

Comparative Genomics

"Know then thyself, presume not God to scan; The proper study of mankind is man."

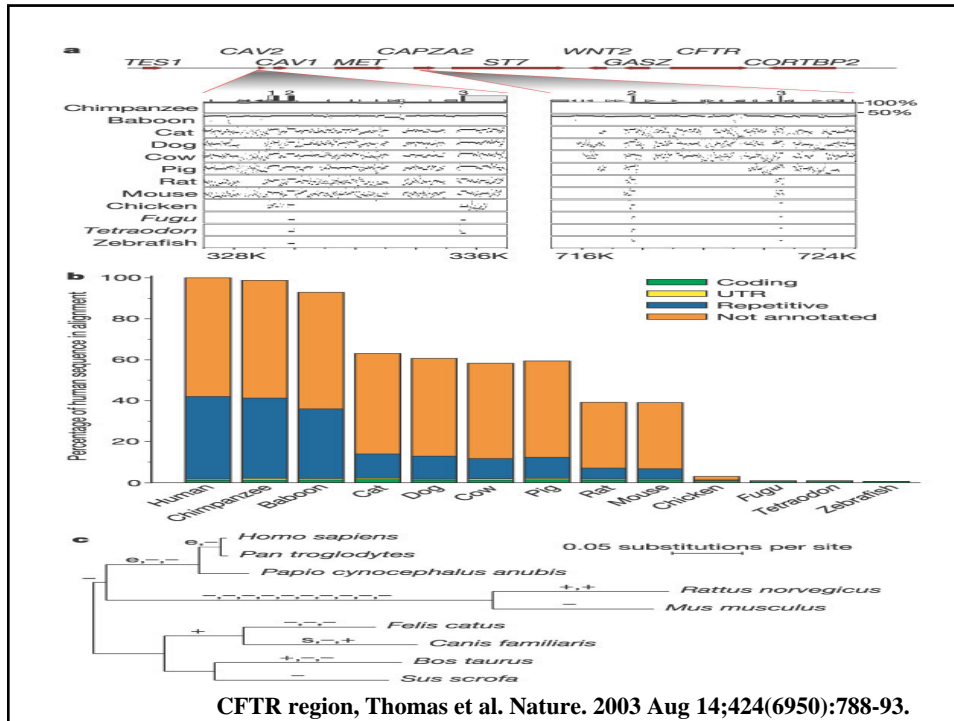
Alexander Pope, 1733

"Nothing in biology makes sense except in the light of evolution."

Theodosius Dobzhansky, 1932

Comparative Genomics

- 1. What is it?**
- 2. Rationale**
- 3. How much is conserved?**
- 4. Applications**
- 5. VISTA tool**
- 6. Genome-scale alignment methods**
 - 1. BLASTZ**
 - 2. AVID**
 - 3. LAGAN**
- 7. Transcription factor binding site identification using conservation**
 - 1. Levy and Hannehalli**
 - 2. rVISTA**
- 8. Evolutionary study via genome rearrangements**



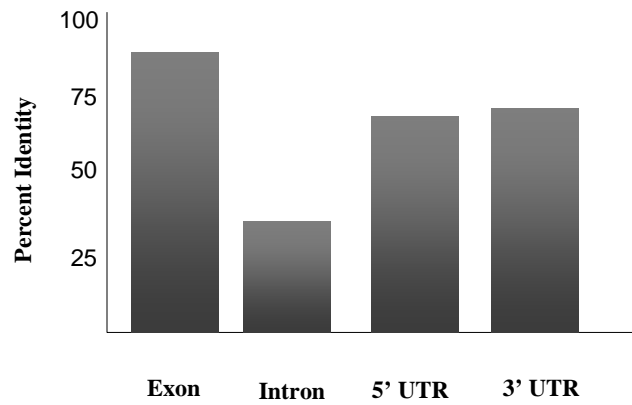
Functional Importance



Preferential Conservation against random mutations

Sometimes in closely related genomes, the ‘differences’ are what we are interested in and not the similarities, eg., Human and Chimp.

