Open questions in bioinformatics and computational biology from an evolutionary and molecular biology perspective

Jason Stajich

Assistant Professor of Bioinformatics & Bioinformaticist
Plant Pathology and Microbiology
University of California, Riverside

Outline

- Where and what is the data?
    - Big data is here, Bigger data is coming.
- Current approaches
- What are the problems?
    - Efficient searching in large data sets
    - Ease of interfacing with data to support genomics research - software, databases, and UI development
    - Finding signal in the datasets - statistical and computational methods
    - Managing the data - TBs of data from the sequencing.
- Other more biologically focused research areas.

Sequencing has been a driving force

- DNA sequencing and analysis has been the bread and butter of bioinformatics and computational biology for genomics and evolutionary research

- Human Genome project - USD $3B (1991 dollars)

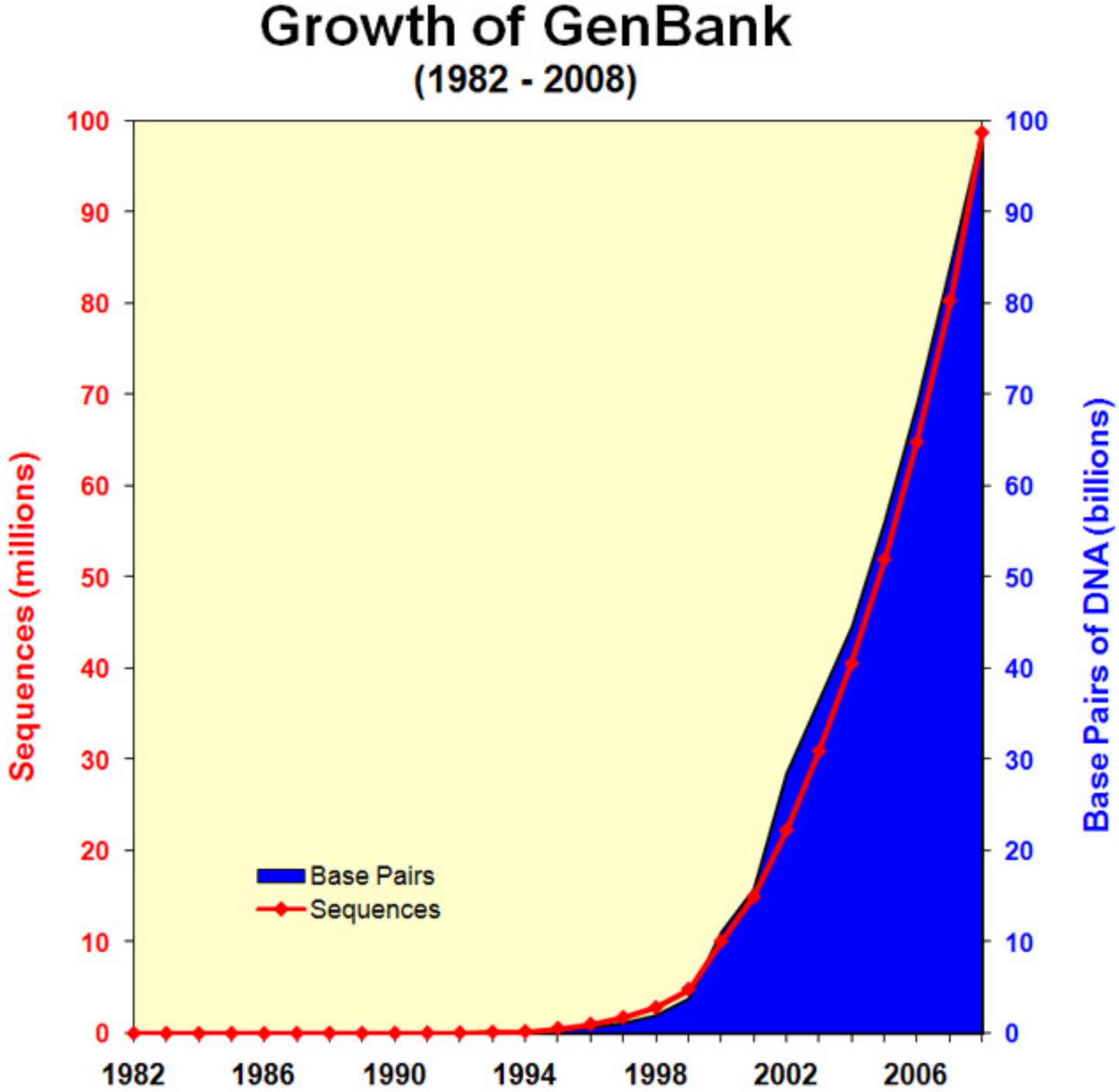
    - Typically sequencing costs have been converging towards $0.01 a nucleotide base - inexpensive but still limiting for large-scale questions

    - New sequencing techniques have dropped the costs by orders of magnitude

- Techniques for searching data - There has always been speed vs sensitivity tradeoffs

    - Initially developed with heuristics tuned towards the 10^3 - 10^6 sequence sized databases

- Previously software, computational resources required specialized centers - less the case now



16 B bases 2007-2008

Human genome = 3B bases

http://www.ncbi.nlm.nih.gov/Genbank/genbankstats.html

Vendor:	Roche			Illumina			ABI		
Technology:	454			Solexa GA			SOLiD		
Platform:	GS20	FLX	Ti	I	II	IIx	1	2	3
Reads: (M)	0.5	0.5	1.25	28	100	150	40	115	320
Fragment									
Read length:	100	200	400	35	50	100	25	35	50
Run time: (d)	0.25	0.3	0.4	3	3	5	6	5	8
Yield: (Gb)	0.05	0.1	0.5	1	5	15	1	4	16
Rate: (Gb/d)	0.2	0.33	1.25	0.33	1.67	3	0.34	1.6	2
Images: (TB)	0.01	0.01	0.03	0.5	1.1	2.8	1.8	2.5	1.9
PA Disk: (GB)	3	3	15	175	300	300	300	750	1200
PA CPU: (hr)	10	140	220	100	70	NA	NA	NA	NA
SRA: (GB)	0.5	1	4	30	50	2.5	100	140	600
Paired-end									
Read length:		200	400	2×35	2×50	2×100	2×25	2×35	2×50
Insert: (kb)		3.5	3.5	0.2	0.2	0.2	3	3	3
Run time: (d)		0.3	0.4	6	10	10	12	10	16
Yield: (Gb)		0.1	0.5	2	9	30	2	8	32
Rate: (Gb/d)		0.33	1.25	0.33	1.67	3	0.34	1.6	2
Images: (TB)		0.01	0.03	1	2.2	5.6	3.6	5	3.8
PA Disk: (GB)		3	15	350	500	500	600	1500	2400
PA CPU: (hr)		140	220	160	120	NA	NA	NA	NA
SRA: (GB)		1	4	60	100	3.5	200	280	1200

15 Gb in 5 days

1-3 Gb / day / machine!
for CURRENT technologies

~$8000 for human genome number of bases

http://www.politigenomics.com/next-generation-sequencing-informatics

Sequencing
Sequencing Strategy
Sequencing Plans
» Statistics
Why Sequence Them?
MyJGI: Information for Collaborators
Protocols

Statistics

Sequencing Progress, Updated Hourly

Date(s)	Total Q20* Bases	Total Lanes**	% Passed†	Ave. Read Length‡
11/11/2009: ABI3730	6.902 Million	10,848	92.91%	683.2
Current month (11/2009)	.186 Billion	282,240	93.89%	702.3
Last month (10/2009)	.549 Billion	944,544	89.71%	646.8
FY to Date (10/1/2009-11/11/2009)	.736 Billion	1,226,784	90.67%	660
Total (3/1999-11/11/2009)	197.868 Billion	313,694,312	92.47%	681.1

*Q20 indicates good confidence in the assignment of a base.
**The number of lanes is the number of samples loaded into a sequencer.
†The % passed is the percentage of lanes with more than 50 bases that meet the Q20 criteria.
‡Read length is the total number of Q20 bases in a read (i.e., in a lane with more than 50 Q20 bases.)

FY 2009 Overall Sequencing Progress, Updated Quarterly

Quarter	Total Q20* Bases (Billions)			Q20* Bases (Billions) by Platform			Operating Hours**		
	Goal	Actual Total	Actual % of Goal	Sanger	454	Illumina	Goal	Actual Total	Actual % Goal
Q1 2009	39.9	124.21	311	6.02	23.01	95.18	2100	2088	99.4
Q2 2009	60.1	196.829	328	5.849	38.48	152.5	2100	2146	102
Q3 2009	71.2	236.566	332	5.251	63.127	168.188	2100	2184	104
Q4 2009	81.8	446.278	546	3.452	45.96	396.866	2100	2208	105
FY 2009 Total	253	1003.887	397	20.58	170.58	812.73	8400	8626	103

*Q20 indicates good confidence in the assignment of a base.
**Number of hours a week that sequencing machines are producing data.

FY 2008 Summary

Quarter	Total Q20* Bases (Billions)			Q20* Bases (Billions) by Platform			Operating Hours**		
	Goal	Actual Total	Actual % of Goal	Sanger	454	Illumina	Goal	Actual Total	Actual % Goal
Q1 2008	9.8	7.2	73.40%	4	3.2	-	2100	1512	72%
Q2 2008	10	9.2	91.19%	2.3	6.9	-	2100	1800	85.7%
Q3 2008	11	41.17	374.27%	5.99	8.28	26.9	2100	2184	104%
Q4 2008	12	67.94	566.17%	8.04	22.8	37.1	2100	2208	105%
FY 2008 Total	42.8	125.51	293%	20.3	41.17	64	8400	7704	92%

*Q20 indicates good confidence in the assignment of a base.

Sequencing
- Sequencing Strategy
- Sequencing Plans
- » Statistics
- Why Sequence Them?
- MyJGI: Information for Collaborators
- Protocols

Statistics

Sequencing Progress, Updated Hourly

Date(s)	Total Q20* Bases	Total Lanes**	% Passed†	Ave. Read Length‡
11/11/2009: ABI3730	6.902 Million	10,848	92.91%	683.2
Current month (11/2009)	.186 Billion	282,240	93.89%	702.3
Last month (10/2009)	.549 Billion	944,544	89.71%	646.8
FY to Date (10/1/2009-11/11/2009)	.736 Billion	1,226,784	90.67%	660
Total (3/1999-11/11/2009)	197.868 Billion	313,694,312	92.47%	681.1

*Q20 indicates good confidence in the assignment of a base.
**The number of lanes is the number of samples loaded into a sequencer.
†The % passed is the percentage of lanes with more than 50 bases that meet the Q20 criteria.
‡Read length is the total number of Q20 bases in a read (i.e., in a lane with more than 50 Q20 bases.)

FY 2009 Overall Sequencing Progress, Updated Quarterly

Quarter	Total Q20* Bases (Billions)			Q20* Bases (Billions) by Platform			Operating Hours**		
	Goal	Actual Total	Actual % of Goal	Sanger	454	Illumina	Goal	Actual Total	Actual % Goal
Q1 2009	39.9	124.21	311	6.02	23.01	95.18	2100	2088	99.4
Q2 2009	60.1	196.829	328	5.849	38.48	152.5	2100	2146	102
Q3 2009	71.2	236.566	332	5.251	63.127	168.188	2100	2184	104
Q4 2009	81.8	446.278	546	3.452	45.96	396.866	2100	2208	105
FY 2009 Total	253	1003.887	397	20.58	170.58	812.73	8400	8626	103

*Q20 indicates good confidence in the assignment of a base.
**Number of hours a week that sequencing machines are producing data.

