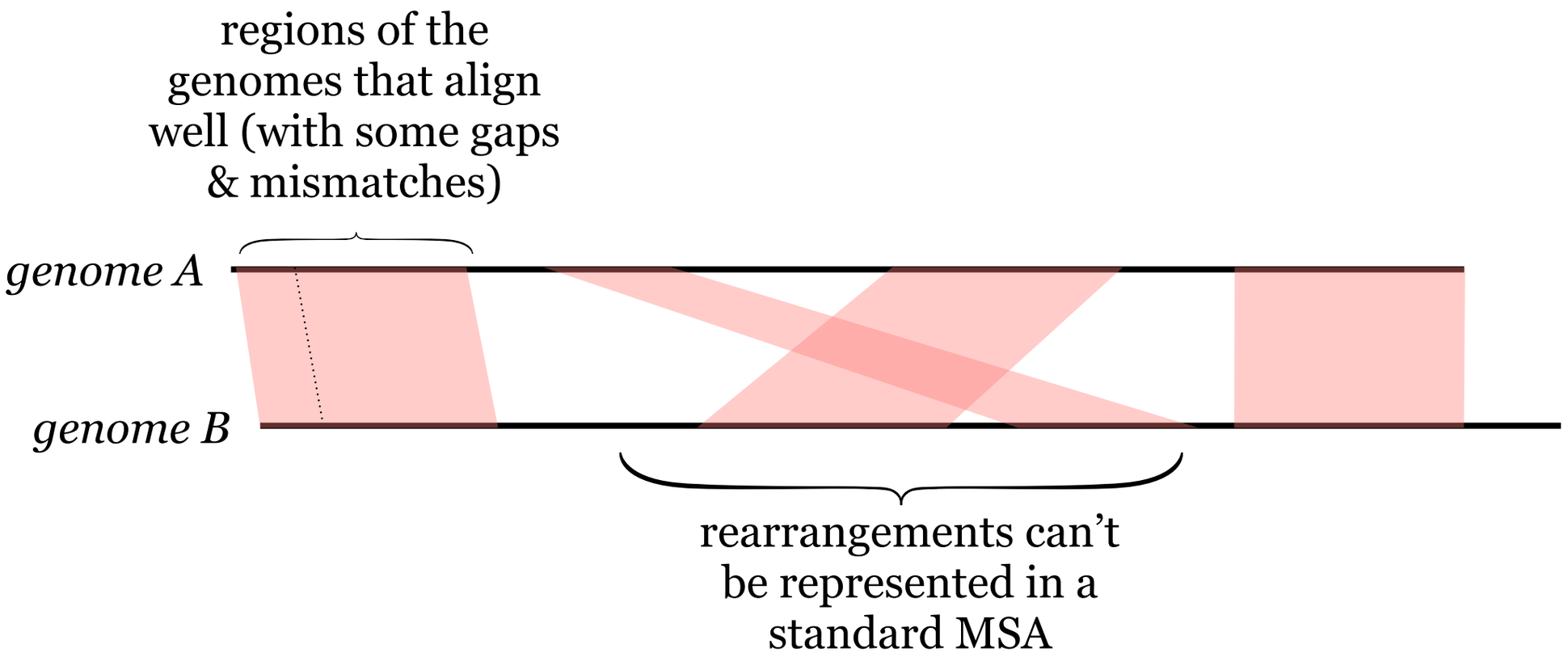
Alignment of Entire Genomes (MUMmer)

