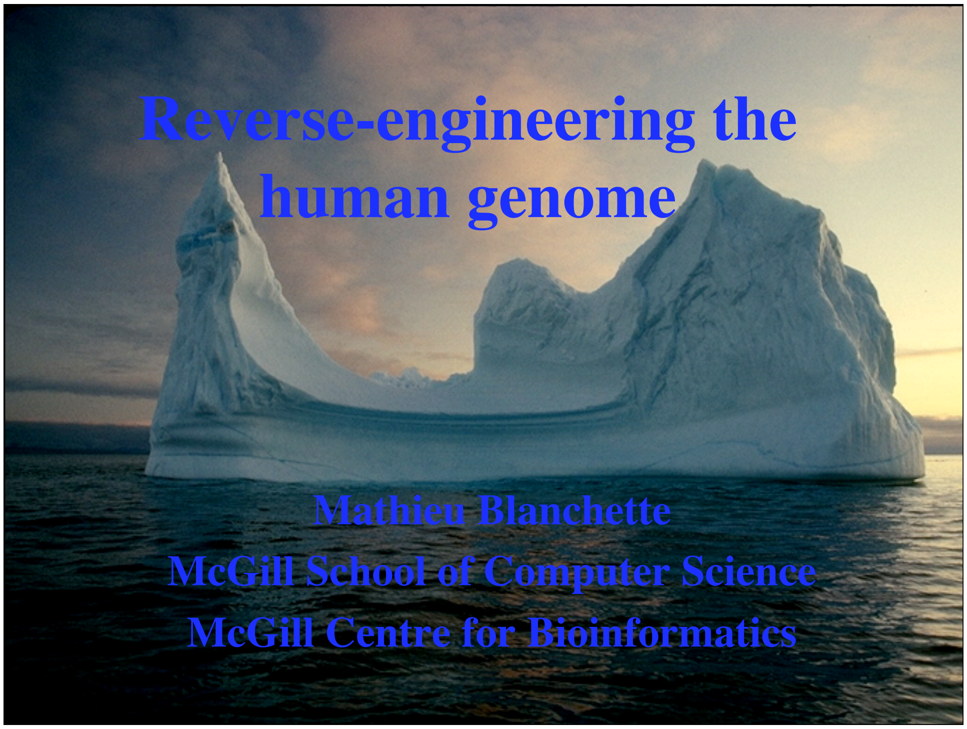
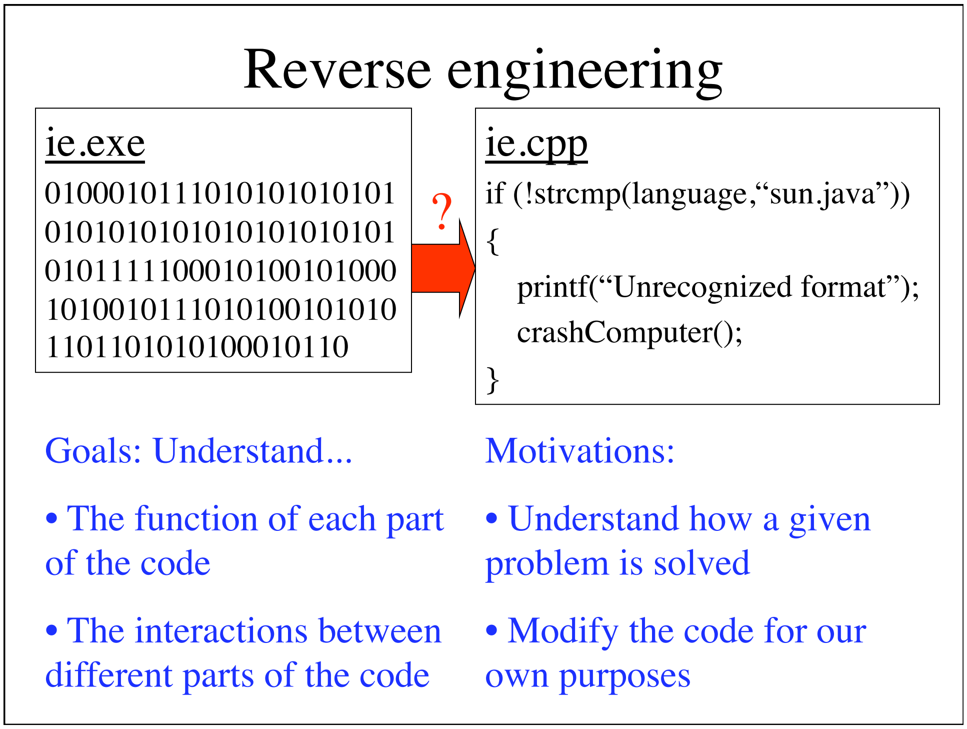


Reverse-engineering the human genome

Mathieu Blanchette
McGill School of Computer Science
McGill Centre for Bioinformatics



Reverse engineering

ie.exe

```
0100010111010101010101
0101010101010101010101
0101111100010100101000
1010010111010100101010
1101101010100010110
```

?

ie.cpp

```
if (!strcmp(language,“sun.java”))
{
   printf(“Unrecognized format”);
   crashComputer();
}
```

Goals: Understand...

