Introduction

02-714 / 02-514

02-714 String Algorithms

Tentative Schedule:

Topic	Topics
1	Exact string matching (Z-algorithm, Knuth-Morris-Pratt, Boyer-Moore, Rabin-Karp)
2	Advanced inexact matching (edit distance, alignment in linear space, Four-Russians' speedup, approximation algorithms for MSA, whole genome alignment)
3	Parallel string search
4	Suffix trees and arrays and their applications; Ukkonen's suffix tree construction algorithm
5	Subword graphs and their construction
6	String compression algorithms
7	Compressed self-indices (data structures that support fast searching and complete reconstruction of the full text in sublinear space). Burrows-Wheeler transform, the FM-index.
8	Read mapping (matching huge collections of substrings to reference strings); tools for doing this such as Bowtie, BWA, TopHat
9	Compressive genomics (i.e. doing analyses directly on compressed data): searching compressed collections of sequences; compariative assembly from compressed read databases
10	Genome assembly; Shortest superstring problem, Celera assember, de Bruijn-graph-based assembly; mixed assembly.
11	Isoform / transcript assembly (e.g. Cufflinks, Trinity)
12	Gene and isoform expression quantification (RSEM, eXpress, Jellyfish, Cufflinks)

String Algorithms in Biology

1. Genome assembly

2. Gene discovery

3. Understanding the origin of swine flu

4. Gene Expression

Genome assembly

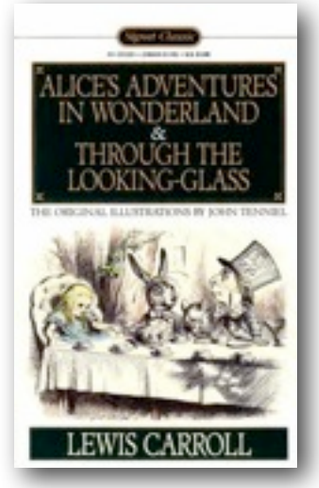
Genome of the Cow

a sequence of 2.86 billion letters

enough letters to fill a million pages of a typical book.

```
TATGGAGCCAGGTGCCTGGGGCAACAAGACTGTGGTCACTGAATTCATCCTTCTTGGTCTAACAGAGAACATAG
AACTGCAATCCATCCTTTTTGCCATCTTCCTCTTTGCCTATGTGATCACAGTCGGGGGCAACTTGAGTATCCTG
GCCGCCATCTTTGTGGAGCCCAAACTCCACACCCCCATGTACTACTTCCTGGGGAACCTTTCTCTGCTGGACAT
TGGGTGCATCACTGTCACCATTCCTCCCATGCTGGCCTGTCTCCTGACCCACCAATGCCGGGTTCCCTATGCAG
CCTGCATCTCACAGCTCTTCTTTTTCCACCTCCTGGCTGGAGTGGACTGTCACCTCCTGACAGCCATGGCCTAC
GACCGCTACCTGGCCATTTGCCAGCCCCTCACCTATAGCATCCGCATGAGCCGTGACGTCCAGGGAGCCCTGGT
GGCCGTCTGCTGCTCCATCTCCTTCATCAATGCTCTGACCCACACAGTGGCTGTGTCTGTGCTGGACTTCTGCG
GCCCTAACGTGGTCAACCACTTCTACTGTGACCTCCCGCCCCTTTTCCAGCTCTCCTGCTCCAGCATCCACCTC
AACGGGCAGCTACTTTTCGTGGGGGCCACCTTCATGGGGGTGGTCCCCATGGTCTTCATCTCGGTATCCTATGC
CCACGTGGCAGCCGCAGTCCTGCGGATCCGCTCGGCAGAGGGCAGGAAGAAAGCCTTCTCCACGTGTGGCTCCC
ACCTCACCGTGGTCTGCATCTTTTATGGAACCGGCTTCTTCAGCTACATGCGCCTGGGCTCCGTGTCCGCCTCA
GACAAGGACAAGGGCATTGGCATCCTCAACACTGTCATCAGCCCCATGCTGAACCCACTCATCTACAGCCTCCG
GAACCCTGATGTGCAGGGCGCCCTGAAGAGGTTGCTGACAGGGAAGCGGCCCCCGGAGTG ...
```

We can only read ~ 1000 characters at a time from a random place:




```
good-natured, she thought: still
                        when it saw Alice. It looked
          ought to be treated
                                 good-natured, she thought, still
                    Cat only
                a greet many
                                                    It looked good-
The Cat only grinned when it saw Alice.
     be treated with respect.
                                 still it had very long claws
          claws and a great many teeth, so she
                                                    so she felt that it ought
```

Fast algorithms are needed to piece the story together.

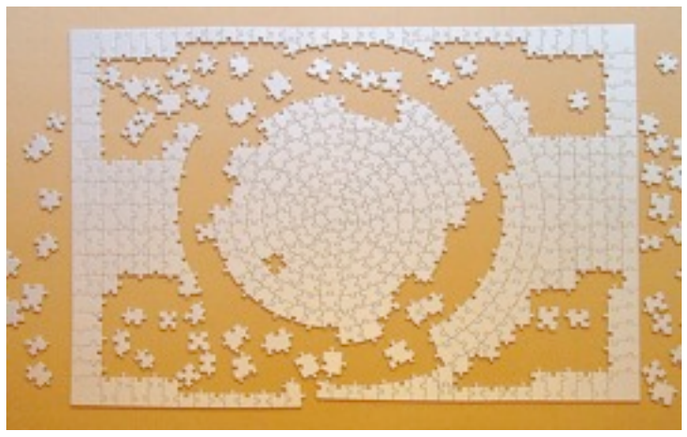
```
The Cat only grinned when it saw Alice.
    Cat only          when it saw Alice. It looked
                                         It looked good-

good-natured, she thought: still
good-natured, she thought, still
                           still it had very long claws

claws and a great many teeth, so she
          a greet many        so she felt that it ought

ought to be treated
          be treated with respect.
```