Comparing 1200 Human-Mouse Orthologs



Makalowski et al. 1996

Overall about 5% of the genome is under selection based on Human-Mouse comparison

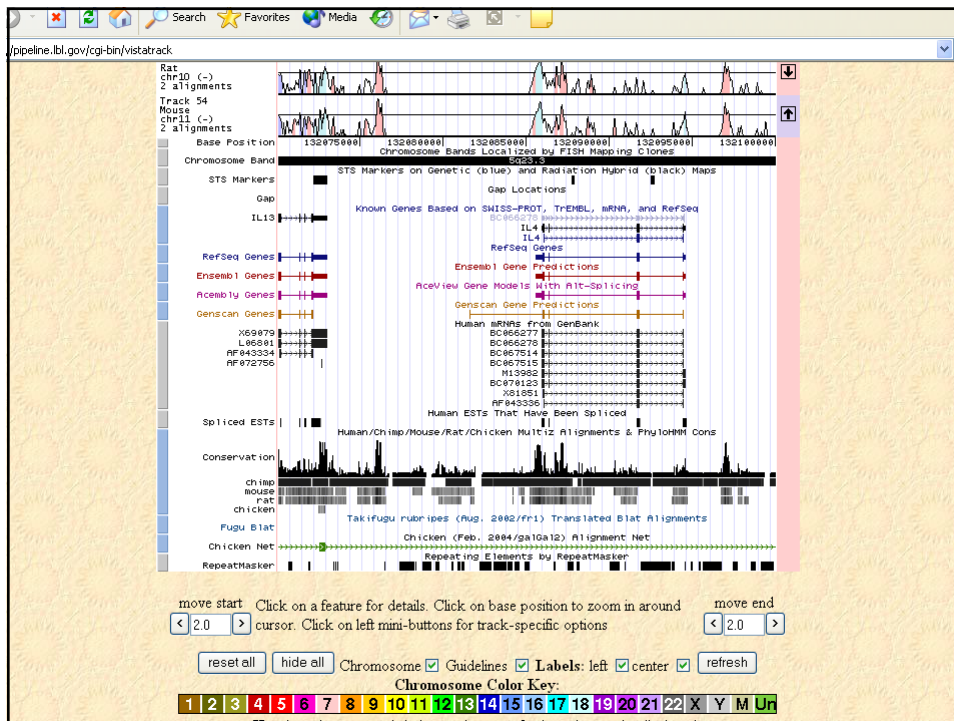
Comparative genomics Applications

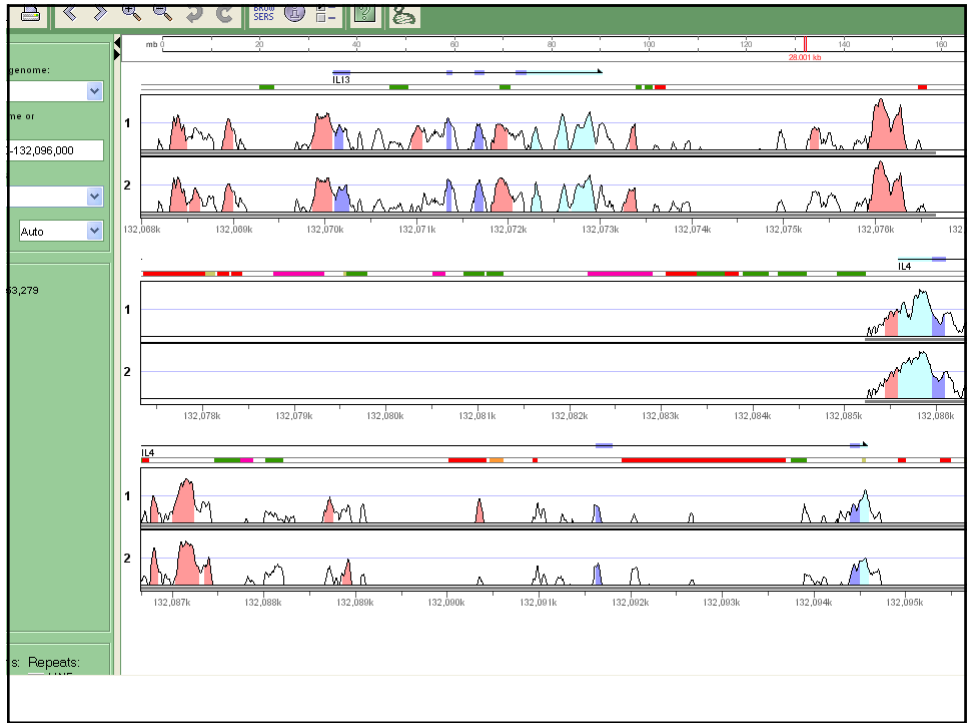
- Gene structure/function prediction
- Discover non-coding (regulatory) functional regions
- Motif prediction
- Study evolution

Identification of a Coordinate Regulator of Interleukins 4, 13, and 5 by Cross-Species Sequence Comparisons
 Loots et al., *Science*, Vol 288, Issue 5463, 136-140 , 7 April 2000

- 1 Mb region around IL cluster in multiple species were sequenced and aligned.
- There are 15 non coding sequence highly conserved in all mammals
- The largest of these was tested positive as a positive regulator for the three IL-4,5,13 which are spread along 120 kb sequence.

