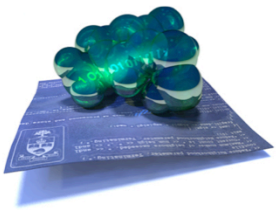
A
Bioinformatics
Course

Genome Annotation



Boris Steipe

department of biochemistry – department of molecular genetics
University of Toronto

Genome Annotation

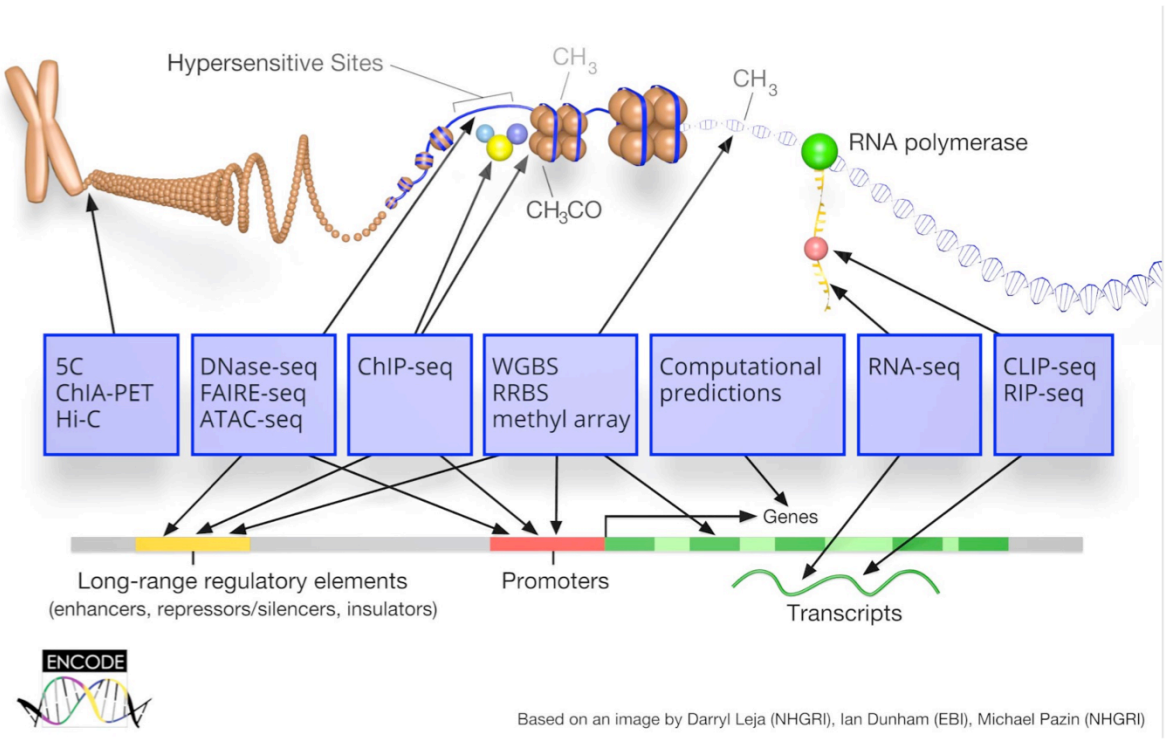
Given a nucleotide sequence that represents a genome, annotate all subsequences with their role (function).

- What are these roles?
- What algorithms can make the annotation?
- What workflow to use, what automated pipeline?
- What other annotations are of interest? (e.g. comparison to other genomes, species)
- How to interpret the results?
- How to disseminate the results?

Also: address meta-questions like how to do this according to the state-of-the-art, in time, on budget ...

The first order of genome annotation is to identify what elements a genome contains in the first place.

ENCODE



https://www.encodeproject.org/

The Encode (Encyclopedia of DNA Elements) project is a large-scale research consortium that aims to annotate all functional aspects of model organism genomes genome through a combination of high-throughput experimentation and bioinformatics. Data is currently avaiable for human, mouse, worm and fly.

ENCODE

ENCyclopedia Of Dna Elements

The aim of the ENCODE project is to identify all functional elements in the human genome sequence through the generation of a diverse collection of high-throughput datasets and mapping these datasets onto the human genome sequence.

Here we describe the production and initial analysis of 1,640 data sets designed to annotate functional elements in the entire human genome. We integrate results from diverse experiments within cell types, related experiments involving 147 different cell types, and all ENCODE data with other resources, such as candidate regions from genome-wide association studies (GWAS) and evolutionarily constrained regions.

The vast majority (80.4%) of the human genome participates in at least one biochemical RNA- and/or chromatin-associated event in at least one cell type. Much of the genome lies close to a regulatory event: 95% of the genome lies within 8 kilobases (kb) of a DNA-Protein interaction (as assayed by bound ChIP-seq motifs or DNase I footprints), and 99% is within 1.7 kb of at least one of the biochemical events measured by ENCODE.

Primate-specific elements as well as elements without detectable mammalian constraint show, in aggregate, evidence of negative selection; thus, some of them are expected to be functional.

Classifying the genome into seven chromatin states indicates an initial set of 399,124 regions with enhancer-like features and 70,292 regions with promoter-like features, as well as hundreds of thousands of quiescent regions.

It is possible to correlate quantitatively RNA sequence production and processing with both chromatin marks and transcription factor binding at promoters, indicating that promoter functionality can explain most of the variation in RNA expression.

Many non-coding variants in individual genome sequences lie in ENCODE-annotated functional regions; this number is at least as large as those that lie in protein-coding genes.