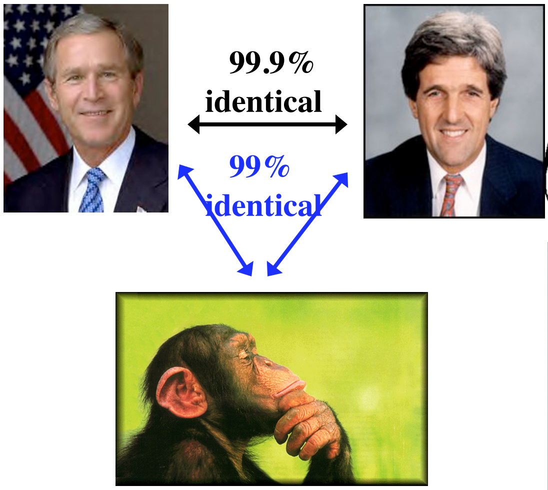
- The function of each part of the code
- The interactions between different parts of the code

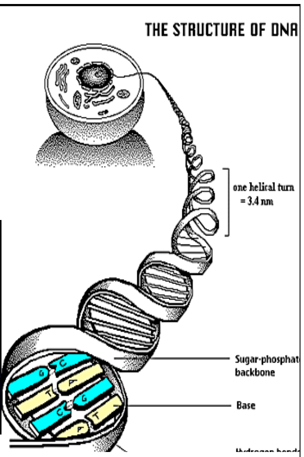
Motivations:

- Understand how a given problem is solved
- Modify the code for our own purposes

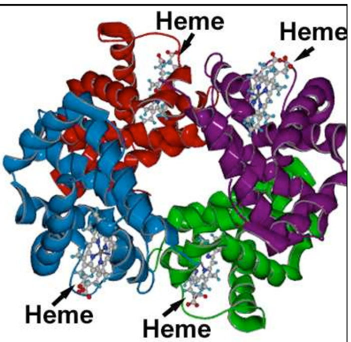
Genomes

- Human genome: $\{A,C,G,T\}^{3 \times 10^9}$
- Each of your 10^{14} cells has two copies