The genomic regions were obtained using PCR from other mammalian species whose genomes are not finished.





VISTA - VISualization Tools for Alignments

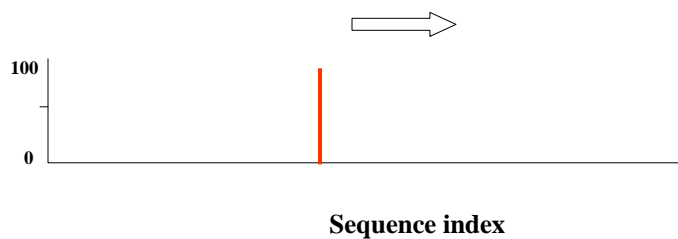
Genome scale alignments

```
aacaaagcgacttagagaatctttaatttcatttcttaactagttagccctatttggttgcttgaatt
caciaagggacttagagaatctttaatttcatttcttaaccagttggccctatttggttgcttgaatc
```

Visualization

85% identical

```
aacaaagcgacttagagaatctttaatttcatttcttaactagttagccctatttggttgcttgaatt
caciaagggacttagagaatctttaatttcatttcttaaccagttggccctatttggttgcttgaatc
```

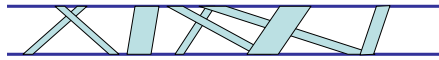


Genome scale alignments

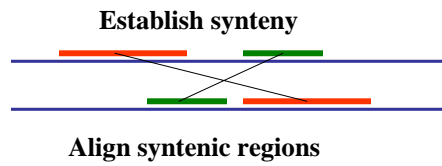
Global Alignment



Local Alignment



Genome Alignment



Main challenges with genome alignment

Speed/Memory

Incomplete Data

What is the right parameter – poor evolutionary model

Lack of gold standards

The approaches are heuristic (not 'algorithmic')

All approaches can be generalized as a hierarchical process

1. Compute anchors

2. Chain/Link the anchors

3. Refinement

Hannenhalli and Levy, 2001
LAGAN, Brudno et al. 2003
AVID, Bray et al., 2003

BLASTZ

1. Find all near perfect 12-mer matches using hashing
2. Do gapless extension
3. Do gapped extension of high scoring matches
4. Repeat from Step 1 for the inter-hit regions with lower match thresholds

AVID

1. Find all exact matches using suffix tree
2. Filter matches(keep the matches at least half as long as the longest match)
3. Compute optimal chain of matches (Needleman-Wunsch)
4. Within each consecutive match pair in the optimal chain
 1. Back to step 2 and repeat OR
 2. Apply Needleman-Wunsch

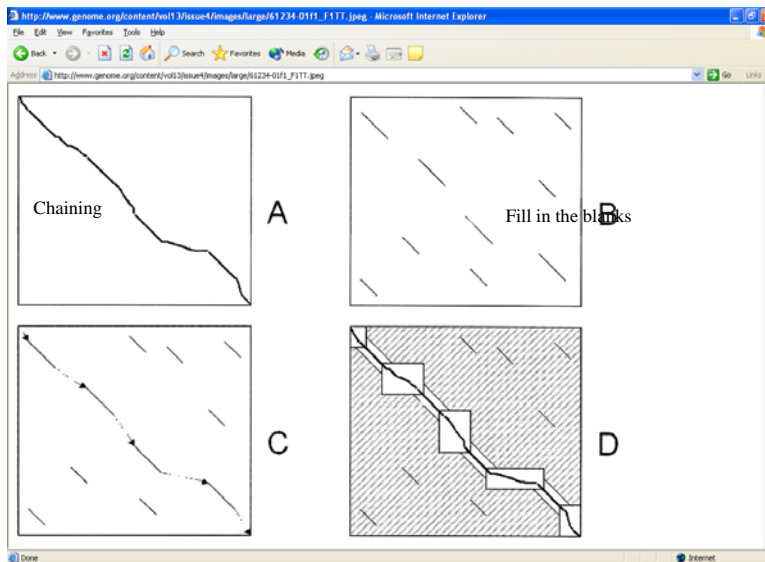
Results on Human versus Rat

The entire finished Rat sequences were mapped to regions in Human, and the corresponding Human annotation was used.