Single nucleotide polymorphisms (SNPs) associated with disease by GWAS are enriched within non-coding functional elements, with a majority residing in or near ENCODE-defined regions that are outside of protein-coding genes.

ENCODE consortium (2012) An integrated encyclopedia of DNA elements in the human genome. *Nature* 489:57-74

ENCODE

Table 1. Experimental assays used by the ENCODE Consortium.

Gene/Transcript Analysis

Region/Feature	Method	Group
Gene annotation	GENCODE	Wellcome Trust
PolyA+ coding regions	RNA-seq; tiling DNA microarrays; PET	CSHL; Stanford/Yale//Harvard; Caltech
Total RNA coding regions	RNA-seq; tiling DNA microarrays; PET	CSHL
Coding regions in subcellular RNA fractions (e.g. nuclear, cytoplasmic)	PET	CSHL
Small RNAs	short RNA-seq	CSHL
Transcription initiation (5′-end) and termination (3-end′) sites	CAGE; diTAGs	RIKEN, GIS
Full-length RNAs	RACE	University of Geneva; University of Lausanne
Protein-bound RNA coding regions	RIP; CLIP	SUNY-Albany; CSHL

Transcription Factors/Chromatin

Elements/Regions	Method(s)	Group(s)
Transcription Factor Binding Sites (TFBS)	ChIP-seq	Stanford/Yale/UC-Davis/Harvard; HudsonAlpha/Caltech; Duke/UT-Austin; UW; U. Chicago/Stanford
Chromatin structure (accessibility, etc.)	DNaseI hypersensitivity; FAIRE	UW; Duke; UNC
Chromatin modifications (H3K27ac, H3K27me3, H3K36me3, etc.)	ChIP-seq	Broad; UW
DNaseI footprints	Digital genomic footprinting	UW

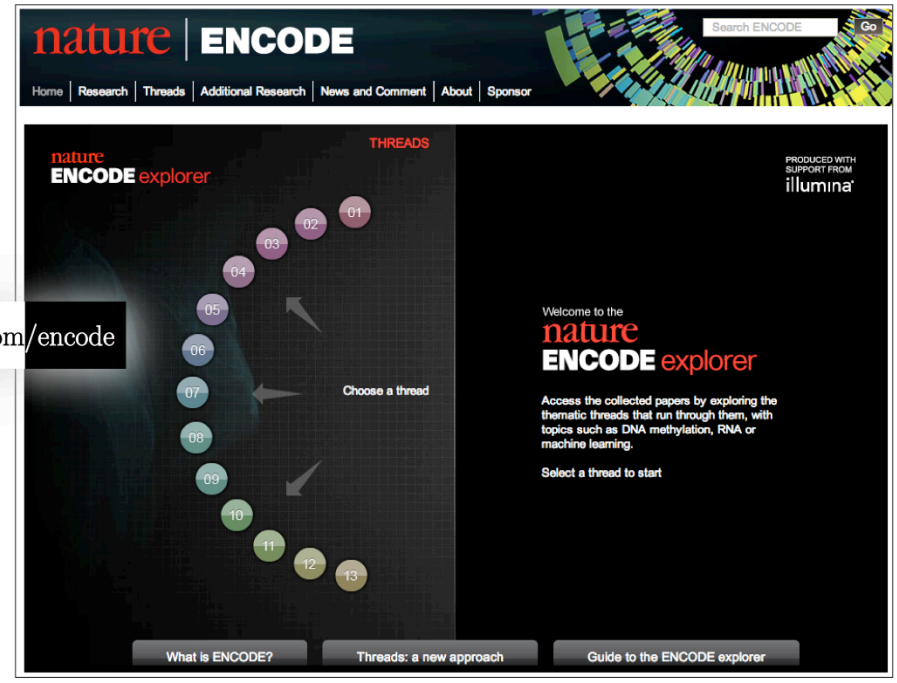
Other Elements/Features

Feature	Method(s)	Group(s)
DNA methylation	RRBS; Illumina Methyl27; Methyl-seq	HudsonAlpha
Chromatin interactions	5C; CHIA-PET	UMass; UW; GIS
Genotyping	Illumina 1M Duo	HudsonAlpha

doi:10.1371/journal.pbio.1001046.t001

Encode Consortium (2011) A User's Guide to the Encyclopedia of DNA Elements (Encode). *PLoS Biology* **9(4)**:e1001046

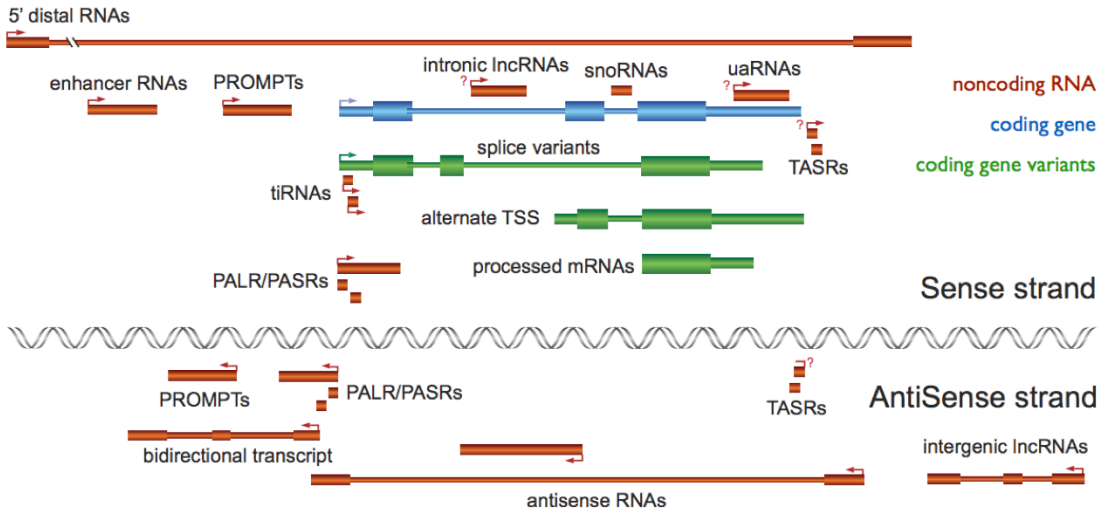
Nature - special online issue (2012)



