BIOINFORMATICS
BIOMOLECULAR ENGINEERING DEPT

Assessing Whole Genome Alignments

Dent A. Earl[1,2], Benedict Paten[2], David Haussler[1,2,3]

1. Bioinformatics Graduate program, UCSC 2. Center for Biomolecular Science and Engineering, UCSC, 3. Howard Hughes Medical Institute

UC SANTA CRUZ

Baskin Engineering UC SANTA CRUZ

What is a whole genome alignment (WGA)?

A whole genome alignment (WGA) is a description of the evolutionary relationships between a set of whole genome sequences (i.e. DNA sequences that describe entire genomes).

Given a set of sequenced genomes a WGA will group together (*align*) positions, or DNA residues, in the genomes that share a common evolutionary history and leave positions that are independent ungrouped.

The problem of creating a WGA using a realistic model of evolution is computationaly intractable and as a result methods of creating WGAs use heuristics.

Why would you want a WGA?

Whole genome alignments can be used to find areas of conservation between genomes. Over long evolutionary time spans, areas of the genome that are not doing anything (i.e. are not under selection pressure) can change or mutate. Areas that do not change are likely under a selective pressure to remain the same and this can be indicative of biologically important regions.

With the ever increasing production of newly sequenced genomes from projects such as Genome 10K (an international collaborative project to sequence 10,000 vertebrate genomes) there will soon be a massive number of sequenced genomes waiting to be analyzed.

The first question a biologist will ask when they sequence a new species will be

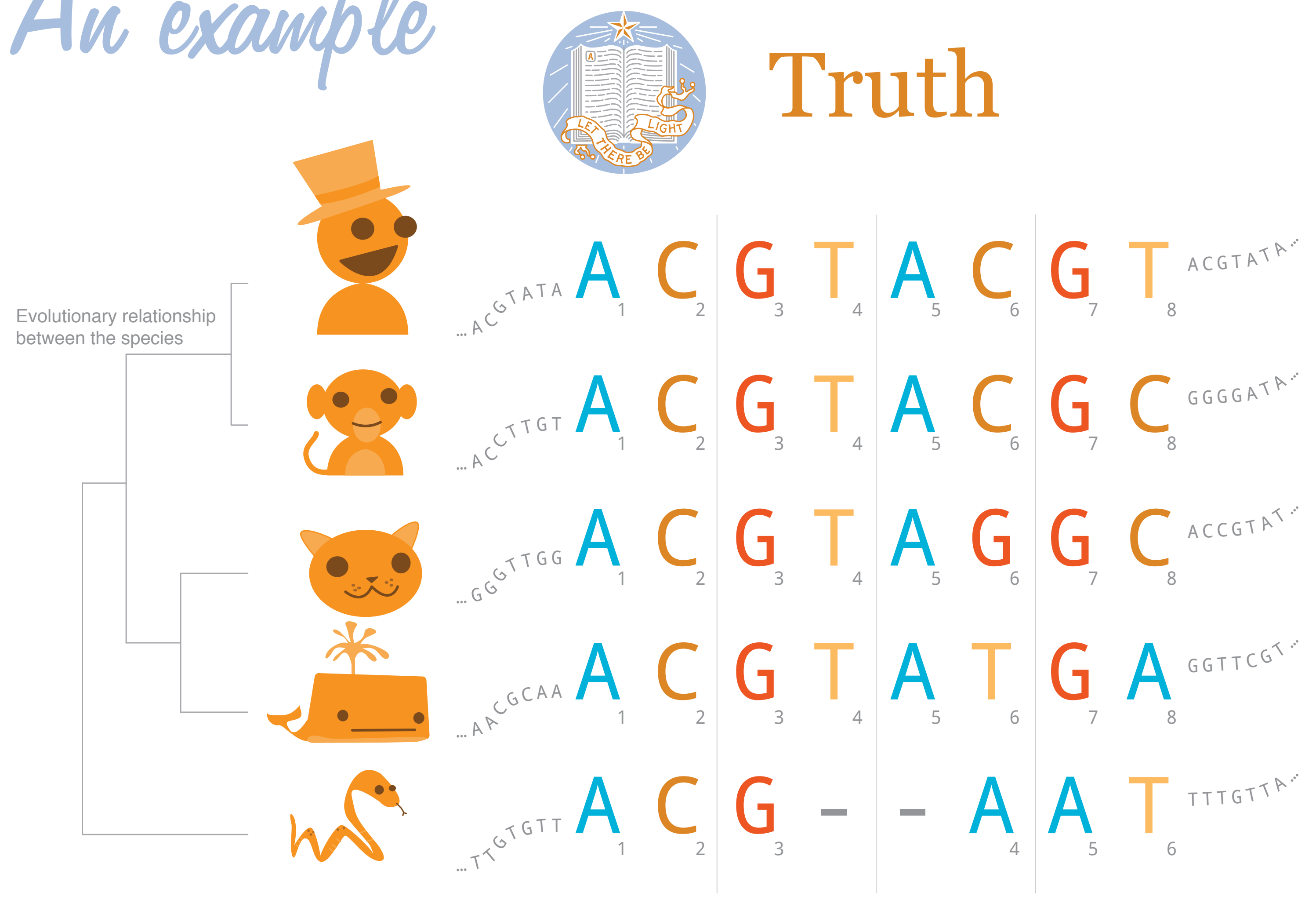
"how is this new species related to this other species we already know about?"

