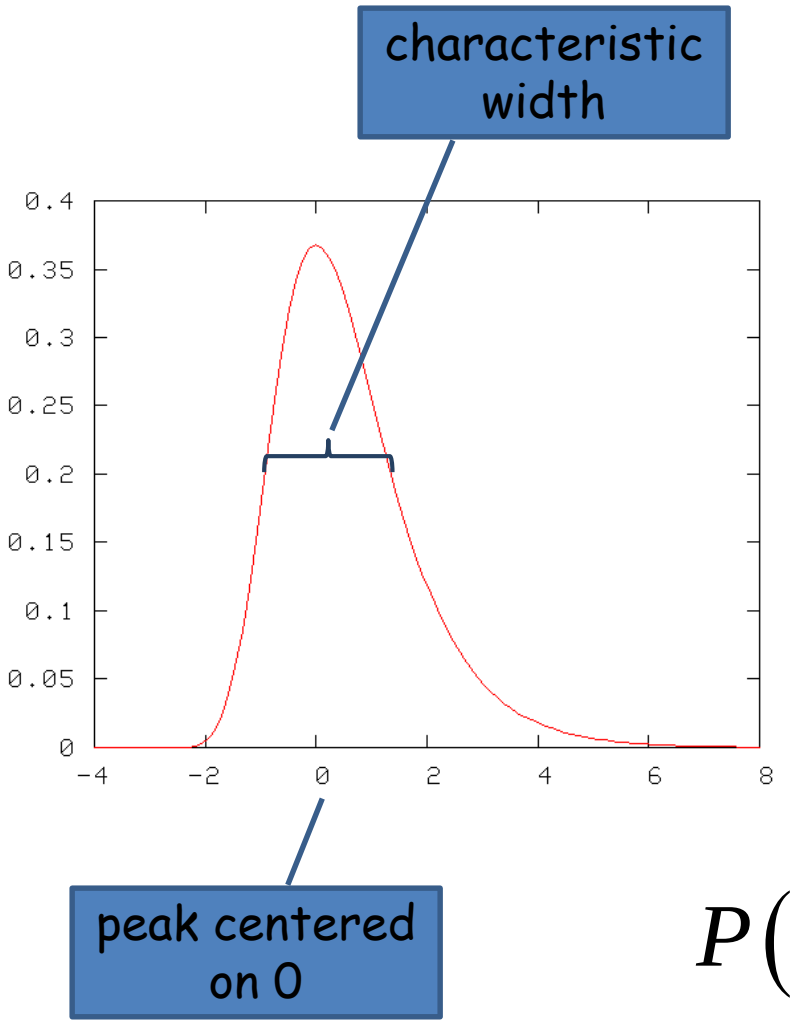
Whole genome alignments

http://faculty.washington.edu/jht/GS559_2017/

Genome 559: Introduction to Statistical and Computational Genomics

Prof. James H. Thomas

Extreme value distribution



$$P(S \geq x) = 1 - e^{(-e^{-x})}$$

S is data score, x is test score

$$P(S \geq x) = 1 - e^{(-e^{-\lambda(x-\mu)})}$$

S is data score, x is test score, μ is mode, λ is width

Summary score significance

- Most statistical tests compare observed data to the expected result according to a null hypothesis.
- Sequence similarity scores of unrelated sequences follow an extreme value distribution, which is characterized by a long tail.
- The p-value associated with an alignment score is the area under the curve to the right of that score.
- The E-value is derived from the p-value accounting for multiple testing. It is the expected number of times that a given score would appear in a randomized database.