396 B

FY 2008 Summary

Quarter	Total Q20* Bases (Billions)			Q20* Bases (Billions) by Platform			Operating Hours**		
	Goal	Actual Total	Actual % of Goal	Sanger	454	Illumina	Goal	Actual Total	Actual % Goal
Q1 2008	9.8	7.2	73.40%	4	3.2	-	2100	1512	72%
Q2 2008	10	9.2	91.19%	2.3	6.9	-	2100	1800	85.7%
Q3 2008	11	41.17	374.27%	5.99	8.28	26.9	2100	2184	104%
Q4 2008	12	67.94	566.17%	8.04	22.8	37.1	2100	2208	105%
FY 2008 Total	42.8	125.51	293%	20.3	41.17	64	8400	7704	92%

*Q20 indicates good confidence in the assignment of a base.

Broad Institute Purchases Another 30 Illumina GAIIx to Bring Total to 89

November 10, 2009

By Julia Karow

This article was originally published Nov. 5.

Illumina said today that the Broad Institute has purchased 30 additional Genome Analyzer IIx systems, bringing its total to 89.

The new instruments will be used in whole-genome, exome, and transcriptome sequencing studies.

"We are happy with the technology's accuracy, ease of use, and scalability, as well as continued system improvements that have recently enabled us to generate multiple runs with more than 50 gigabases of high-quality sequence data," said Rob Nicol, director of sequencing operations at the Broad, in a statement issued by Illumina.

This is the second time this year that the Broad Institute has increased the number of Genome Analyzers at its genome center — in April, it said it was installing 22 additional instruments (see *In Sequence* 4/21/2009).

As of this spring, the institute also had 10 Roche/454 Genome Sequencers FLX, eight Applied Biosystems SOLiD systems, one Helicos Genetic Analysis System, and one Polonator installed.

In addition, the institute is currently testing Complete Genomics' human genome sequencing service as an early-access customer (see *In Sequence* 2/6/2009).

"...bringing its total to 89 [GA II machines]"

Related Stories

Korea's Genomic Medicine Institute Adds Seven Illumina Genome Analyzers
July 22, 2009 / GenomeWeb Daily News

Life Tech's IP-infringement Suit Against Illumina to Go to Trial in July 2011
November 12, 2009 / In Sequence

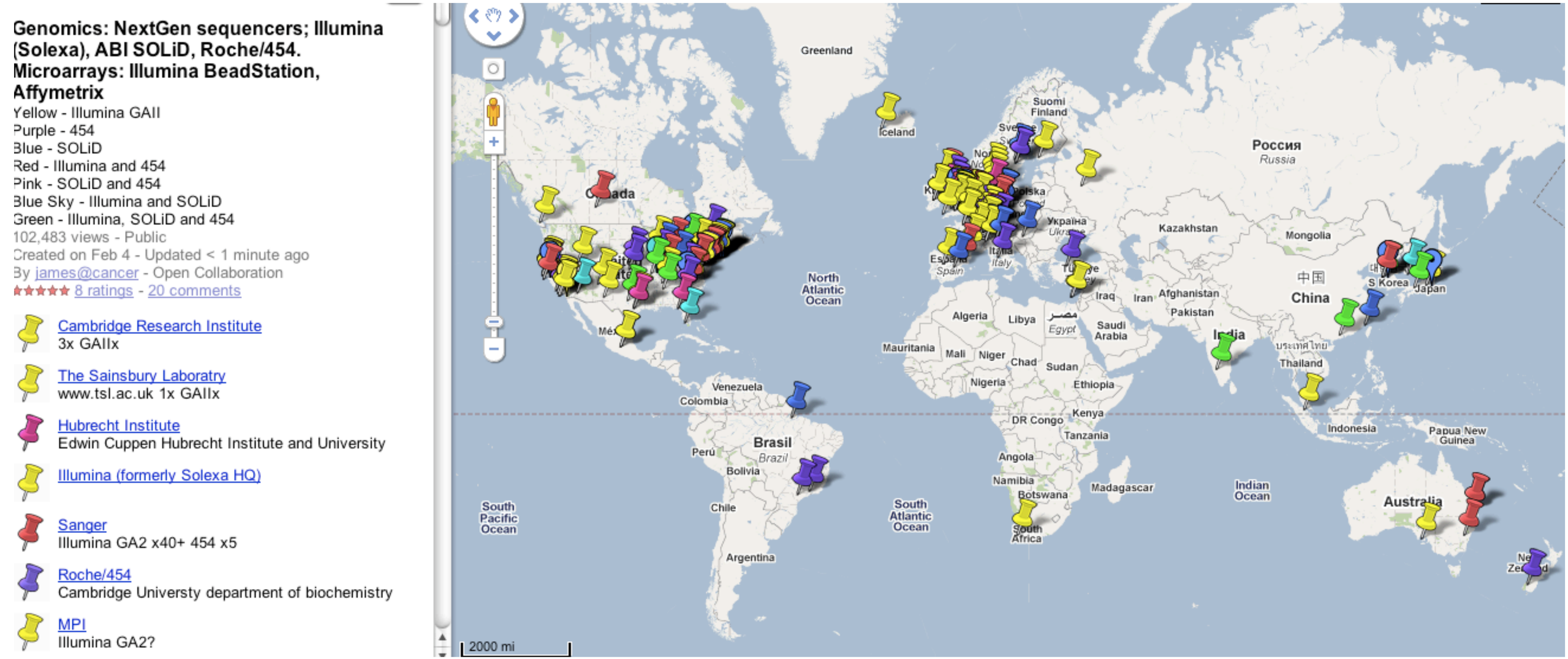
Complete Genomics Details Low-Cost Sequencing Tech in Paper; Collaborators Encouraged by Results
November 10, 2009 / In Sequence

Illumina Sells 30 More Genome Analyzers to Broad Institute
November 5, 2009 / GenomeWeb Daily News

Illumina Blames 'Disappointing' Q3 on GWAS Decline, Order Delays, Reagent Kit QC Issues
November 3, 2009 / BioArray News

director of scientific affairs, flow cytometry, and anatomic pathology at **US Labs** and scientific director of the division of immunology at **Ameripath Specialty Laboratories**, and he helped pioneer **Luminex**'s xMAP technology, said CombiMatrix.

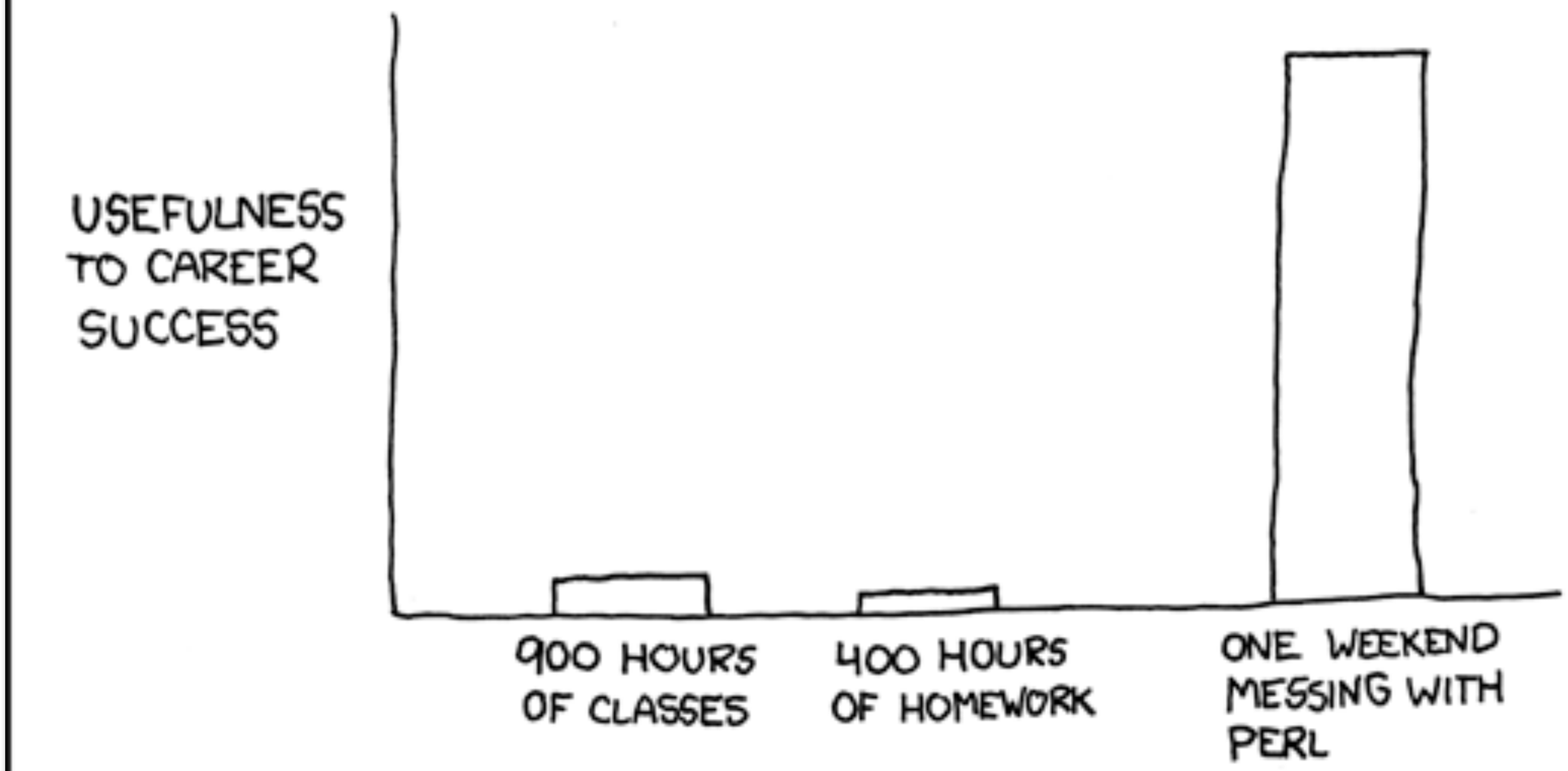
Cheaper sequencing is leading to democratized genome science. Everyone* can do it



