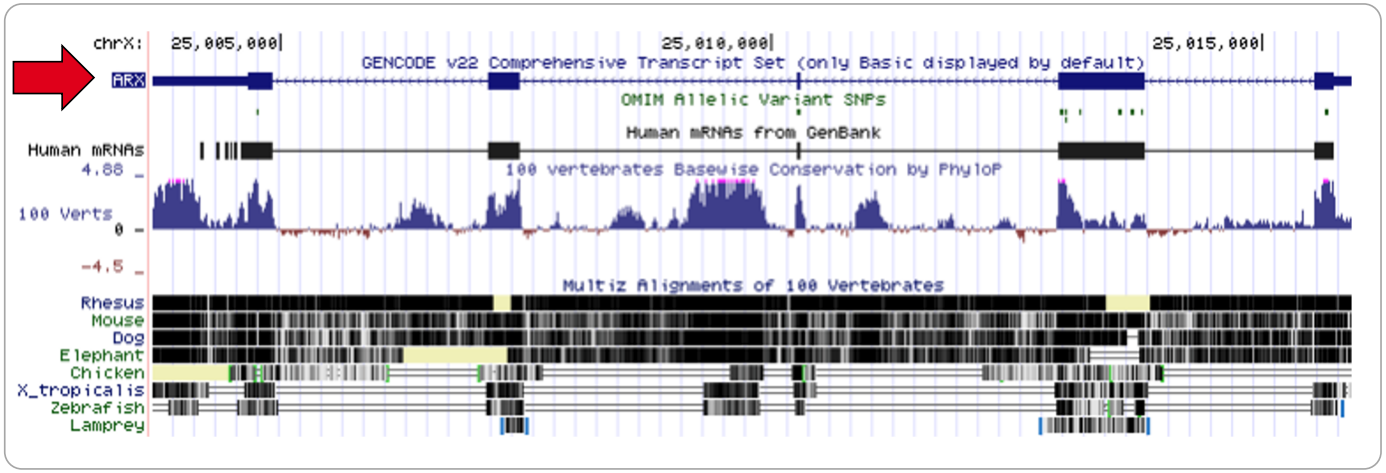
GENCODE human annotation = Ensembl human annotation



GENCODE also annotate mouse

GENCODE is the default UCSC human gene annotation

UCSC Genome Bioinformatics



Currently UCSC update for every other release

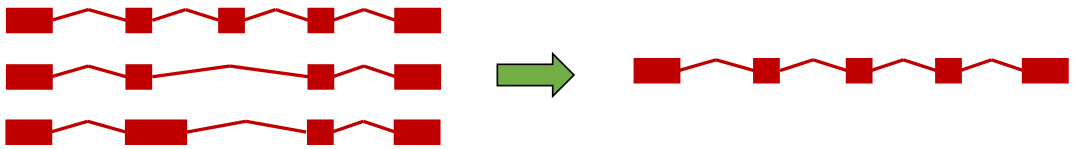
We offer ‘filtered’ transcript sets

‘**Comprehensive**’: the complete set of GENCODE annotations

‘**Basic**’: a smaller set based on filtering

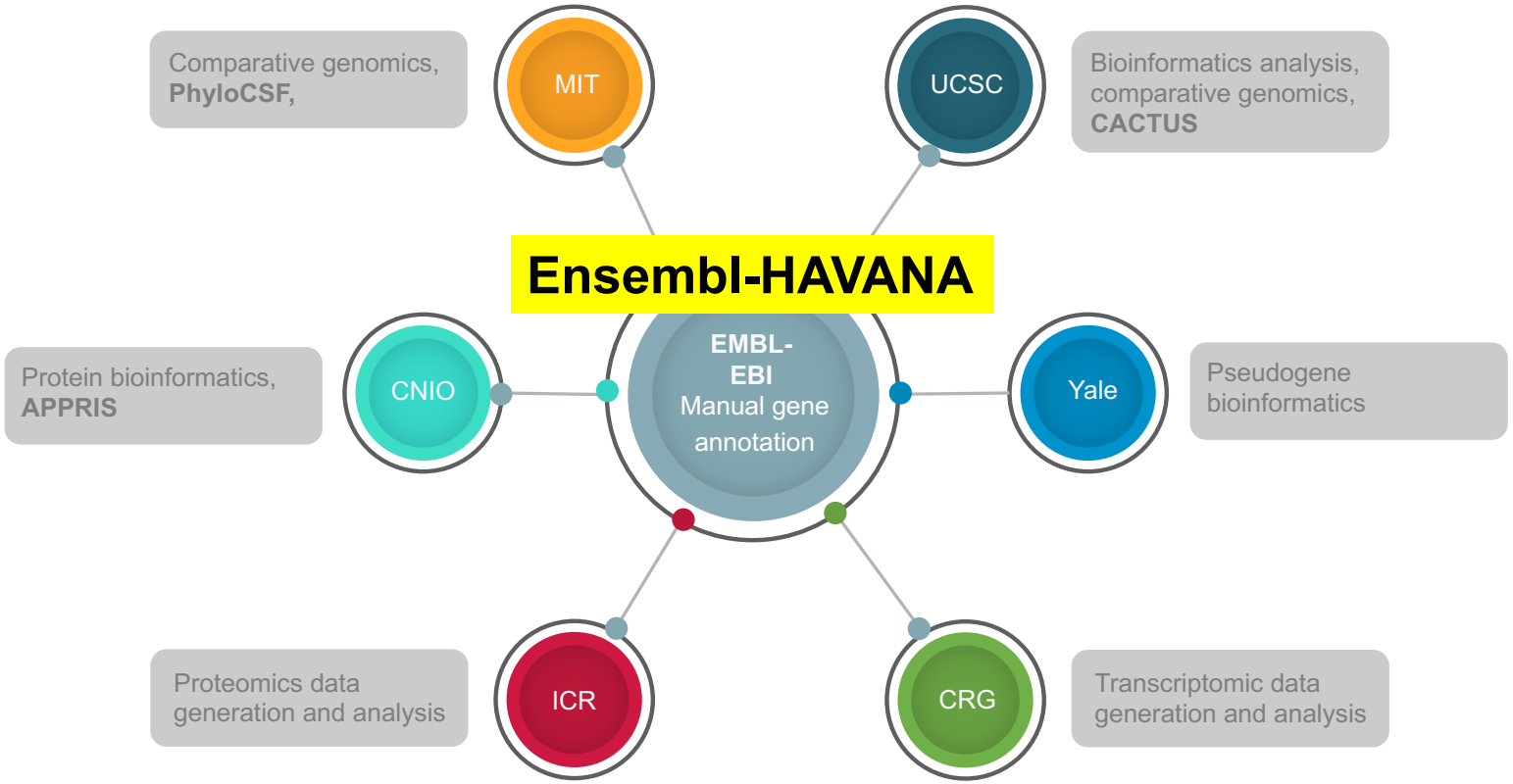
Protein-coding genes contain only models with full-length CDS

lncRNA genes contain minimum number of transcripts that provide 80% of the exonic coverage



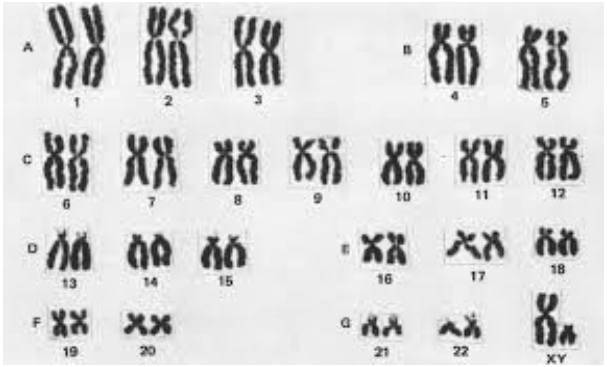
Now: working on transcript choice with RefSeq as **MANE project**

GENCODE is a multifaceted consortium



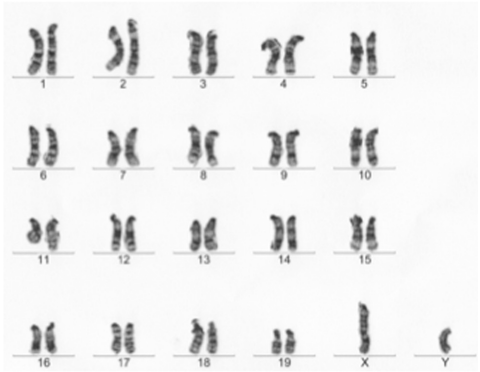
The ‘first pass’ annotation was completed