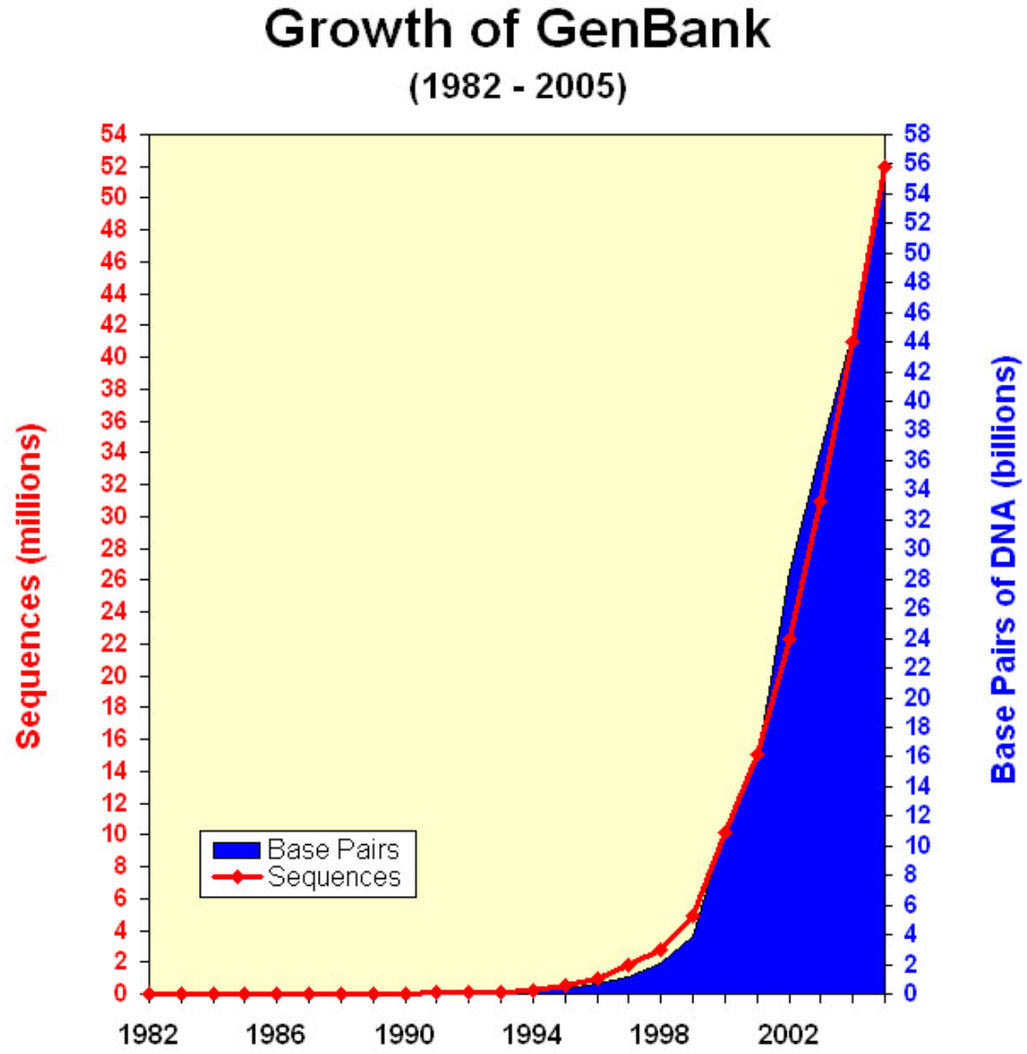
It's a jigsaw puzzle ...



...except with 35 million pieces

Recent Genomics (DNA)

- First genome sequenced in 1995 (the bacteria *H. influenzae* with a genome of 1,830,140 letters).
- 1st draft of human genome finished in 2001 (~ 3 billion letters)
- Now: Over 1100 bacterial genomes
- Hundreds of higher-order genomes done or in progress.
- Several complete human genomes finished.



Researchers at many institutions are putting together the genomes of many animals.

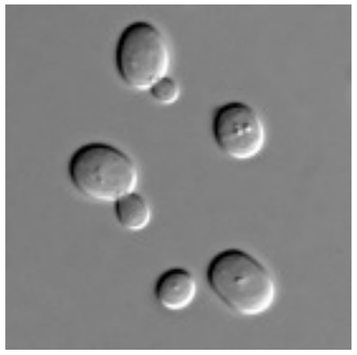
Help understand how to make animals and plants more hardy, resistant to disease, and understand their biology.

New technologies and larger genomes require new algorithms and faster computers.

Other Sequenced Genomes

Arabidopsis thailiana

Callithrix jacchus
(marmoset)



Saccharomyces cerevisiae
(baker's yeast)

Canis lupus familiaris
(dog)

Apis mellifera
(honey bee)

Drosophila melanogaster
(fruit fly)

Bos torus
(cow)

Equus caballus
(horse)

Sequenced Eukaryotic Chromosomes

and many more...