*With $USD 500K

http://bit.ly/5zNBk

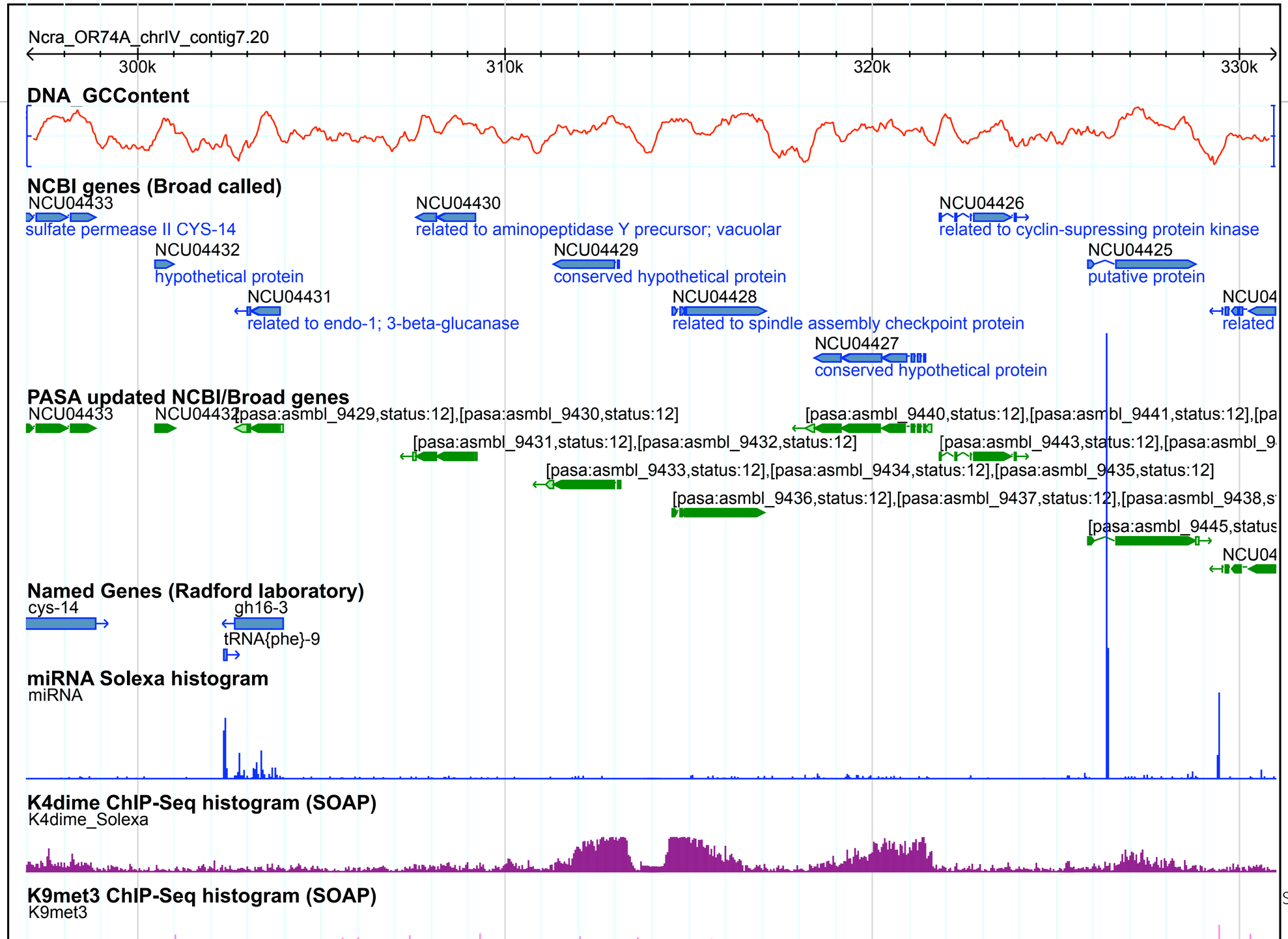
11TH-GRADE ACTIVITIES:



Tools for Bioinformatics and Comparative Genomics

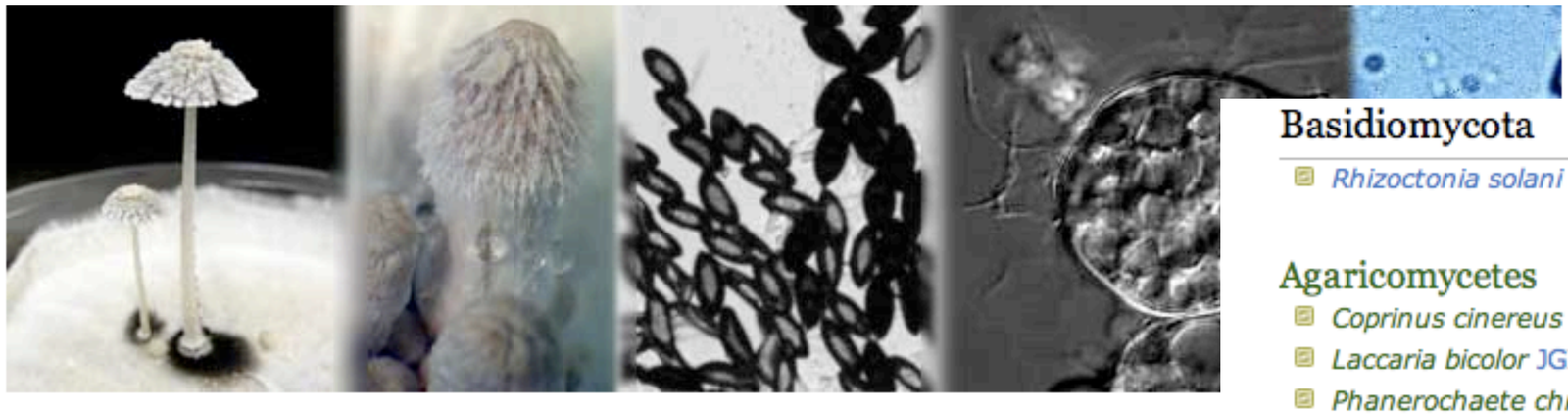
- Interfacing with the data (flat text and binary files, databases, web resources)
- BioPerl is a toolkit for bioinformatics data processing
    - Parsers for common file formats
    - Visualization of some kinds of genomic data - basis for genome browser
    - Genome Browsers to see genomic context information, important for visualizing high density data like 2nd-generation sequencing (RNA-Seq, ChIP-Seq)
- Community web resources for information sharing
    - Social networking challenges to get scientists to share information with attribution

Genome Browser data integration - Gbrowse



Gbrowse - Stein et al 2002

Stajich et al, unpublished
Smith, Freitag, et al unpublished



Links and references for the currently available fungal genome sequences or proposed (and austensibly funded) fungal genomes.

Contents [hide]

Basidiomycota [edit]

- *Rhizoctonia solani* JCVI rsolani.org

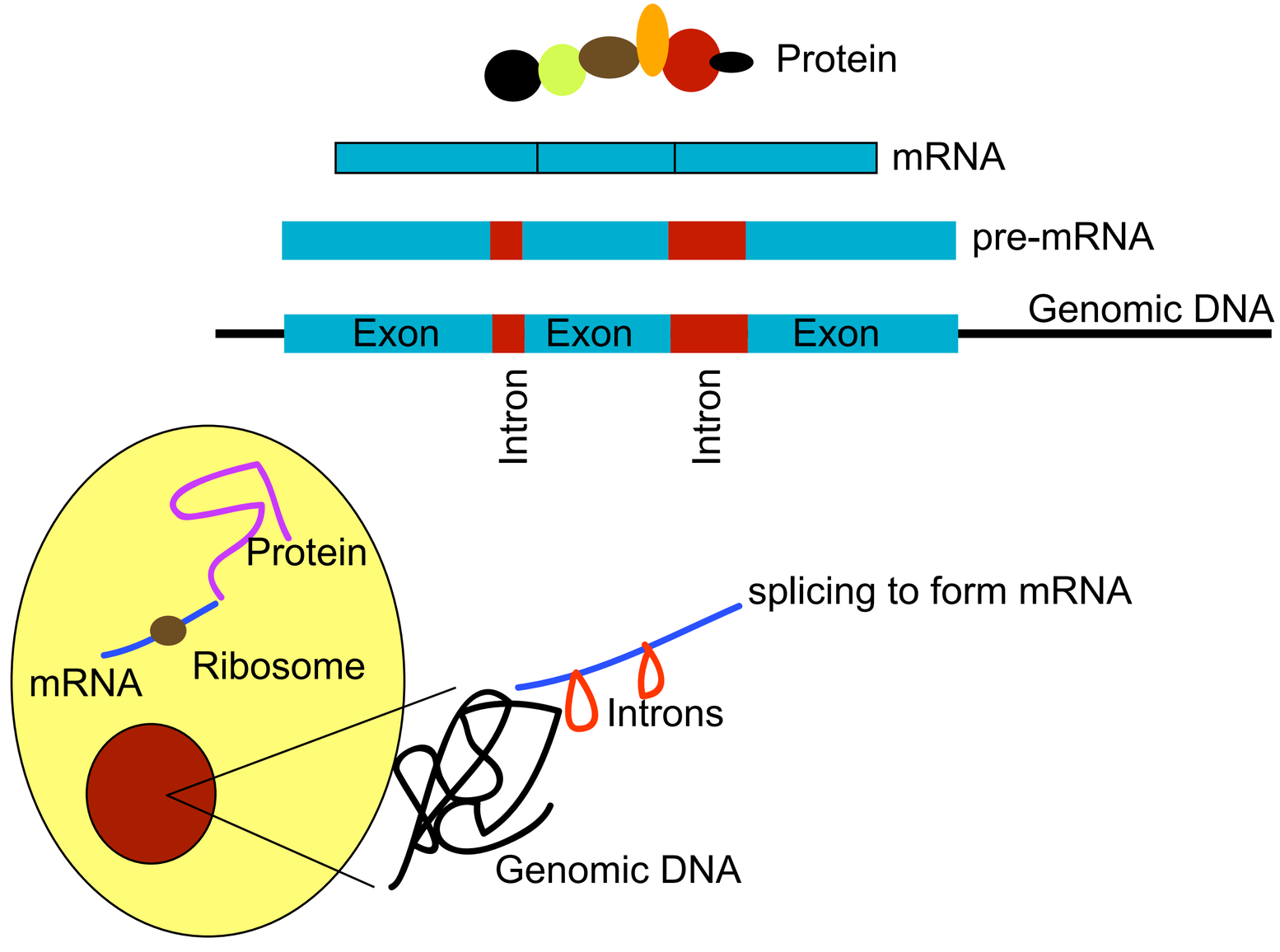
Agaricomycetes [edit]

- *Coprinus cinereus* (*Coprinopsis cinerea*) - FGI, Duke, SEMO
- *Laccaria bicolor* JGI [7]
- *Phanerochaete chrysosporium* (white rot) - JGI [8, 9]
- *Phanerochaete carnosa* (white rot) - JGI
- *Amanita bisporigera* - Michigan State/Hallen-Walton Labs 454 sequencing
- *Heterobasidion annosum* - JGI
- *Pleurotus ostreatus* - Announcement JGI (8X assembly)
- *Postia placenta* (brown rot) - JGI [10]. Note a heterokaryon was sequenced so the assembly and annotation has 2 copies of many loci.
- *Paxillus involutus* - in progress, JGI and through MycorWeb)
- *Agaricus bisporus* (button mushroom) - JGI Warwick
- *Moniliophthora perniciosa* (causes cacao disease) (*Crinipellis perniciosa*) Unicamp Witches' Broom project page. 1.9X coverage survey sequence [11]
- *Ganoderma lucidum* (polypore) National Yang-Ming Bioinformatics Research Center
- *Hebeloma cylindrosporum* (ESTs) MycorWeb
- *Pisolithus microcarpus* (ESTs) MycorWeb
- *Schizophyllum commune* JGI (8X sequencing completed)
- *Serpula lacrymans* JGI
- *Ceriporiopsis subvermispora* JGI
- *Rhizopogon salebrosus* JGI

Coprinopsis cinerea

fungalgenomes.org/genomes

Central dogma of eukaryotic biology



Information flow in the cell - Central Dogma

- DNA (4 bases, {A,C,G,T}) transcribed into

- RNA (4 bases, {A,C,G,U}) translated into

- Protein (20 amino acid residues, {A,C,D,E,F,G,H,I,K,L,M,N,P,Q,R,S,T,V,W,Y}) by triplets (codons) of RNAs
  UCA -> Serine (S)
  AUG -> Methionine (M)

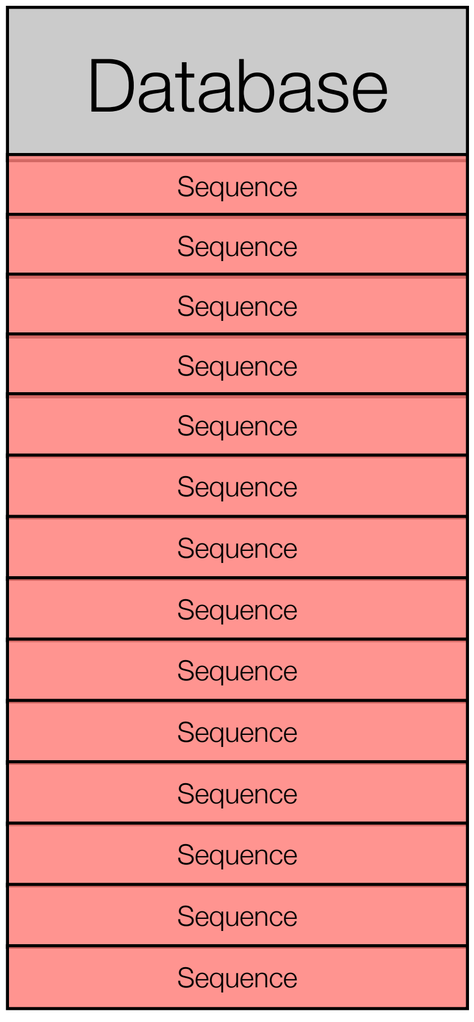
- 3 stop codons (UGA, UAA,UAG) in most species

- As always in Biology, there are exceptions!
  Some species use different stop codons. The codon table (codon -> AA) is not the same for all species, the mitochondria has different codon table.

