Distinguishing protein-coding and noncoding genes in the human genome

Michele Clamp, Ben Fry, Mike Kamal, Xiaohui Xie, James Cuff, Michael F. Lin, Manolis Kellis,Kerstin Lindblad-Toh, and Eric S. Lander.

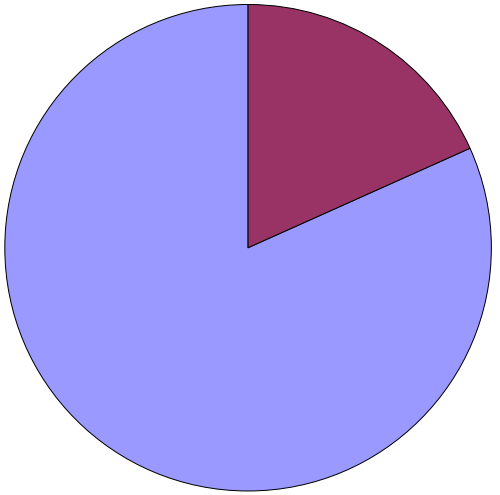
BIIT Journal Club presentation 12.12.2007
by Konstantin Tretjakov

How are genes detected on the DNA?

- An ORF exceeding ≈300bp?
  - But human noncoding regions are GC-rich:
    - Lower probability of a stop-codon (UAG/UAA/UGA)
    - Long ORFs occur by chance

Spurious genes currently in databases

- Out of ≈**24 500** human genes in *Ensembl, RefSeq* and *Vega*, only ≈**20 000** show conservation with dog.



- How can you prove that a given ORF is **not** a protein-coding gene?

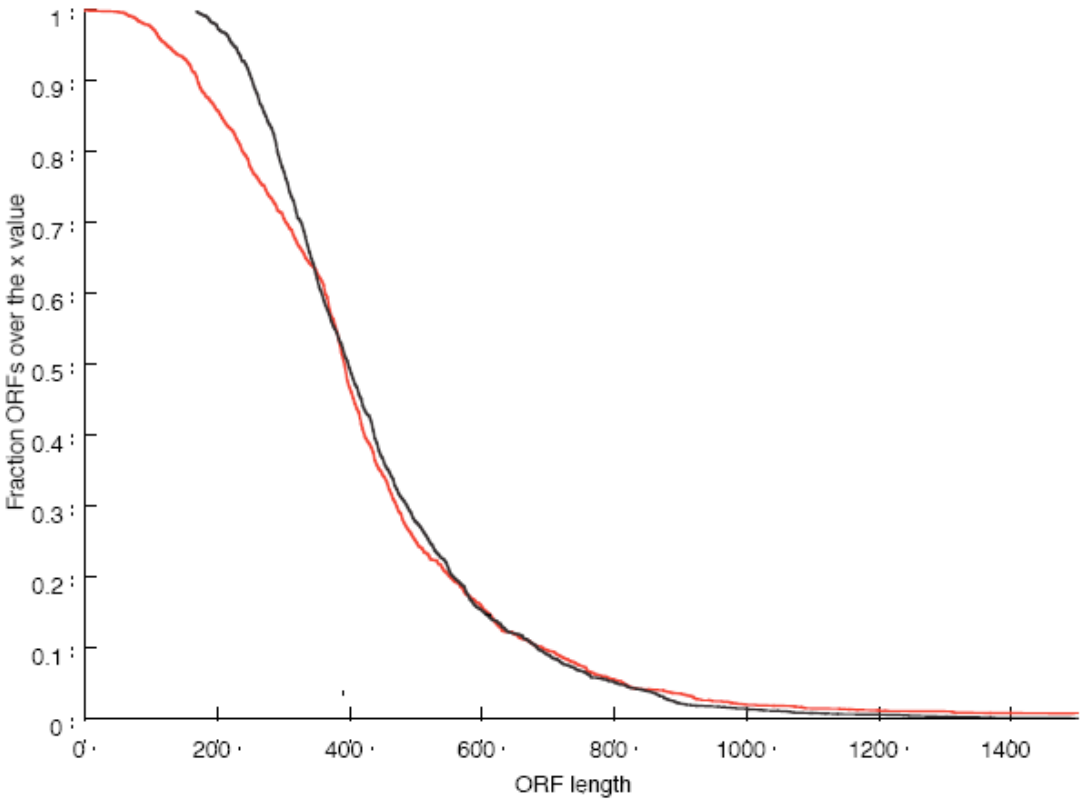
Revision of human genes

- Ensembl 35: **22 218** protein-coding genes
- Considers **21 895** of these:
  - Transposons & pseudogenes: **1 538**
  - Genes with orthologs in dog/mouse: **18 752**
  - Genes with paralogs orthologous to dog/mouse: **155**
  - Genes with human-only paralogs: **51** + 17
  - Genes with Pfam domains: **36** + **40** + 21
  - Orphans: **1 177** + 68 + 40