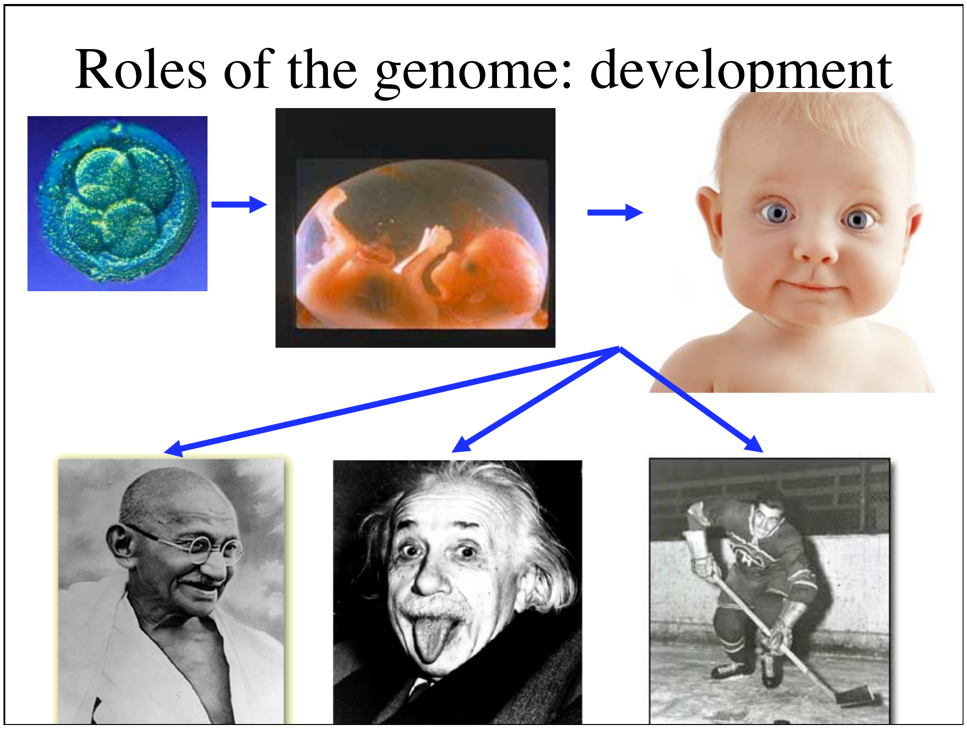




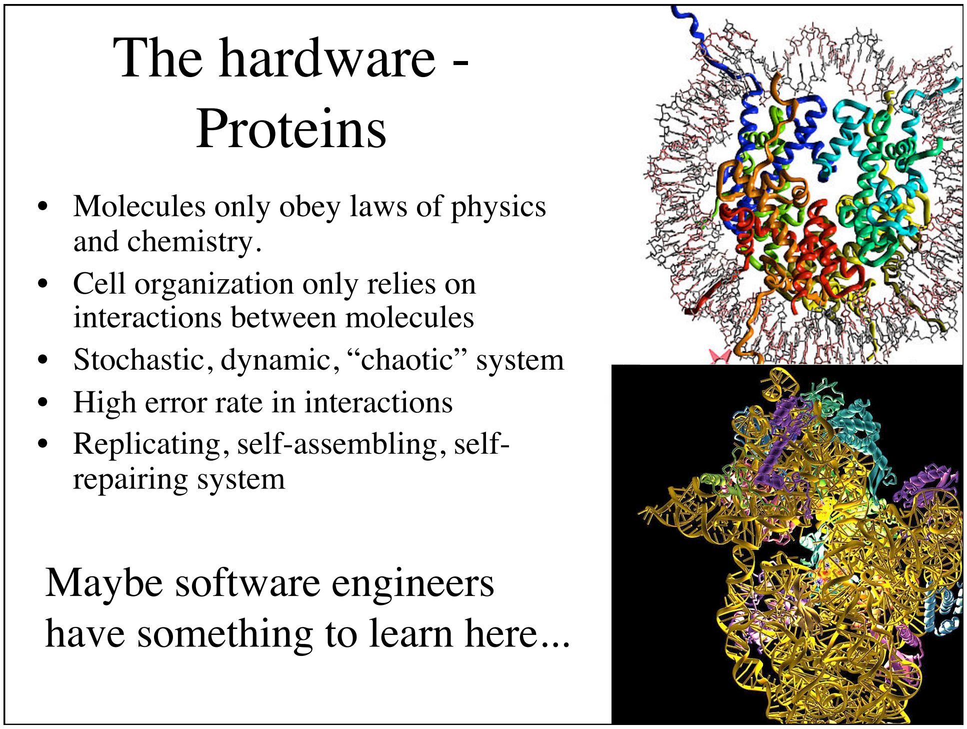
Roles of the genome



- Genome is a blue print for a cell
- Describes *how* to build proteins
  - 25,000 genes --> 25,000 proteins (+variations)
  - Each protein has its biochemical function
- Describes *when* to build each protein
  - Under which situations should a gene be expressed?
- Proteins allow:
  - Cell administration and maintenance
  - Reaction to stimuli
  - Protocols for communication between cells



Roles of the genome: development



The hardware - Proteins

- Molecules only obey laws of physics and chemistry.
- Cell organization only relies on interactions between molecules
- Stochastic, dynamic, “chaotic” system
- High error rate in interactions
- Replicating, self-assembling, self-repairing system

Maybe software engineers have something to learn here...