The Sequence DB Search problem

Query

Pairwise alignment For each sequence



Generates Scores

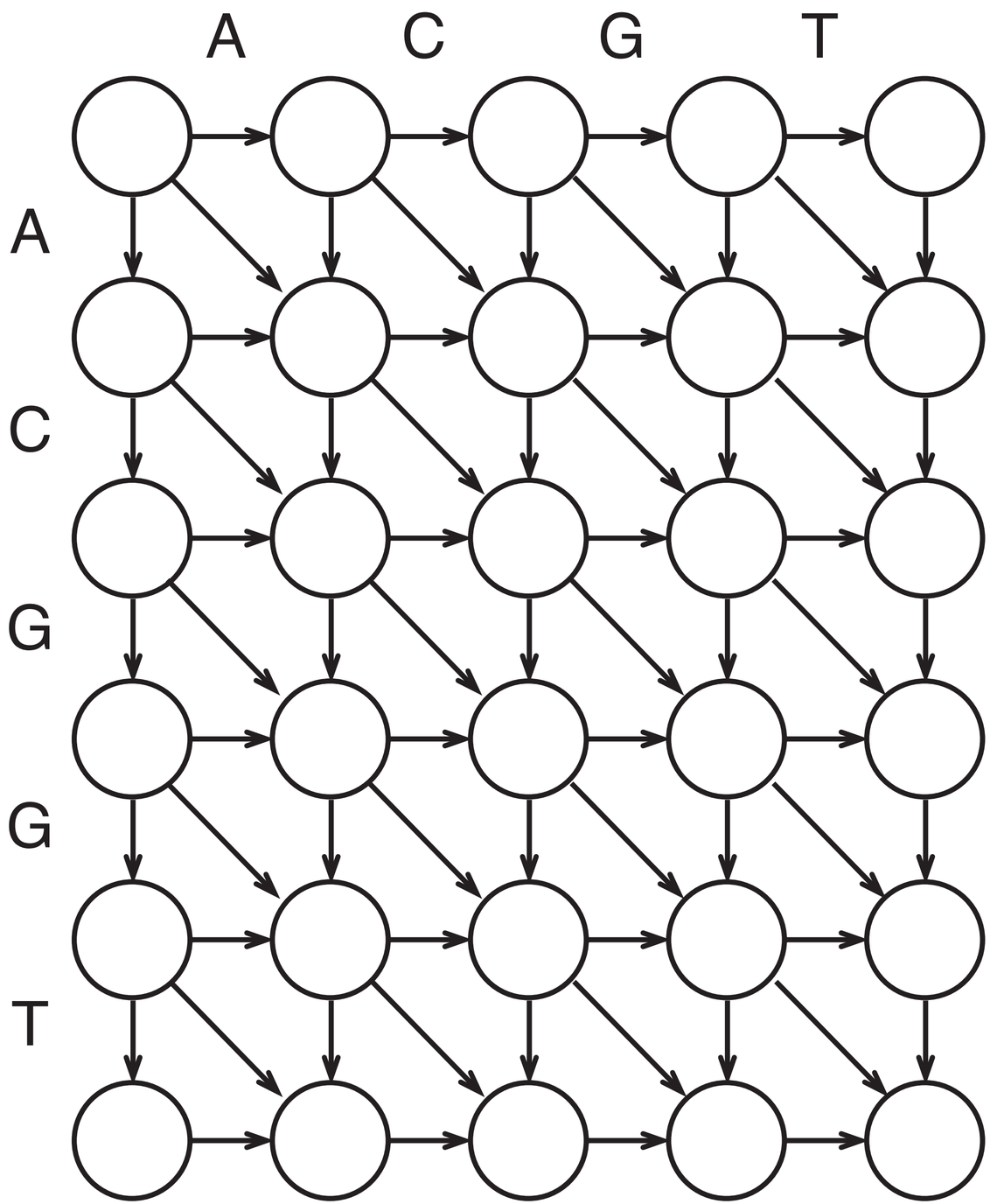
```
      opt      E()
< 20   17      0:=           one = represents 22 library sequences
  22    0      0:
  24    0      0:
  26    2      0:=
  28    7      3:*
  30    7     18:*
  32   45     68:===*
  34  166    184:=========*
  36  337    379:================ *
  38  581    626:========================== *
  40  869    873:=========================================*
  42 1009   1067:============================================  *
  44 1276   1177:=================================================*====
  46 1253   1198:=================================================*==
  48 1199   1147:================================================*==
  50 1032   1047:============================================*
  52  949    920:=========================================*==
  54  838    786:===================================*===
  56  578    657:==========================  *
  58  467    539:======================  *
  60  393    437:================== *
  62  339    350:================*
  64  276    278:=============*
  66  214    220:=========*
  68  188    173:========*=
  70  140    136:=======*
  72  131    106:====*=
  74   88     83:====*
  76   71     64:==*=
  78   48     50:==*
  80   43     39:=*
  82   38     30:=*
  84   27     24:=*
  86   21     18:*
  88   15     14:*
  90   17     11:*
  92    7      8:*         :=======*    = represents 1 library sequence
  94   22      7:*         :======*===============
  96    3      5:*         :=== *
  98    8      4:*         :===*====
 100    6      3:*         :==*===
 102    5      2:*         :=*===
 104    9      2:*         :=*=======
 106    4      1:*         :*===
 108    5      1:*         :*====
 110    4      1:*         :*===
 112    4      1:*         :*===
 114    4      1:*         :*===
 116    6      0:=         *======
 118    1      0:=         *=
>120   32      0:==        *================================
```

E-value

Sequence alignment algorithms

- String matching to find common substrings
  - Allowing for mismatches (mutations) and gaps (insertions/deletions)
- Dynamic Programming solution
- Needleman-Wunsch - Global Alignments; Smith-Waterman - Local alignments (SSEARCH, GGSEARCH; water, needle)
- Speedups by looking for common words (KTUPLES - 2-7) to seed the alignment start
- Basically what happens in BLAST, FASTA
- Heursitics typically employ a fixed insertion penalty and cutoffs on low scores so that DP matrix is sparse.

Figure 11: An alignment path matrix

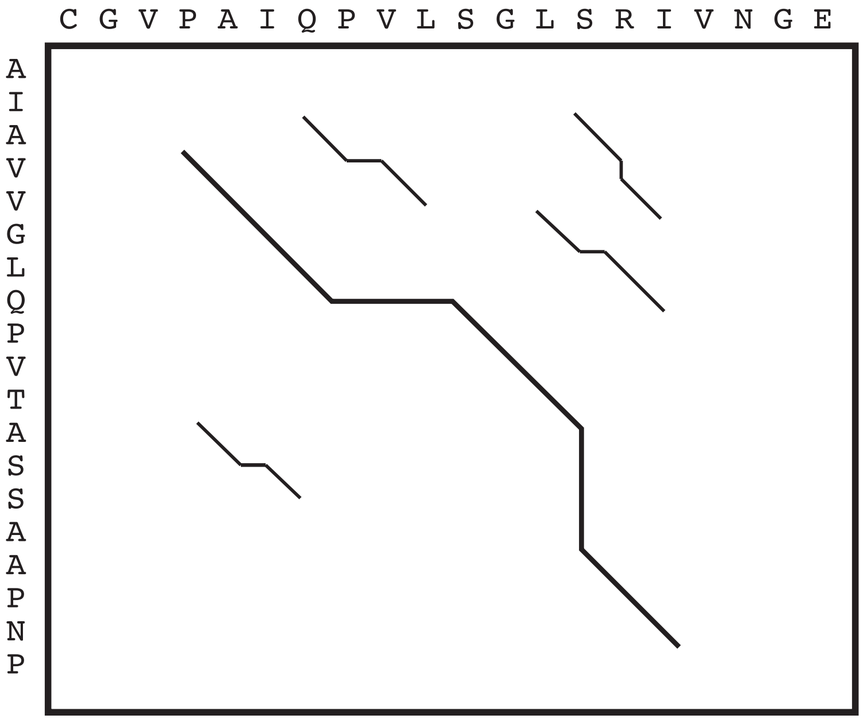


Pearson, ISMB 2000

Sequence alignment algorithms and complexity

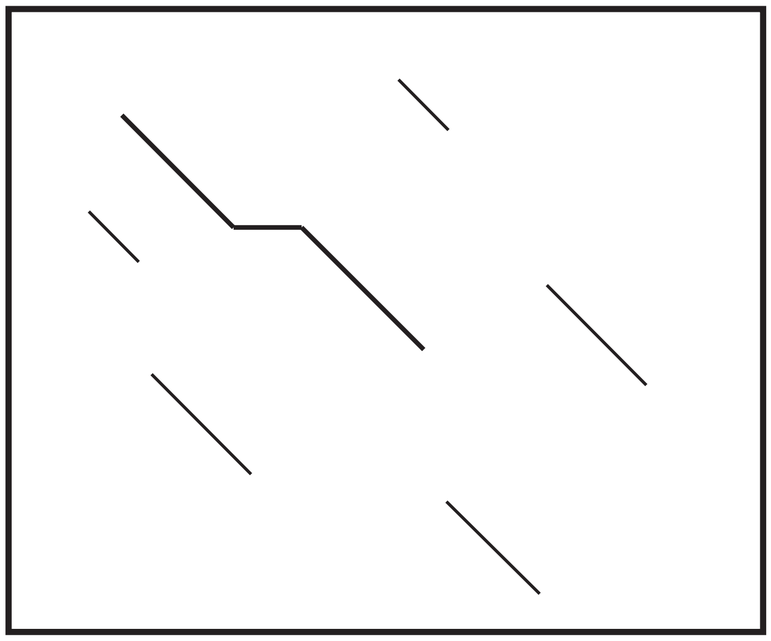
Table 9: Algorithms for comparing protein and DNA sequences

algorithm	value calculated	scoring matrix	gap penalty	time required	
Needleman-Wunsch	global similarity	arbitrary	penalty/gap q	$O(n^2)$	Needleman and Wunsch, 1970
Sellers	(global) distance	unity	penalty/residue rk	$O(n^2)$	Sellers, 1974
Smith-Waterman	local similarity	$\hat{S}_{ij} < 0.0$	affine $q + rk$	$O(n^2)$	Smith and Waterman, 1981 Gotoh, 1982
FASTA	approx. local similarity	$\hat{S}_{ij} < 0.0$	limited gap size $q + rk$	$O(n^2)/K$	Lipman and Pearson, 1985 Pearson and Lipman, 1988
BLASTP	maximum segment score	$\hat{S}_{ij} < 0.0$	multiple segments	$O(n^2)/K$	Altshul et al., 1990