Whole genome alignments

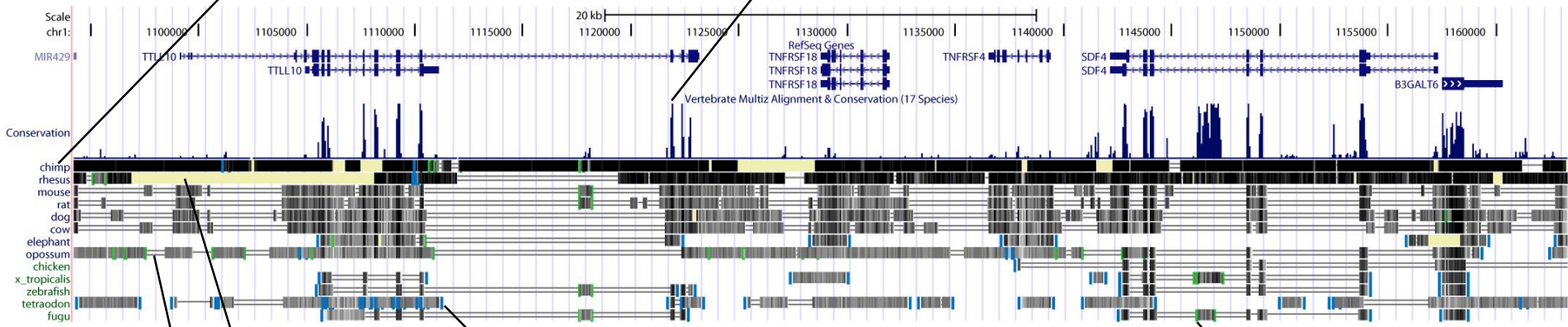
Why?

- genome-wide alignment data (efficient)
- inference of shared (orthologous) genes across species
- genome evolution
- curiosity (an under-appreciated motivation)