Human



‘finished’ ~2015

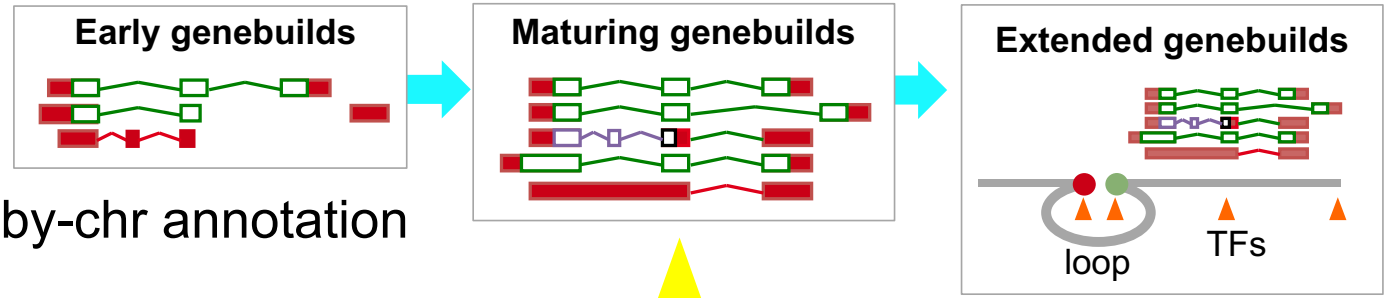
Mouse



‘finished’ ~2018

i.e. systematic chromosome-by-chromosome annotation

Gene annotation will continue for years



chr-by-chr annotation

e.g. linking enhancers to genes

Massive amounts of new data

Discovery of new features

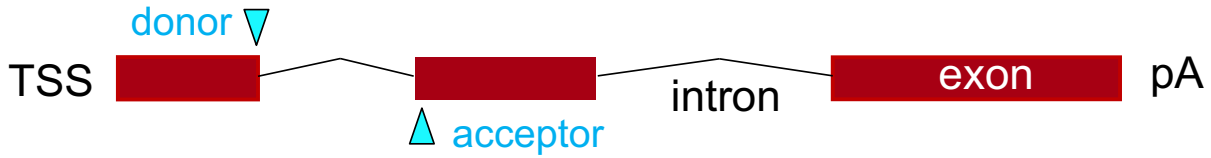
Extending incomplete models

Correct ‘functional’ annotation

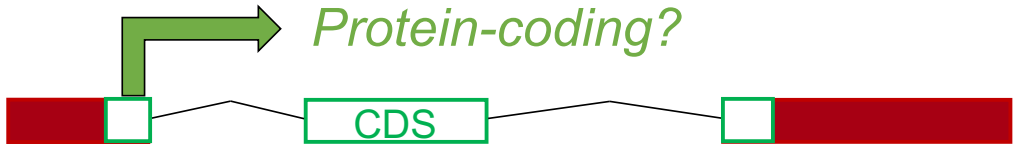
Manual annotation and analysis

The two principles of annotation

1. '**Structural**': making the exons and introns of the model



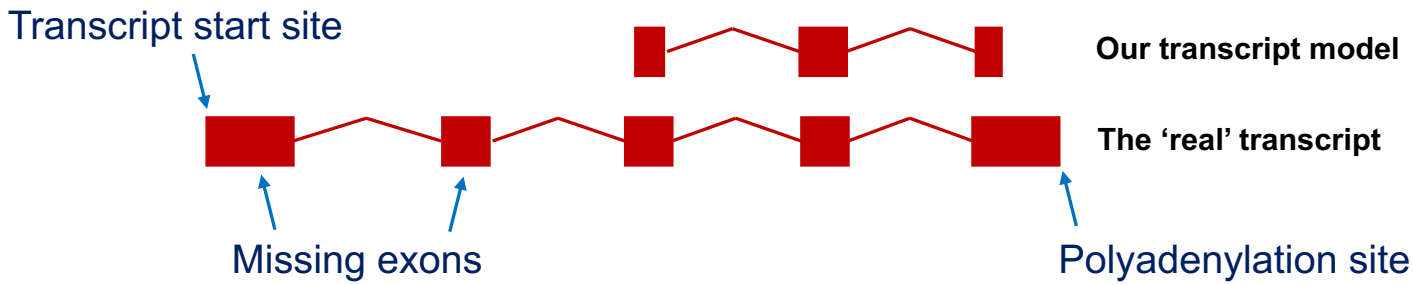
2. '**Functional**': assessing the biological nature of the model



Challenges in structural annotation

Transcript models are commonly ‘incomplete’, i.e. too short

… because the RNA evidence used to construct was not ‘full-length’

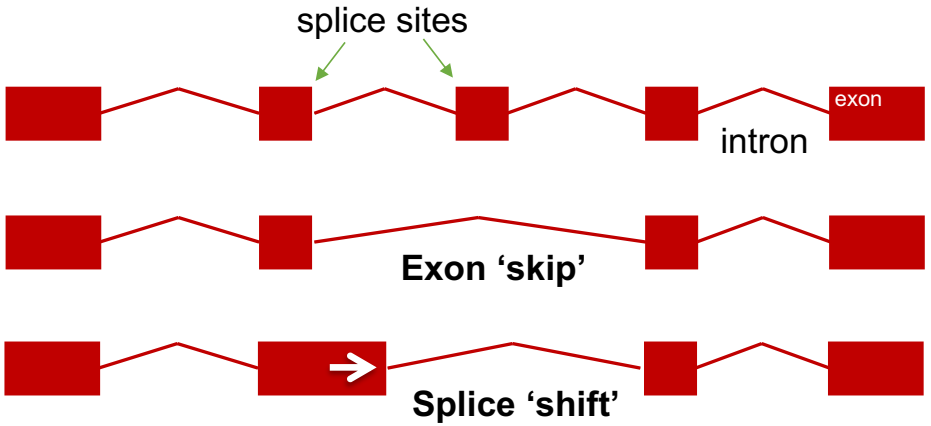


Incomplete models can lack correct biological features (e.g. CDS)

Challenges in structural annotation

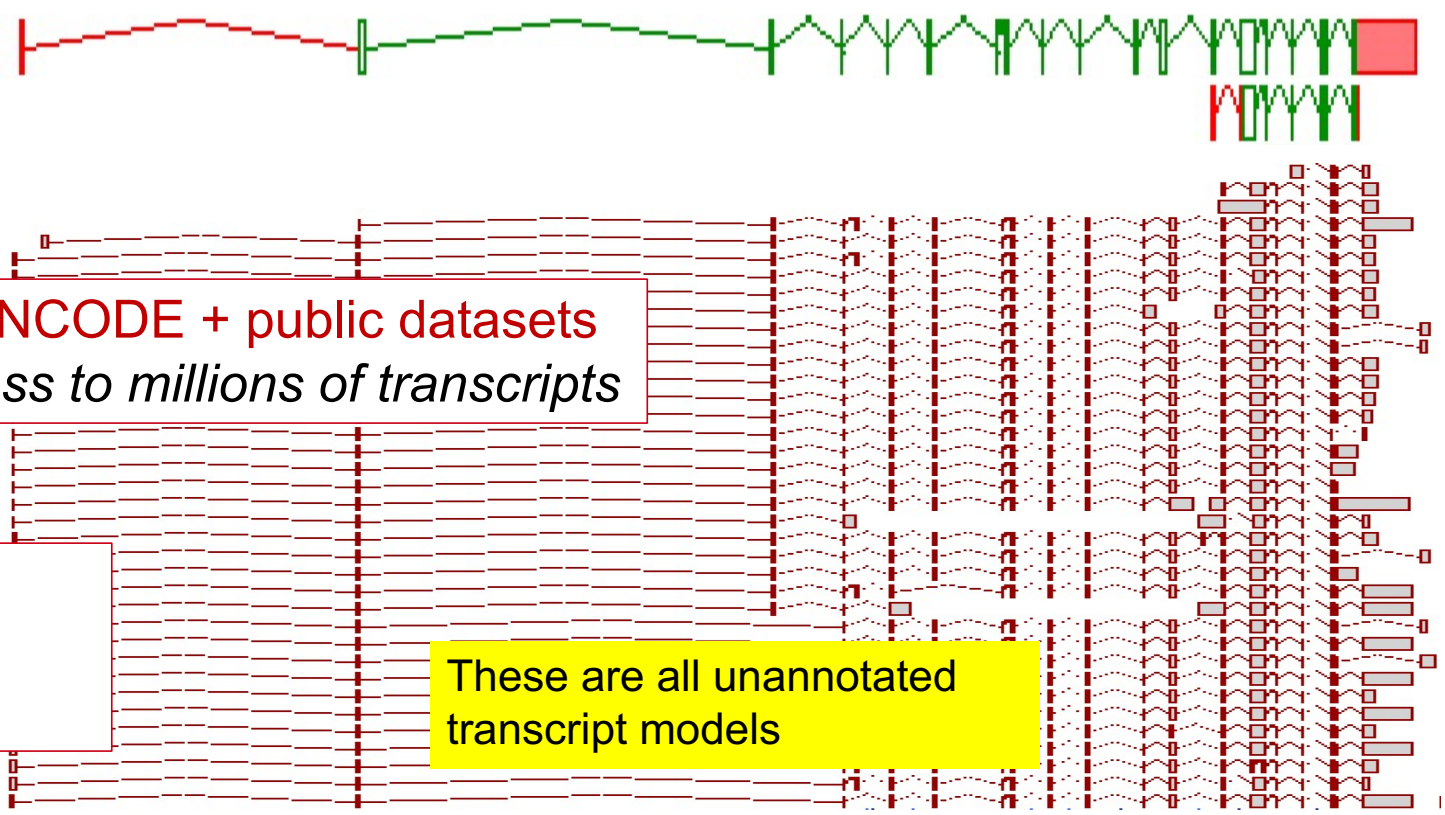
We can find novel transcripts within existing genes

i.e. due to **alternative splicing**



New models may contain additional biological features

Long-read sequencing data at *IL16*



PacBio data from GENCODE + public datasets
… *we now have access to millions of transcripts*

