Genome 559
Intro to Statistical and Computational Genomics
2009

Lecture 16a:
Computational Gene Prediction, II

Larry Ruzzo

1 Minute Reflections

Gene finding:

- cool; codons or nucleotides?

Python:

- review, extra time were good, still need more, I'm lost too.
- a.b or b.a or...?

hw:

- verbose. hard (~9hrs avg), large variance (3-25)
- relevance was pleasant. LLR scores?

Future:

- biopython, more on multiple alignment

Today:

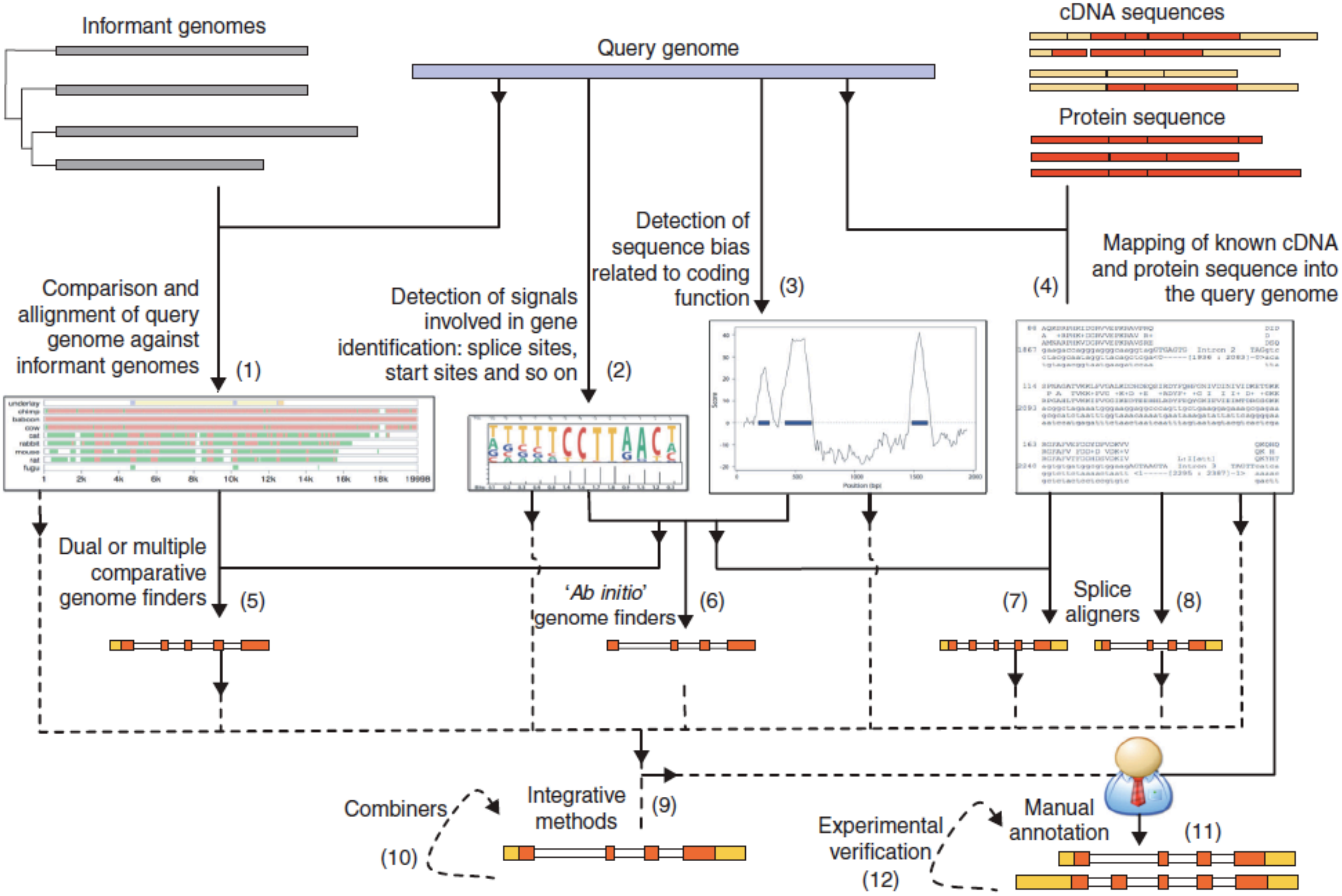
More on Gene Finding

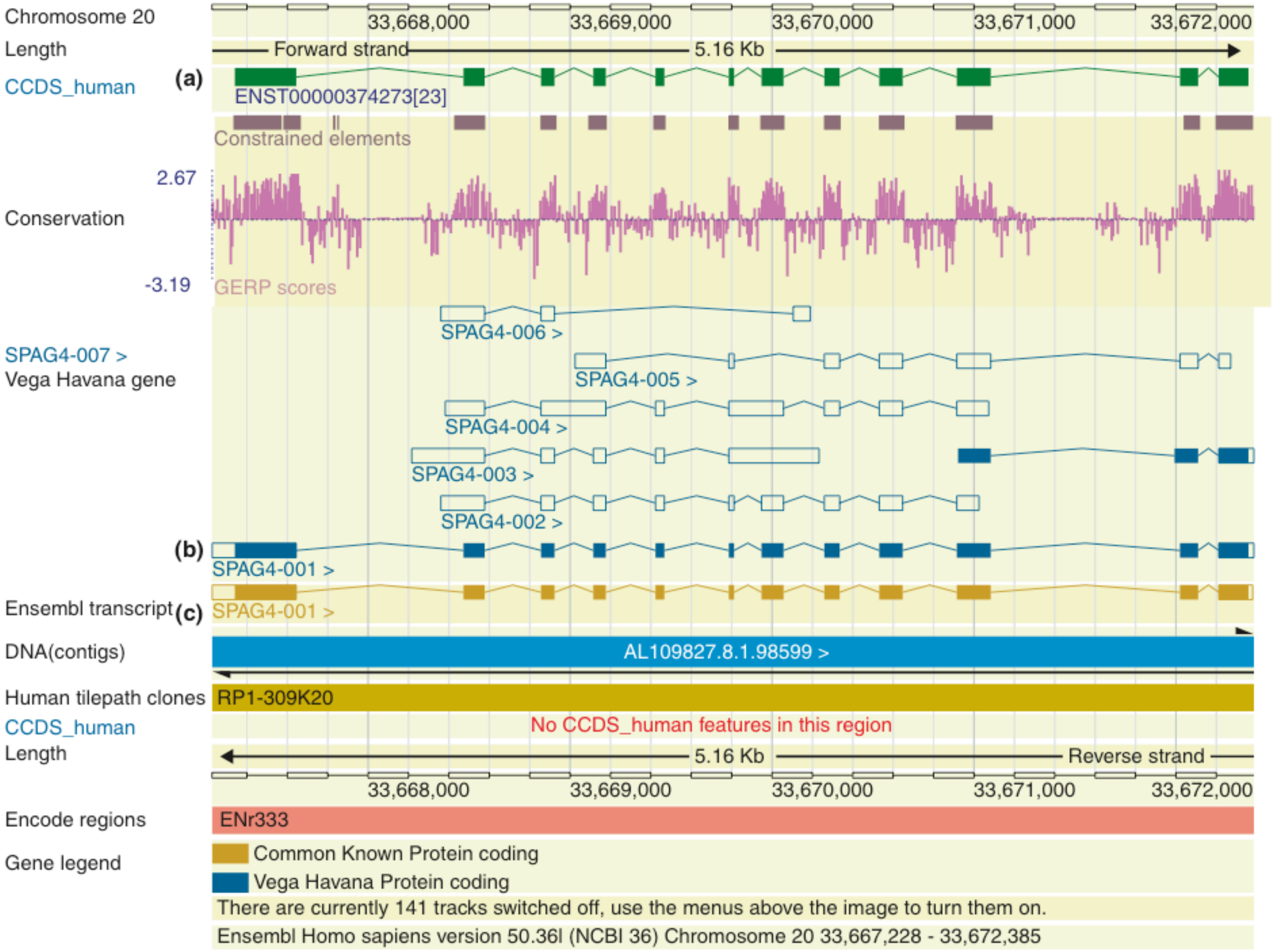
Harrow, *et al.* Identifying protein-coding genes in genomic sequences. Genome Biol. 2009, 10(1):201.

More on classes

Again, my goal is not to make you instant experts at this, but to acquaint you with the issues so you can use "object-oriented" tools, e.g., BioPython, and won't be intimidated by these features.

More practice


Informant genomes
Query genome
cDNA sequences
Protein sequence
Comparison and allignment of query genome against informant genomes
(1)
Detection of signals involved in gene identification: splice sites, start sites and so on
(2)
Detection of sequence bias related to coding function
(3)
(4)
Mapping of known cDNA and protein sequence into the query genome
Dual or multiple comparative genome finders
(5)
'Ab initio' genome finders
(6)
(7)
Splice aligners
(8)
Combiners
(10)
Integrative methods
(9)
Experimental verification
(12)
Manual annotation
(11)



There are currently 141 tracks switched off, use the menus above the image to turn them on.

Ensembl Homo sapiens version 50.36l (NCBI 36) Chromosome 20 33,667,228 - 33,672,385

Summary

Integrate many sources of information

Many tools you’ve seen:

- BLAST, pairwise alignment, multiple alignment, sequence profiles/weight matrix models, phylogenetic modeling

And extensions:

- *Hidden* Markov models, *spliced* alignment, ...

Assessment:

- purely computational predictions – ~80% accurate on exons, ~60% on genes (e.g., often extra/missing exons)
- So, manual curation still valuable