Content of the genome - Genes

- Gene: region of DNA that encodes one protein

DNA sequences $\xrightarrow{?}$ Protein sequence

$\{A,C,G,T\}^L$

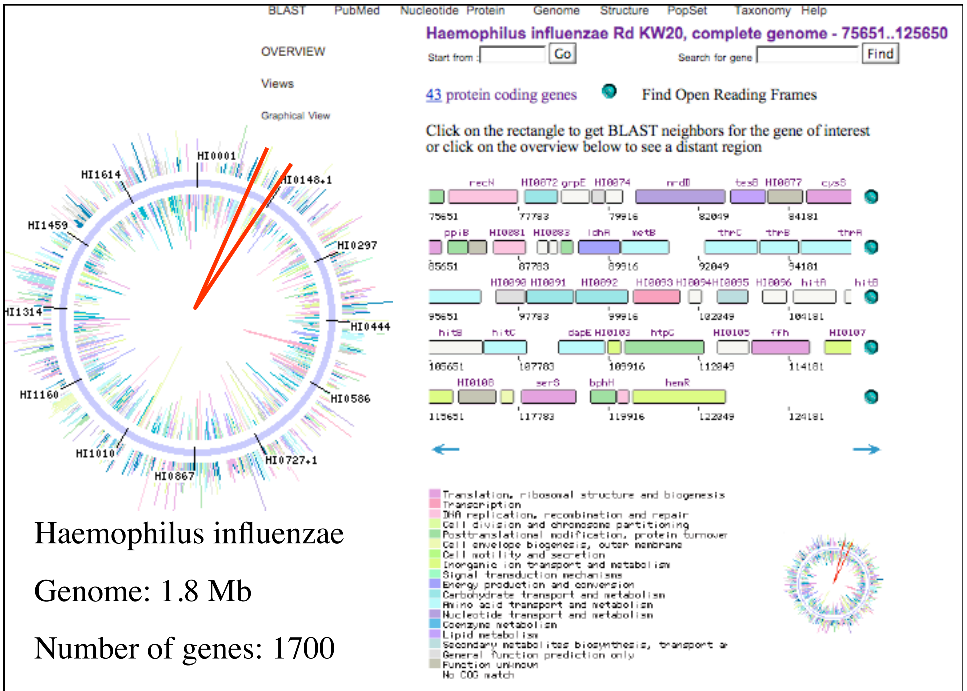
$\{A,C,D,E,F,G,H,I,K,L,M,N,P,Q,R,S,T,V,W,Y\}^K$

Second Letter of Codon

First letter of Codon (5' end)	U	C	A	G
U	UUU Phe UUC Phe UUA Leu UUG Leu	UUC Ser UCC Ser UCA Ser UCG Ser	UAU Tyr UAC Tyr UAA Stop UAG Stop	UGU Cys UGC Cys UGA Stop UGG Trp
C	CUU Leu CUC Leu CUA Leu CUG Leu	CCU Pro CCC Pro CCA Pro CUG Pro	CAU His CAC His CAA Gin CAG Gin	CGU Arg CGC Arg CGA Arg CGG Arg
A	AUU Ile AUC Ile AUA Ile AUG Met	ACU Thr ACC Thr ACA Thr ACG Thr	AAU Asn AAC Asn AAA Lys AAG Lys	AGU ser AGC ser AGA Arg AGG Arg
G	GUU Val GUC Val GUA Val GUG Val	GCU Ala GCC Ala GCA Ala GCG Ala	GAU Asp GAC Asp GAA Glu GAG Glu	GGU Gly GCG Gly GGA Gly GGG Gly