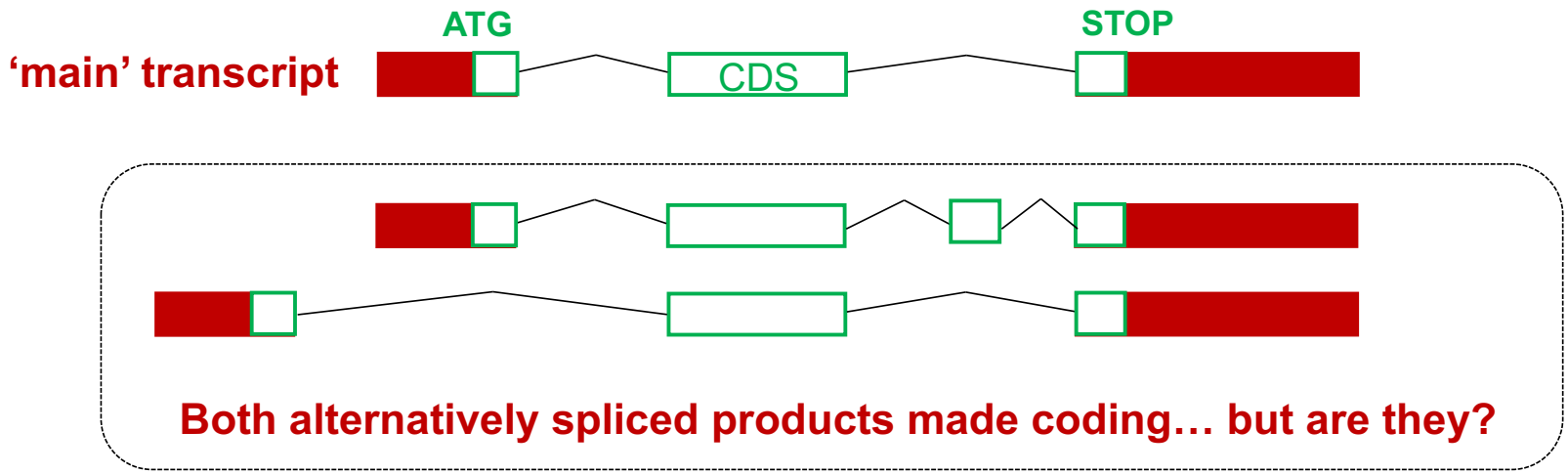
‘**TAGENE**’ pipeline for computational incorporation

These are all unannotated transcript models

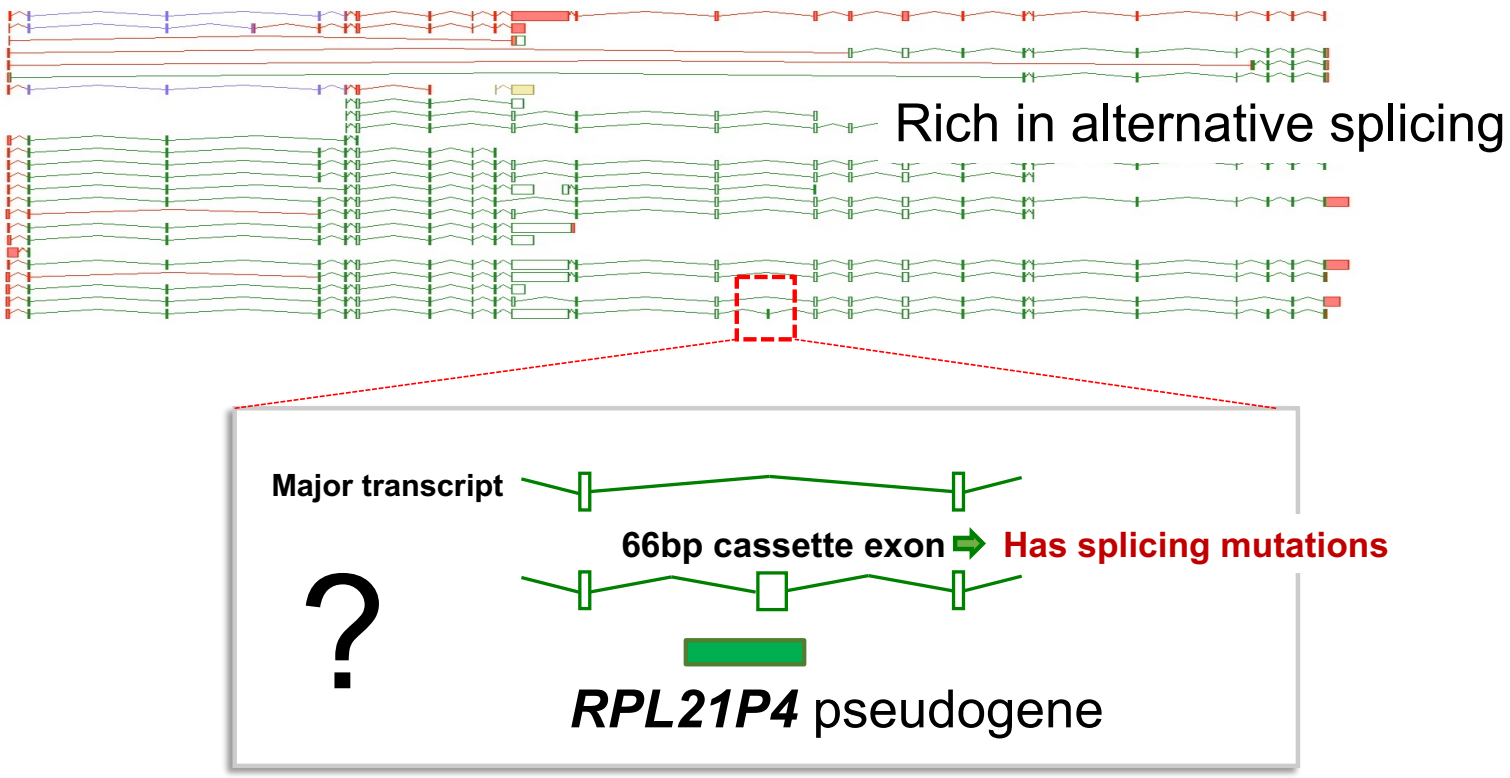
‘Functional annotation’

Major question: which transcripts encode proteins?

Traditionally: CDS annotation is predictive, based on first principles



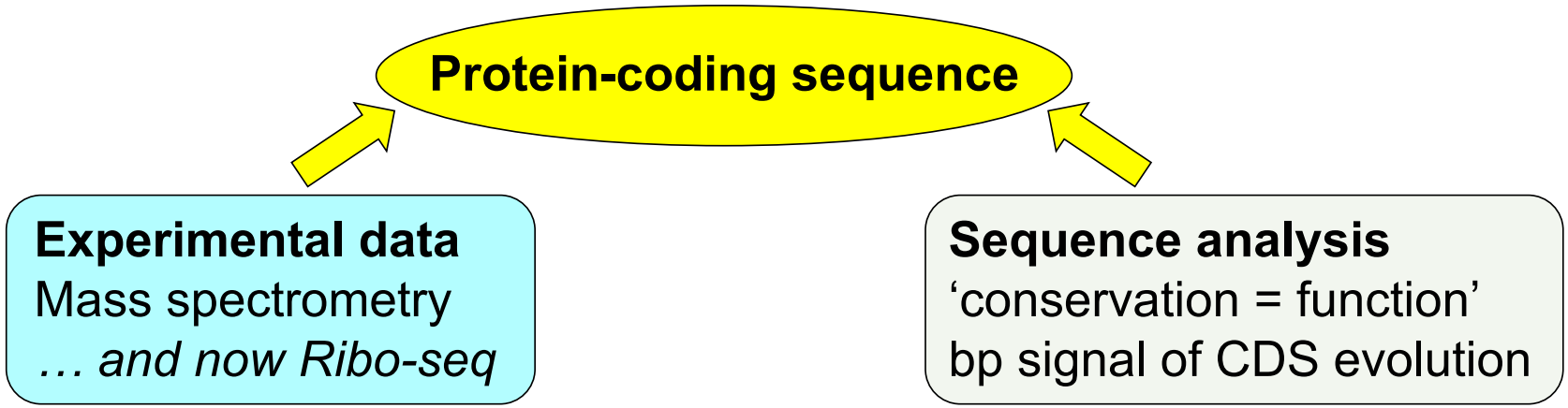
BRCA1 putative coding exon



Confirming protein-coding transcripts

Fundamental problem: sequencing protein remains very challenging

… protein-coding annotation is highly predictive



Both used for **discovery** as well as **validation**

Inverted formin 2 (INF2)

Existing model

PhyloCSF: identifies protein-coding evolution

new

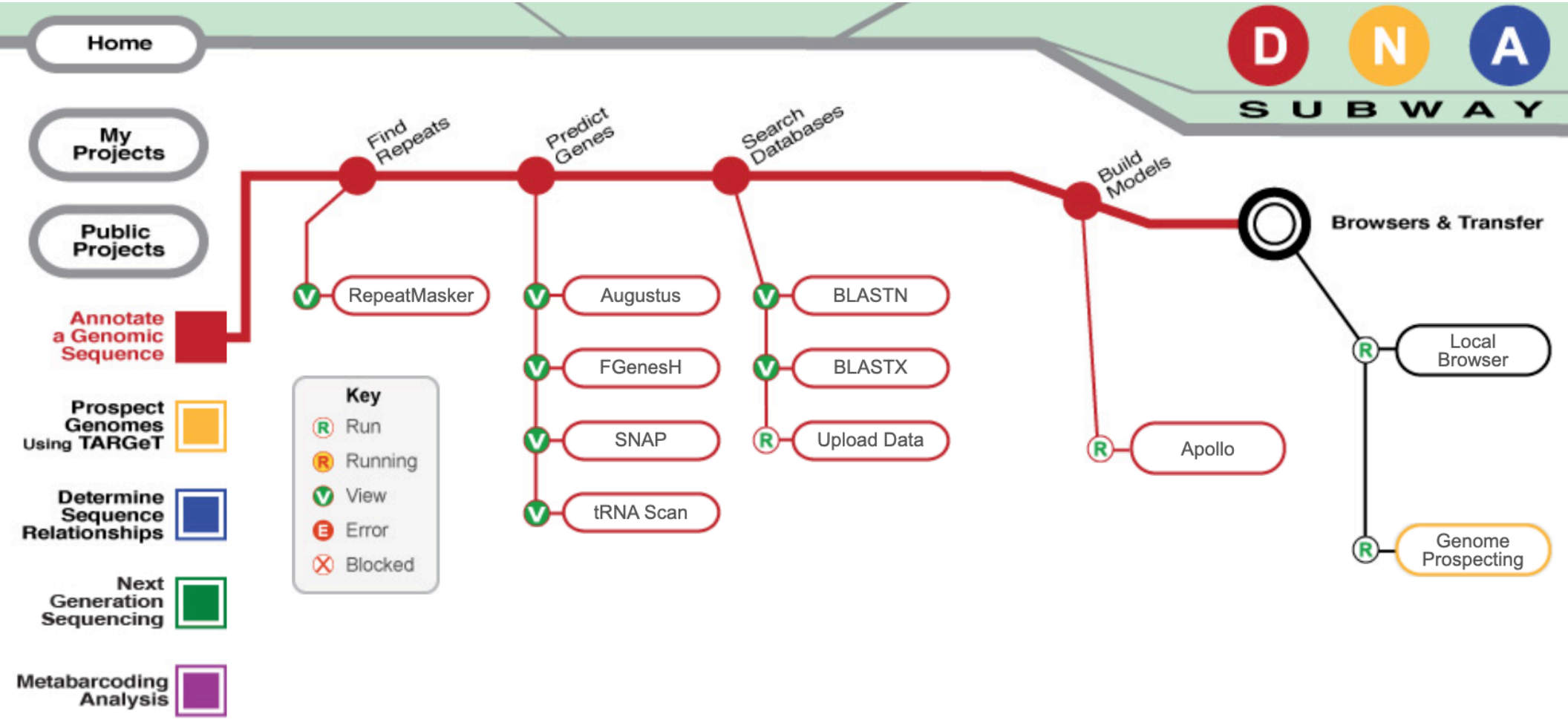
A gene linked to focal segmental glomerulosclerosis