WGAs provide the answer to that question.

But which whole genome aligners are the most accurate? How do you assess a WGA?

GENOME 10K
http://www.genome10k.org/

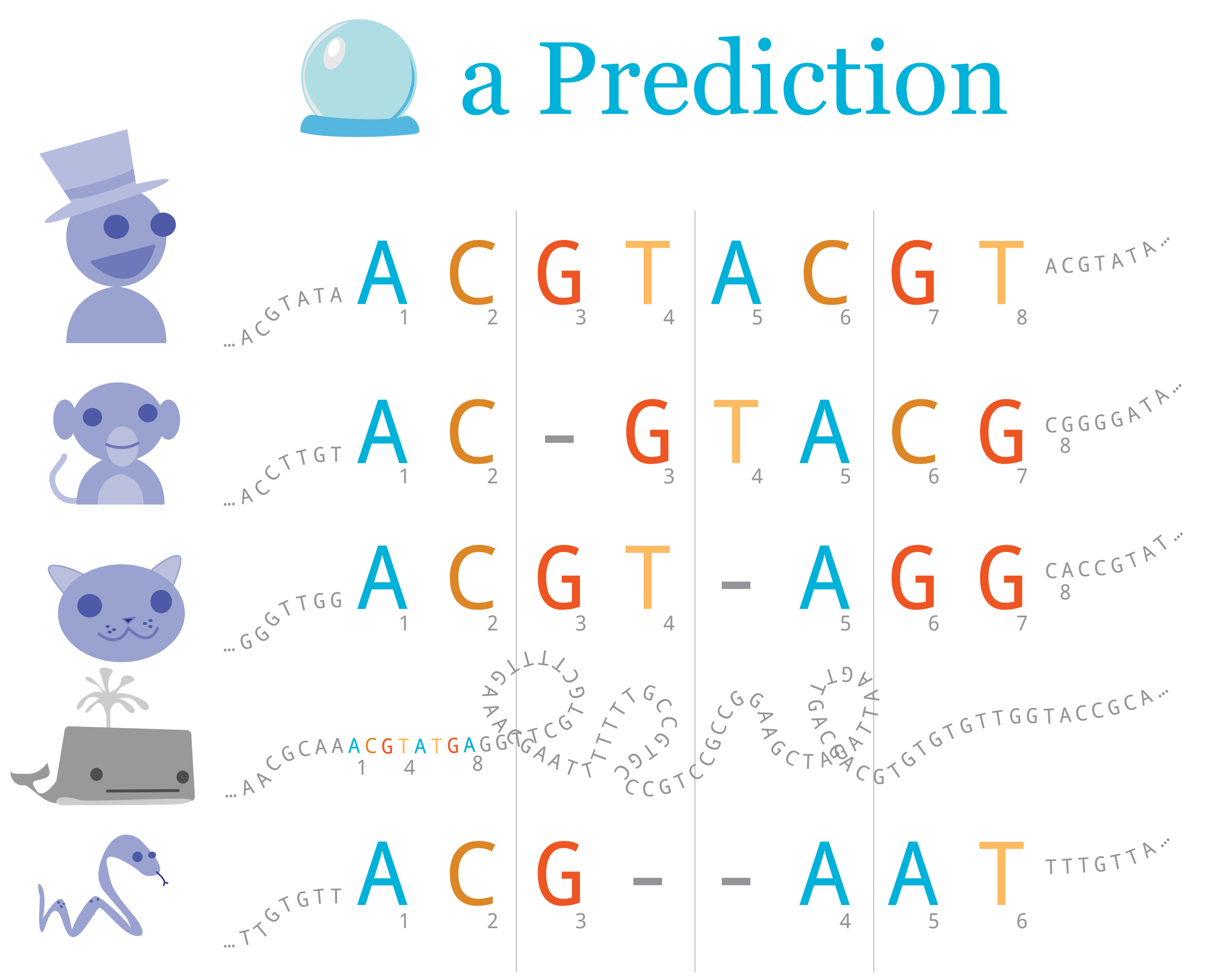
An example


Truth
Evolutionary relationship between the species
ACGTATA A C G T A C G T ACGTATA
1 2 3 4 5 6 7 8
ACCTTGT A C G T A C G C GGGGATA
1 2 3 4 5 6 7 8
GGGTTGG A C G T A G G C ACCGTAT
1 2 3 4 5 6 7 8
AACGCAA A C G T A T G A GGTTCGT
1 2 3 4 5 6 7 8
TTGTGTT A C G - - A A T TTTGTTA
1 2 3 4 5 6

The alignment on the left is the known True alignment, generated by realistic simulation. Each row shows the DNA sequence of a species.