	Coverage (bp)	Refseq	UTR	Time (sec)
AVID	1686932	82920	38973	209
BLASTZ	2115917	90662	42267	447
BLASTZ (chaining)	1618195	83338	39099	101
CHAOS	456152	59918	14415	403
GLASS	1076219	62978	29376	28993
MUMmer	850710	65361	31006	258

LAGAN

The initial matches are based on multiple short/inexact matches than 1 long match (AVID)

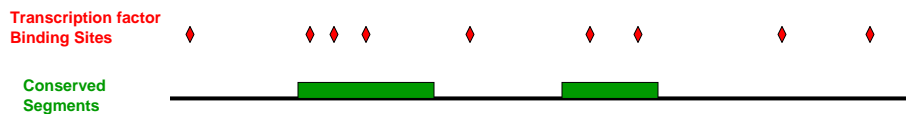


Results on 230 Mouse exons in CFTR region

70-100% of exon length aligned correctly

Program	#Exons	Time(sec)	Memory(Mb)
MUMmer	14	7	40
AVID	100	215	581
BlastZ	100	46	202
LAGAN	100	78	90

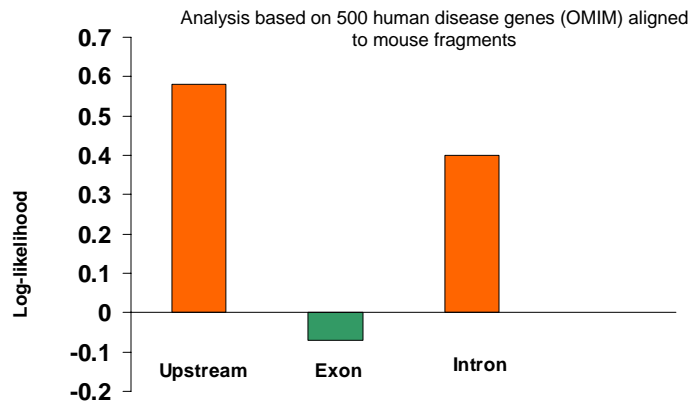
Enrichment of TF binding sites in conserved non-coding regions



$$\text{Relative abundance of TF sites in Conserved Segments} = \log \frac{\text{fraction of TF sites in Conserved Segments}}{\text{fraction Conserved Segments}}$$

+ ve = over-representation
- ve = under-representation

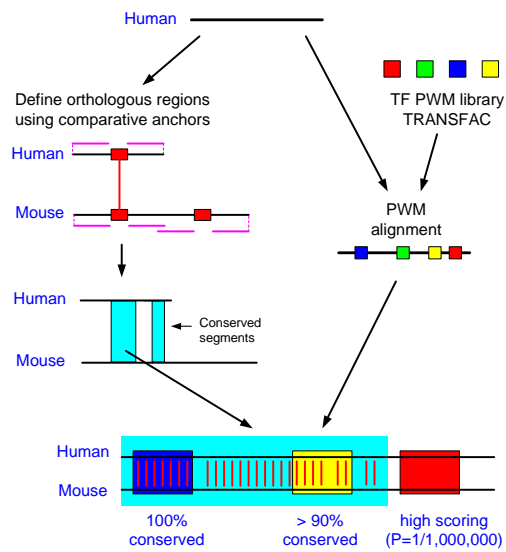
Enrichment of transcription factor binding sites in conserved regions in non-coding sequence



Levy and Hannenhalli, 2000

TF binding site annotation based on conservation

Levy and Hannenhalli, 2000



Regulatory VISTA - rVISTA

- TFBS identification independently in multiple species
- Select only the ones that occur in multiple species
- Apply additional window-based conservation criterion around the TFBS

Analysis of 1 Mb region around IL cluster

Number of GATA hits on Human sequence 20654

Aligned with same site predicted in Mouse 2618

Conserved \geq 80% in 24 bp window 731

How surprising is this?