http://www.nature.com/encode

Read more about the projects' goals, procedures and results in this special issue of *nature*.

Genome Contents



Mike Clark (2013) http://www.genome.gov/Multimedia/Slides/GWASWebinar2013/02_Clark.pdf

Transcription is complex and about 75% of the genome may have functional significance when transcribed. Moreover, need to account for non-transcribed elements! (Promoters, enhancers, silencers, TF-binding sites, origins, telomeres, centromeres ...); also need to account for repeats.

Name	Size	Location	Number in humans	Functions	Illustrative examples	Refs
Short ncRNAs						
miRNAs	19–24 bp	Encoded at widespread locations	>1,424	Targeting of mRNAs and many others	miR-15/16, miR-124a, miR-34b/c, miR-200	3–8
piRNAs	26–31bp	Clusters, intragenic	23,439	Transposon repression, DNA methylation	piRNAs targeting *RASGRF1* and LINE1 and IAP elements	13–19
tiRNAs	17–18bp	Downstream of TSSs	>5,000	Regulation of transcription?	Associated with the *CAP1* gene	37
Mid-size ncRNAs						
snoRNAs	60–300 bp	Intronic	>300	rRNA modifications	U50, SNORD	20–22
PASRs	22–200 bp	5′ regions of protein-coding genes	>10,000	Unknown	Half of protein-coding genes	10
TSSa-RNAs	20–90 bp	−250 and +50 bp of TSSs	>10,000	Maintenance of transcription?	Associated with *RNF12* and *CCDC52* genes	35
PROMPTs	<200 bp	−205 bp and −5 kb of TSSs	Unknown	Activation of transcription?	Associated with *EXT1* and *RBM39* genes	36
Long ncRNAs						
lincRNAs	>200 bp	Widespread loci	>1,000	Examples include scaffold DNA–chromatin complexes	*HOTAIR*, *HOTTIP*, *lincRNA-p21*	2,28–30
T-UCRs	>200 bp	Widespread loci	>350	Regulation of miRNA and mRNA levels?	uc.283+, uc.338, uc160+	31–34
Other lncRNAs	>200 bp	Widespread loci	>3,000	Examples include X-chromosome inactivation, telomere regulation, imprinting	*XIST*, *TSIX*, TERRAs, *p15AS*, *H19*, *HYMAI*	2,23–25

*There is not necessarily a clear delineation between classes of non-coding RNA (ncRNA); for example, X-inactivation specific transcript (*XIST*) and its antisense transcript *TSIX* could be considered as large intergenic non-coding RNAs (lincRNAs). In the 'Location' column, '−' represents the number of base pairs upstream of the transcription start site (TSS) and '+' represents the number of base pairs downstream of the TSS. *CAP1*, CAP, adenylate cyclase-associated protein 1; *CCDC52*, coiled-coil domain containing 52 (also known as *SPICE1*); *EXT1*, exostosin 1; *HOTAIR*, homeobox (HOX) transcript antisense RNA; *HOTTIP*, HOXA distal transcript antisense RNA; *HYMAI*, hydatidiform mole associated and imprinted; IAP, intracisternal A-particle; lncRNA, long non-coding RNA; miRNAs, microRNAs; piRNAs, PIWI-interacting RNAs; PASRs, promoter-associated small RNAs; PROMPTs, promoter upstream transcripts; *RASGRF1*, RAS-protein-specific guanine nucleotide-releasing factor 1; *RBM39*, RNA-binding motif protein 39; *RNF12*, ring finger protein 12 (also known as *RLIM*); snoRNAs, small nucleolar RNAs; TERRAs, telomeric repeat containing RNAs; tiRNAs, transcription initiation RNAs; TSSa-RNAs, TSS-associated RNAs; T-UCRs, transcribed ultraconserved regions.

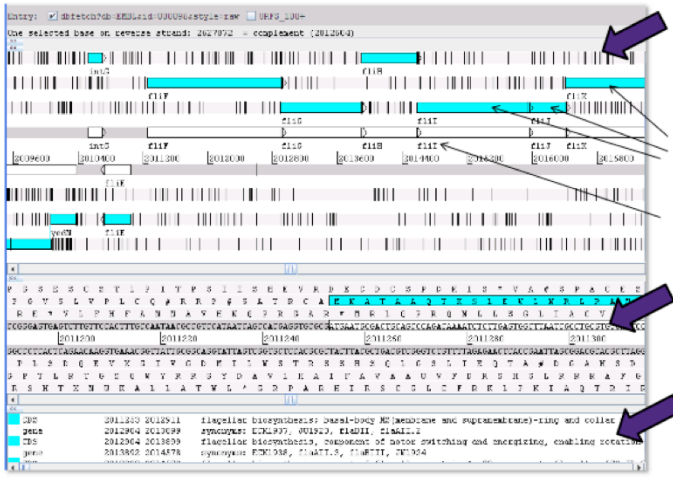
Esteller M. (2011) Non-coding RNAs in human disease. *Nature Reviews Genetics* **12**:861-874

An overview of the recognized functional classes of RNA. This list is likely to be in need of significant update every three or four years.