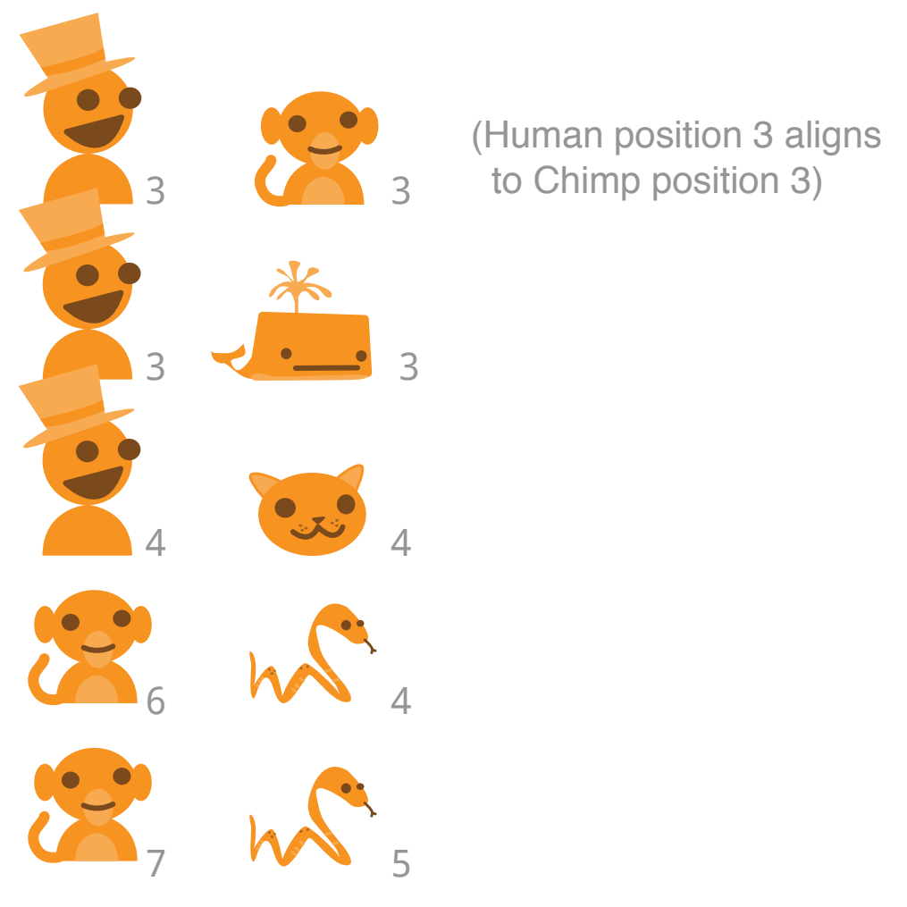
Letters that fall into the same column (large, in color, numbered) are said to be aligned. Positions that are aligned have shared evolutionary history and are useful for answering many questions in Biology.

The alignment on the right is a predicted alignment generated by software.


a Prediction
ACGTATA A C G T A C G T ACGTATA
1 2 3 4 5 6 7 8
ACCTTGT A C - G T A C G CGGGGATA
1 2 3 4 5 6 7 8
GGGTTGG A C G T - A G G CACCGTAT
1 2 3 4 5 6 7 8
AACGCAAACGTATGAGG
TTGTGTT A C G - - A A T TTTGTTA
1 2 3 4 5 6