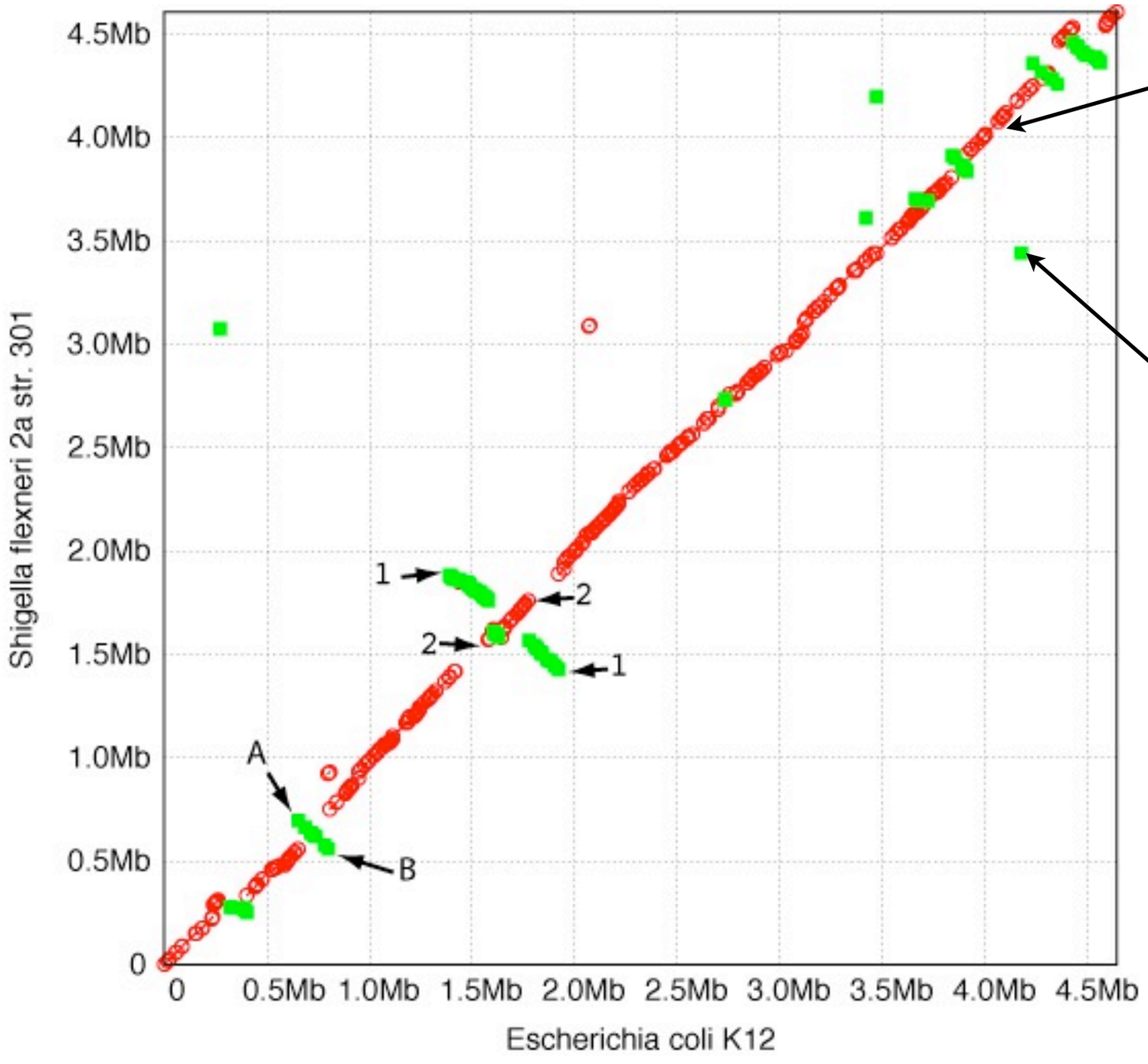
CMSC 423
Carl Kingsford

Challenges aligning whole genomes



- Aligning two sequences of 130 million letters isn't feasible using an $O(n^2)$ algorithm.