Characterizing the orphans: ORF Lengths

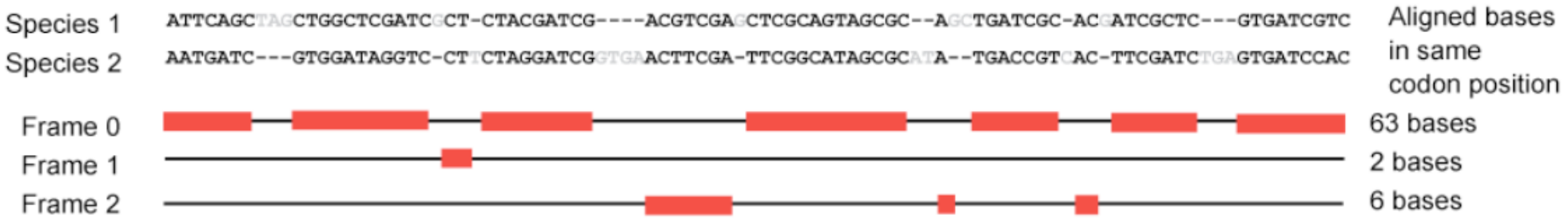
- Short (median length = 393bp)
- Distribution of lengths resembles random

red – random
black – orphans



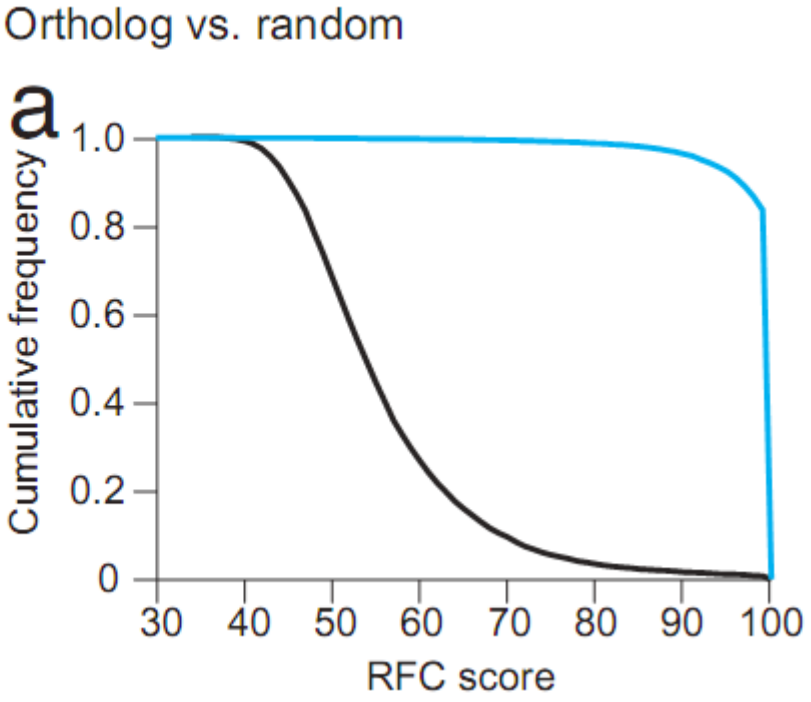
Characterizing the orphans: Conservation