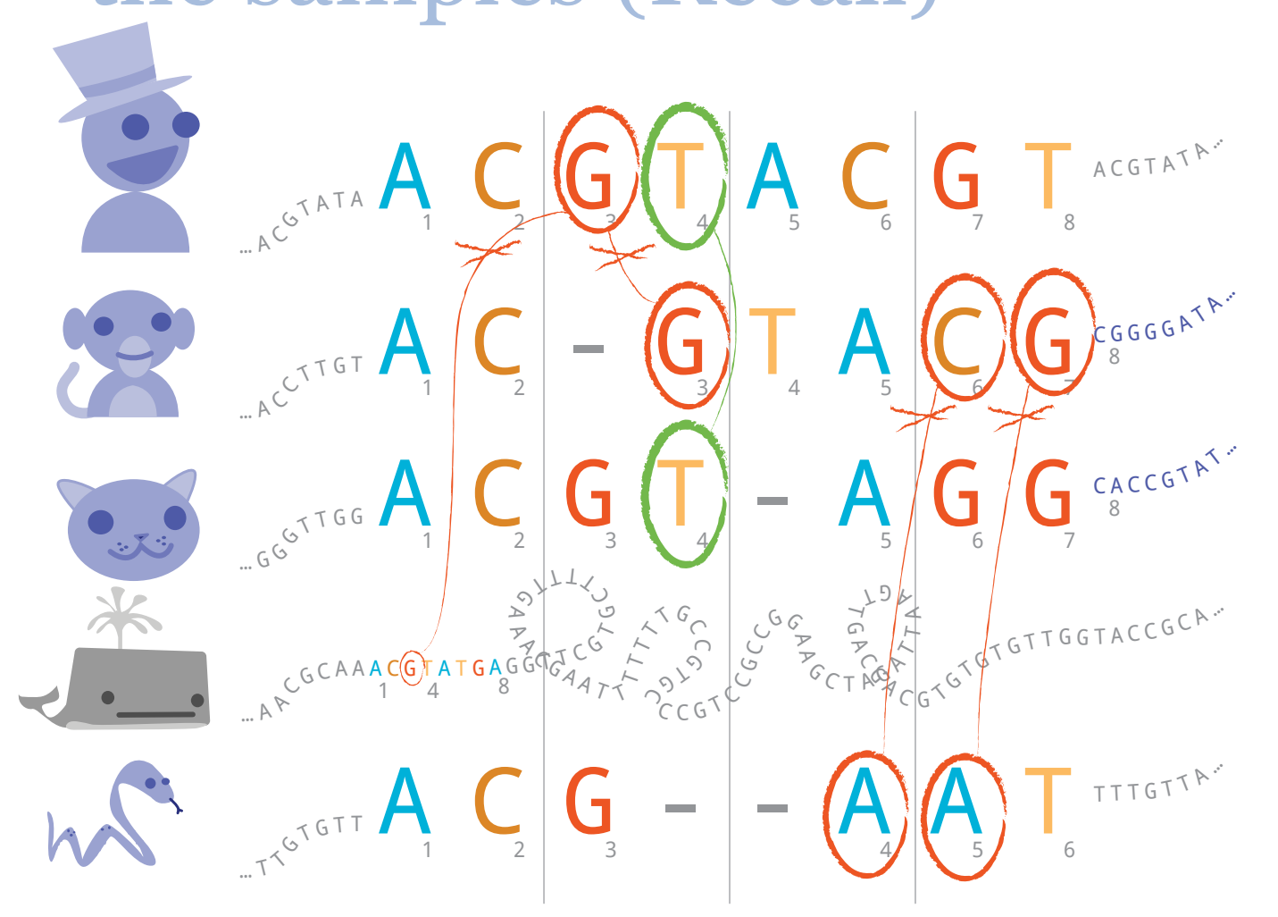
One assessment: Calculating Precision and Recall

1) Sample pairs from the truth



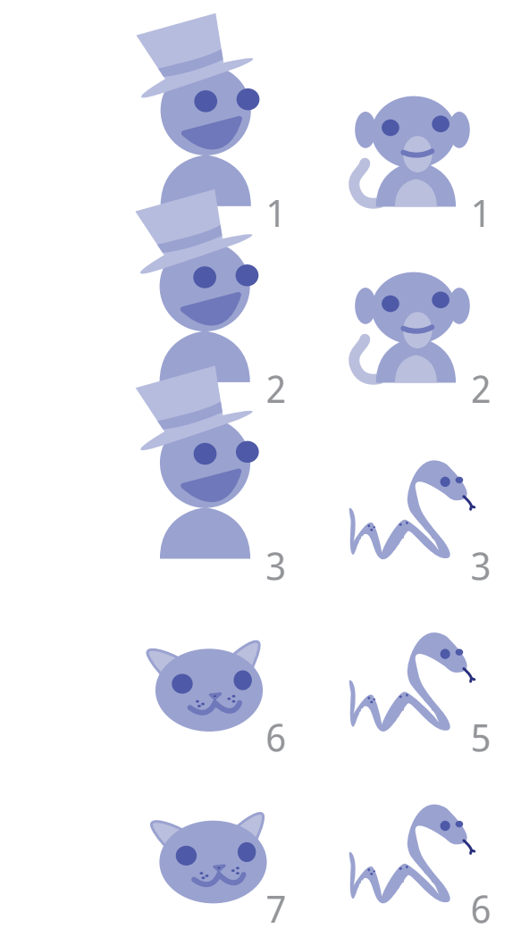
(Human position 3 aligns to Chimp position 3)

2) Check the prediction for the samples (Recall)