UCSC Browser track

individual genome alignments, darker = higher scoring

averaged conservation for 17 genomes

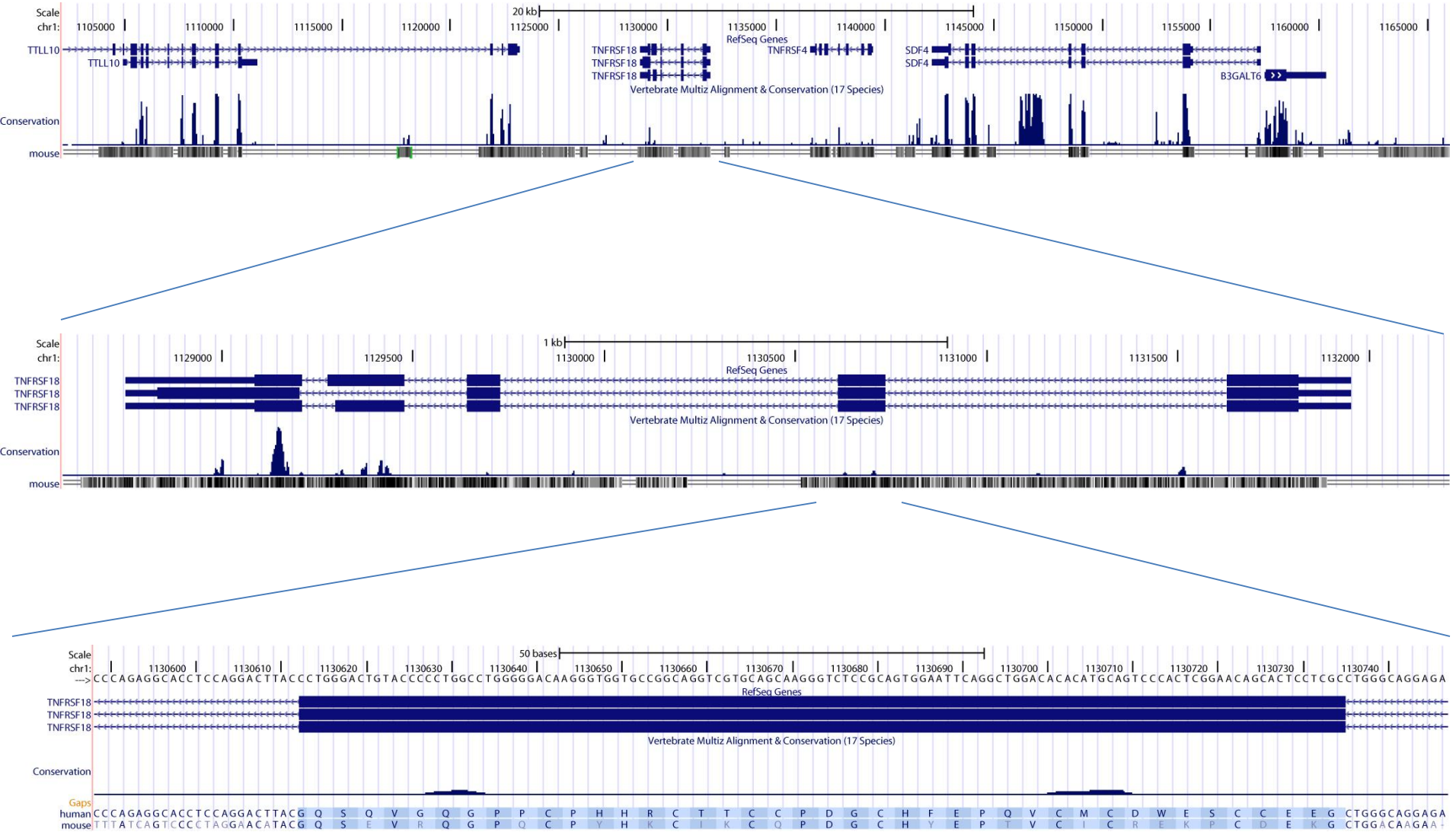


known gap in assembly

alignment discontinuity (e.g. translocation break point)

questionable alignment segment

= sequence present but unalignable



GQSQVGQGPPCPHHRCTTCCPDGCHFEPQVCMCDWESCCEEG

GQSEVRQGPQCPYHKCIKCQPDGCHYEPTVCICREKPCDEKG

How are genome-wide alignments made?

- mouse and human genomes are each about 3×10^9 nucleotides.

- how many calculations would a dynamic programming alignment have to make?

- at a minimum - 3 integer additions and 3 inequality tests for each DP matrix position

- DP matrix size is 3×10^9 by 3×10^9

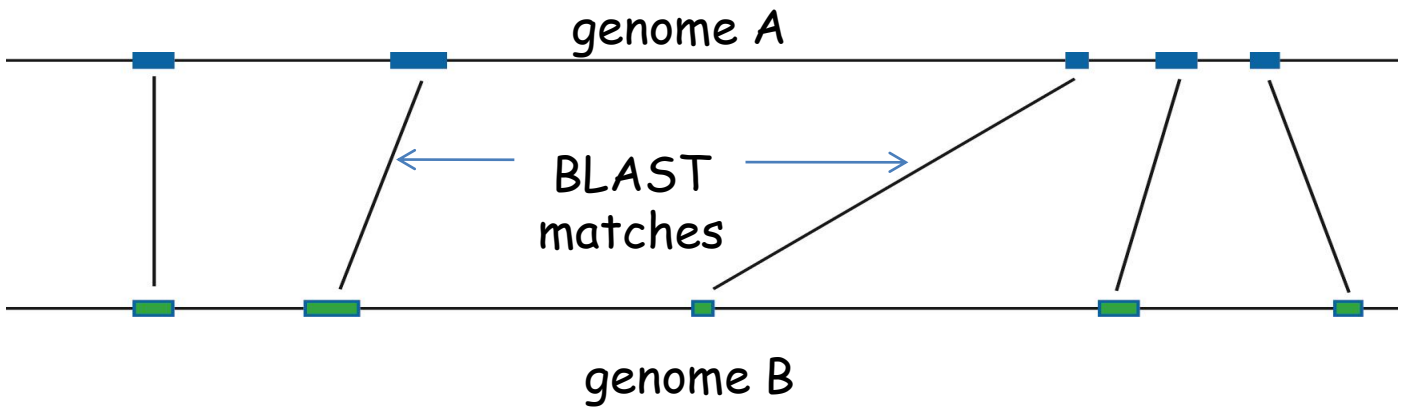
- about $6 \times (3 \times 3 \times 10^{18}) = 5.4 \times 10^{19}$ calculations!