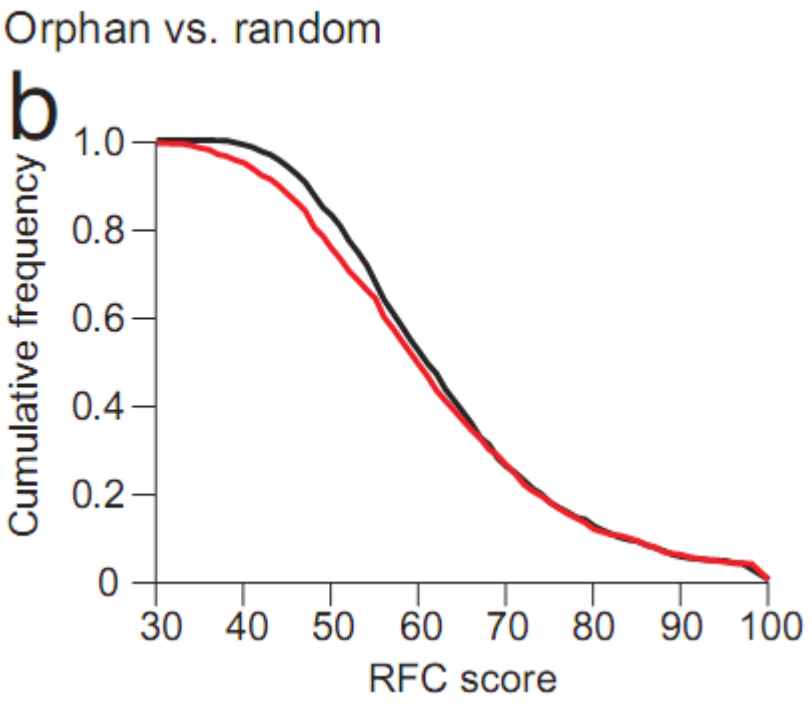
- Compare a set of 5 985 “well-studied” genes to random controls
  - Indel density < 10 per kb
    - 97.3% of well-studied genes and only 2.8% of controls
  - Reading-frame conservation (RFC) > 90
    - 98.2% of well-studied genes and only 1% of controls



Reading frame conservation

JOINT
(MOUSE, DOG)





Codon substitution frequency (CSF)

- Idea:
  - Align a gene codon-by-codon to an ortholog
  - For each mutated codon, consider how much more likely is this change to occur in a coding region than in a noncoding region. Sum log-likelihoods.
- Results:
  - Among 16 210 genes with orthologs 99.2% yielded results consisted with the expected evolution.
  - Among the orphans there were 2, and these turned out to be just misannotated/mistranslated genes.

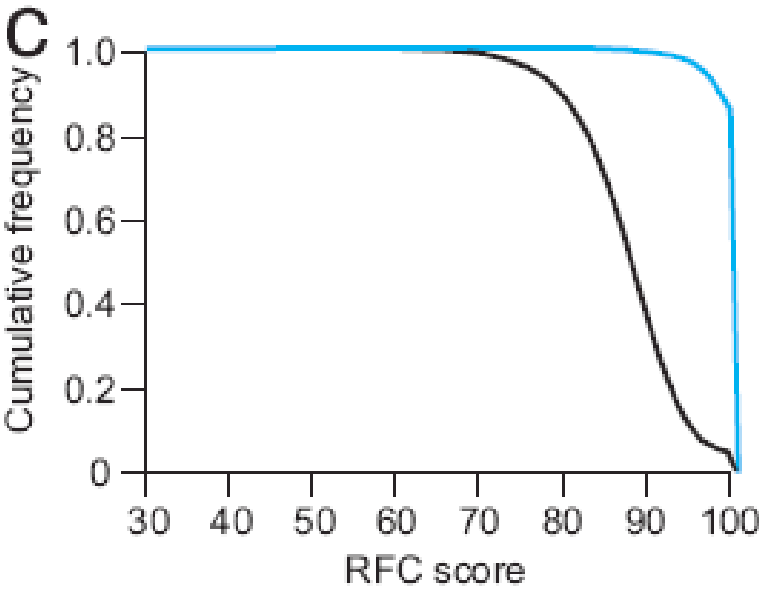
Orphans are not protein-coding?

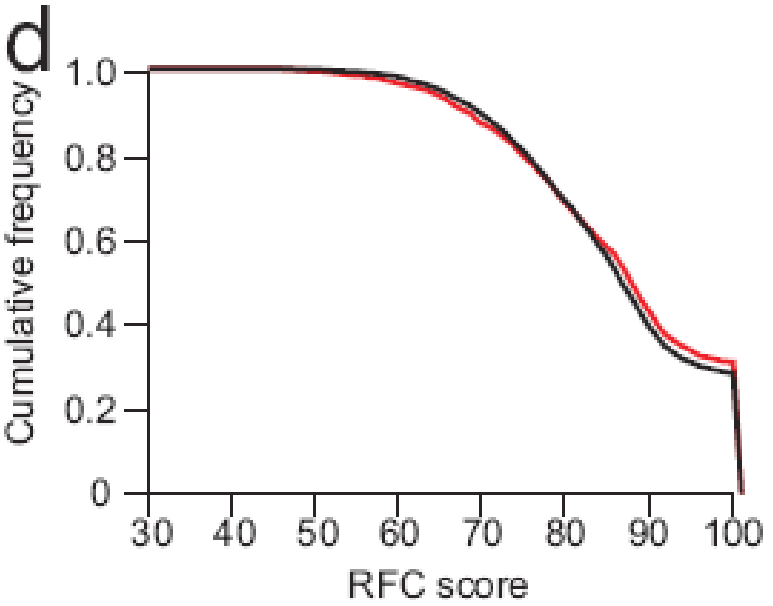
- Suppose they are protein-coding, then
  - They might be “old” genes (lost in mouse and dog).
    - Then they should be functional in macaque and chimp.
  - They might be “young” genes.
    - Then the majority of them should be functional in macaque and chimp.