Gene discovery

```
   1 atactataaa tccacctctc attttattca cttcatacat gctattacac actctgtgcc
  61 atcatagtat gttttcatac atcctccctt ctttcacacc ctatgtatat cgtacattaa
 121 tggtgtaccc cccctccccc tatgtatatc gtgcattaat ggcgtgcccc atgcatataa
 181 gcatgtacat actgtgcttg gctttacatg aggatactca ttacaagaac ttatttcaag
 241 cgatagtcta tgagcatgta tttcacttag tccaagagct tgatcaccaa gcctcgagaa
 301 accagcaatc cttgcgagta cgtgtacctc ttctcgctcc gggcccataa tttgtggggg
 361 tttctatact gaaactatac ctggcatctg gttcttacct cagggccatg ttagcgtcaa
 421 ctcaatccta ctaacccttc aaatgggaca tctcgatgga ctaatgacta atcagcccat
 481 gatcacacat aactgtggtg tcatgcattt ggtatttttt aattttaggg ggggaacttg
 541 ctatgactca gctatgaccg taaaggtctc gtcgcagtca aatcagctgt agctgggctt
 601 attcatcttt cgaggctcct catggacacc cataaggtgc aattcagtca atggtcacag
 661 gacataacac tatagatcac ccggactggc gttacgtgta cgtacgtgta cgtacgtgta
 721 cgcacgtgta cgtacgtgta cgcacgtgta cgtacgtgta cgcacgtgta cgcacgtgta
 781 cgtacgtgta cgcacgtgta cgtacgtgta cgtacgtgta cgcacgtgta cgcacgtgta
 841 cgtacgtgta cgcacgtgta cgtacgtgta cgcacgtgta cgcacgtgta cgcacgtgta
 901 cgtacgtgta cgcacgtgta cgtacgtgta cgcacgtgta cgcacgtgta cgtacgtgta
 961 cgcacgtgta cgcacgtgta cgcacgtgta cgcacgtgta cgcacgtgta cgtacgtgta
1021 cgcgtacgta ttttagatac taagttagct tagacaaacc ccccttaccc cccgtaactt
1081 caagaagctt acatatactt atggatgtcc tgccaaaccc caaaaacaag actaaatata
1141 tgcgcaaaca tgaagtcact tacacctaaa cccatataat taagctaacc ccccagccaa
1201 tgttgcaaca actacggaca tgggactcta aattttaatt tatctataga tatttttctt
1261 ttactgtgtc tccccagcat tgatttttta attatcatta ttccacacca ccaatttcca
1321 ttgagctatt tcacatgagt tccaaatcaa ttatgttcat gtagcttaac gaataaagca
1381 aggtactgaa aatgcctaga tgggtcacgc taccccatag acataaaggt ttggtcctag
1441 ccttcctatt agccattaac aagattacac atgtaagtct ccacgctcca gtgaaaatgc
1501 cccttaagtc ctcttagacg acctaaagga gcgggtatca agcacacctt atggtagctc
1561 acaacgcctt gcttagccac acccccacgg gaaacagcag tgataaaaat taagctatga
1621 acgaaagttc gactaagcta tgttaatact agggttggta aatctcgtgc cagccaccgc
1681 ggtcatacga ttaactcgag ttaataggcc tacggcgtaa agcgtgtaaa agaaaaaatc
1741 tcctctacta aagttaaagt atgattaagc tgtaaaaagc taccattaat actaaaataa
1801 actacgaaag tgactttaaa atttctgatt acacgatagc tagggcccaa actgggatta
1861 gataccccac tatgcctagc tctaaacata gatattttac taaacaaaac tattcgccag
1921 agaactacta gcaacagctt aaaactcaaa ggacttggcg gtgctttata tccccctaga
1981 ggagcctgtt ctgtaatcga taaaccccga tagacctcac catcccttgc taattcagtt
2041 tatataccgc catcttcagc aaacccttaa aaggaaaaaa agtaagcata actaccctac
2101 ataaaaaagt taggtcaagg tgtaacctat gggctgggaa gaaatgggct acattttcta
2161 ttcaagaaca acttctacga aaacttttat gaaactaaaa gctaaaggcg gatttagtag
2221 taaattaaga atagagagct taattgaaca gggcaatgaa gcacgcacac accgcccgtc
2281 accctcctcg agtgatataa tttaattata acctatttaa actaagcaaa gcataagagg
2341 agacaagtcg taacaaggta agcatactgg aaagtgtgct tggatgagcc aaagtgtagc
2401 ttaaacaaag cgtctggctt acatccagaa gatttcatta atatatgact actttgaacc
2461 caaagctagc ccaagcaaca atgactagta aaaccattat gaaacattca aacaaaacat
2521 ttagtagcat gactagagta taggagatag aaatttttaa ctggagctat agagagagta
2581 ccgcaaggga atgatgaaag attacctaaa gtgataaaca gcaaagattg ccccttctac
2641 cttttgtata atgagttagc tagaaataac ttaacaaaga gaacttaagc taagtccccc
2701 gaaaccagac gagctaccta tgaacaatcc actgggatga actcatctat gttgcaaaat
2761 agtgagaaga tccataggta gaggtgaaag gcctaacgag cctggtgata gctggttgcc
2821 cagaatagaa ttttagttcg actttaaacc tgcctacaaa actaataatt ctaatgcaga
2881 tttaaaatat attctaaaaa ggtacagctt tttagagtta aggatacagc cttacttaga
2941 gagtaaatat ttatataagc catagtaggc ctagaggcag ccatcaatta agaaagcgtt
3001 aaagctcaac atctctatta acttaatacc aagaatattt aatcaactcc taatgtatta
3061 ctgggtcaat ctatttaaat atagaagtga taatgctaat atgagtaaca agaaatattt
3121 ctcccatgca taagcttata acagcaacgg ataaccactg atagttaaca acaacataga
3181 aataacctaa tgataaaaca cctattaaat caattgttag tccaacacag gcatgcaatc
3241 agggaaagat taaaagaagt gaaaggaact cggcaaatac aaaccccgcc tgtttaccaa
3301 aaacatcacc tccagcattt ccagtattgg aggcactgcc tgcccggtga catcagttaa
3361 acggccgcgg tattctgacc gtgcaaaggt agcataatca tttgttctct aaataaggac
3421 ttgtatgaat ggccacacga gggtttaact gtctcttact tccaatcagt gaaattgacc
3481 tccccgtgaa gaggcgggga taagacaata agacgagaag accctatgga gctttaatta
3541 actaattcaa aaagaaacta ctaacgaccc aacaggaata atatatctct tttatgaatt
3601 agcaatttag gttggggcga cctcggagga caaaatagcc tccgagtgat tataaatcta
3661 gacttaccag tcaaaatgct taatcactta ttgatccaaa aattcttttg atcaacggaa
3721 caagttaccc tagggataac agcgcaatcc tatccgagag tccatatcga caatagggtt
3781 tacgacctcg atgttggatc aggacatcct aatggtgcag cagctattaa aggttcgttt
3841 gttcaacgat taaagtccta cgtgatctga gttcagaccg gagcaatcca ggtcggtttc
3901 tatctattca aataatttct cccagtacga aaggacaaga gaaataaggc ctacttctct
3961 gaagcgcctt aagaccaata gatgaattta tctaaatcta gtaaatctaa ctccaatatt
4021 gcccaagaga cagggctttg ttagggtggc agagcccggt aattgtgcaa aacttaaact
4081 cttgtgtcca gaggttcaat tcctctccct agcatatgtt tataattaac atcttctcac
4141 taattgtacc tattcttctt gctgtagcct ttctgactct agtagaacga aaagtactag
```

Example Genomic Sequence

⇐ Giant Panda (*Ailuropoda melanoleuca*) mitochondrion sequence [Peng et al, Gene 397:76-83 (2007)]

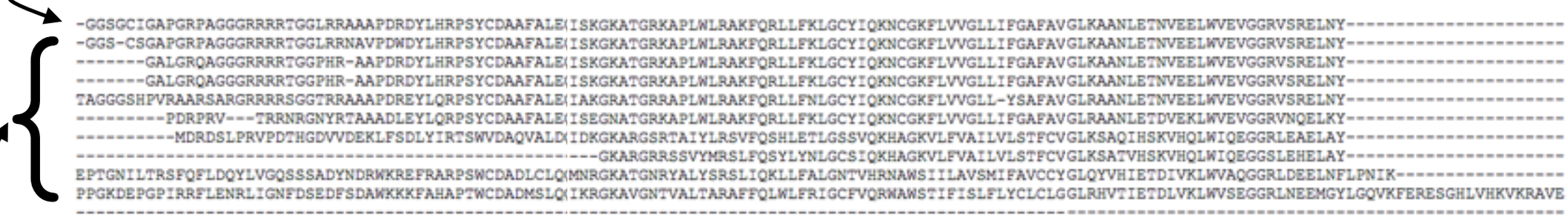
Obviously, computers are needed to understand what this means.

- Where are the genes encoded in this sequence?
- What causes each gene to be turned on or off?
- How does the genome produce observed traits?

Two ways to find genes...

1. Search for sequence of DNA similar to known gene:

Human version of gene



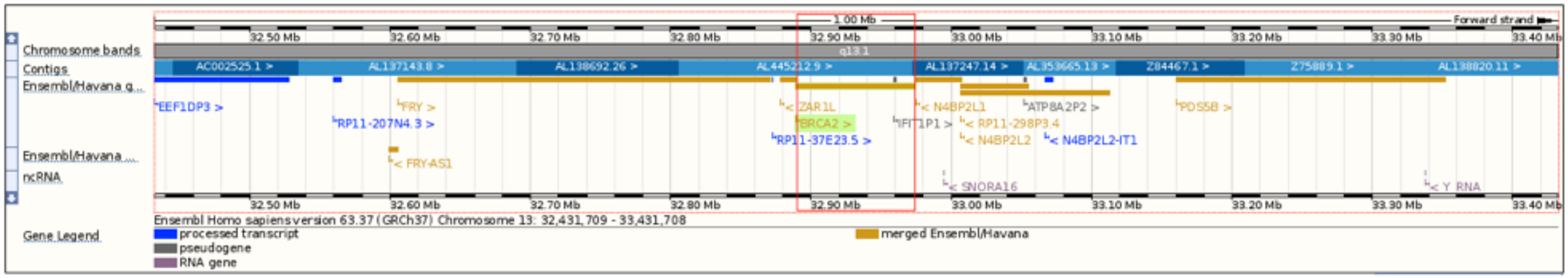
Similar sequences in other organisms

⇒ need to search for strings really fast

(allowing for errors)