STOP

START



Haemophilus influenzae Rd KW20, complete genome - 75651..125650

Start from : Go

Search for gene Find

43 protein coding genes

Find Open Reading Frames

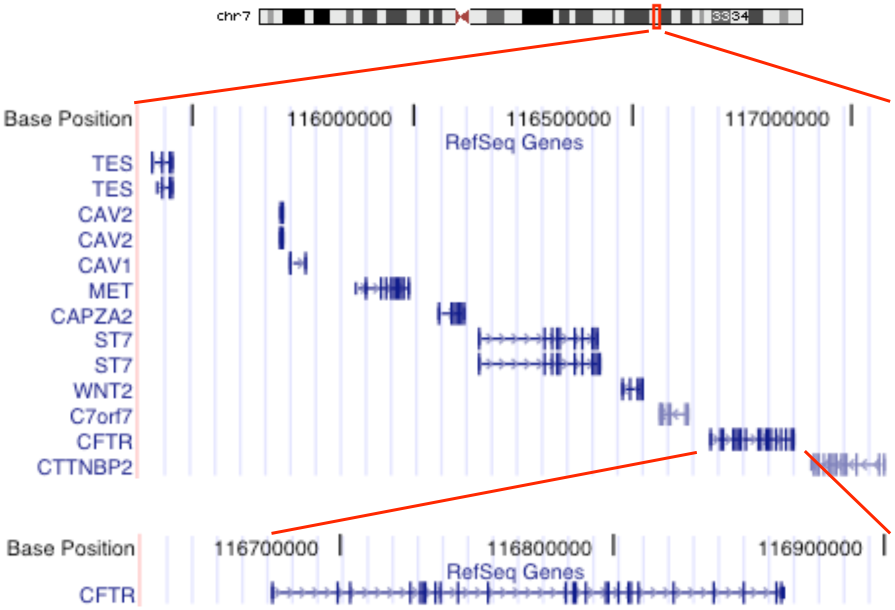
Click on the rectangle to get BLAST neighbors for the gene of interest or click on the overview below to see a distant region

- Translation, ribosomal structure and biogenesis
- Transcription
- DNA replication, recombination and repair
- Cell division and chromosome partitioning
- Posttranslational modification, protein turnover
- Cell envelope biogenesis, outer membrane
- Cell motility and secretion
- Inorganic ion transport and metabolism
- Signal transduction mechanisms
- Energy production and conversion
- Carbohydrate transport and metabolism
- Amino acid transport and metabolism
- Nucleotide transport and metabolism
- Coenzyme metabolism
- Lipid metabolism
- Secondary metabolites biosynthesis, transport a
- General function prediction only
- Function unknown
- No COG match