Structural and **functional** annotation combine to identify a new biological feature of presumed importance

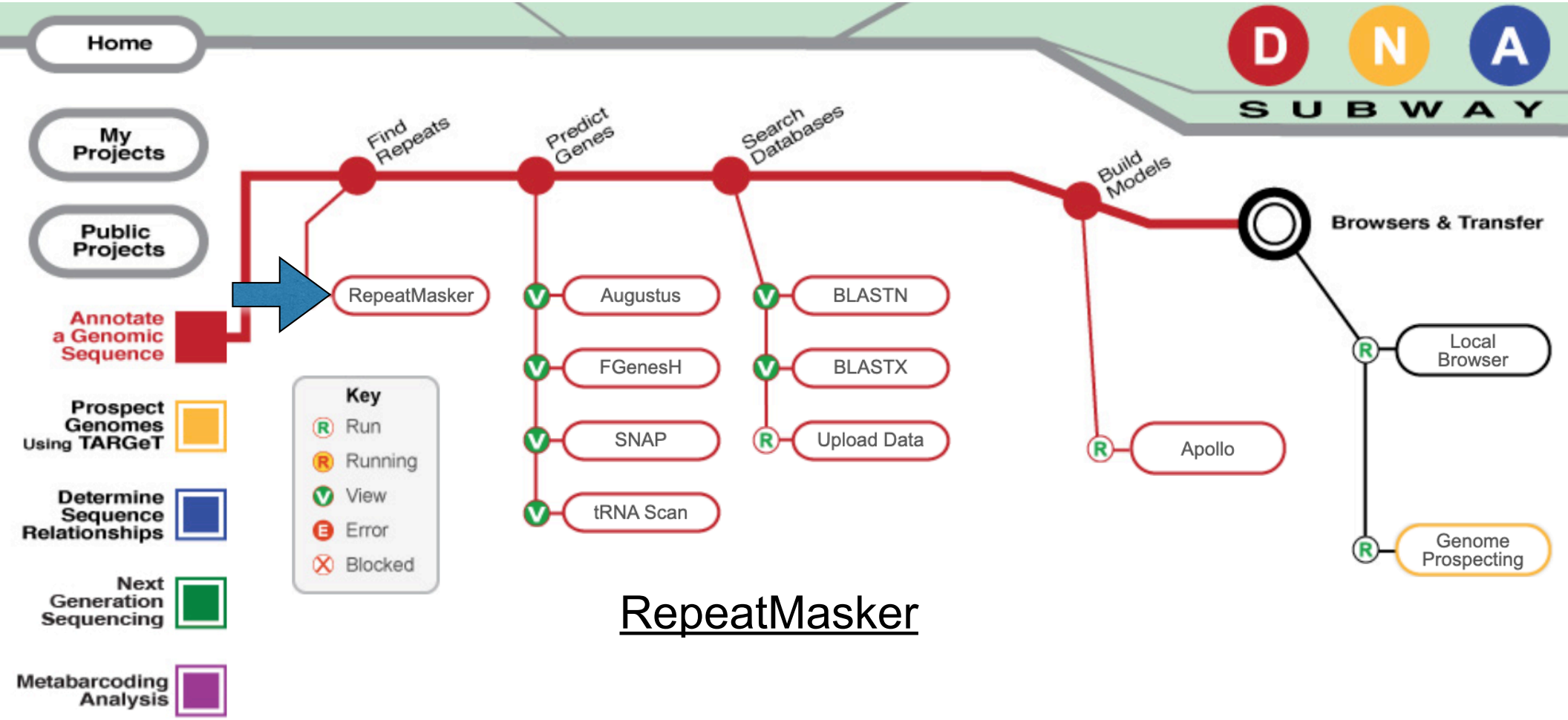
CDS
UTR / non-coding

Cyverse User Portal

DNA SUBWAY


Home
My Projects
Public Projects
Annotate a Genomic Sequence
Prospect Genomes Using TARGeT
Determine Sequence Relationships
Next Generation Sequencing
Metabarcoding Analysis
D N A
SUBWAY
Find Repeats
Predict Genes
Search Databases
Build Models
Browsers & Transfer
RepeatMasker
Augustus
FGenesH
SNAP
tRNA Scan
BLASTN
BLASTX
Upload Data
Apollo
Local Browser
Genome Prospecting
Key
Run
Running
View
Error
Blocked

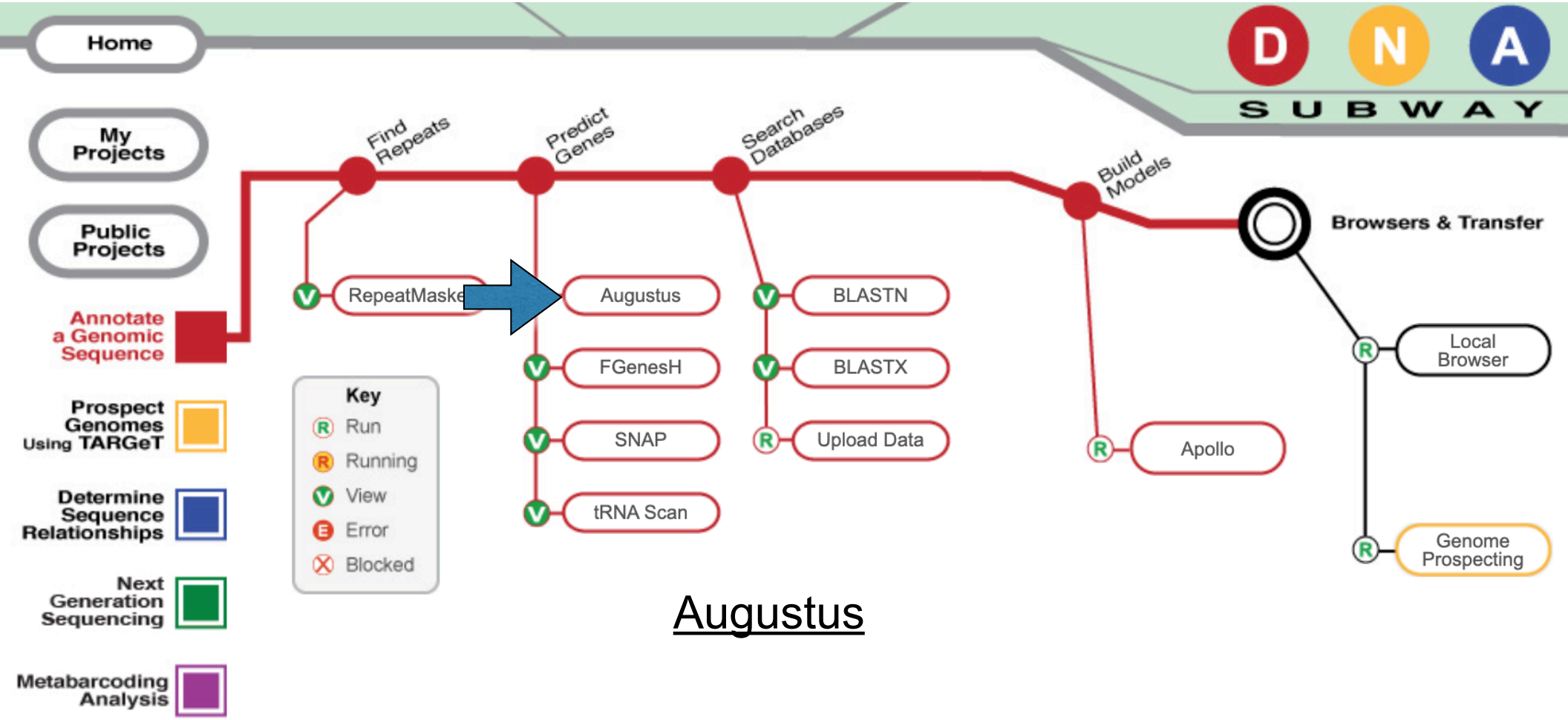
DNA SUBWAY



RepeatMasker

RepeatMasker is a program that screens DNA sequences for interspersed repeats and low complexity DNA sequences. The output of the program is a detailed annotation of the repeats that are present in the query sequence as well as a modified version of the query sequence in which all the annotated repeats have been masked (default: replaced by Ns). Currently over 56% of human genomic sequence is identified and masked by the program. Sequence comparisons in RepeatMasker are performed by one of several popular search engines including nhmmer, cross_match, ABBlast/WUBlast, RMBlast and Decypher. RepeatMasker makes use of curated libraries of repeats and currently supports Dfam (profile HMM library derived from Repbase sequences) and Repbase, a service of the Genetic Information Research Institute.