2. Search the genome sequence for patterns of letters that "look like genes".

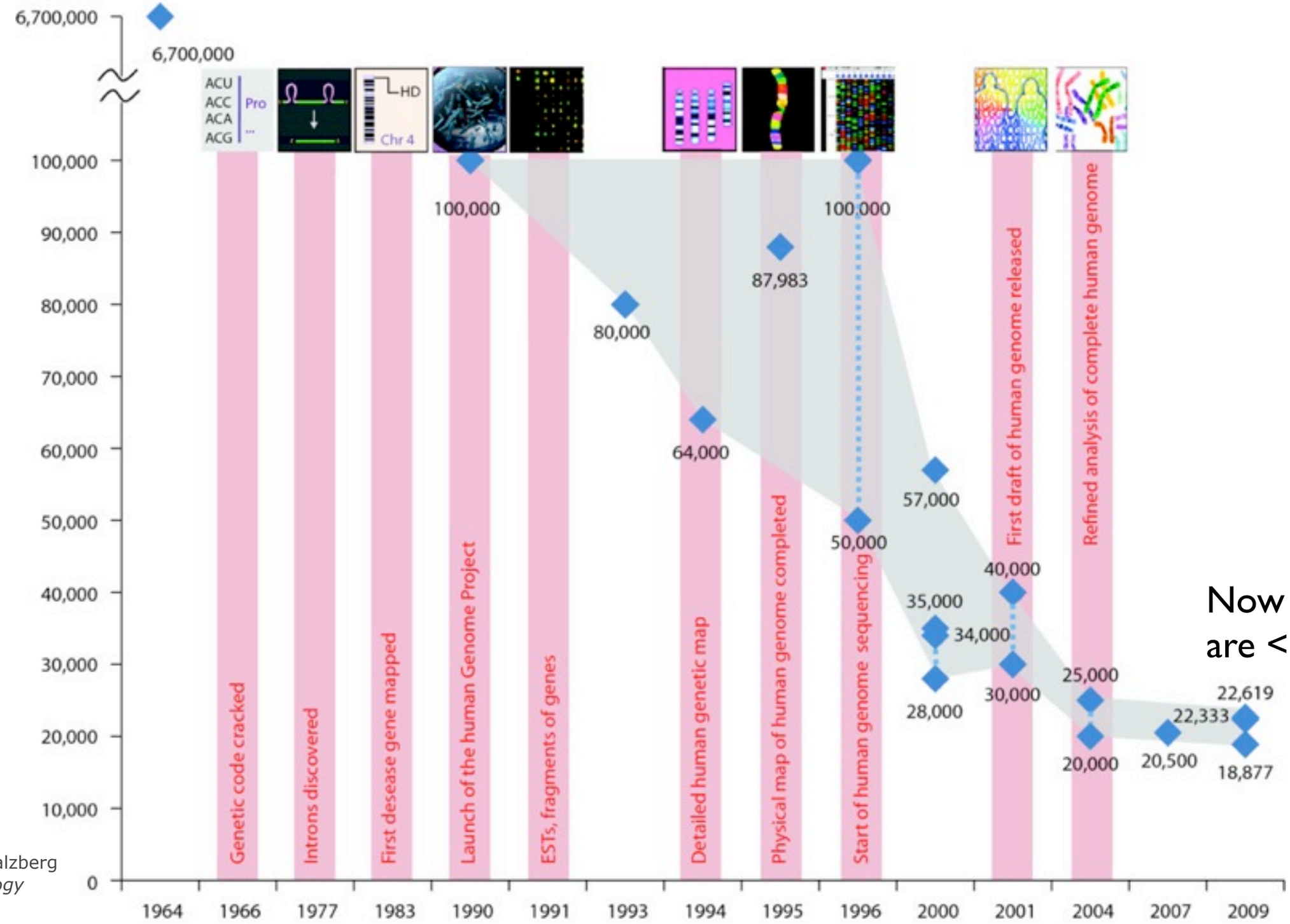
Some Human Genes



http://useast.ensembl.org/Homo_sapiens/Location/View?db=core;g=ENSG00000139618;r=13:32889611-32973805

Estimates for the # of Human Genes

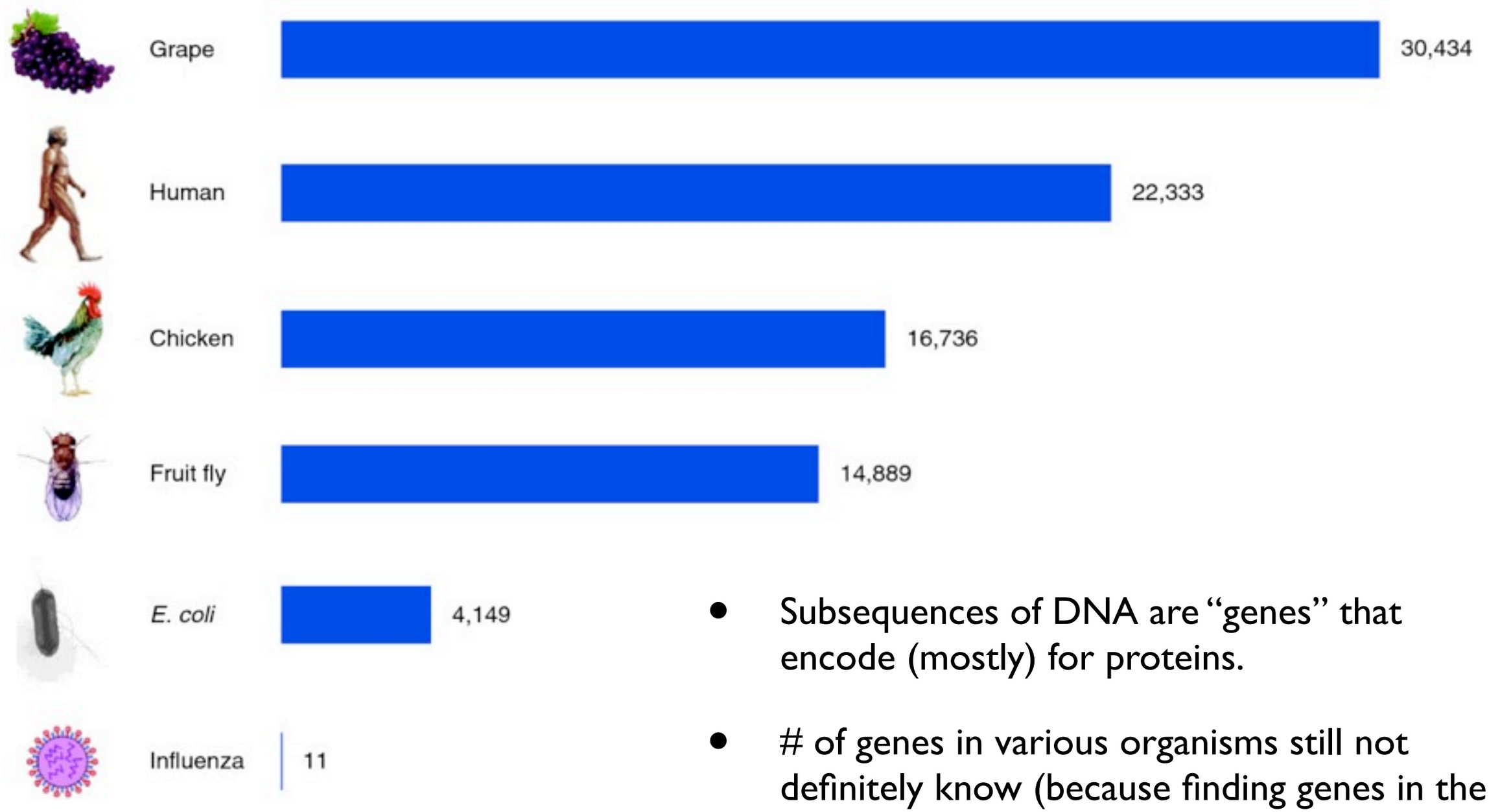
Before human genome sequence was available, many
(but not all) estimates for # of genes were high (> 80,000).