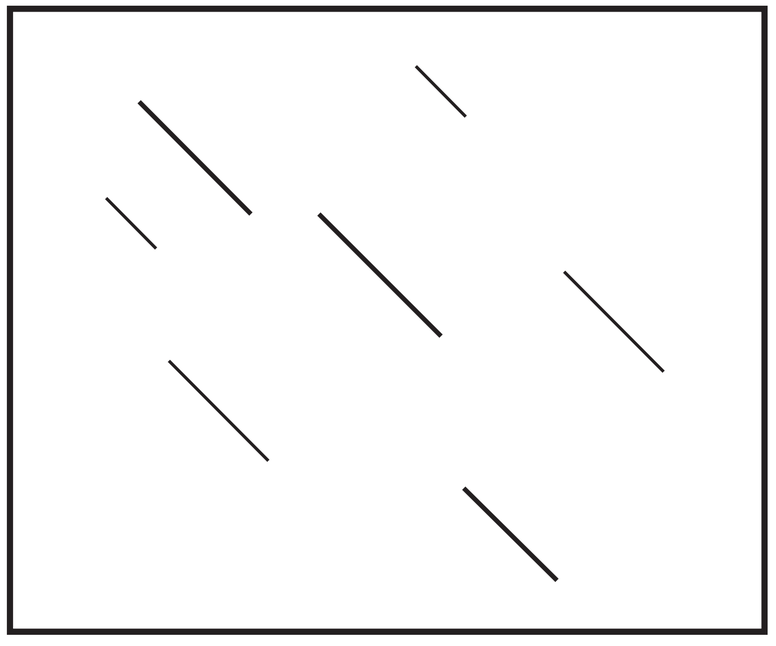
Smith-Waterman

time: 10:00 min



FASTA

time: 2:00 min



BLAST

time: 20 sec

Heuristic algorithms for DP searching for sequence similarities

Pearson, ISMB 2000

Variants on the alignment theme

- HMMER (v1, v2) - Profile Hidden Markov Model where sequences are compared to a profile of a multiple sequence alignment and probability that HMM could have emitted query sequence is assessed.

- MUMMER - suffix trees, longest increasing subsequence, and Smith-Waterman DP

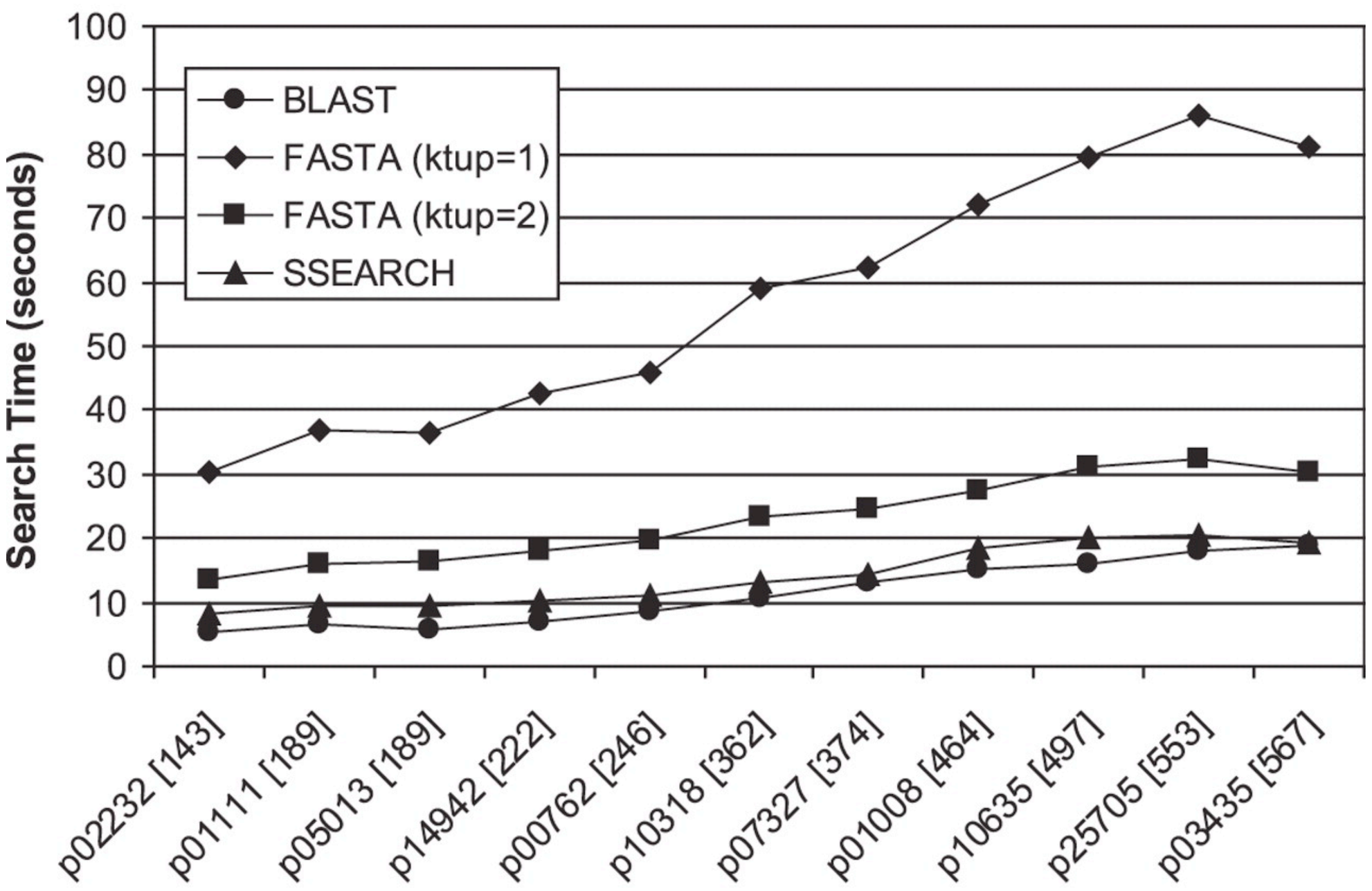
- BLAT - big hash tables to find exact matching or patterns (ie 24 out a 28bp word) in O(1) and string them together. For nearly identical sequences

- For protein sequence alignments - HMMER3 is a game changer.

    - Speed is now basically as fast as BLAST in some cases. New (today!) HMMER3.0b3 is multithread and MPI enabled

The need for speed

- Acceleration of current algorithms
    - Hardware acceleration with GPUs and FPGAs
        - TimeLogic, Paracel (defunct)
    - SIMD acceleration
        - SSEARCH using striped Smith-Waterman on x86 (SSE2)
- Embarrassingly Parallel problem!
    - MPI and PVM enabled to split up searches on a cluster, and recombined the score distribution at the end to assess E-value

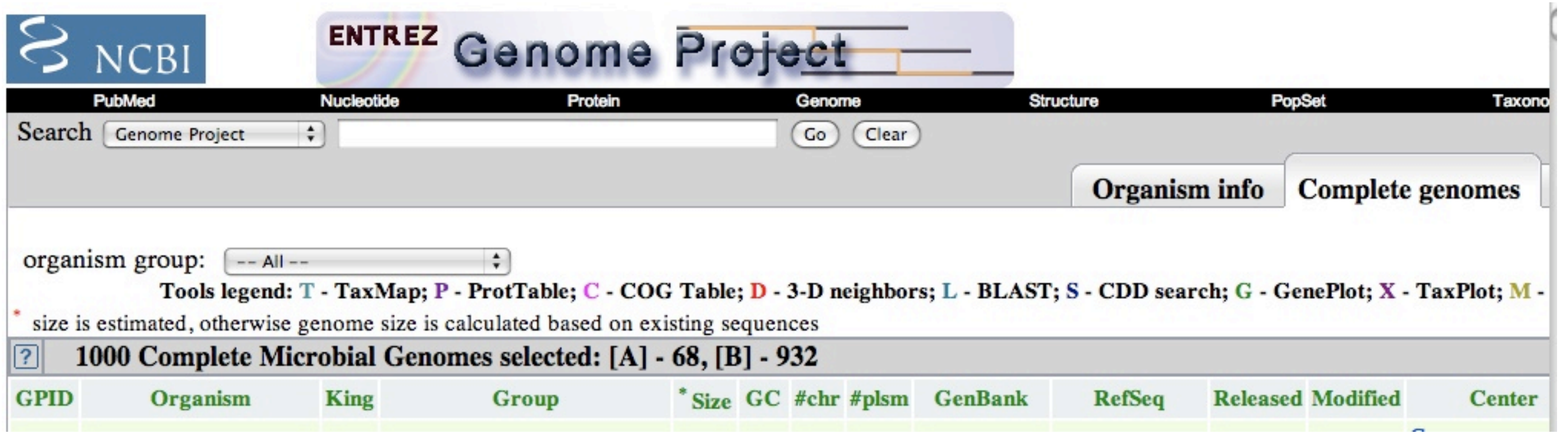


SSEARCH on par with BLAST!

Farrar, Bioinformatics 2007

Large numbers of genome sequences

- 1000+ Bacterial and Archeal genomes
- **100s of fungal genomes**
- 10s of animal and plant genomes
- 10s of other eukaryotes
- Coming online
  - 1000 human genome project - http://1000genomes.org
  - 1001 *Arabidopsis* genomes - http://1001genomes.org
  - 1000 *Drosophila* genomes - http://dpgp.org
  - 15,000 vertebrate genome project (propsed)



http://bit.ly/2ruuhM

Metagenomics

The coming storm: what to do with all this data?