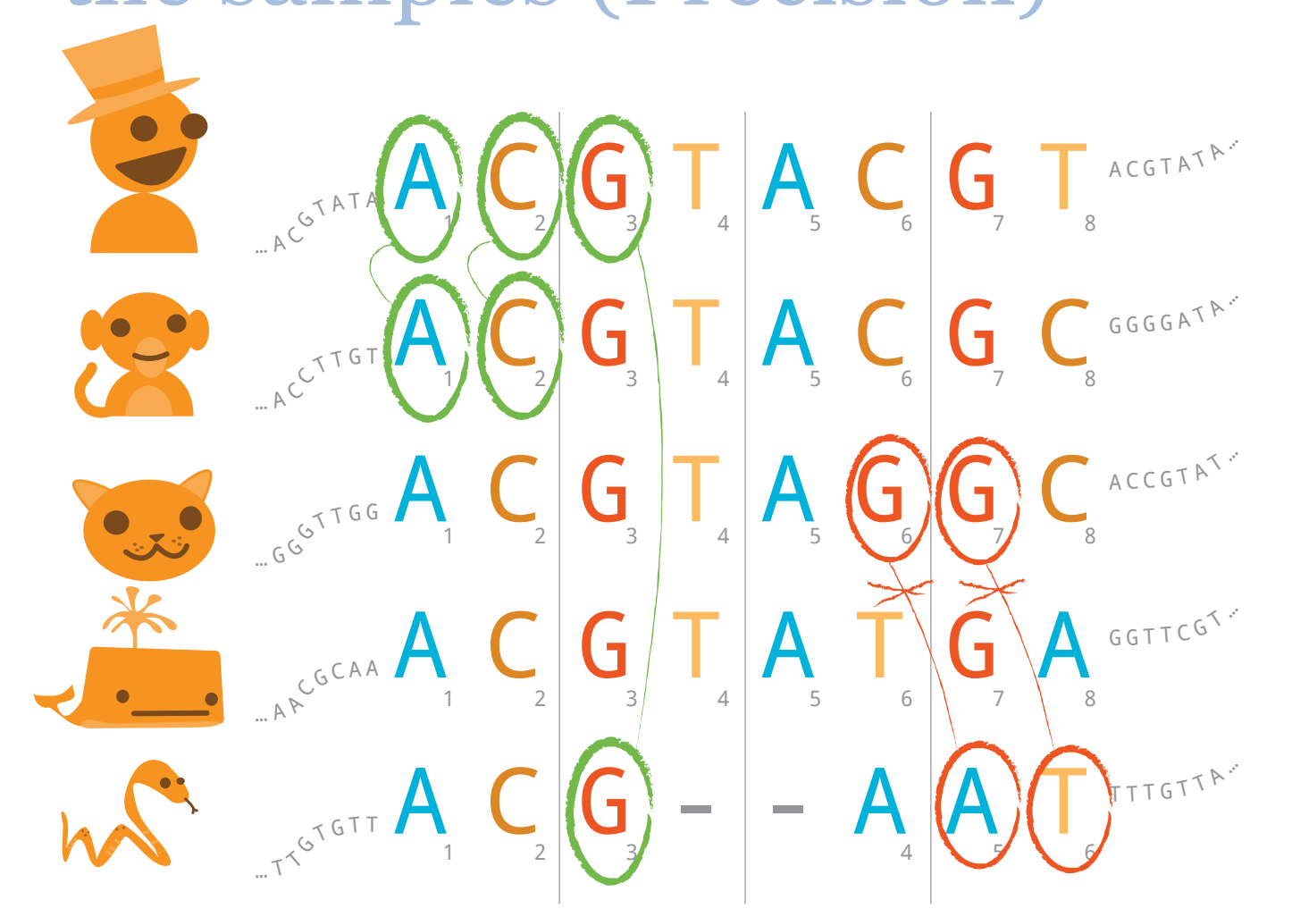


True Positives: 1, False Negatives: 4

3) Sample pairs from the prediction



4) Check the truth for the samples (Precision)