Genome annotation – Gene finding

Artemis

Java based genome visualization and annotation tool - particularly suitable for analysing microbial genomes



6 frames of the DNA sequence – vertical lines represent stop codons

ORFs (arrows show direction)

Gene annotations

DNA and 6 frame AA translations

Feature list

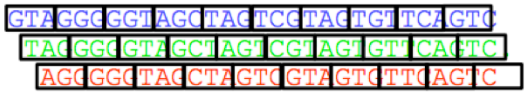
Slide courtesy of John Parkinson

Gene finding is a major task for genome annotation. There are four principal methods to identify genes: analysis by signal, analysis by contents, analysis by homology, and interpretation of the transcriptome.

De novo Gene Finding – Content sensors: exploiting codon bias

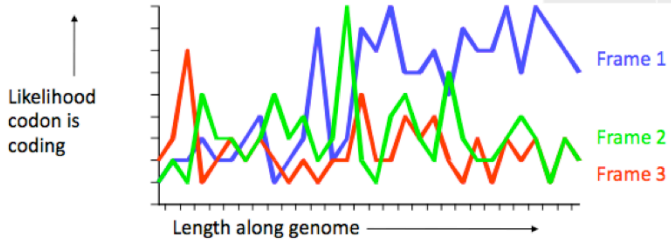
Genome sequence

GTAGGGGGTAGCGTAGTCGTAGTGTCAGTC



Codon bias table for Gly

Codon	E. coli	Human
GGG	2	25
GGA	0	25
GGU	59	16
GGC	39	34



Can identify regions where there is a high frequency of 'preferred' codons
Applied methods typically look at dicodon statistics as these provide better predictions than single codons

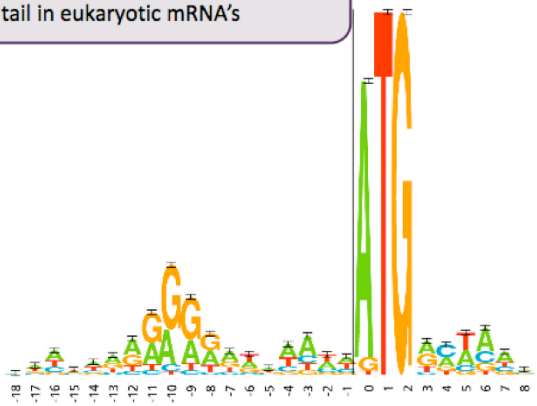
Slide courtesy of John Parkinson

Analysis by contents looks for trinucleotide patterns that are characterisitc of transcribed and tranlated sequence.

Gene Finding – Signal Sensors

An alternate method to composition based methods for gene finding are signal sensors which look for signal sequences e.g. AATAAA is a signal for the addition of the poly(A) tail in eukaryotic mRNA's

ORPHEUS
Uses codon statistics to determine likely ORFs, but also includes homology searches and probabilistic models of ribosome-binding sites

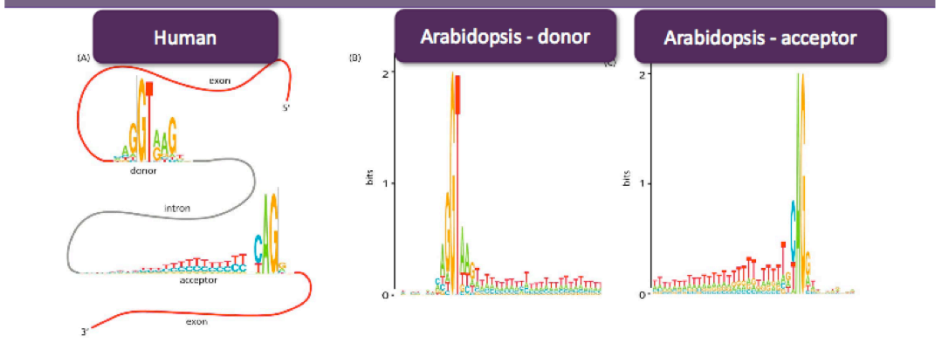


This sequence logo of RBS was created from an alignment of 149 sequences (aligned on their start codon).

Slide courtesy of John Parkinson

Analysis by signal looks for translation start sites, ploy(A) attachment sites, exon-intron boundary signals etc.

Gene Finding in eukaryotes – finding splice sites



Again eukaryotic gene finders can exploit signal sensors

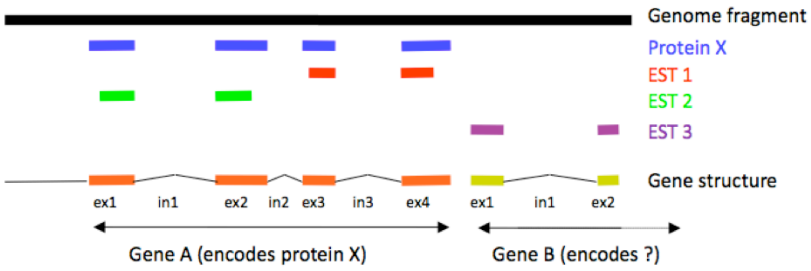
Most introns start with a GT and end with AG – the extended splice site recognition motif includes less well defined sequence up and downstream and can vary between species