Dot Plots

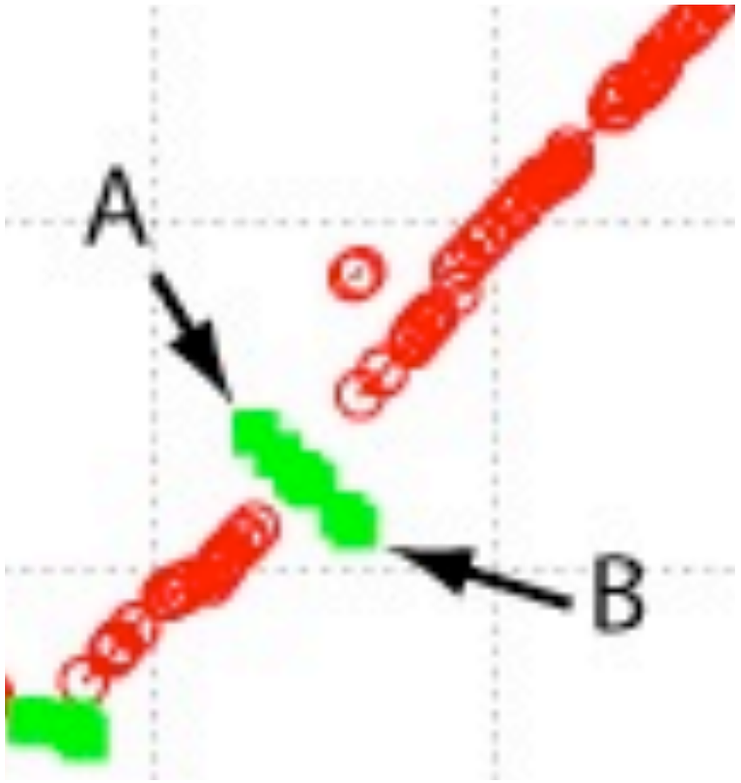


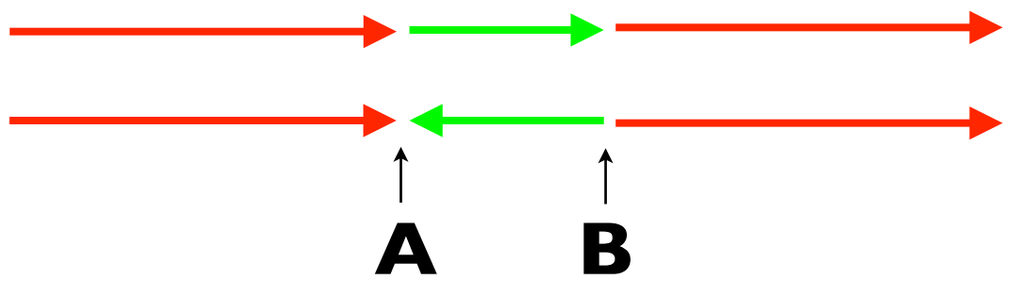
short region that is equal in both genomes

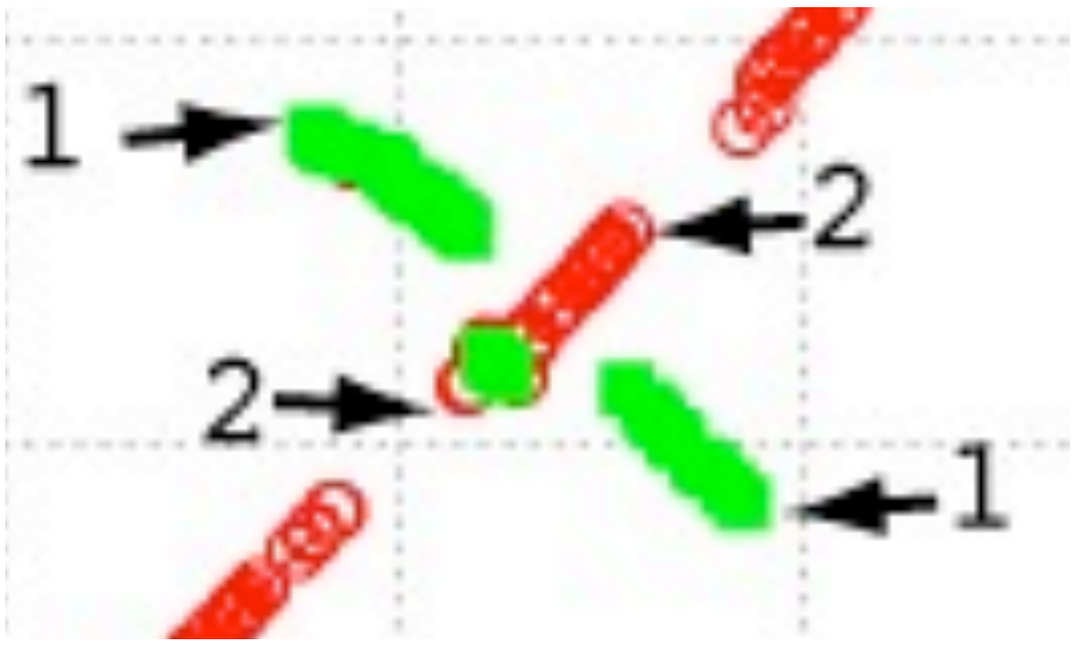
green means the sequence in one genome matches the *reverse complement of the sequence in the other*