True Positives: 3, False Positives: 2

5) Compute

True Positives: 1
False Negatives: 4
Recall = 0.2

$$\frac{1}{1 + 4}$$

True Positives: 3
False Positives: 2
Precision = 0.6

$$\frac{3}{3 + 2}$$

F-score = 0.3

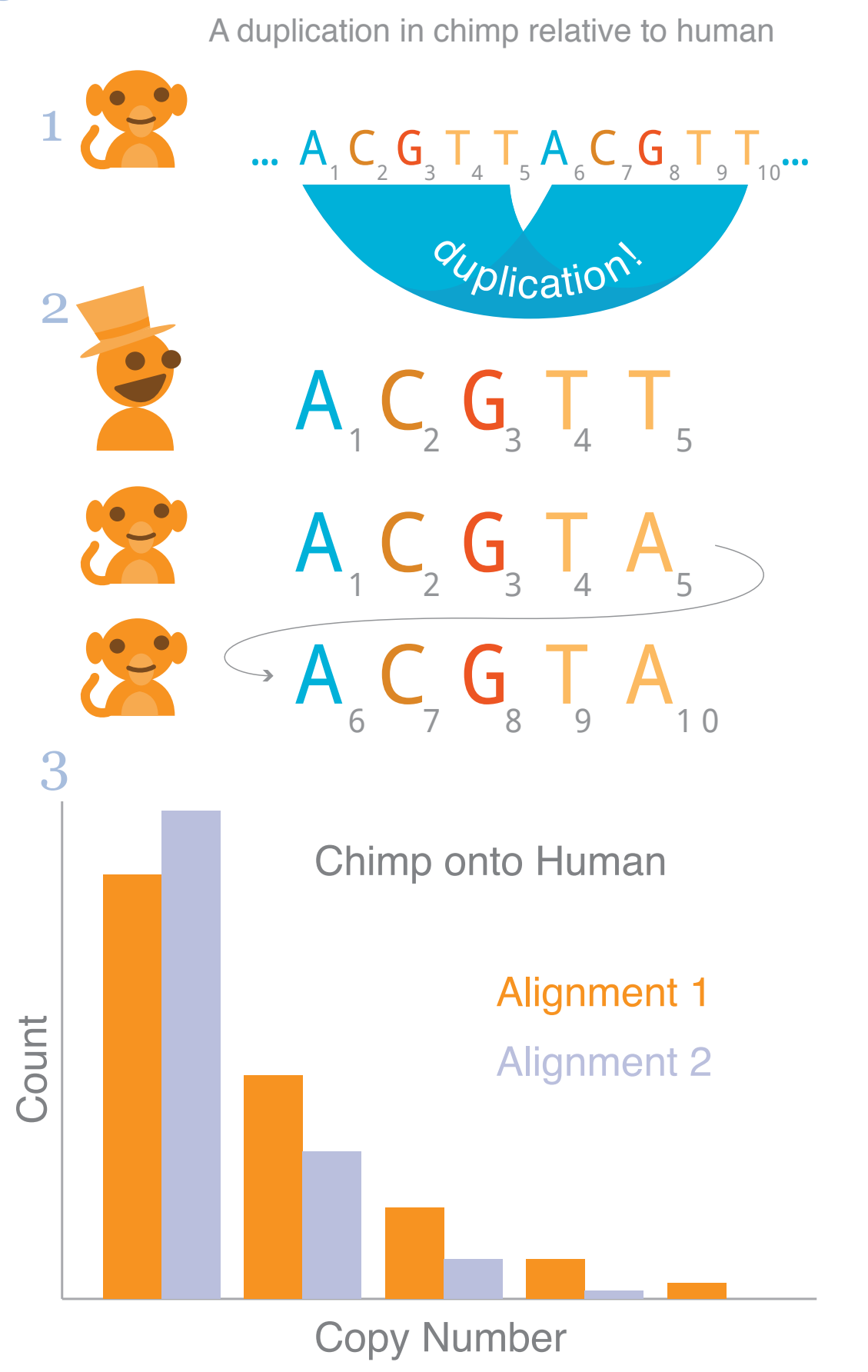
$$\frac{2 * 0.6 * 0.2}{0.6 + 0.2}$$

Other assessments

Coverage and copy number

An individual alignment contains information about how the different species are related to one another. One summary level assessment is to look at the *coverage* of one species onto anthor. Coverage is the number of bases of species B that align to a region of species A.

Occassionally DNA sequences are duplicated in genomes. These duplications can be observed in alignments. This property is sometimes refered to as *copy-number* because there are different numbers of a region (copies) in different species.

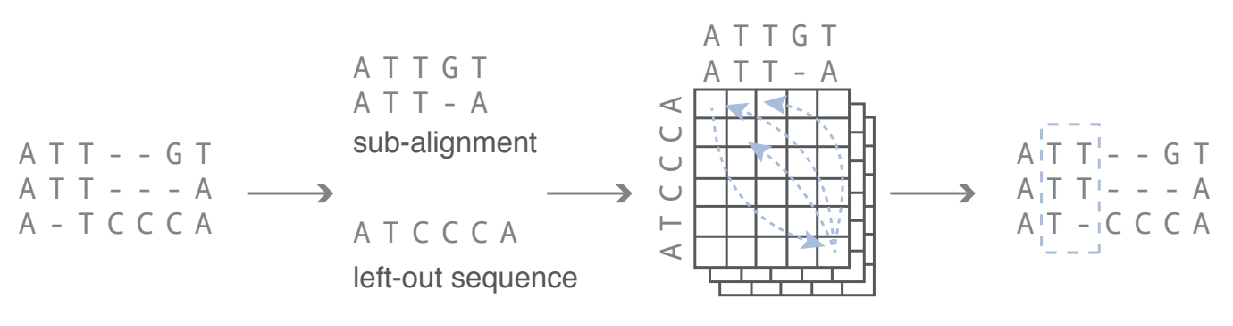


Statistical assessments

If we make assumptions about the nature of evolution it is possible to estimate how well a given alignment (or region within an alignment) conforms to the assumptions.