- Can use the more sensitive tools on larger datasets.
- Datasets are of course getting larger. Challenge Moore's law? Producing data faster.
    - Need to get more efficient in how the data is processed, organized, and accessed
- Not discussing the storage and data management challenges

The Sequence DB Search problem

1000Xs more

Query

Pairwise alignment For each sequence

Generates Scores

```
      opt      E()
< 20   17     0:=          one = represents 22 library sequences
  22    0     0:
  24    0     0:
  26    2     0:=
  28    7     3:*
  30    7    18:*
  32   45    68:===*
  34  166   184:========*
  36  337   379:================ *
  38  581   626:========================== *
  40  869   873:=========================================*
  42 1009  1067:=============================================  *
  44 1276  1177:=================================================*====
  46 1253  1198:=================================================*==
  48 1199  1147:================================================*==
  50 1032  1047:==============================================*
  52  949   920:=========================================*==
  54  838   786:=================================*===
  56  578   657:=========================  *
  58  467   539:====================  *
  60  393   437:================= *
  62  339   350:===============*
  64  276   278:============*
  66  214   220:=========*
  68  188   173:========*=
  70  140   136:=======*
  72  131   106:====*=
  74   88    83:====*
  76   71    64:==*=
  78   48    50:==*
  80   43    39:=*
  82   38    30:=*
  84   27    24:=*
  86   21    18:*
  88   15    14:*
  90   17    11:*
  92    7     8:*         :=======*    = represents 1 library sequence
  94   22     7:*         :======*===============
  96    3     5:*         :=== *
  98    8     4:*         :===*====
 100    6     3:*         :==*===
 102    5     2:*         :=*===
 104    9     2:*         :=*=======
 106    4     1:*         :*===
 108    5     1:*         :*====
 110    4     1:*         :*===
 112    4     1:*         :*===
 114    4     1:*         :*===
 116    6     0:=         *======
 118    1     0:=         *=
>120   32     0:==        *================================
```

E-value

Next Generation Sequencing data challenges

- **Where did you come from?**
  Map all the reads (10-30M tags of length 25 - 120 bp) against the genome
  With mismatch/gaps.
  Soln: Hashing, Suffix trees, Burrows-Wheeler Transformation. BowTie, MAQ, SOAP, BWA

- **Find the errors, mutations, insertions or deletions, or copy number variation.**
  Align reads to the genome, but confidently call the differences between the sample and reference as real variations.
  Soln: Some bayesian approaches that factor quality and number of sequences that support it. Shortcut is heuristic: 3+ reads that agree = believable. Can condition on number of agreeing reads. Mosaik, MAQ

- **Genomic Assembly.**
  Make contiguous sequence stretches by putting jigsaw back together (30 - 120bp reads, sometimes paired).
  Soln: Best solutions are graph based -but many paths through the data so lots of memory required. Velvet, ABYSS

- **Show me your genes!**
  Align transcript pieces (short-reads from RNA-Seq) to the genome, but with large gaps for introns.
  Soln: Allow partial alignments and stitch together those that are paired, extra points for occurring at a splice-site and having multiple independent reads that agree. BowTie+TopHat/CuffLinks

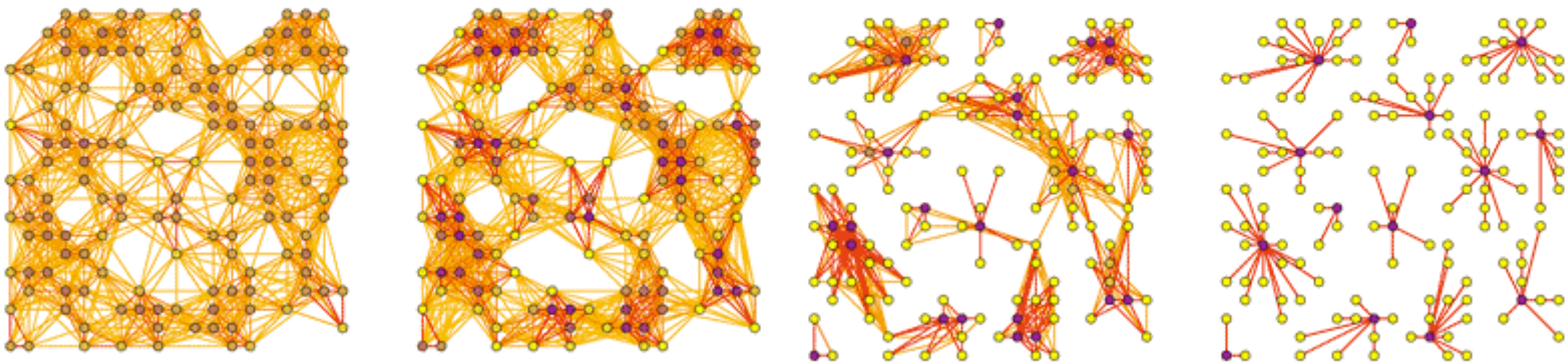
Jacob Appelbaum & Donald Knuth Demonstrate The Recursive Homeboys Principle

Evolutionary questions

- Gene content changes
- Gene order evolution
- Evolutionary relationships between organisms (branching order)
- Evolutionary Rates

Comparing gene content among species

- Identification of the same genes found in each species
- Use similar searches to find significant ones - pairwise problem is fairly straightforward
- Combining these for multiple species where missing data is allowed (gene loss) or additional data (gene duplication) makes the problem harder.
- Typical methods cluster into Orthologous groups by similarity cutoff, but transitive property is not always met (A similar to B, B similar to C, not always the case that A and C are significant similar or homologous).
    - Variations on theme using Markov Clustering (MCL) have been applied with some success.
    - Other approaches include phylogenetic (tree building) for putative clusters to sort out relationships, but is also fraught with false positives and computation complexity.



Clustering with MCL

http://micans.org/mcl

Ortholog and paralog finding still active areas

- De novo identification identification of families still has many false positives. Dynamic adjustment of parameters based on sequence length, evolutionary rates, etc may help refine clusters better.

- As more genomes are produced, how can we re-build ortholog sets incrementally?

- Use HMMs of ortholog sets as searching database to improve sensitivity

Gene order evolution

- Comparing order of the genes in different organisms.

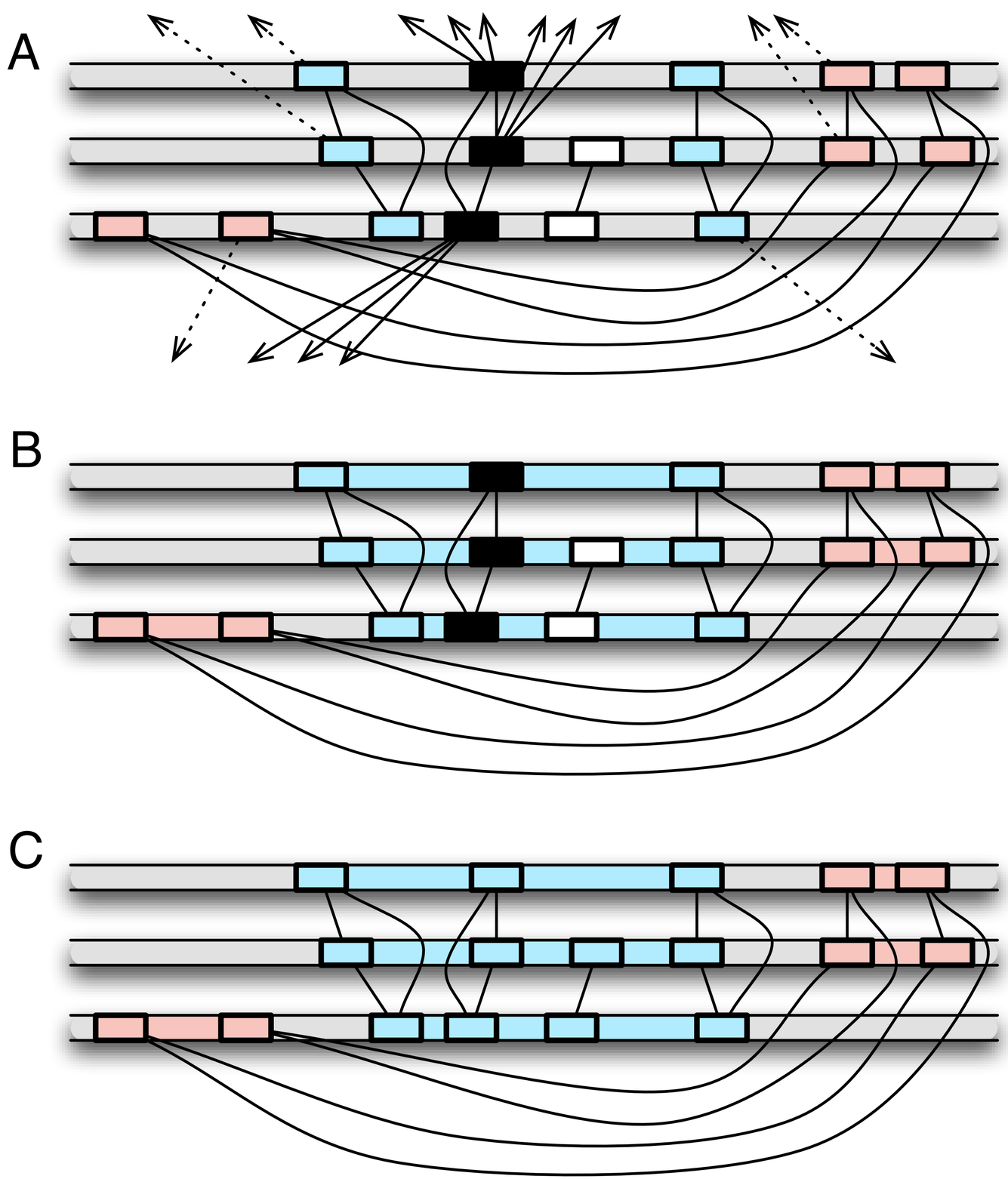
- Conservation of order is referred to as synteny

- Why is it important? Shared synteny suggests between different organisms suggests it is the ancestral state. If it has been preserved over time, maybe a functional reason

- Also useful for inferring corresponding (orthologous) genes in different species

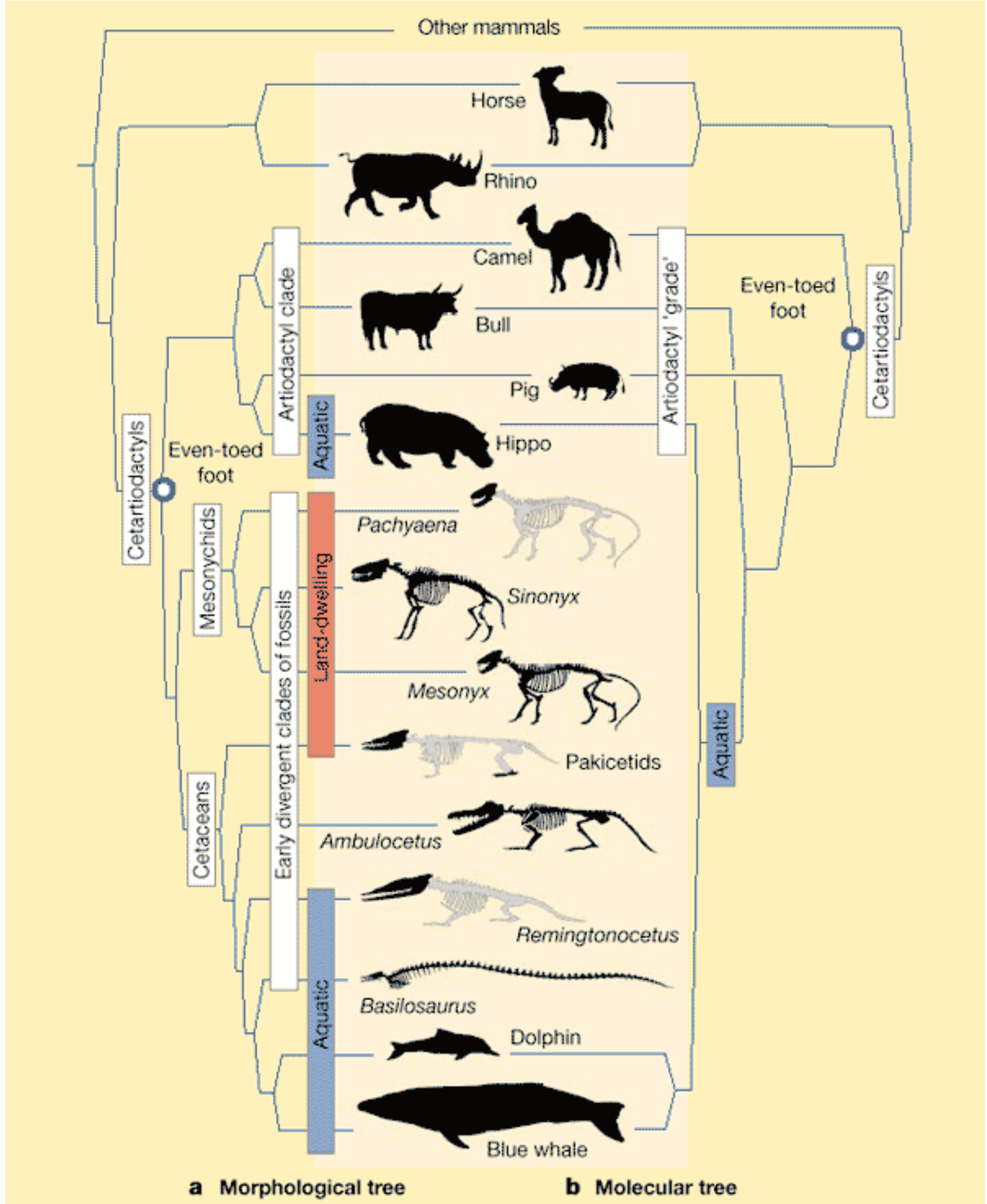
Synteny blocks

- Identifying regions that have shared history
- Identify anchors, consider ordering of anchors in each organism
- Allow for insertions and deletions
- Find longest common “substring” of genes.
- Genes in a block found in different species may be orthologous