Now estimates are < 23,000.

Pertea and Salzberg
Genome Biology
2010 **11**:206

of Genes in Various Organisms

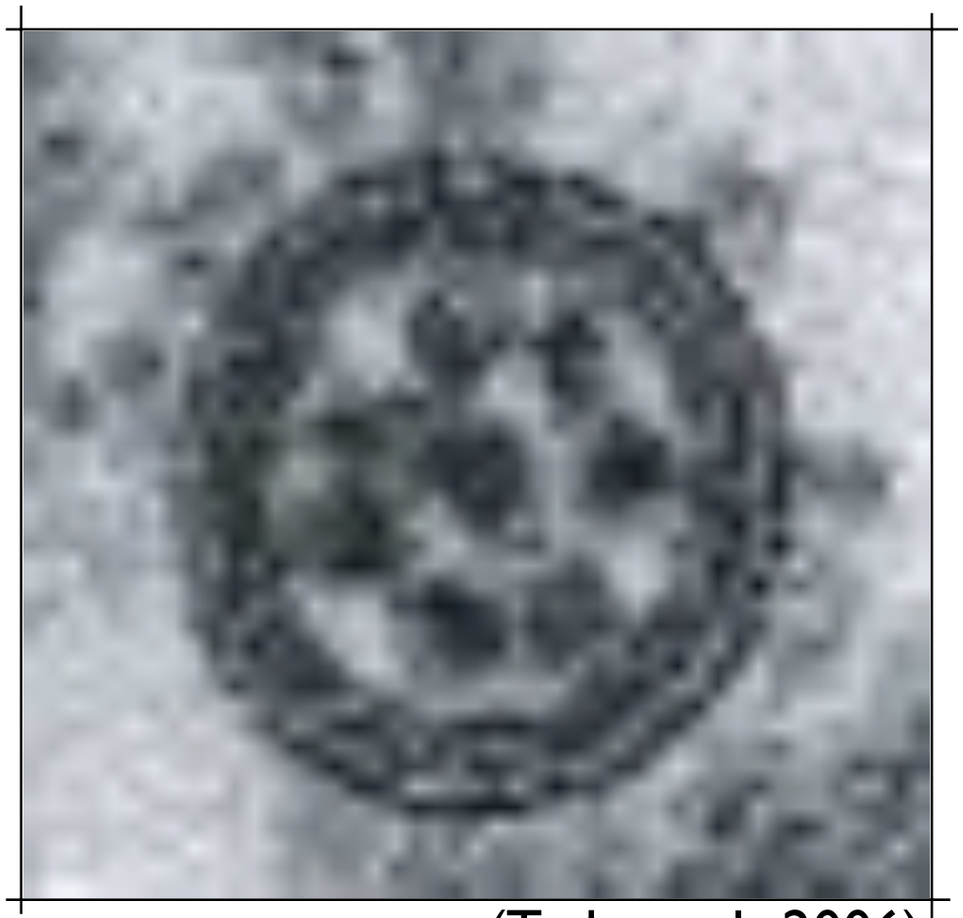


Pertea and Salzberg
Genome Biology 2010
11:206

- Subsequences of DNA are "genes" that encode (mostly) for proteins.
- # of genes in various organisms still not definitely know (because finding genes in the sequence is a hard problem that we will talk about).
- But there are reasonably good estimates.

Tracking the evolution of influenza

Influenza Virus



(Toda et al., 2006)

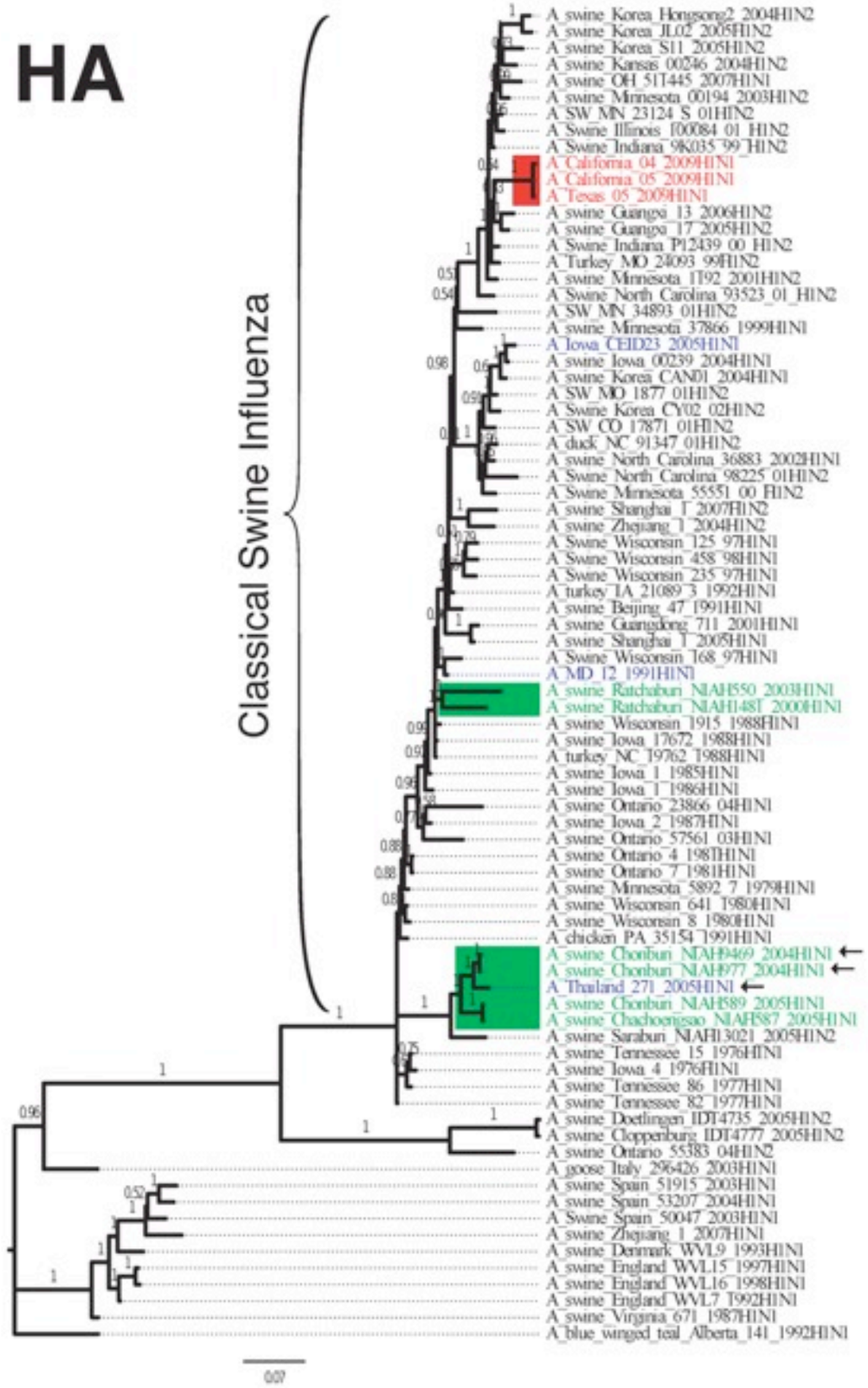
Rapidly evolving (it's genome is mutating):