Haemophilus influenzae

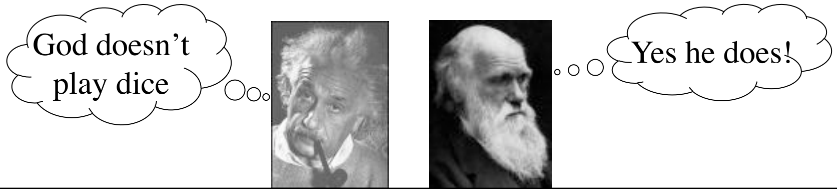
Genome: 1.8 Mb

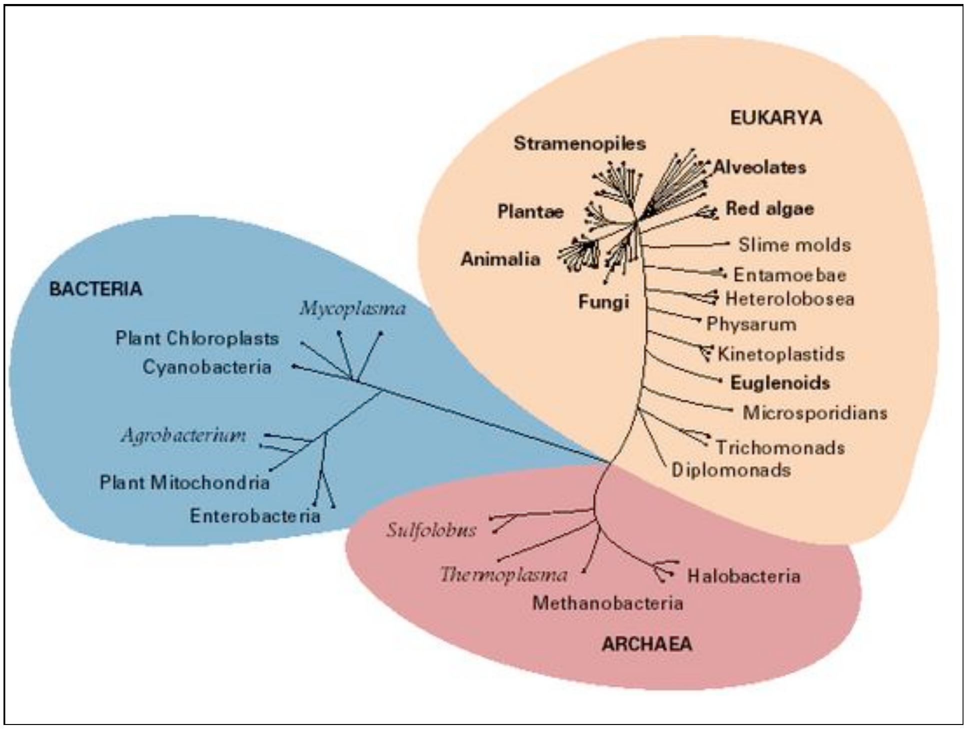
Number of genes: 1700



The Programmer - Evolution

- Design principles:
  - Random modifications (variation)
  - Survival of the fittest (natural selection)
- 3 Billion years of evolution
- Today's species are the current solution of the fitness optimization problem

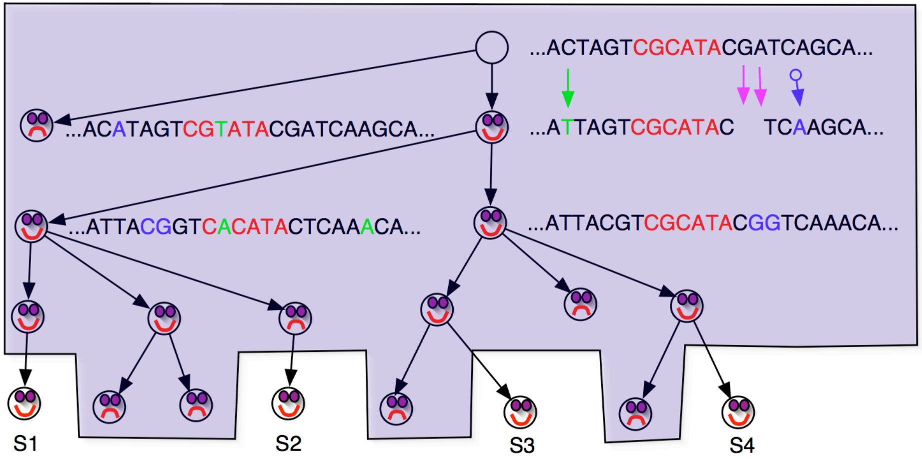




Central dogma of comparative genomics

...ACTAGTCGCATACGATCAGCA...

Unavailable!



...ACTAGTCGCATACGATCAGCA...

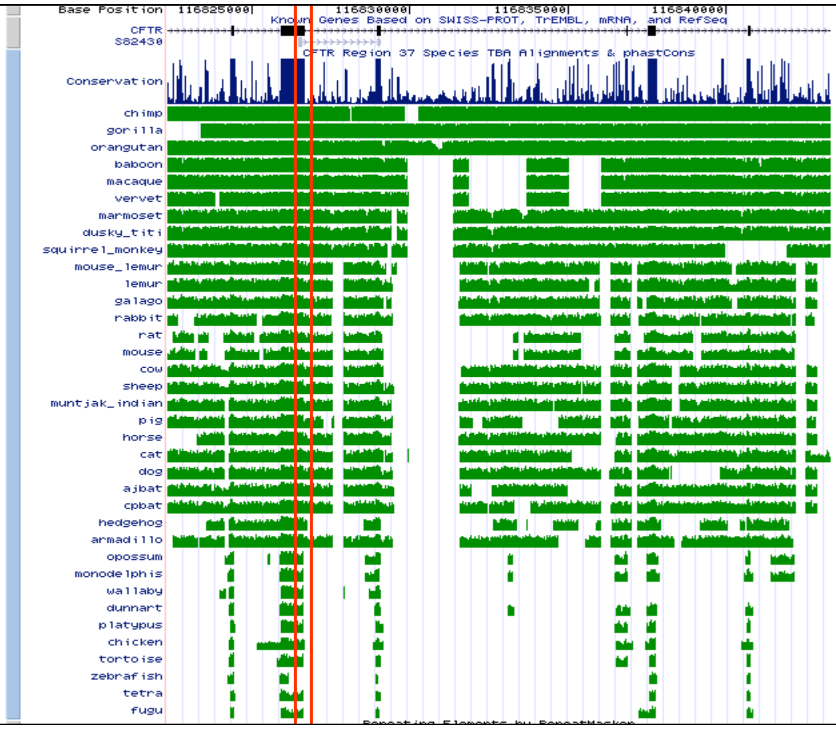
...ATTAGTCGCATAC TCAAGCA...