Phylogenetic Trees

- Represents evolutionary history of organism
- Can be used with binary character traits or DNA or protein sequences
- Constructed by a calculation of the distance between the taxa
- Bayesian and maximum likelihood approaches as well as simpler distance-only (neighbor-joining) and parsimony approaches to represent the data.



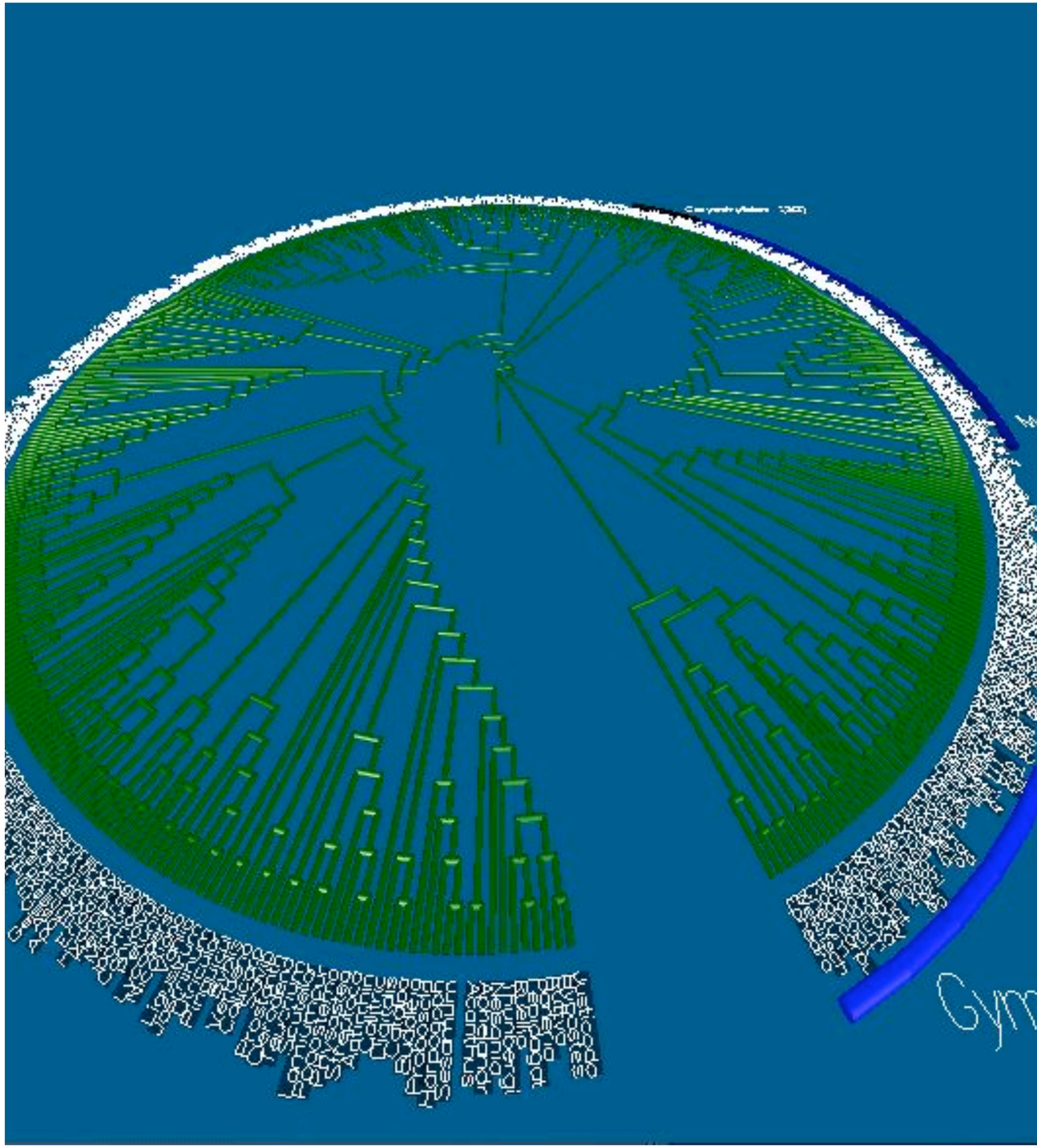
Tree building future problems

- How do we assemble very large trees?
  In pieces and stitch together?
  All in one go with huge distributed systems?

- Very much an active area of research
  Need more efficient algorithms and hardware acceleration approaches.

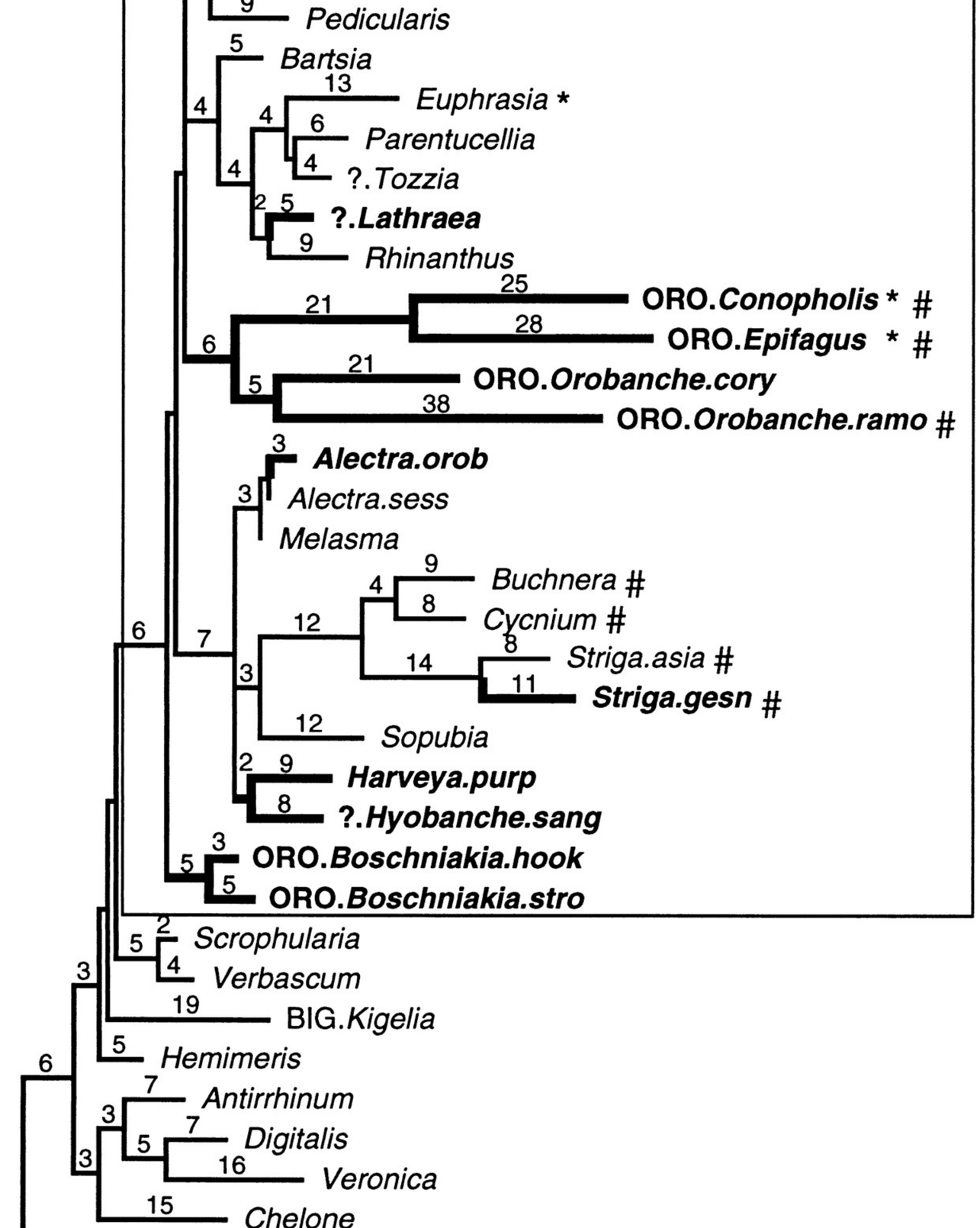
- NSF funded CIPRES have been on algorithmic improvements -for large number of taxa from Tree of Life project data.

- Tree Viz is also important area - how do we represent the large amount of data? Dynamically and interactively?