These complex patterns can be represented by models that use the relative likelihood of each base at each position to identify genome features

Slide courtesy of John Parkinson

Gene Finding – Exploiting sequence similarity

To identify exons and reveal the structure of a gene within a genome, sequence similarity searches of genomic DNA against protein / EST databases



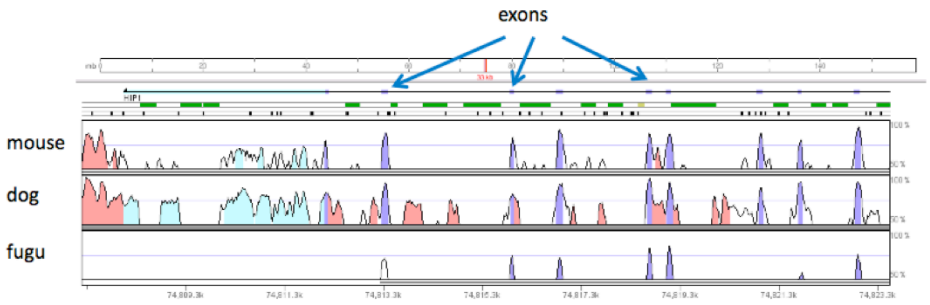
Slide courtesy of John Parkinson

Analysis by homology is the most accurate method – if it is available. Not all genes in an organism have homologues in other species.

Finding exons with phylogenetic footprinting

Vista genome browser (http://pipeline.lbl.gov/vistabrowser/)

Below is a comparison of a section of human the human genome with equivalent sections from mouse, dog and fugu genomes showing % identity



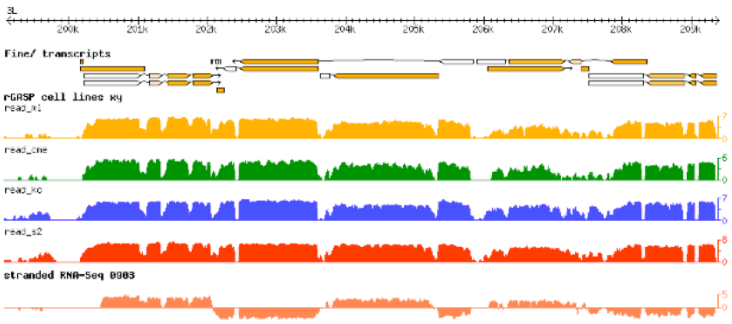
Using just mouse and dog genomes difficult to identify exons in the human genome (too much similarity in non-exonic regions)
Comparison with the *fugu* genome led to the identification of 1,000 previously unidentified human genes

Slide courtesy of John Parkinson

The alignment of syntenic regions can confirm the existence of genes, and their intron-exon structure.

Gene Finding – Exploiting RNASeq data

With the availability of cheap high throughput sequencing, many genome sequencing projects now perform RNASeq in parallel to help identify genes



Not all genes will be expressed in your RNASeq sample

Slide courtesy of John Parkinson

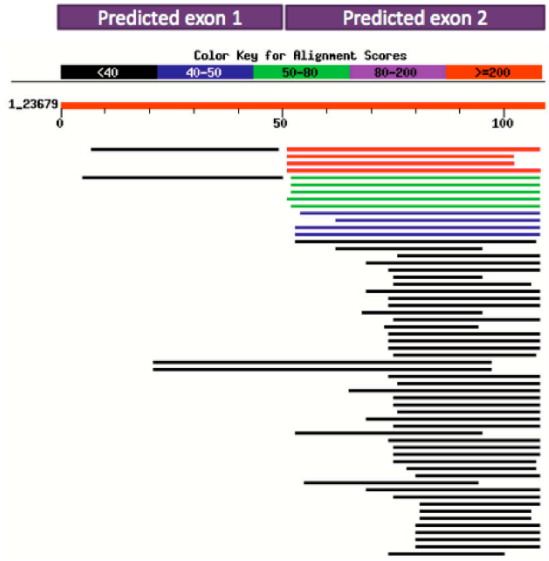
Transcriptome analysis obviously can contribute greatly to identify genes. However, not all genes are expressed at a level sufficient for confidently observing a gene's transcript in the transcriptome. mRNA levels have been reported to have a range of 5-7 orders of magnitude.

Gene Finding in Eukaryotes – Errors

Frameshift errors can be identified through homology searches

The first exon of ALDH10 gene is incorrectly predicted by one 1 base – resulting in a wrong reading frame.