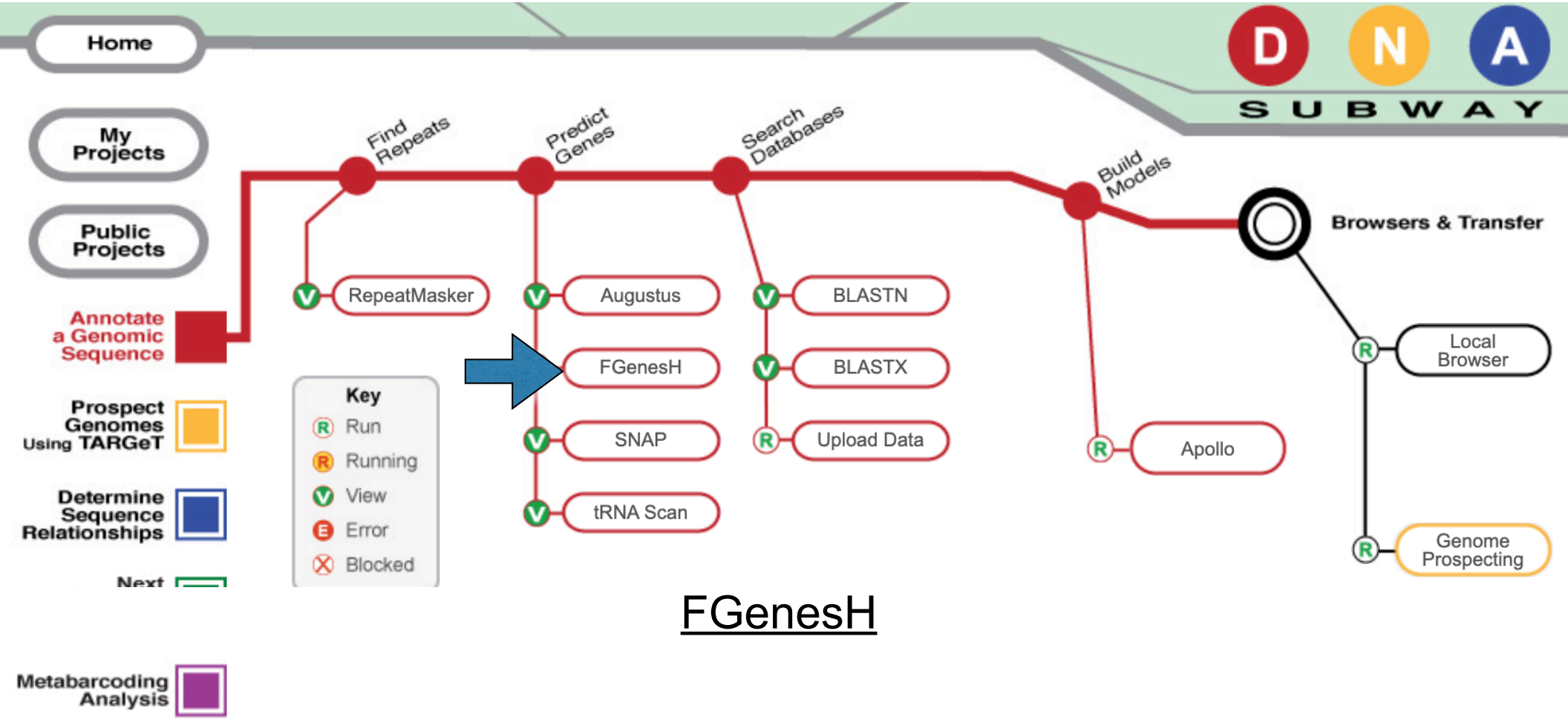
DNA SUBWAY



Augustus

AUGUSTUS is a program that predicts genes in eukaryotic genomic sequences. It can be run on this web server, on a new web server for larger input files or be downloaded and run locally. It is open source so you can compile it for your computing platform. You can now run AUGUSTUS on the German MediGRID. This enables you to submit larger sequence files and allows to use protein homology information in the prediction. The MediGRID requires an instant easy registration by email for first-time users.