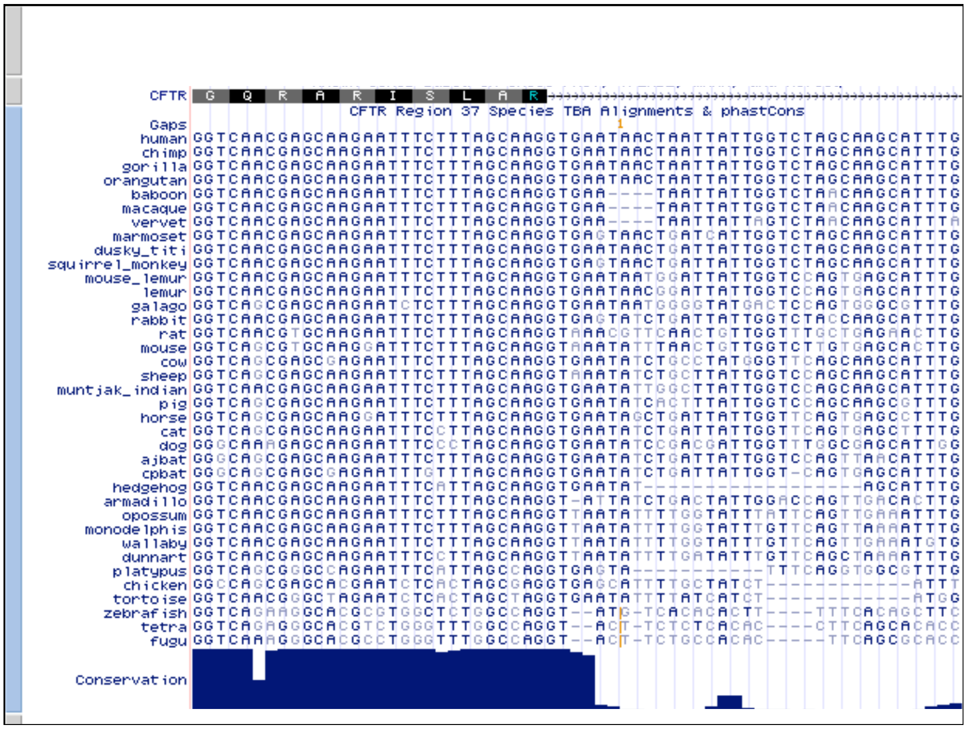
...ACATAGTCGTATACGATCAAGCA...

...ATTACGGTCACATACTCAAACA...

...ATTACGTCGCATACGGTCAAACA...

```
S1:  GTTACGGTCACATACTGAAACA
S2:  GTTATGGTCACATACTGAAACTGA
S3:  ATTACTCGCATACGGTCTAACA
S4:  ATTTACTCGCATACGGTCTAGCAC
```





```
           CFTR  G  Q  R  A  R  I  S  L  A  R
                 CFTR Region 37 Species TBA Alignments & phastCons
                                                  1
           Gaps
          human GGTCAACGAGCAAGAATTTCTTTAGCAAGGTGAATAACTAATTATTGGTCTAGCAAGCATTTG
          chimp GGTCAACGAGCAAGAATTTCTTTAGCAAGGTGAATAACTAATTATTGGTCTAGCAAGCATTTG
        gorilla GGTCAACGAGCAAGAATTTCTTTAGCAAGGTGAATAACTAATTATTGGTCTAGCAAGCATTTG
      orangutan GGTCAACGAGCAAGAATTTCTTTAGCAAGGTGAATAACTAATTATTGGTCTAGCAAGCATTTG
         baboon GGTCAACGAGCAAGAATTTCTTTAGCAAGGTGAA----TAATTATTGGTCTAACAAGCATTTG
        macaque GGTCAACGAGCAAGAATTTCTTTAGCAAGGTGAA----TAATTATTGGTCTAACAAGCATTTG
         vervet GGTCAACGAGCAAGAATTTCTTTAGCAAGGTGAA----TAATTATTAGTCTAACAAGCATTTA
       marmoset GGTCAACGAGCAAGAATTTCTTTAGCAAGGTGAGTAACTGATCATTGGTCTAGCAAGCATTTG
     dusky_titi GGTCAACGAGCAAGAATTTCTTTAGCAAGGTGAATAACTGATTATTGGTCTAGCAAGCATTTG
squirrel_monkey GGTCAACGAGCAAGAATTTCTTTAGCAAGGTGAGTAACTGATTATTGGTCTAGCAAGCATTTG
    mouse_lemur GGTCAACGAGCAAGAATTTCTTTAGCAAGGTGAATAATGGATTATTGGTCCAGTGAGCATTTG
          lemur GGTCAACGAGCAAGAATTTCTTTAGCAAGGTGAATAACGGATTATTGGTCCAGTGAGCATTTG
         galago GGTCAGCGAGCAAGAATCTCTTTAGCAAGGTGAATAATGGGGTATGACTCCAGTGGGCGTTTG
         rabbit GGTCAACGAGCAAGAATTTCTTTAGCAAGGTGAGTATCTGATTATTGGTCTACCAAGCATTTG
            rat GGTCAACGTGCAAGAATTTCTTTAGCAAGGTAAACGTTCAACTGTTGGTTTGCTGAGAACTTG
          mouse GGTCAGCGTGCAAGGATTTCTTTAGCAAGGTAAATATTTAACTGTTGGTCTTGTGAGCACTTG
            cow GGTCAGCGAGCGAGAATTTCTTTAGCAAGGTGAATATCTGCCTATGGGTTCAGCAAGCATTTG
          sheep GGTCAGCGAGCAAGAATTTCTTTAGCAAGGTAAATATCTGCTTATTGGTCCAGCAAGCATTTG
 muntjak_indian GGTCAACGAGCAAGAATTTCTTTAGCAAGGTGAATATTGGCTTATTGGTCCAGCAAGCATTTG
            pig GGTCAGCGAGCAAGAATTTCTTTAGCAAGGTGAATATCACTTTATTGGTCCAGCAAGCGTTTG
          horse GGTCAGCGAGCAAGGATTTCTTTAGCAAGGTGAATAGCTGATTATTGGTTCAGTGAGCCTTTG
            cat GGTCAGCGAGCAAGAATTTCCTTAGCAAGGTGAATATCTGATTATTGGTTCAGTGAGCTTTTG
            dog GGGCAAAGAGCAAGAATTTCCCTAGCAAGGTGAATATCCGACGATTGGTTTGGCGAGCATTGG
          ajbat GGGCAGCGAGCAAGAATTTCTTTAGCAAGGTGAATATCTGATTATTGGTCCAGTTAACATTTG
          cpbat GGGCAGCGAGCGAGAATTTGTTTAGCAAGGTGAATATCTGATTATTGGT-CAGTGAGCATTTG
       hedgehog GGTCAACGAGCAAGAATTTCATTAGCAAGGTGAATAT------------------AGCATTTG
      armadillo GGTCAACGAGCAAGAATTTCTTTAGCAAGGT-ATTATCTGACTATTGGACCAGTTGACACTTG
        opossum GGTCAACGAGCAAGAATTTCTTTAGCAAGGTTAATATTTTGGTATTTATTCAGTTGAAATTTG
    monodelphis GGTCAACGAGCAAGAATTTCTTTAGCAAGGTTAATATTTTGGTATTTGTTCAGTTAAAATTTG
        wallaby GGTCAACGAGCAAGAATTTCTTTAGCAAGGTTAATATTTTGGTATTTGTTCAGTTGAAATGTG
        dunnart GGTCAACGAGCAAGAATTTCCTTAGCAAGGTTAATATTTTGATATTTGTTCAGCTAAAATTTG
       platypus GGTCAGCGGGCCAGAATTTCATTAGCCAGGTGAGTA-----------TTTCAGGTGGCGTTTG
        chicken GGCCAGCGAGCACGAATCTCACTAGCGAGGTGAGCATTTTGCTATCT------------ATTT
       tortoise GGTCAACGGGCTAGAATCTCACTAGCTAGGTGAATATTTTATCATCT------------ATGG
      zebrafish GGTCAGAAGGCACGCGTGGCTCTGGCCAGGT--ATG-TCACACACTT----TTTCACAGCTTC
          tetra GGTCAGAGGGCACGTCTGGGTTTGGCCAGGT--ACT-TCTCTCACAC----CTTCAGCACACC
           fugu GGTCAAAGGGCACGCCTGGGTTTGGCCAGGT--ACT-TCTGCCACAC-----TTCAGCGCACC

   Conservation
```