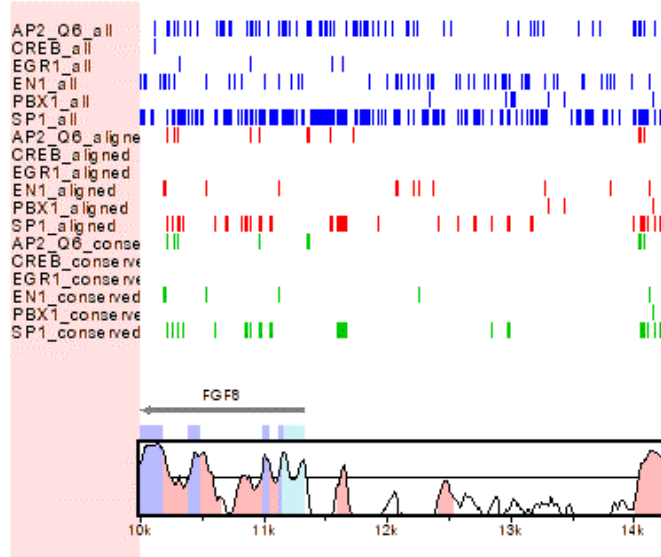
Given 5-10% conservation between Human and Mouse 2618 out of 20654 is not surprising

Given that we only consider regions with \geq 70% identity

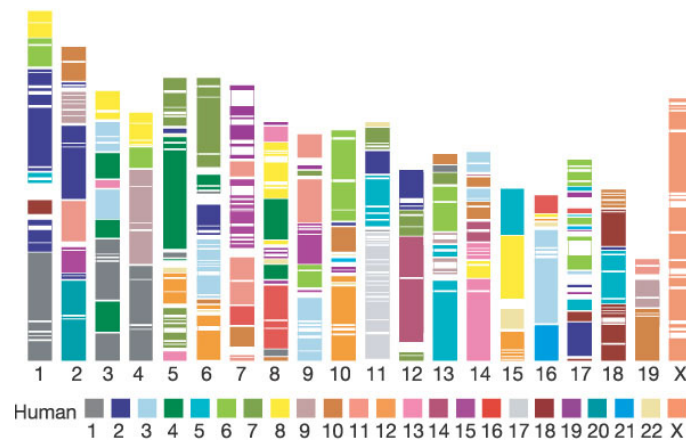
For a 10 bp site, Prob(0, 1, or 2 mismatches) =

$$0.7^{10} + 10 \cdot 0.3 \cdot 0.7^9 + 45 \cdot 0.3^2 \cdot 0.7^8 = 0.38$$

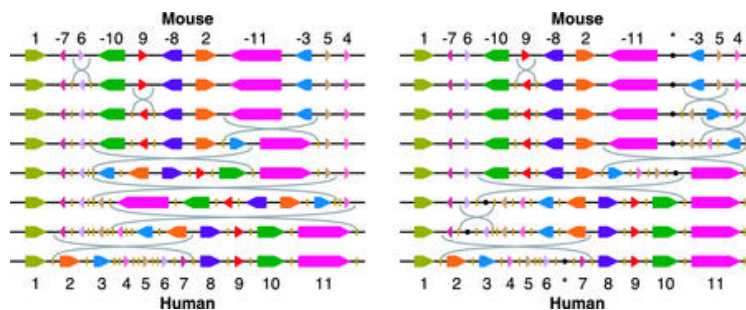
rVISTA images



Relative genome organization between Human and Mouse

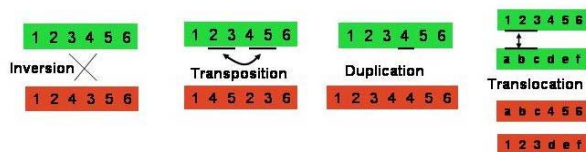


Combinatorial puzzle



Evolution and Genome Rearrangement

- Large scale genome rearrangement events like inversions, transpositions, translocations and duplications present alternative modes of evolution.
- Due to relative rarity of these events, the resulting phylogenetic trees are robust.
- Whole genome duplication followed by gene loss seems to be the most common mode of evolution in Yeast.
- Segmental duplications are a major source of increase in gene repertoire



Comparative Genomics

- 1. What is it?**
- 2. Rationale**
- 3. How much is conserved?**
- 4. Applications**
- 5. VISTA tool**
- 6. Genome-scale alignment methods**
 - 1. BLASTZ**
 - 2. AVID**
 - 3. LAGAN**
- 7. Transcription factor binding site identification using conservation**
 - 1. Levy and Hannehalli**
 - 2. rVISTA**
- 8. Evolutionary study via genome rearrangements**