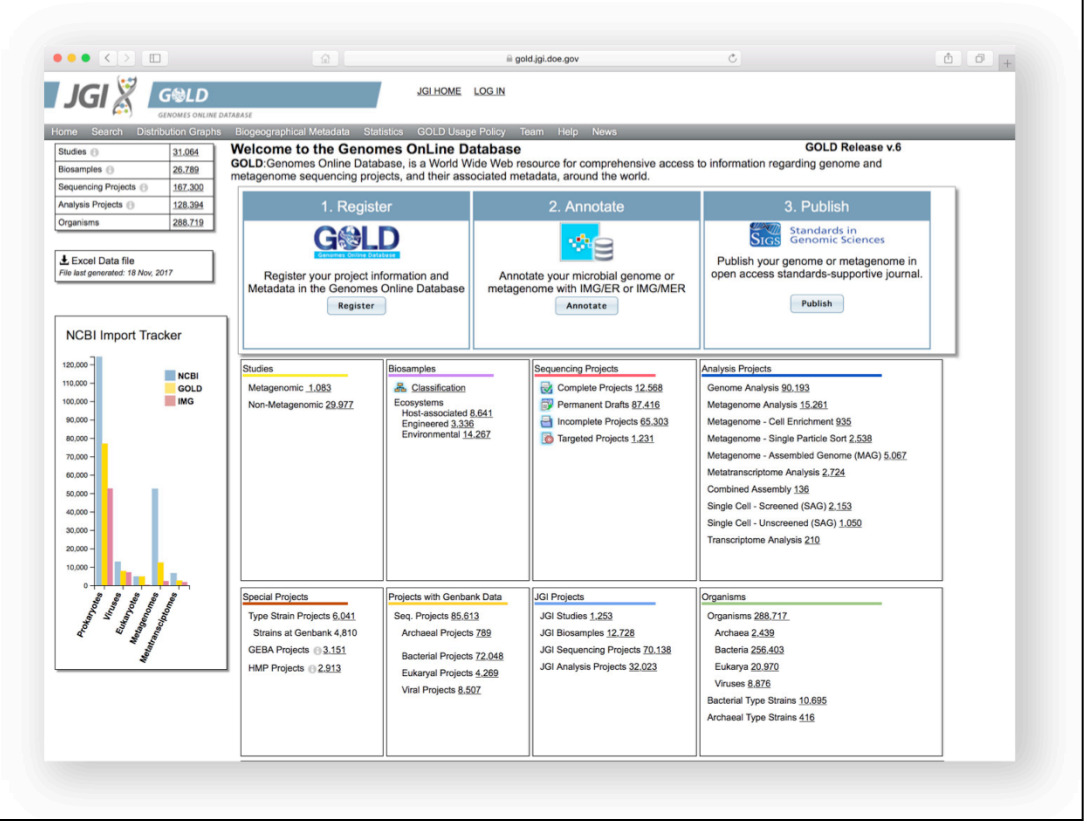
By translating the predicted coding sequence in all 6 frames and performing a BLAST search, we identify that exon 2 has a good match that does not extend to exon 1



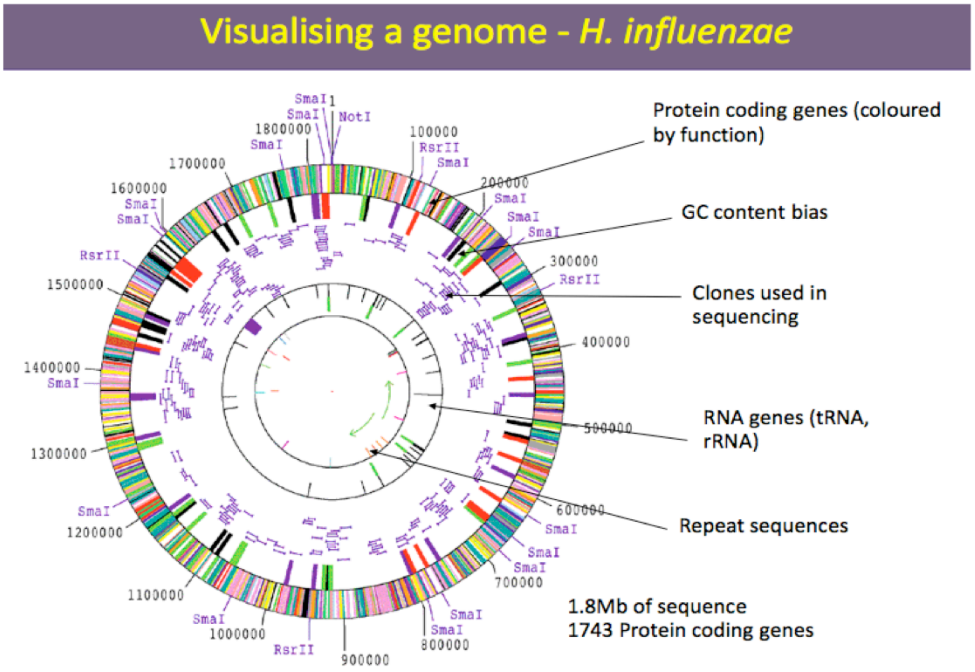
Slide courtesy of John Parkinson

Early genome drafts have many errors – even if your gene finder algorithm is 99 % correct, you will have 60 errors in a 6,000 gene yeast genome. If a sequence alignment shows areas of missing sequence, or unexpectedly high diversity, it is worthwhile to attempt a local amino acid alignment with sequence from alternate reading frames to identify possible sequencing frameshift errors. Another typical error is misidentifying the correct start codon, and if the sequence has a truncated N-terminus, it is worthwhile to check whether an alternate, upstream start codon can define the missing sequence fragment.