Mammalian evolution

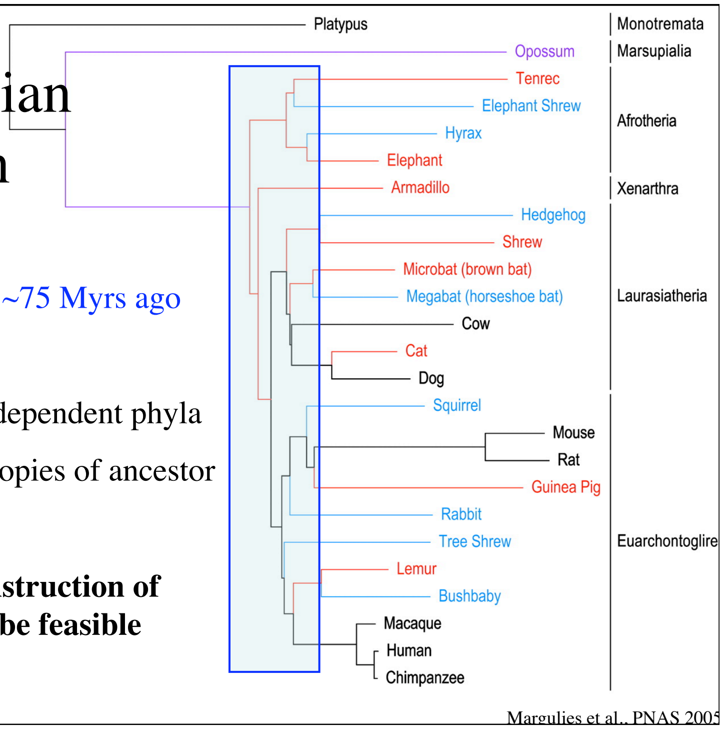
- Rapid radiation ~75 Myrs ago

⬇

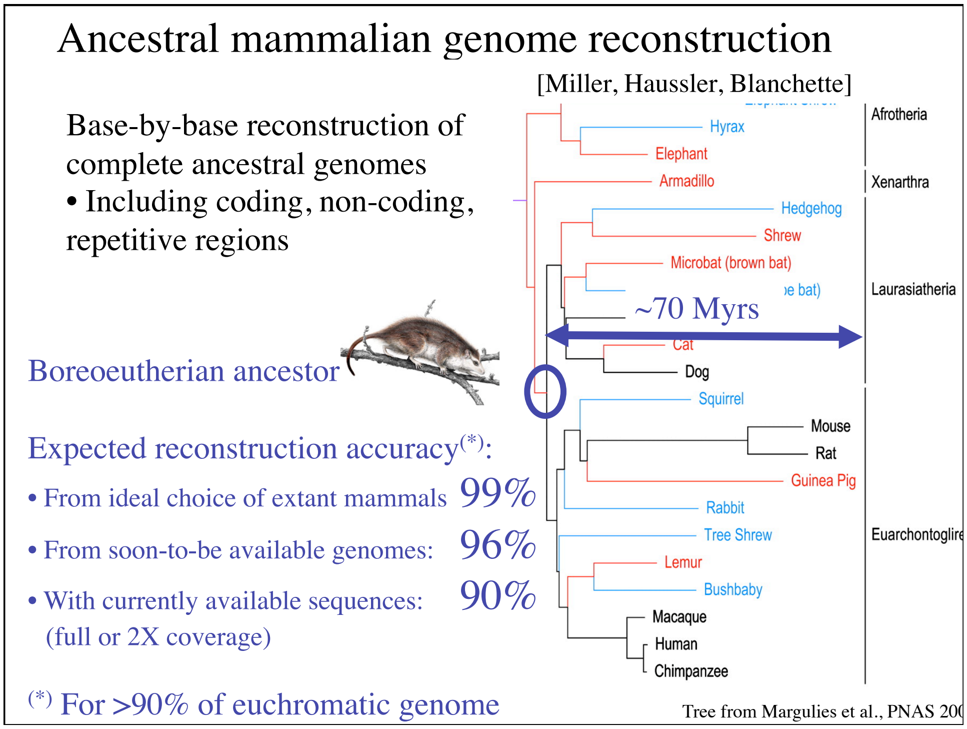
- Many nearly independent phyla
- Many “noisy” copies of ancestor

⬇

- **Accurate reconstruction of ancestors may be feasible**



Margulies et al., PNAS 2005



Ancestral mammalian genome reconstruction

[Miller, Haussler, Blanchette]

Base-by-base reconstruction of complete ancestral genomes
- Including coding, non-coding, repetitive regions

Boreoeutherian ancestor

Expected reconstruction accuracy(*):
- From ideal choice of extant mammals 99%
- From soon-to-be available genomes: 96%
- With currently available sequences: 90%
  (full or 2X coverage)

(*) For >90% of euchromatic genome

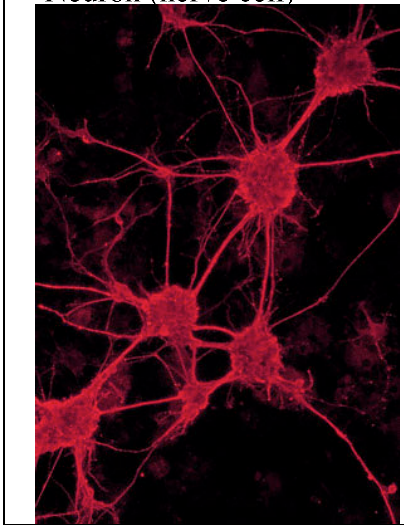
Tree from Margulies et al., PNAS 200

Why should we care?

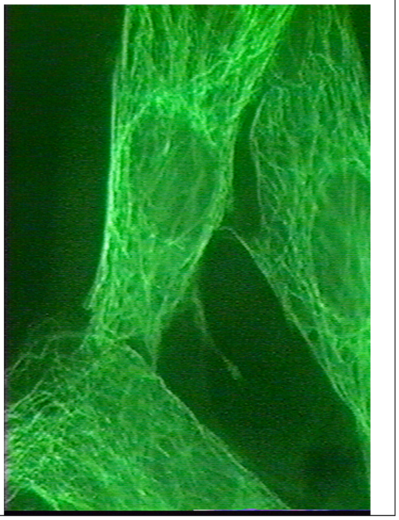
- Se st see its
  –
  –
- Bo ian stuff
  –
  –
  –

One program, many functions

Neuron (nerve cell)



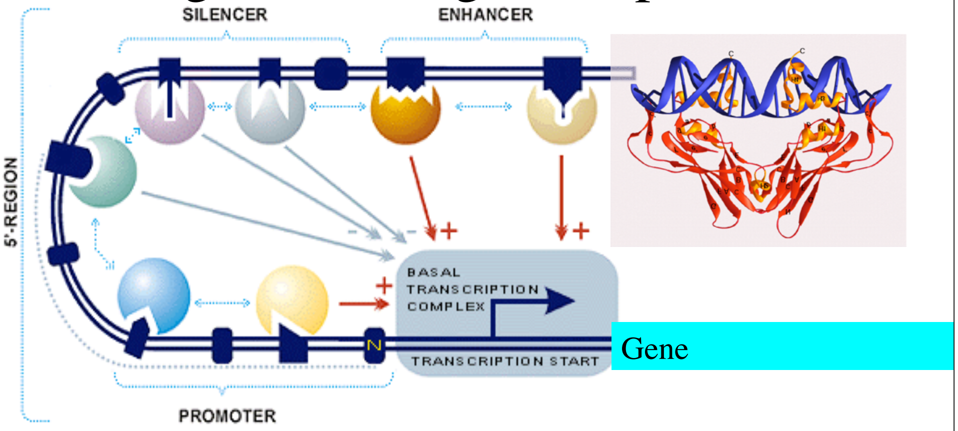
Muscle cell



Same genome!

↔

Regulation of gene expression



Transcription factor binding sites:

☹ Short: 6 to 20 nucleotides

☹ No specific signature; each TF has different binding site

☹ Can be up to 1 million nucleotides upstream of gene regulated

☺ Often clustered with other binding sites, forming modules