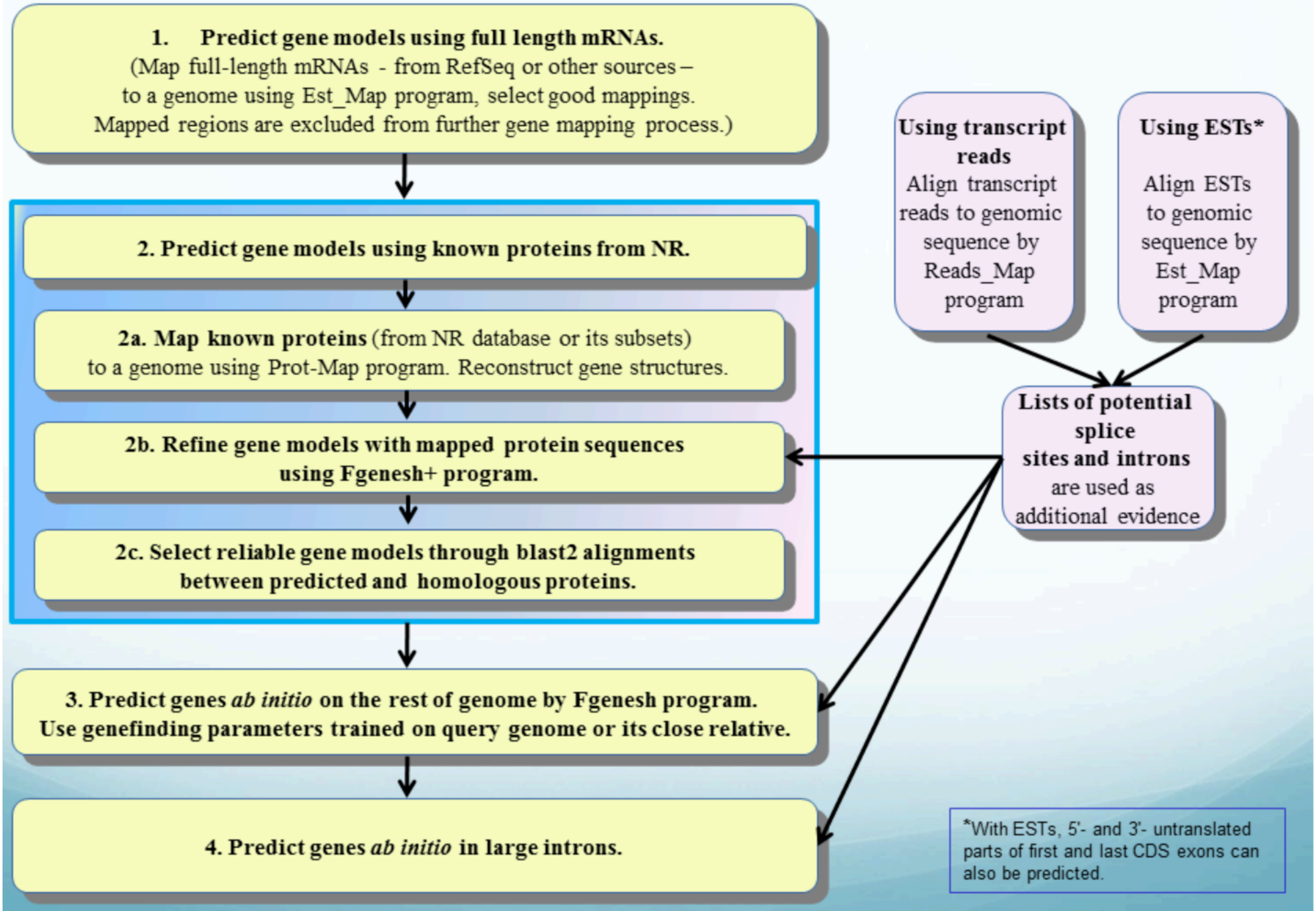
DNA SUBWAY



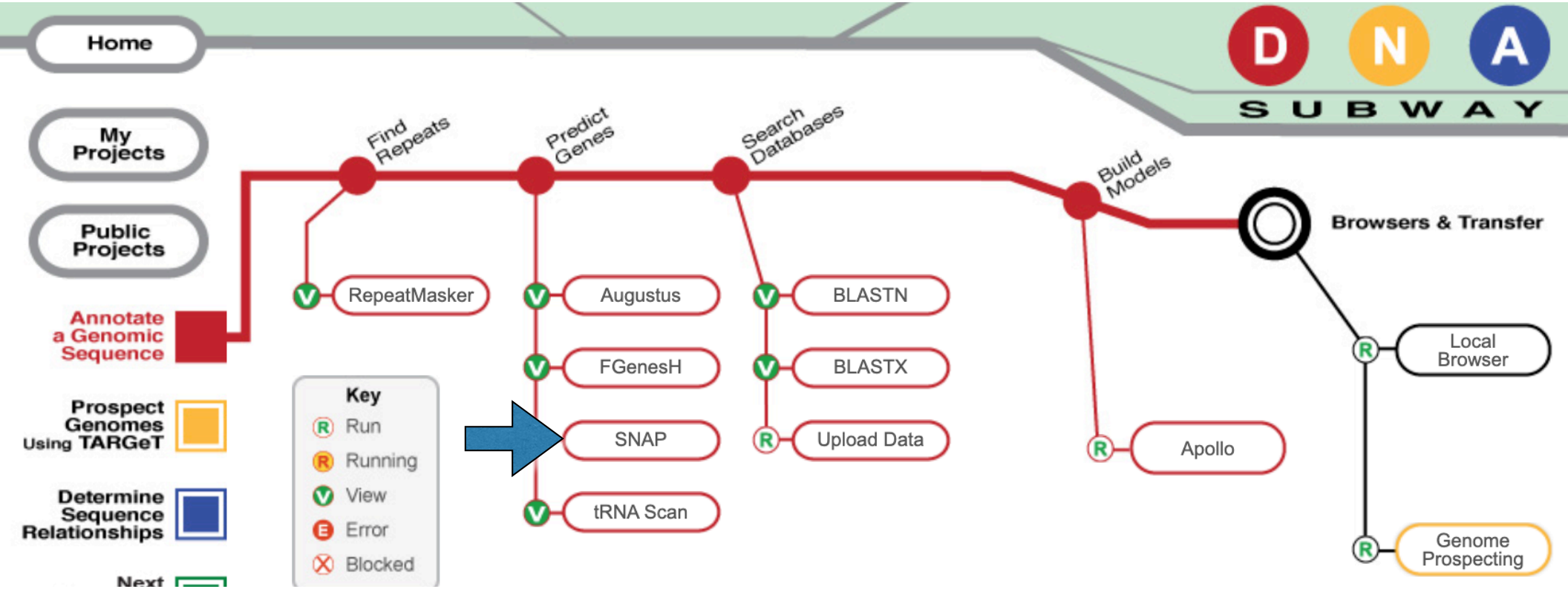
FGenesH

- Fgenesh – most accurate and fastest HMM-based gene prediction program;
- Fgenesh+ – gene prediction program that uses similar protein information;
- gene finding parameters for Fgenesh and Fgenesh+;
- est_map – program for mapping known mRNAs to genome (genome alignment with splice sites identification) and mapping a set of available ESTs to improve the gene prediction accura and add 5' and 3'– noncoding sequences;
- prot_map – mapping protein database to genomic sequences;