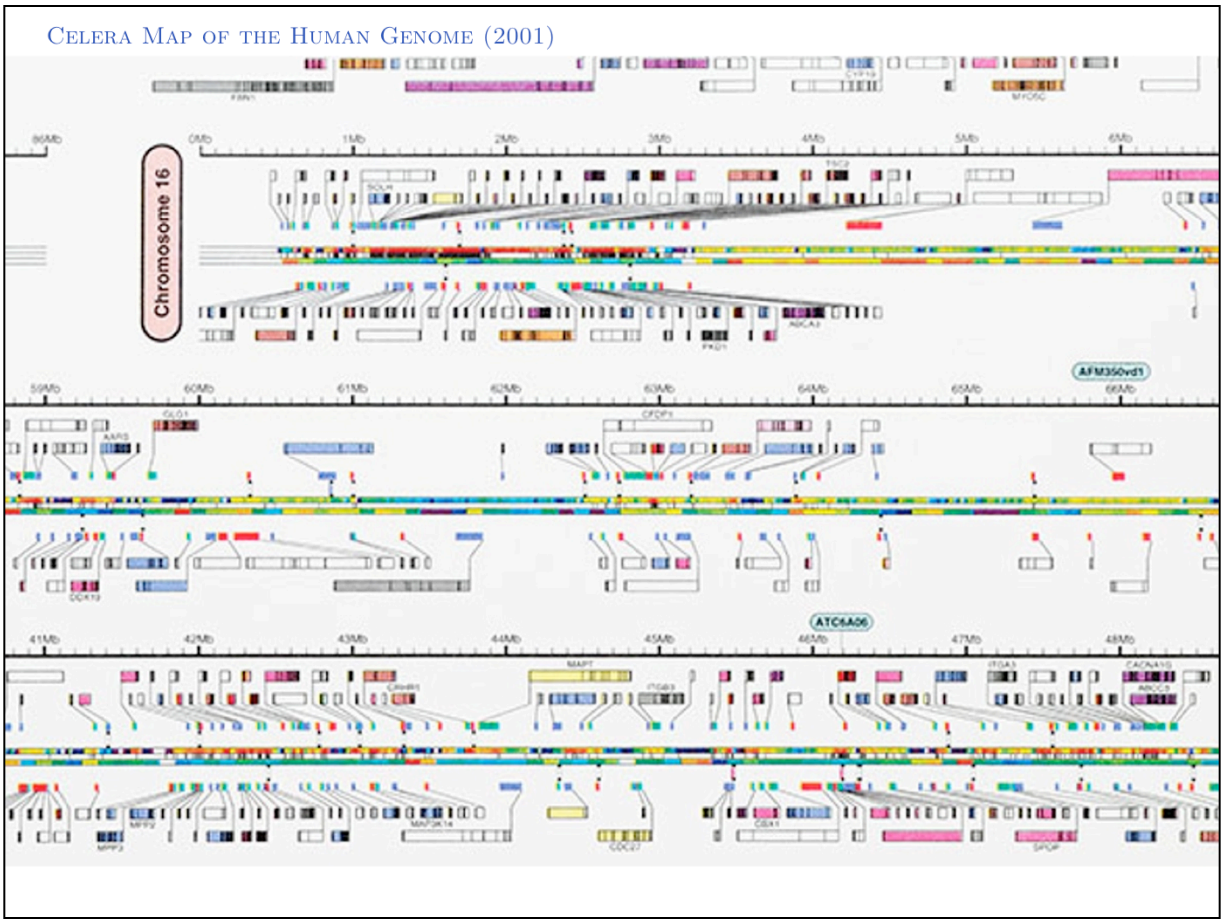
Genome Online Database



The GOLD genome database hosts information about the status of current genome projects and links to datasources.



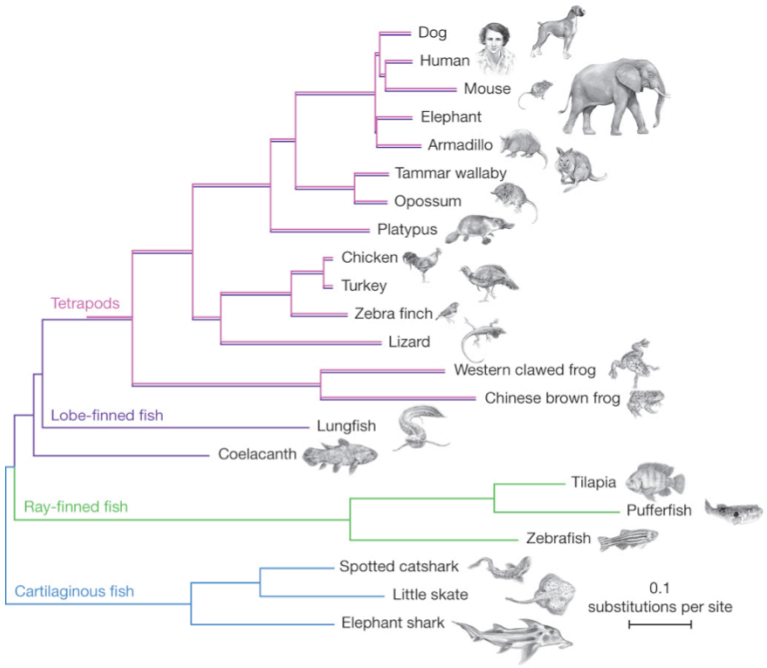
Bacterial genome information is sometimes depicted in a circular map. Not that there is only weak clustering, if any, of functionally related genes (similar colours). In fact, cross-species comparisons of syntenic regions show that the organization of the genome is in generally undergoing **random** fluctuations, with many rearrangements, duplications, and inversions, and without apparent global, organizing principles.



This is an excerpt from a (now historic) poster by Celera with annotations of their first draft human genome.