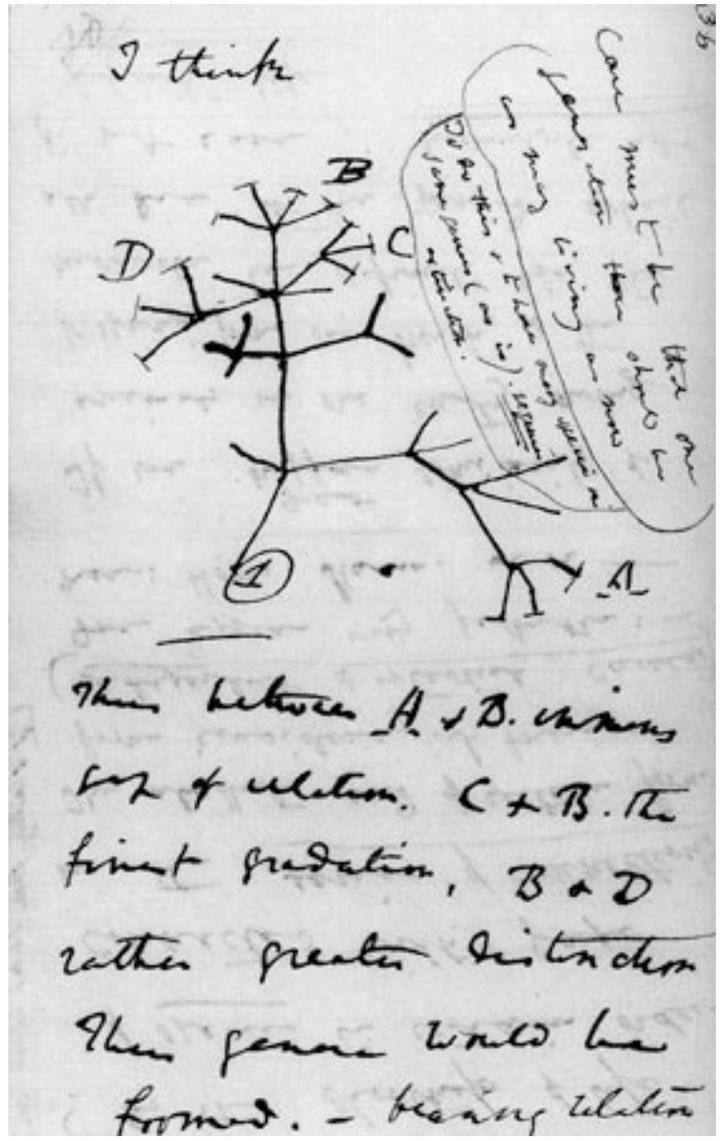
that's why you have to get a different flu shot every year

3 strains must be selected each year to include in the vaccine.

So, the evolution of the virus is must be predicted.