PSAR

PSAR (Probabilistic sampling-based alignment reliability) samples suboptimal alignments and computes an agreement score based on how often the suboptimal alignments match the original



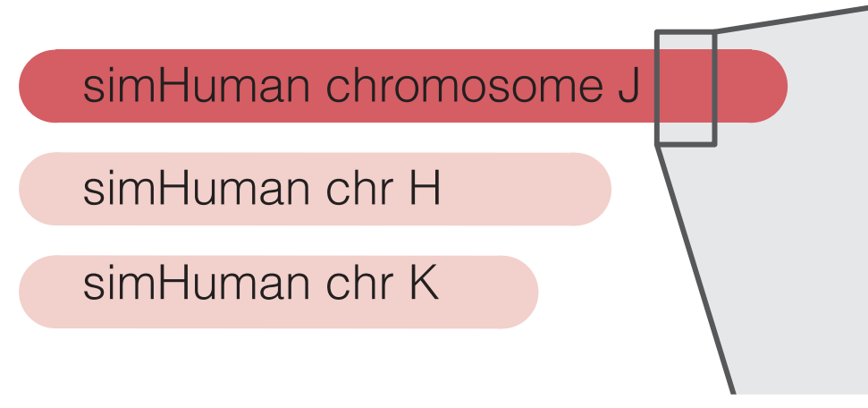
From Figure 1, Kim and Ma. Nucleic Acids Res (2011) vol. 39 (15) pp. 6359-68

Alignathon: assessing the state of the art in WGA

A collaborative project

The Alignathon is a collaborative project to assess whole genome aligners and promote development of the field of whole genome alignment. Alignathon is insprired by the Assemblathon project, (an iterative collaborative competition to assess the state of the art in de novo genome assembly.)

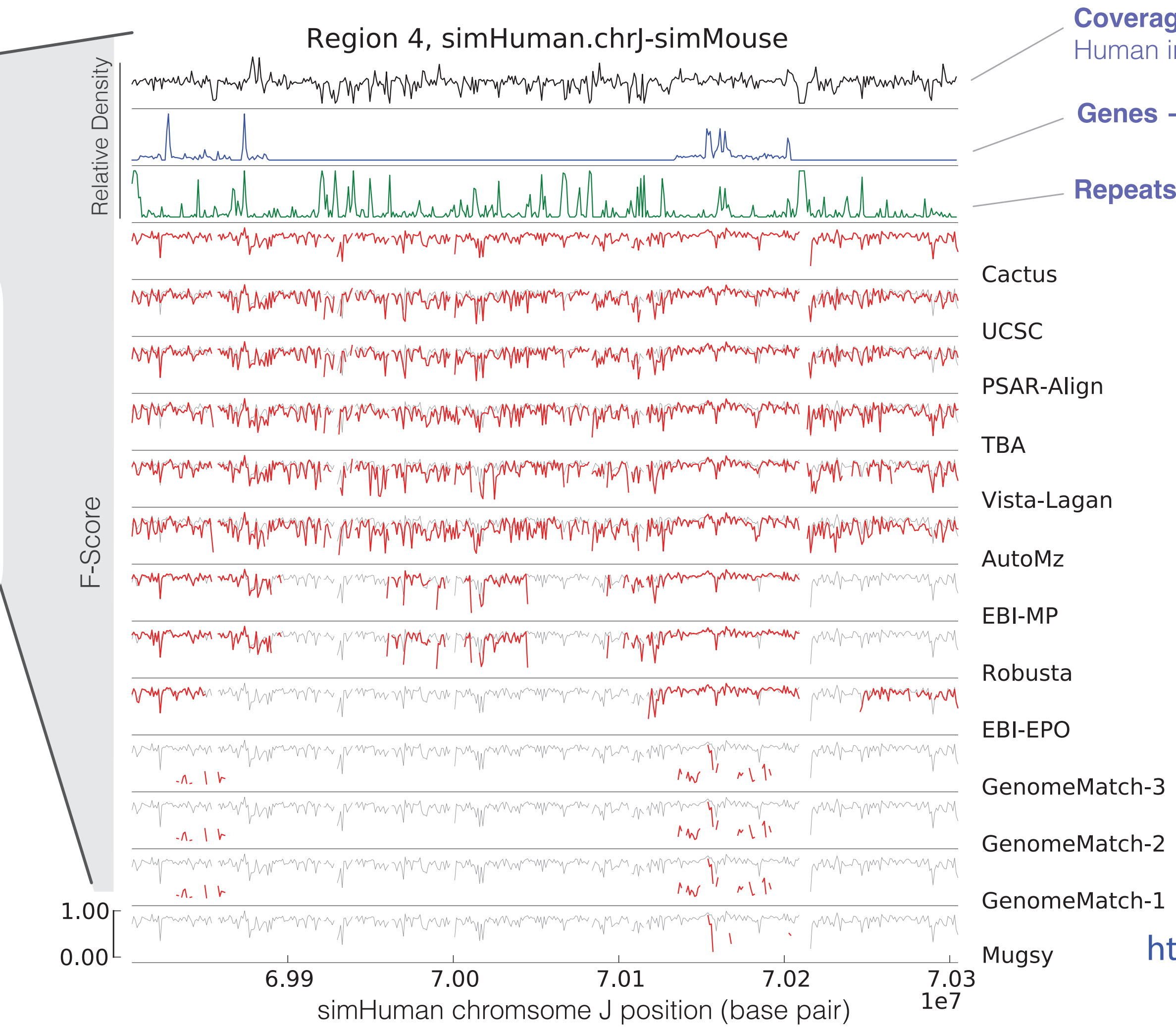
As more and more genomes are sequenced to fill out the phylogeny of vertebrates in small independent projects and in larger coordinated efforts like **Genome 10K**, there is going to be a strong desire to know the evolutionary relationships between the genomes. There is coming a time quickly when whole genome alignment is going to be very much in demand.