EXAMPLE PAPER: 2013

The African coelacanth genome provides insights into tetrapod evolution



A phylogenetic tree of a broad selection of jawed vertebrates shows that lungfish, not coelacanth, is the closest relative of tetrapods.

nature

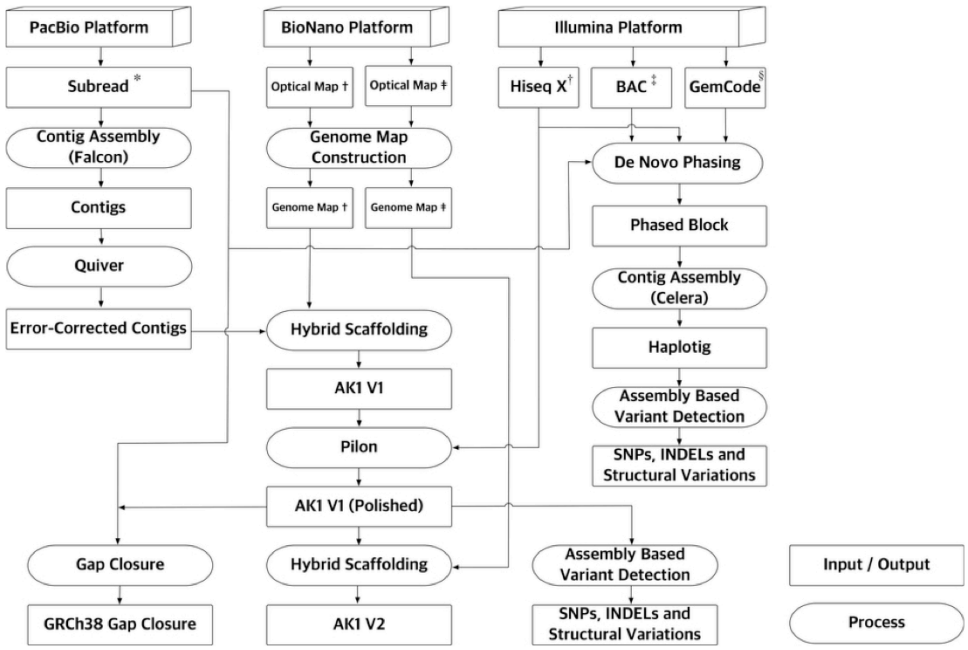
CT Amemiya *et al. Nature* **496**, 311-316 (2013) doi:10.1038/nature12027

Genome anotation methods and paradigms are rapidly evolving.

To get a better sense of **current** methods in this field, get a recently published high-impact genome publication and carefully study the methods section. The *coelacanth* genome is an excellent example for 2013.

Example Paper: 2016

De novo assembly and phasing of a Korean human genome



Flowchart of the data generation, processing and analysing for the de novo assembly and haplotype phasing of the AK1 diploid genome.

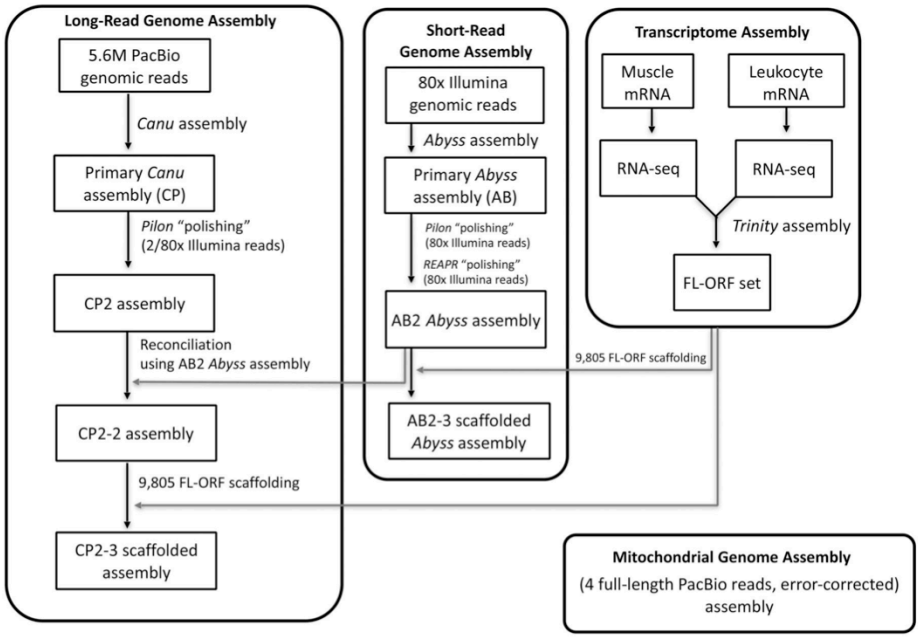
nature

JS Seo *et al. Nature* 538, 243-247 (2016) doi:10.1038/nature20098

A good 2016 paper is the haplotype resolved sequencing of a Korean genome from single-molecule sequences. Note in particular the workflow diagram. "Phasing" means assigning the individual reads to specific haplotypes.

Example Paper: 2017

De Novo Genome and Transcriptome Assembly of the Canadian Beaver (*Castor canadensis*)



Assembly pipeline

Lok *et al. Genes Genomes Genetics* **7(2)**, 755-773 (2017) doi:10.1534/g3.116.038208

A 2017 example paper is the genome sequence and transcriptome of the Canadian beaver (*Castor canadensis*) by a Toronto-based consortium on the occasion of Canada's 150th birthday. This is a genome that was primarily assembled from long-read data; short-read RNAseq technology was used primarily to characterize the transcriptome.

http://steipe.biochemistry.utoronto.ca/abc

B O R I S . S T E I P E @ U T O R O N T O . C A

DEPARTMENT OF BIOCHEMISTRY & DEPARTMENT OF MOLECULAR GENETICS
UNIVERSITY OF TORONTO, CANADA