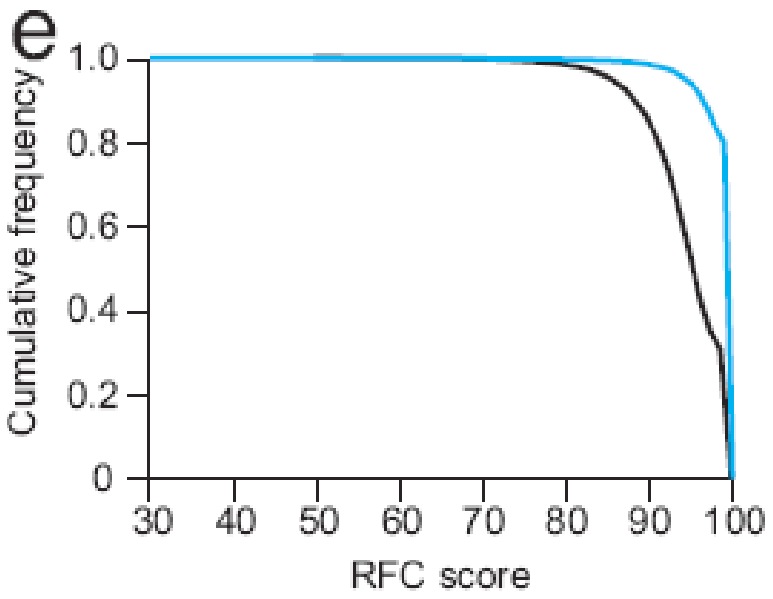
RFC with macaque and chimp

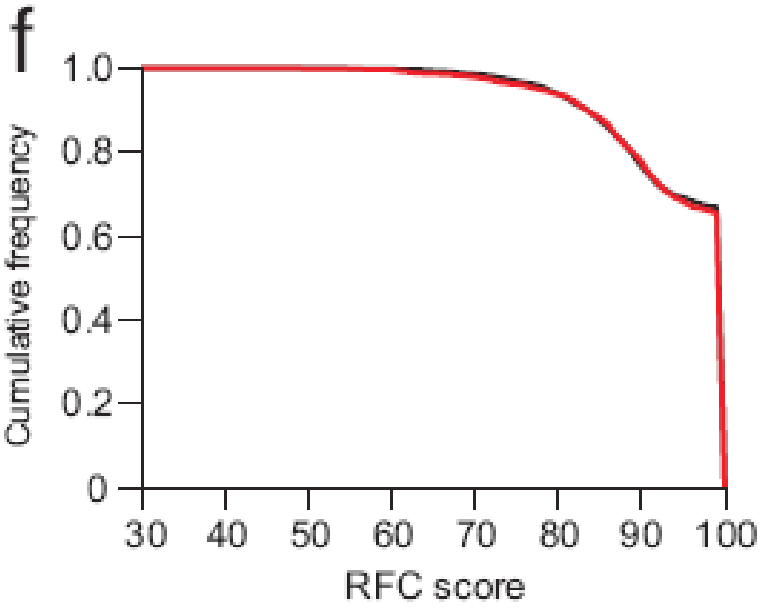
MACAQUE

CHIMP









Experimental evidence

- Out of 1 177 orphans only 12 are reported to encode a protein in vivo, some of the reports are equivocal.

Summary

- Analysis reports **19 108** valid genes.
- Ensembl 35 has therefore ≈**19 199** genes.
- Ensembl 38 + RefSeq + Vega =>
  20 470 valid genes (out of **24 551**)

- Bonus: Gene report cards
  http://www.broad.mit.edu/mammals/alpheus

Discussion points

- Use stringent conservation-based filters before reporting an ORF as a protein-coding gene.
- The largest problem is with short ORFs.
- Novel protein-coding genes arise rarely in mammalian lienages (≈168 human-specific genes, which are not all very novel)
- Need proper catalog reviews for other species.

End of talk. Thank you. Questions please.