  Age of the universe is about 4.3×10^{17} seconds

(there are other big problems too, including assuming colinearity)

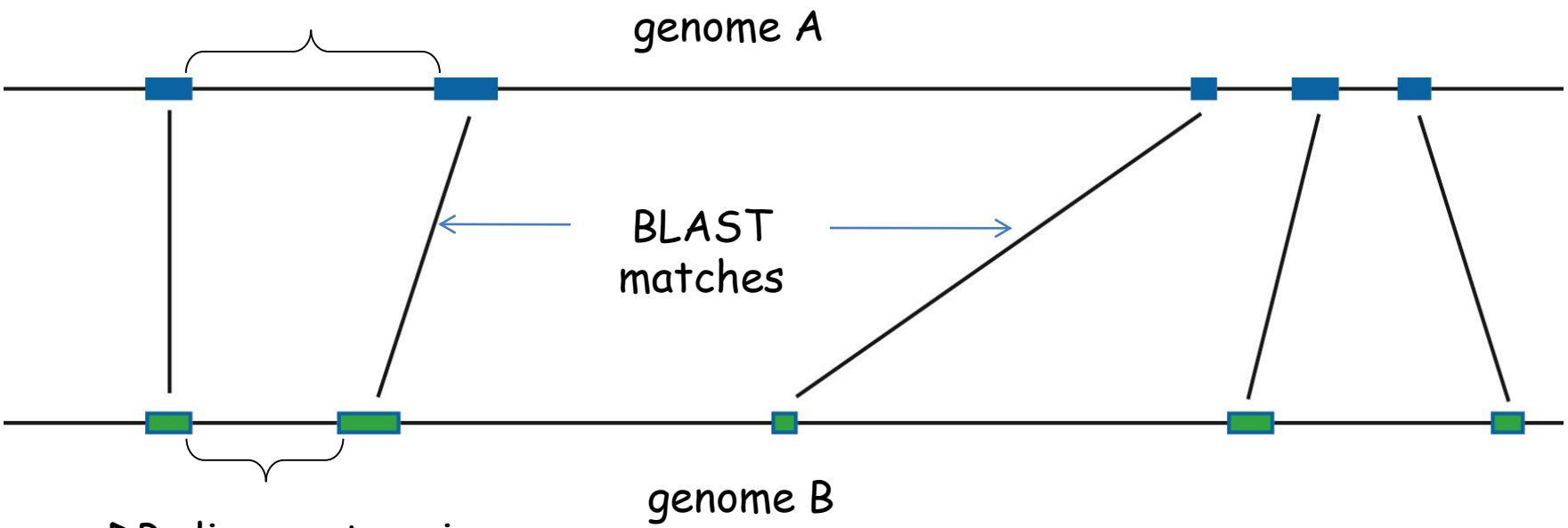
BLAST whole genome against another

- Runtime (my desktop) for mouse vs. human, about 24 hours*
- Extract best match segments, reverse blast
- Keep reciprocal best match regions as anchors (the full process is a bit more involved and includes anchor collinearity).
- Schematic of part of results:

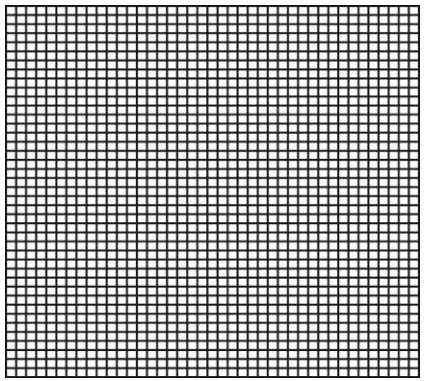


* megablastn with repeat-masked human genome

Dynamic programming after BLAST matching



DP alignment region



M x N manageable

Anchored DP alignment: if two reciprocal best blast matches are nearby and in the same orientation, DP align everything between them.