Fgenesh++ pipeline main steps



DNA SUBWAY



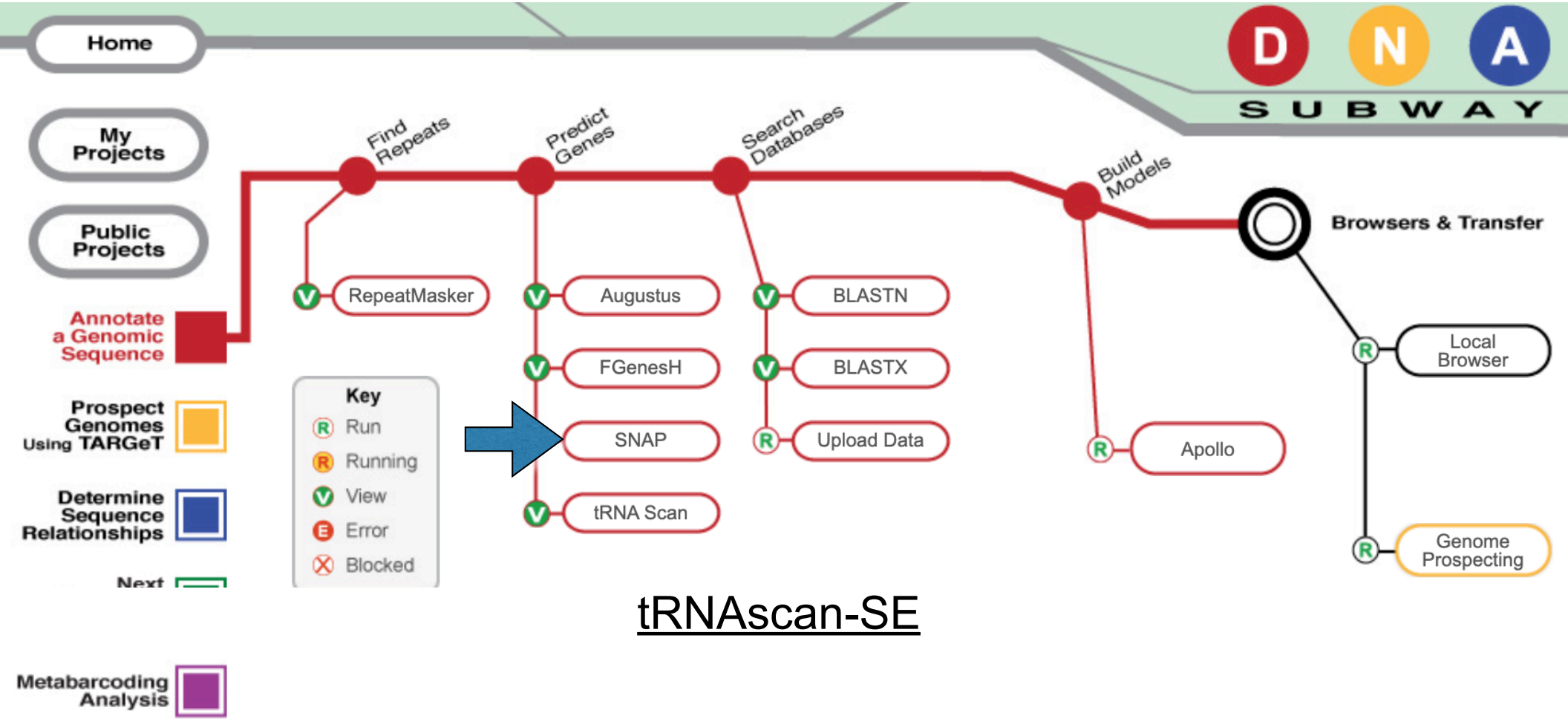
SNAP

The Korf Lab

ATAGCGAAT
TATCGCTTA

Making sense of sequences

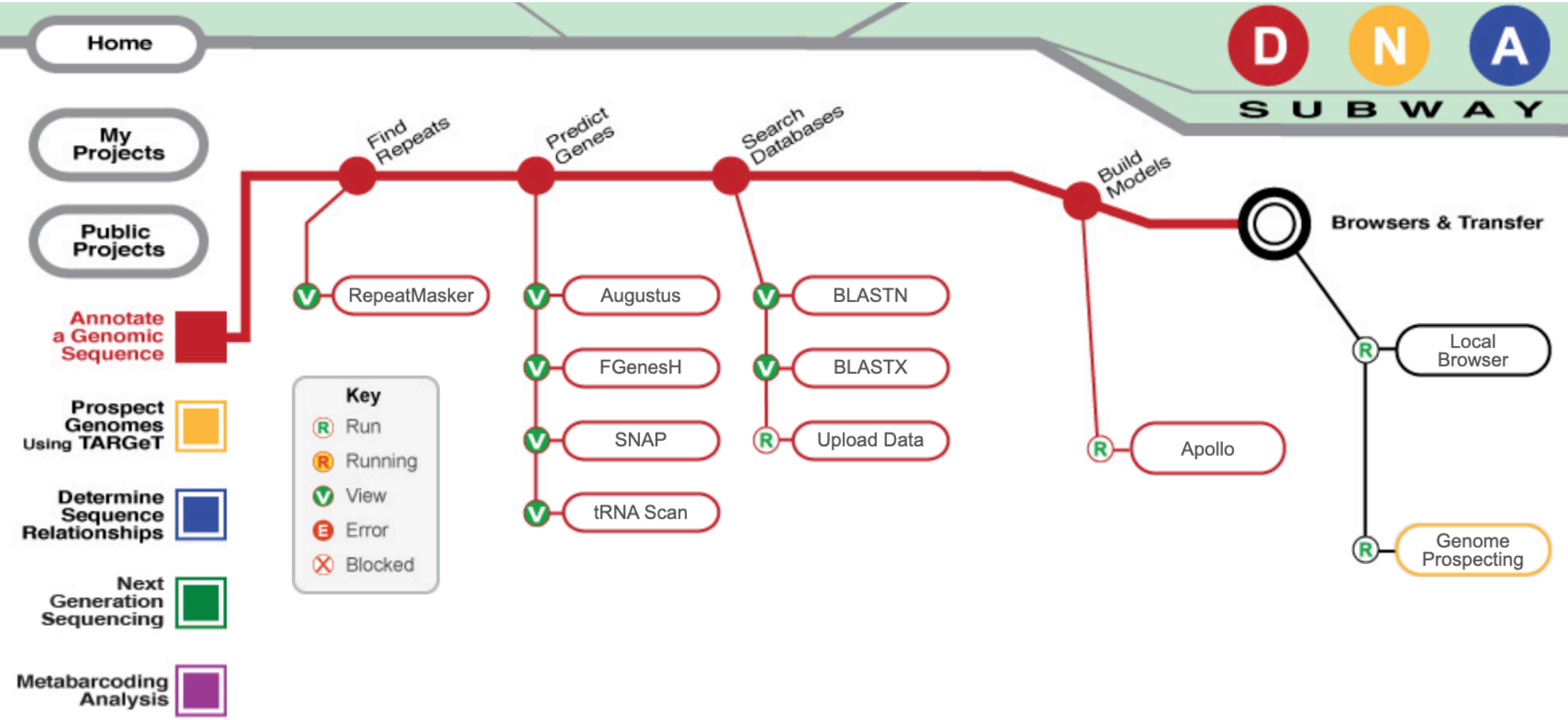
DNA SUBWAY



tRNAscan-SE

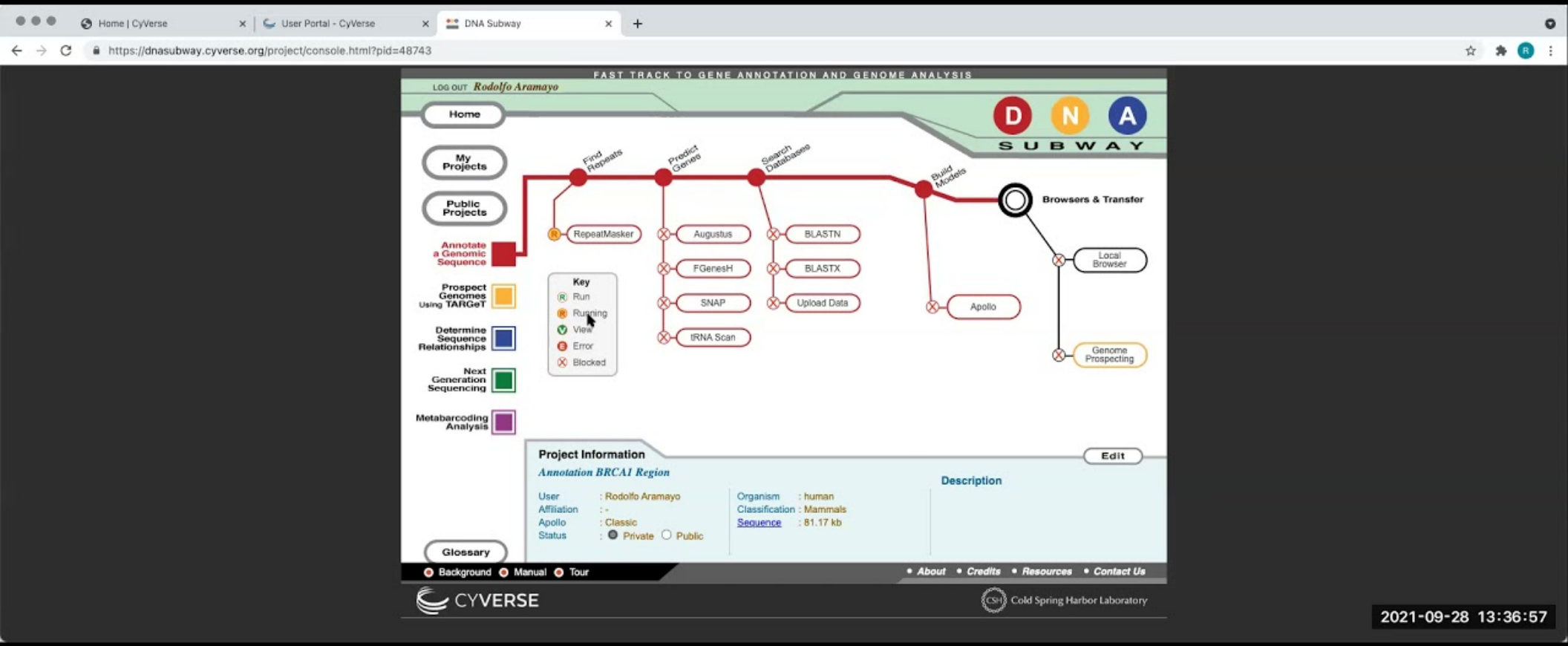
tRNAscan-SE detects ~99% of eukaryotic nuclear or prokaryotic tRNA genes, with a false positive rate of less than one per 15 gigabases, and with a search speed of about 30 kb/second. It was implemented for large-scale human genome sequence analysis, but is applicable to other DNAs as well. It applies our COVE software (see below) with a carefully built tRNA covariance model, while getting around COVE's speed limitations by using two tRNA finding programs from other research groups as fast first-pass scanners (Fichant and Burks', and an implementation of an algorithm from A. Pavesi's group). It runs on any UNIX system with Perl and a C compiler installed.

DNA SUBWAY

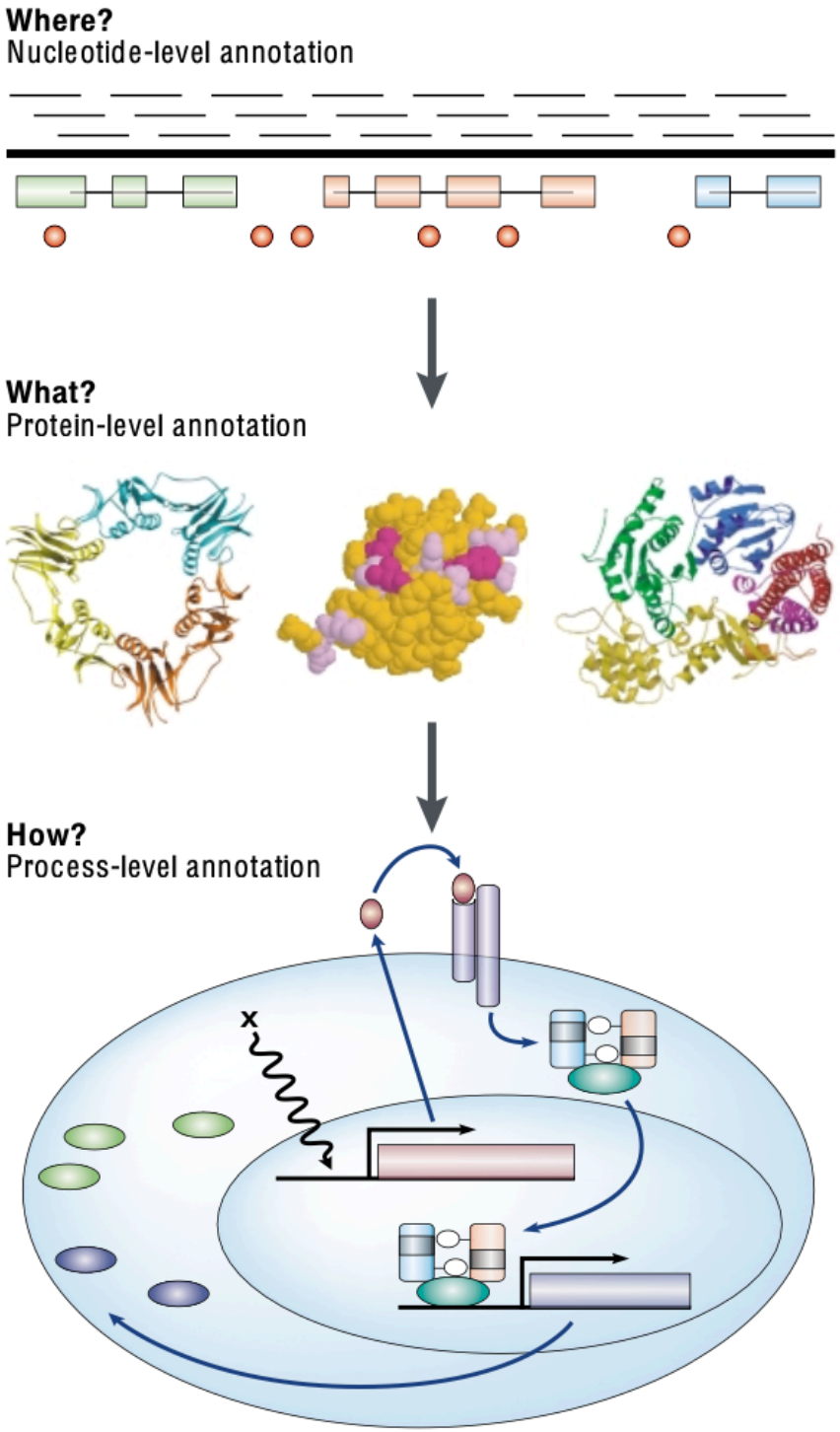

Home
My Projects
Public Projects
Annotate a Genomic Sequence
Prospect Genomes Using TARGeT
Determine Sequence Relationships
Next Generation Sequencing
Metabarcoding Analysis
D N A
SUBWAY
Find Repeats
Predict Genes
Search Databases
Build Models
Browsers & Transfer
RepeatMasker
Augustus
FGenesH
SNAP
tRNA Scan
BLASTN
BLASTX
Upload Data
Apollo
Local Browser
Genome Prospecting
Key
Run
Running
View
Error
Blocked

DNA SUBWAY

BRACA1 DNA Region



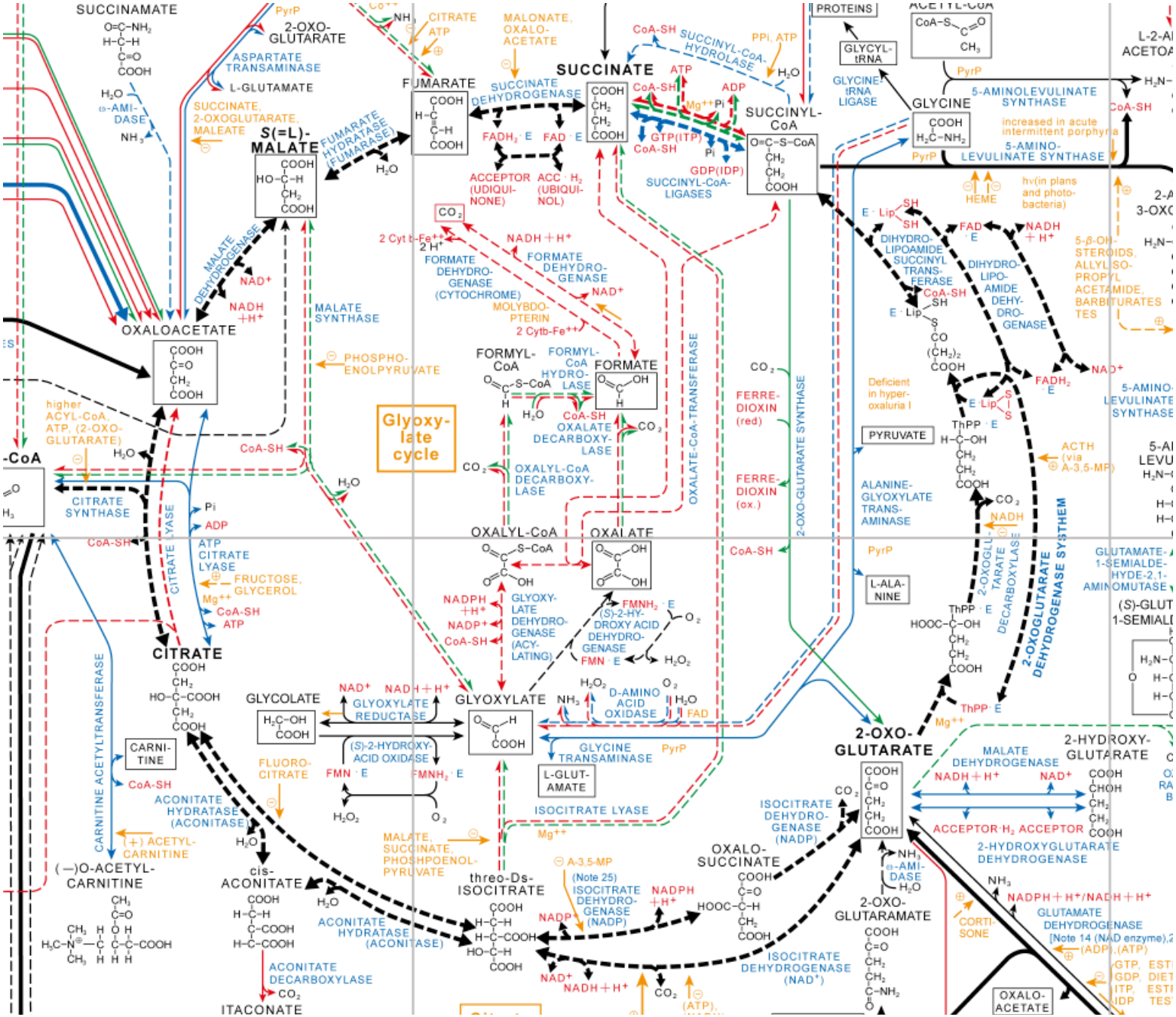
Defining Pathways: Process-Level Annotation