With that near future in mind we organized a collaborative project to assess whole genome aligners. We designed a test suite comprised of three problems. We created two sets of simulated genomes and one real data set comprised of the 20 fly genomes.

The combination of using the most advanced genome evolution simulator available along with a set of real genomes provides a test suite of similar sized problems for aligners and a way to compare metrics that don't require a known truth to those that do.

http://compbio.soe.ucsc.edu/alignathon/

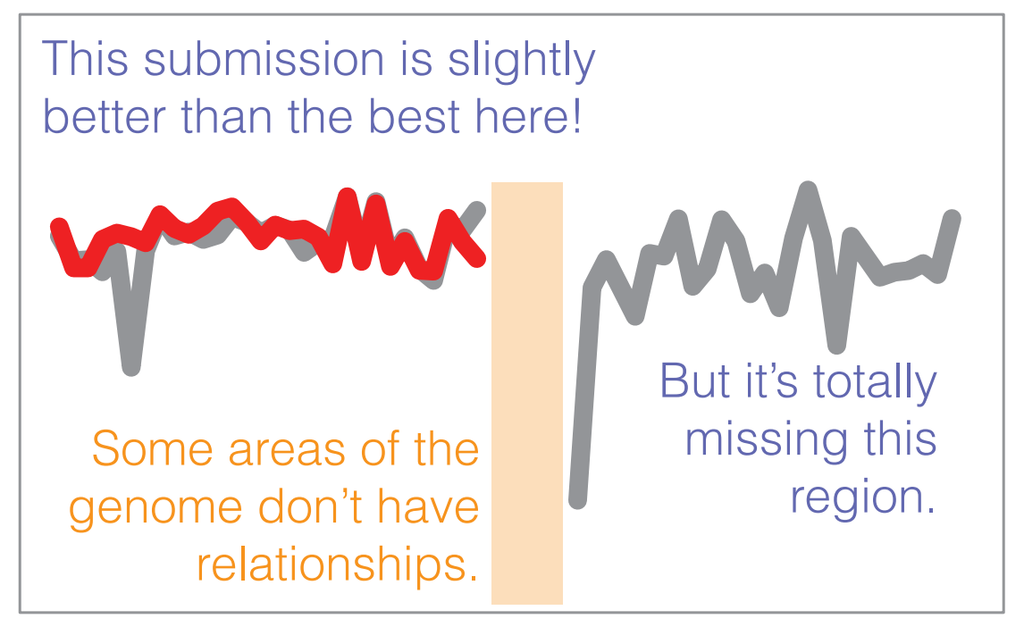


Coverage — How much of Mouse aligns to Human in this region?

Genes — Where are the gene dense areas?

Repeats — Where are the repetitive areas of DNA?

Submisisons — Ordered by overall F-Score in the region. The RED line is the submission. The GREY line is the BEST submission (Cactus).



http://users.soe.ucsc.edu/~dearl/posters/
earld@google.com
dearl@soe.ucsc.edu

7 July 2013 PhD Intern Poster Session @ Google