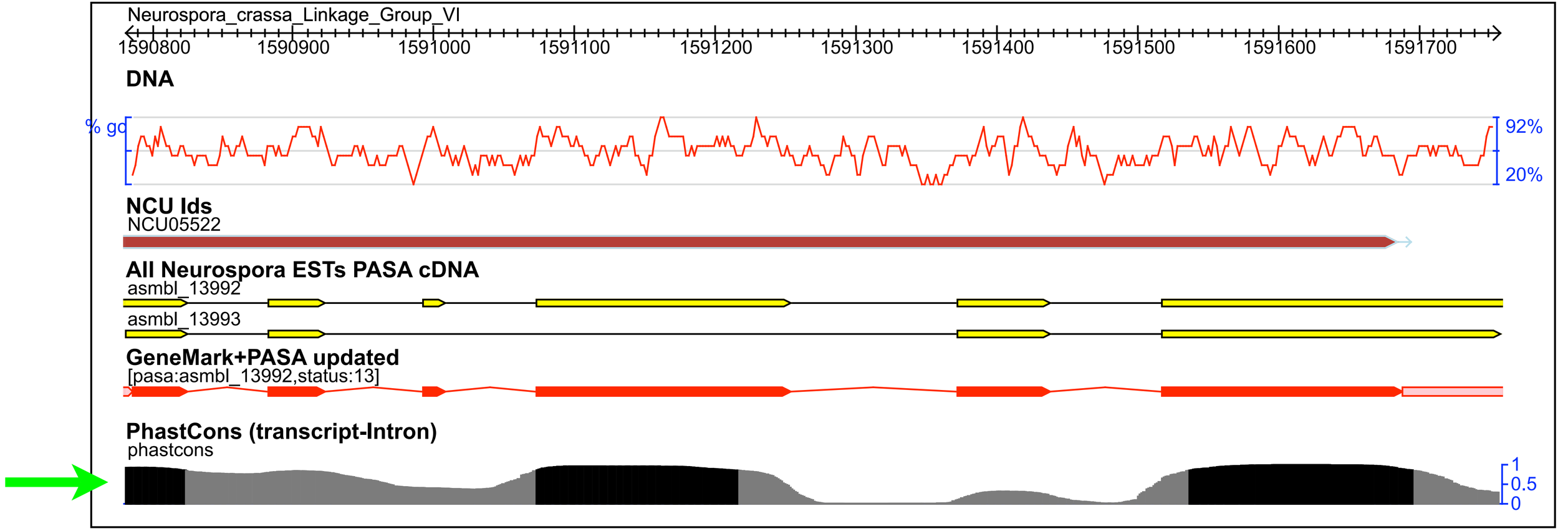
Evolutionary rates

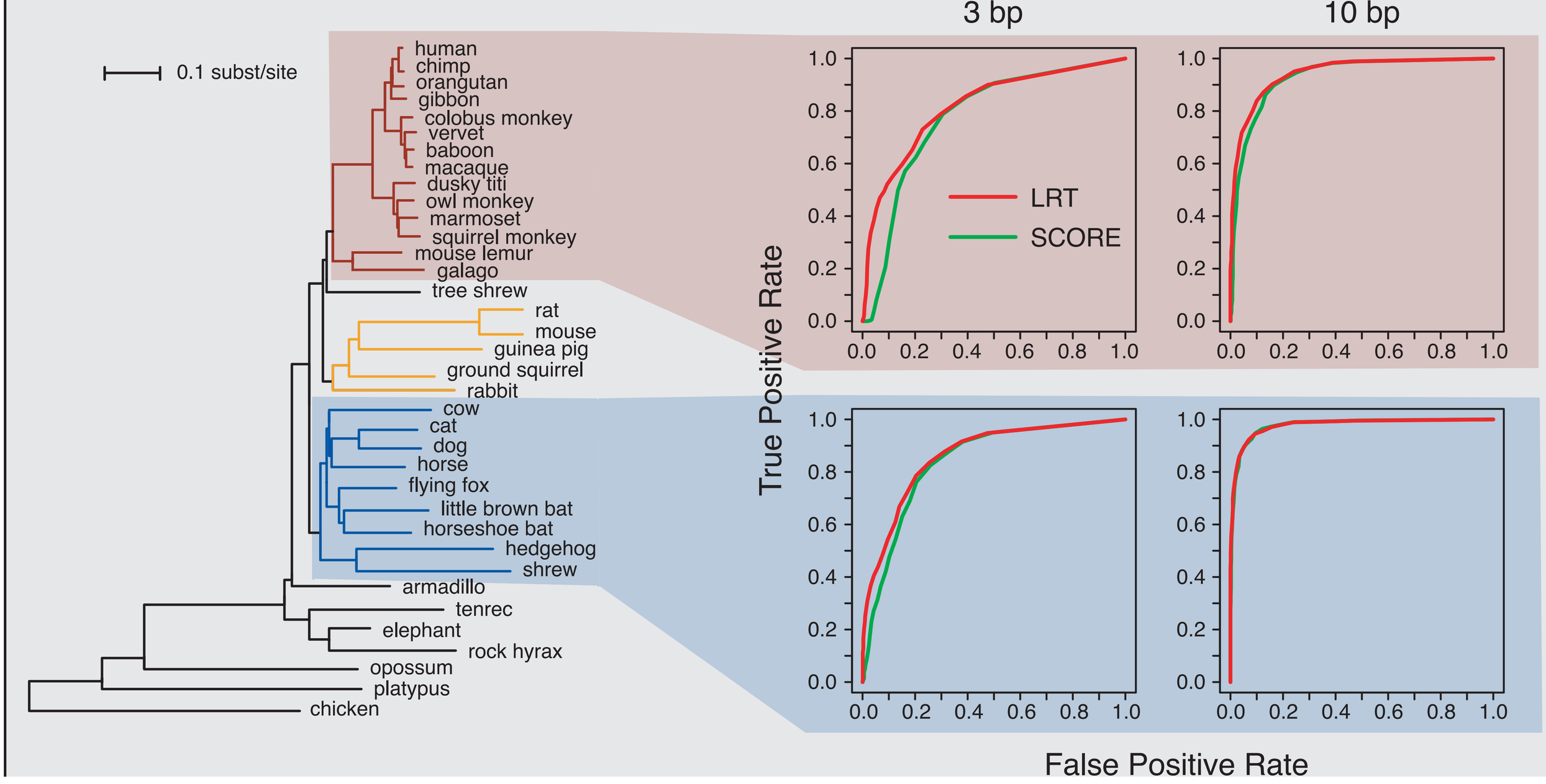
- Calculate evolutionary rates as amount of change in a sequence as compared to others
- If the amount of change is greater in some branches suggests interesting phenomenon -identify by comparing the relative rates.
- Statistical tests for whether rates are significantly different. Can apply to protein loci and genome alignments.
- Open questions: inferring rates in the first place still a hard problem and quite slow. Methods that efficiently scan whole genome of alignments still need work.



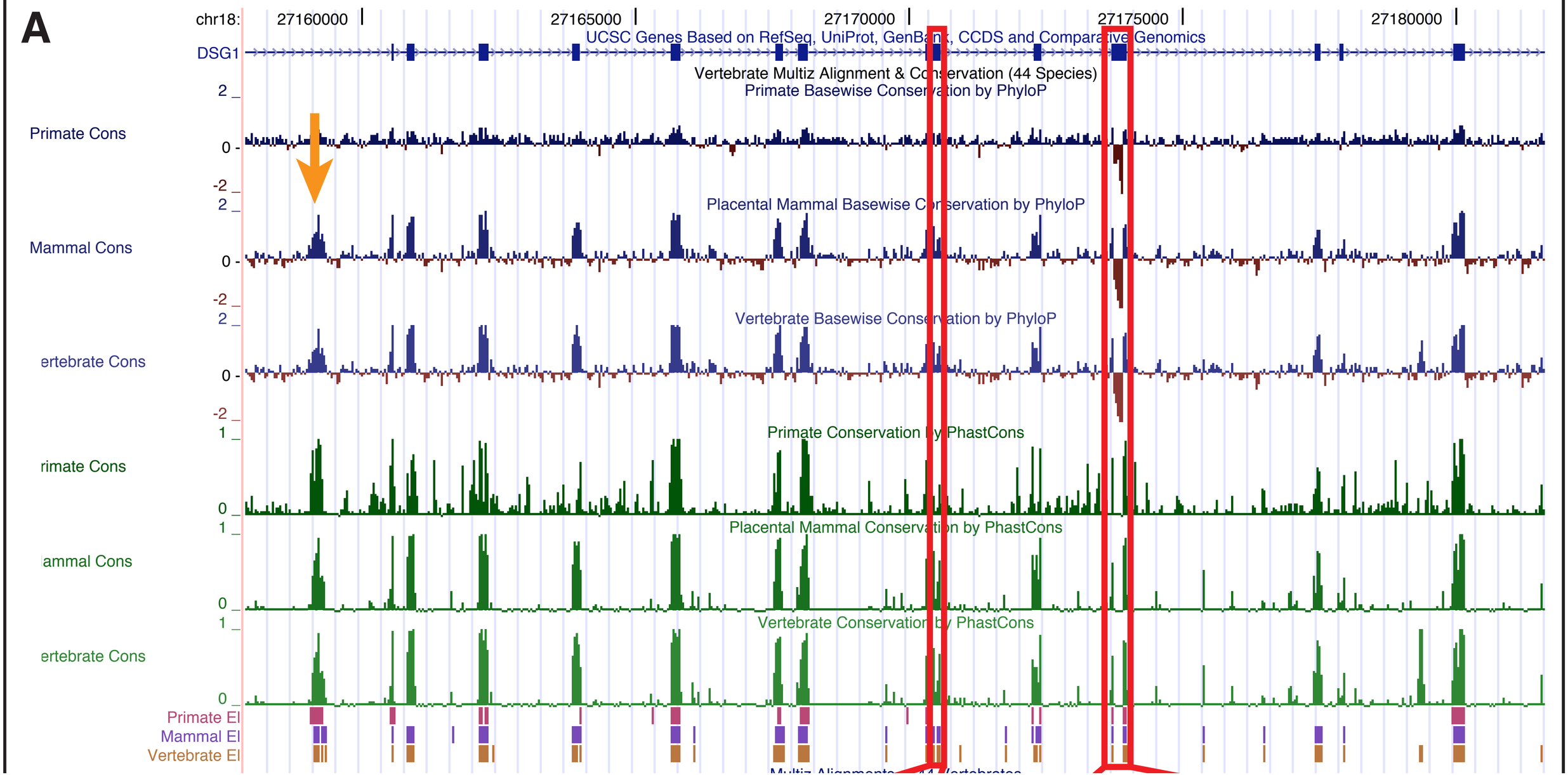
Comparing conservation profiles

```
Nc   TTCCTTACCAGCAGCCCCGGCCGTCGCCCCGCTCTTTCGACCCACGATGTTGTC---GAT
Nt   TTCCTTACCAGCAGCCCCGGCCGTCGCCCCGCTCTTTCGACTCACGATGTTGTC---GAT
Nd   TTCCTGACCAGCAGCCCCGGCCGTCGCCCCGCTCTTTCGACCCACGATGTTCTC---GAT
Sm   TTCCTTACCAGCAGCCCCGGCCGTCGCCCCGCGCTCTCGACTCACGATGTTGTC---GAC
Pa   TTCTTGACCAGCAGCCCAGGACGCCATCCAGCACT---GACCCACCTTGATGTCCAAGAT
Cg   TTCCTCACCAGCAGCCCTGGGCGCCACCCGGCTCTGTCGTCGCA---TGACGTCAAAGAT
     *** * ********** ** ** *  ** ** **   * * **   **   **   **
```





Pollard et al, Genome Res, 2009



Simultaneously identifying slow and rapidly evolving regions in the human genome

Pollard et al, Genome Res, 2009

More open areas in computational biology

- **Image processing - decoding visual images into computable data**
  Microscopy images for phenotypes (trait) analyses:
  How big is that neuron cell? After I apply this drug?

  Patterns of gene expression through staining and visualization - what is the pattern?

- **Protein 3-D structure analysis and prediction**
  Molecular dynamic simulations to take 1D protein sequence to 2D and 3D sequence structure predictions. Predictions are still very poor - using hints of known sub-sequences and their folding pattern. {Folding@Home}

- **RNA structure prediction**
  Identification of the secondary structure of sequences via thermodynamic and base-pairing rules. Finding real versus artifact and comparing this to sequence conservation.

Still more

- **Gene network reconstruction from time series data**
  How do genes function together and are dependent on each other.

- **Gene identification and annotation**
  Use of computation and experimental data to define the regions of the genome which are transcribed into genes

- **Gene function prediction**
  Using the known annotation of function of some genes can we infer the function of others using ontologies of gene function and phylogenetic relatedness of the genes and species

For more information

http://bioinfo.ucr.edu

http://lab.stajich.org

http://bioperl.org

Center for Plant Cell Biology at UC Riverside

UC RIVERSIDE UNIVERSITY OF CALIFORNIA | Institute for Integrative Genome Biology