https://pubmed.ncbi.nlm.nih.gov/11433356/

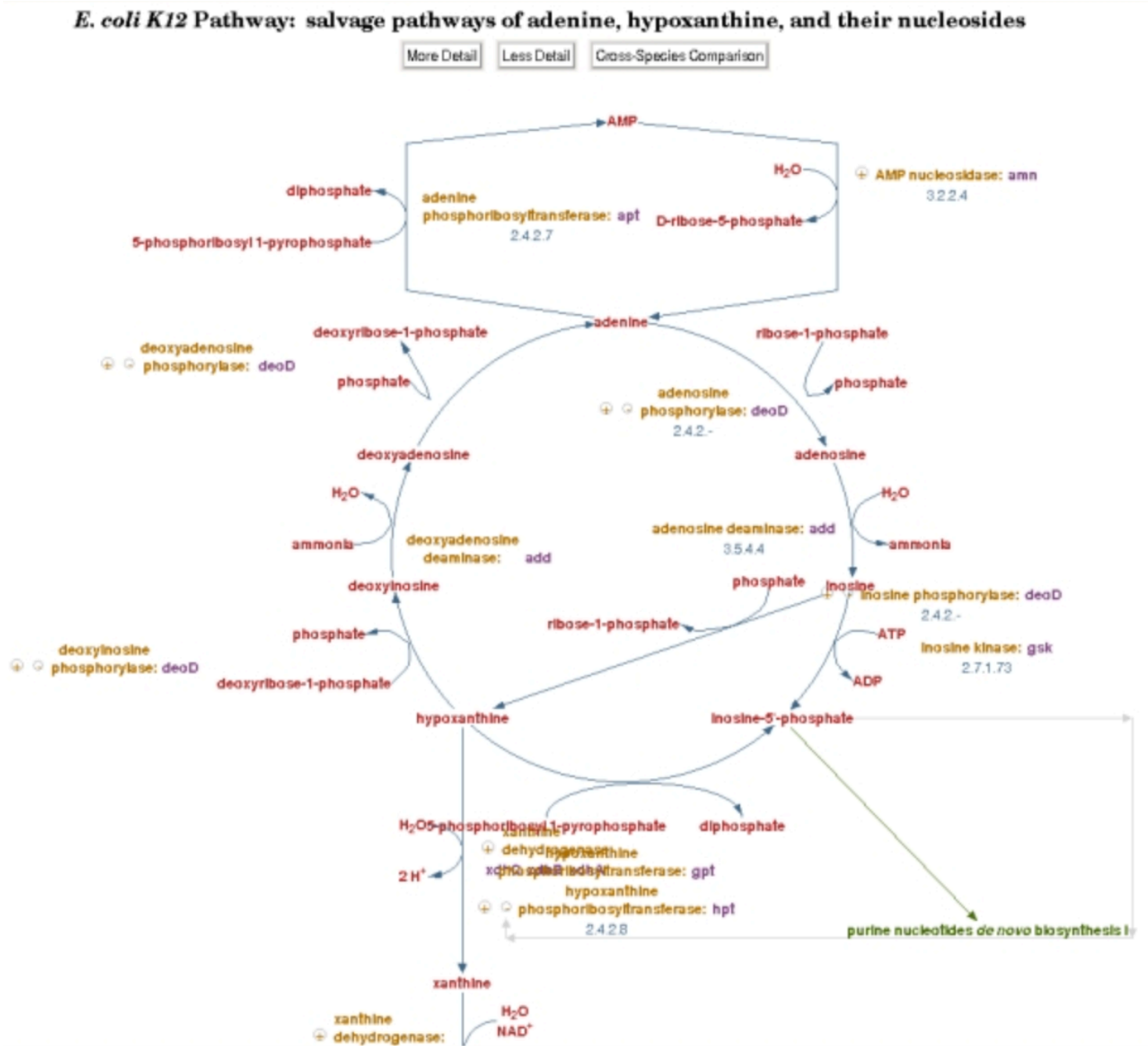
Defining Pathways: Process-Level Annotation

Metabolic Pathways



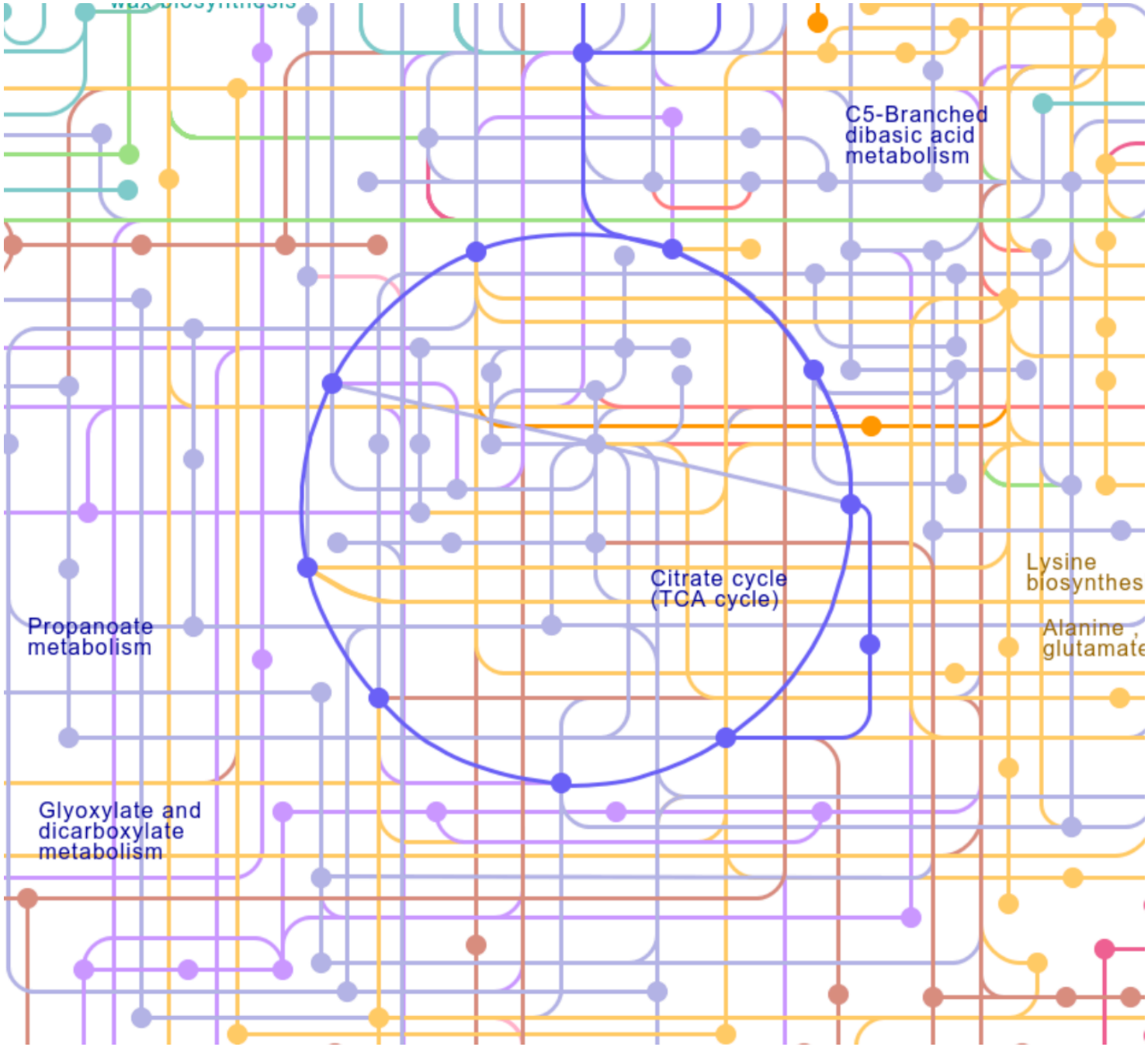
Defining Pathways: Process-Level Annotation

Pathways Tools



Defining Pathways: Process-Level Annotation

The KEGG Pathway Database



Defining Pathways: Process-Level Annotation

The Reactome Database