Evolutionary Trees



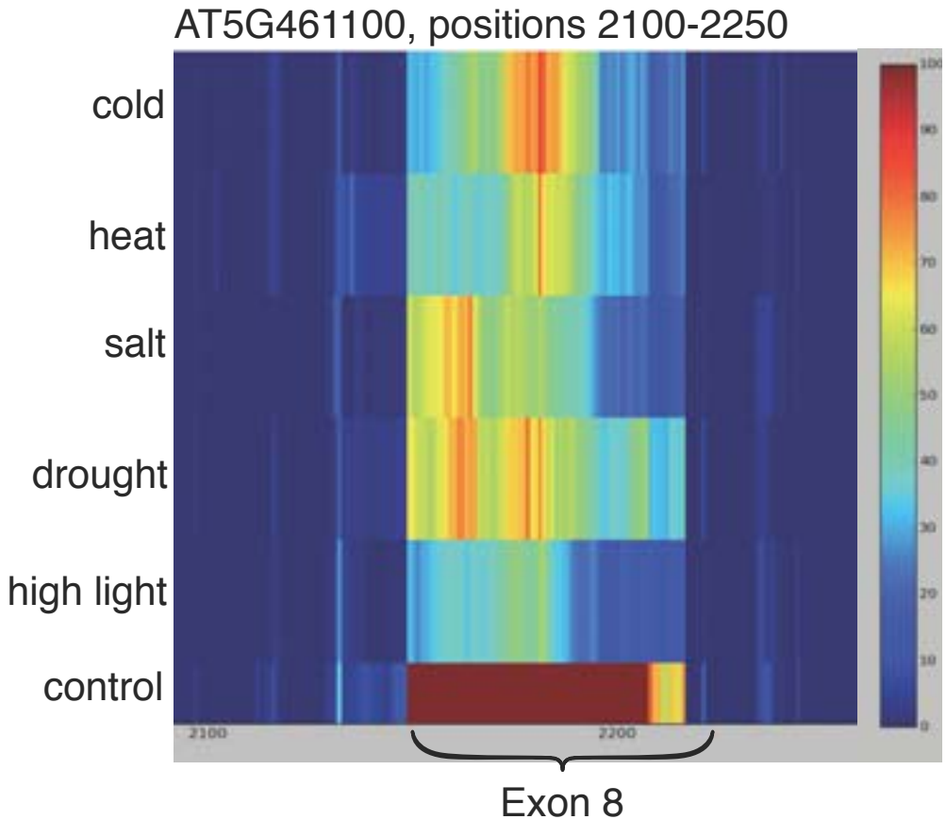


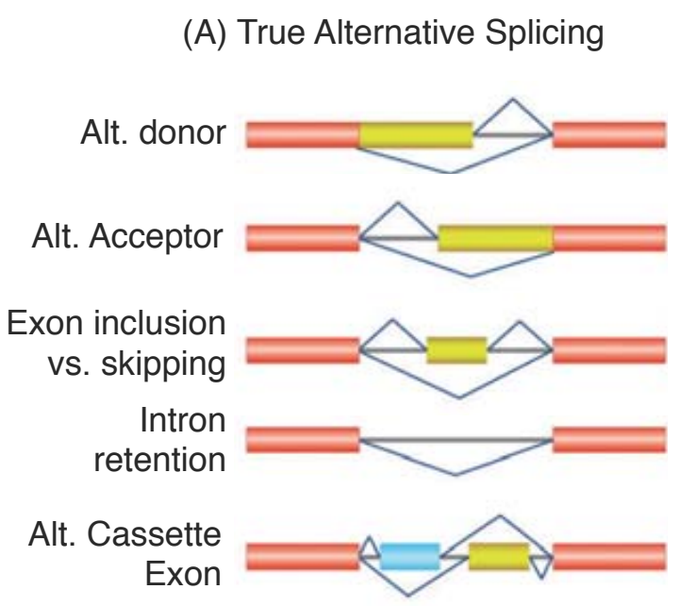
http://www.amnh.org/exhibitions/darwin/idea/treelg.php

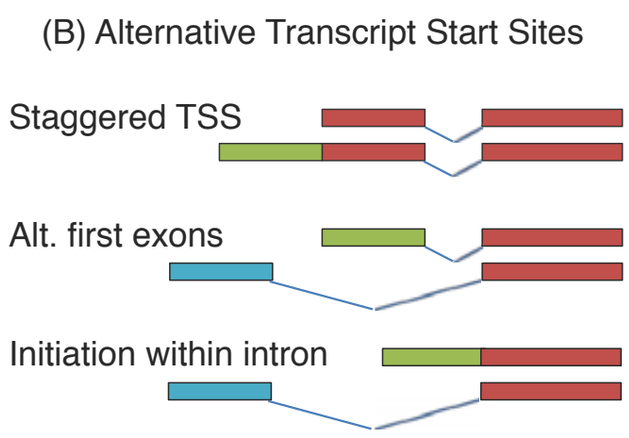
Gene Expression

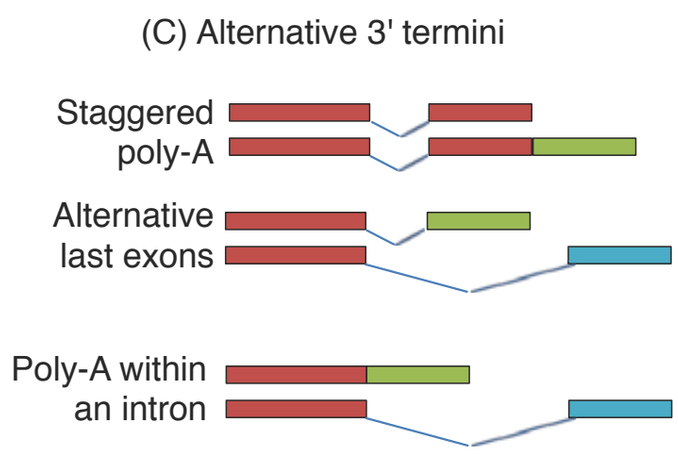
Alternative Splicing & Isoform Expression

- Expression of genes can be measured via RNA-seq (sequencing transcripts)
- Sequencing gives you short (35-300bp length reads)









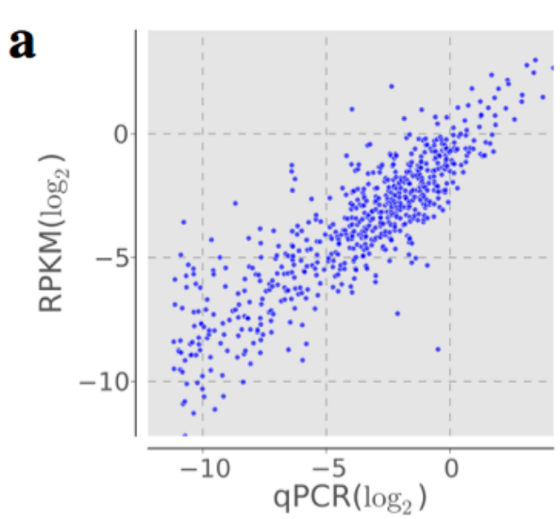
The Isoform Expression Estimation Problem

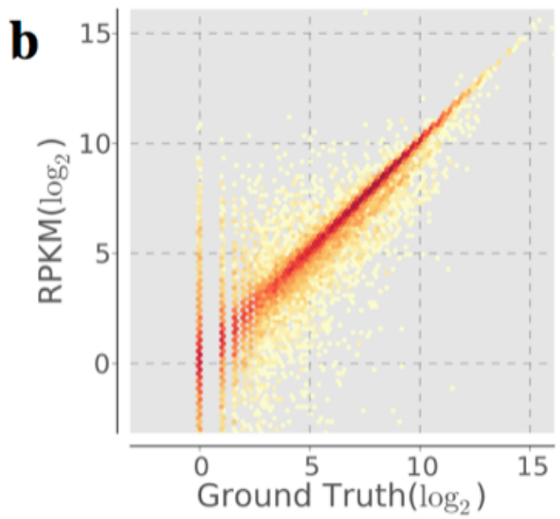
- RNA-Seq now standard for gene and isoform expression estimation.
- A main use for transcriptome sequencing is estimating gene and isoform abundance.
- This leads to the following computational problem:

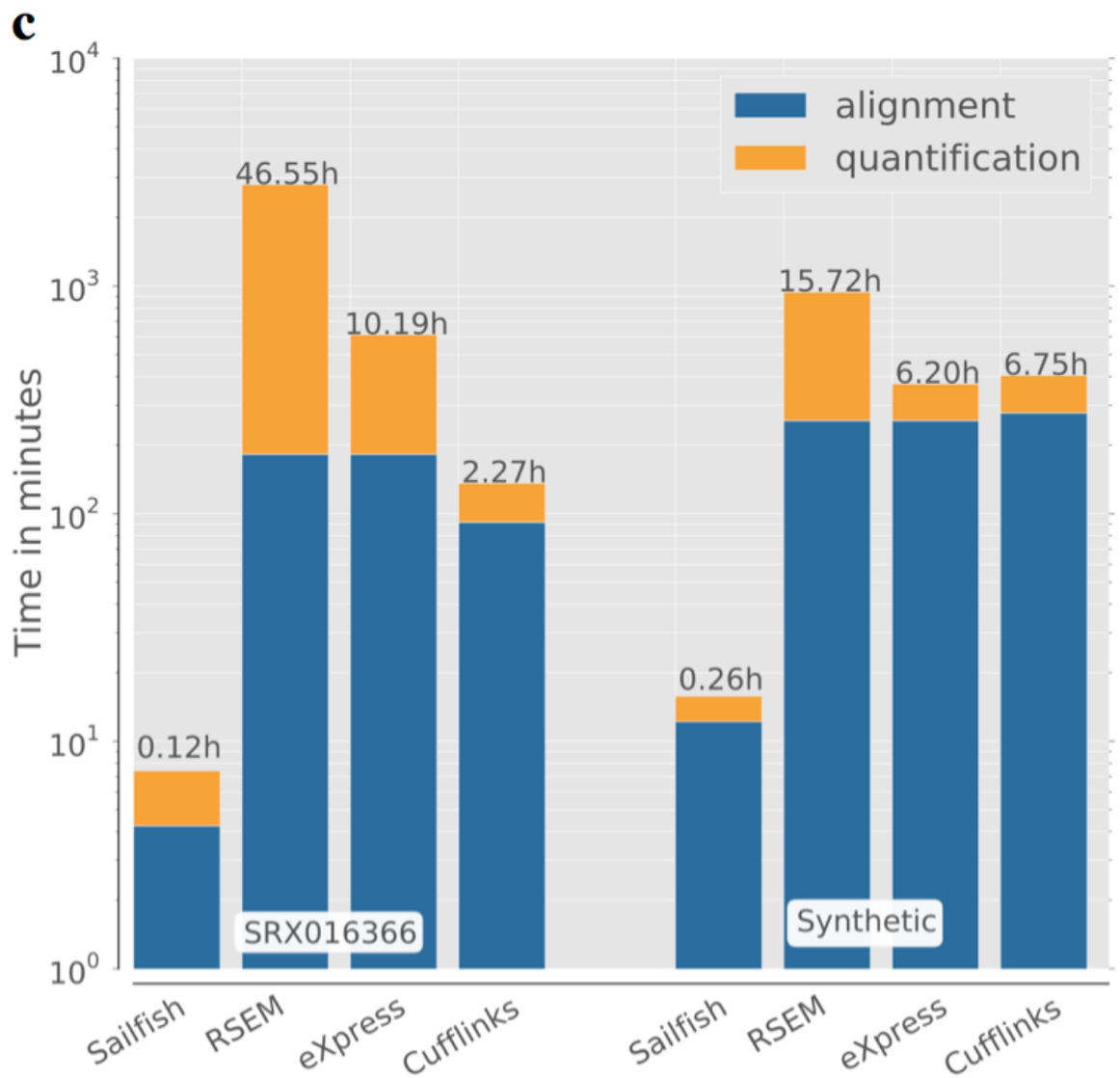
Given:
- Collection of RNA-Seq reads
- A set of known transcript sequences

Estimate:
- The relative abundance of each transcript

Performance on Universal Human Brain Tissue







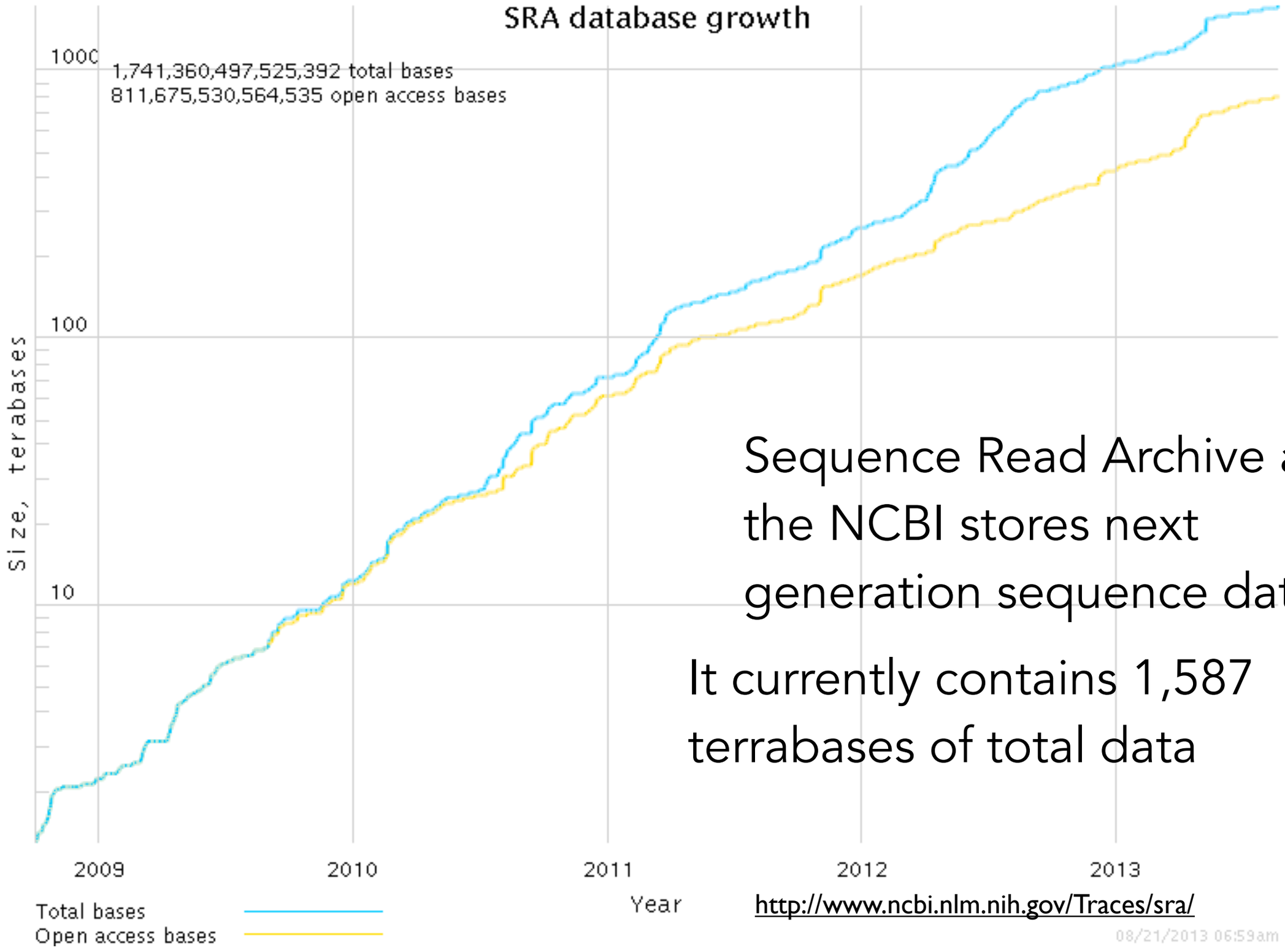
d

	Human Brain Tissue				Synthetic			
	Sailfish	RSEM	eXpress	Cufflinks	Sailfish	RSEM	eXpress	Cufflinks
Pearson	0.86	0.83	0.86	0.86	0.92	0.92	0.64	0.91
Spearman	0.85	0.81	0.86	0.86	0.94	0.93	0.66	0.93
RMSE	1.69	1.86	1.69	1.67	1.26	1.24	2.80	1.31
medPE	31.60	36.63	32.73	30.75	4.24	5.97	26.44	6.76

92,524,365 reads, each 35bp long

Bullard et al, 2010; MACQ Consortium, 2006

Big Genomic Data



Sequence Read Archive at the NCBI stores next generation sequence data

It currently contains 1,587 terrabases of total data

http://www.ncbi.nlm.nih.gov/Traces/sra/