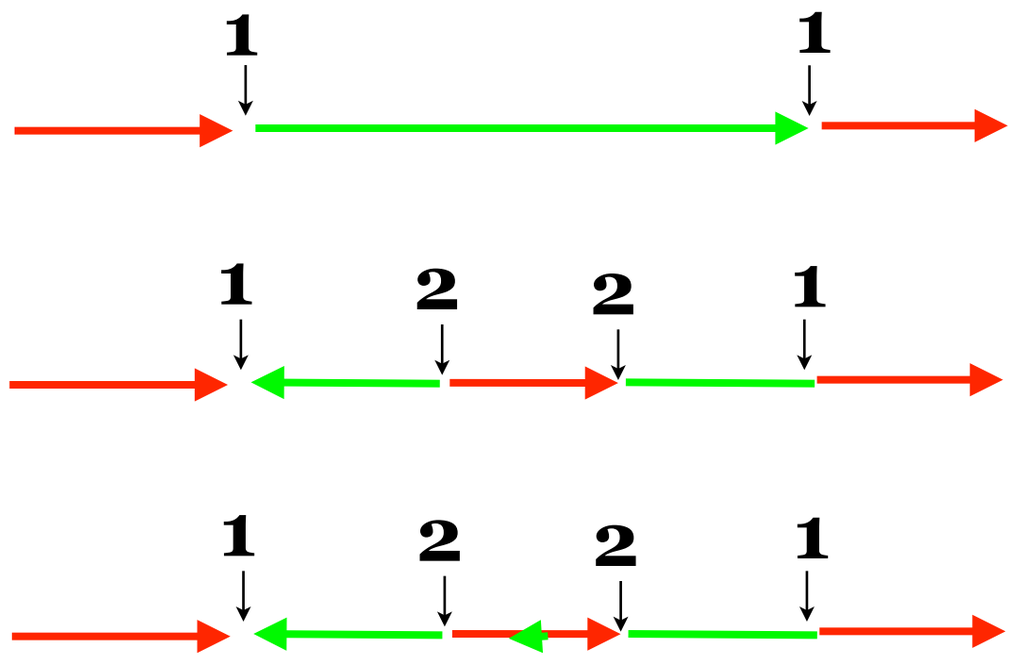
s = ACCGGTG

rc(s) = CACCGGT

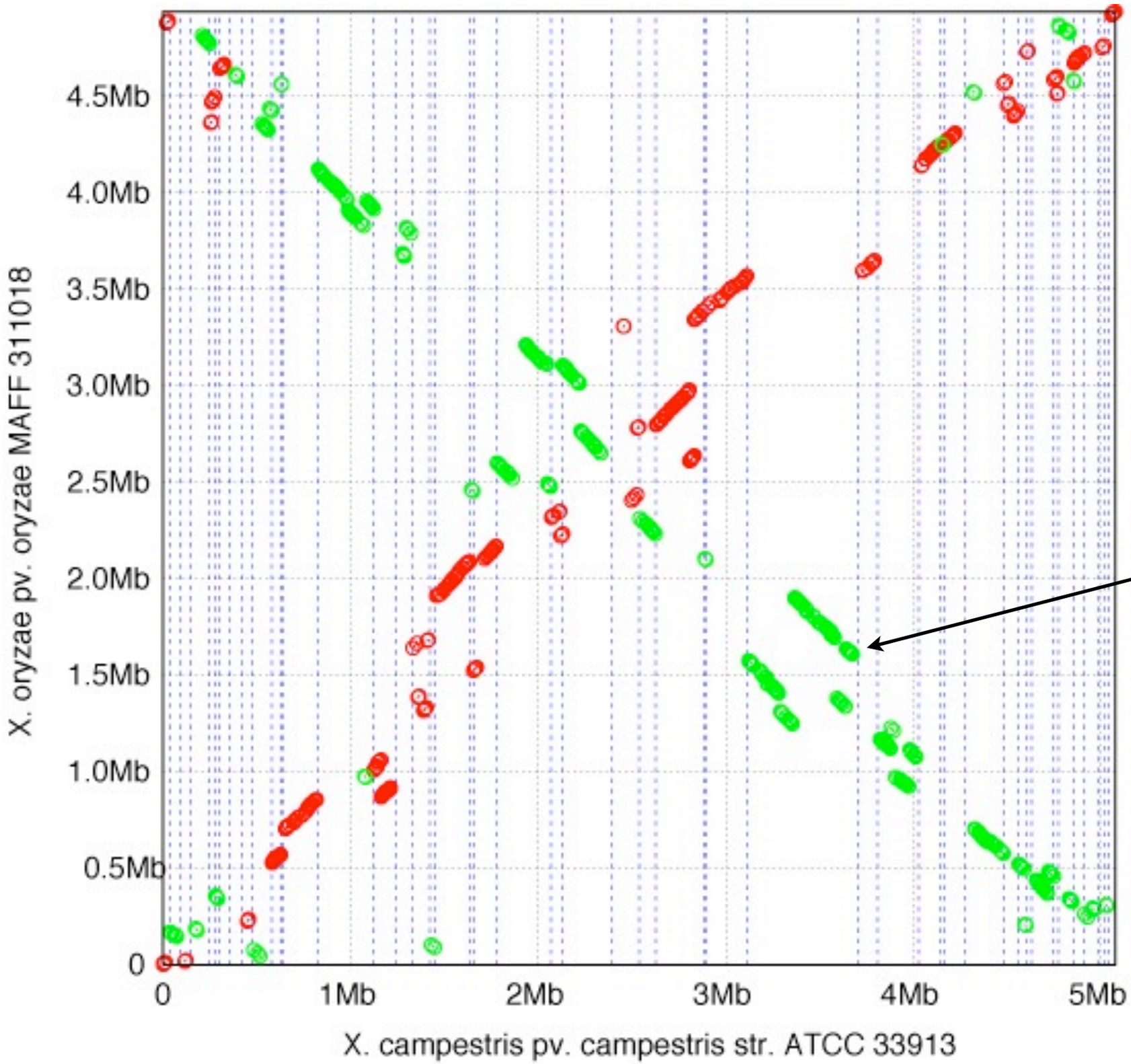

A
B


A
B


1
2
2
1

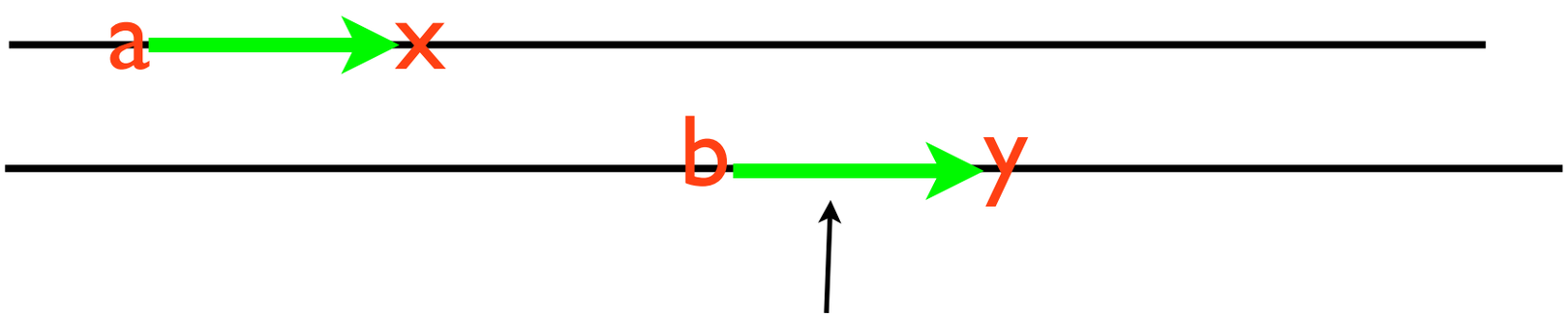

1
1
1
2
2
1
1
2
2
1

More Distant Organisms



Shifts from the diagonal correspond to regions that have moved.

Alignment Anchors



Maximum Unique Match (MUM): a region that

- matches exactly between the two genomes, and
- exists exactly once in each genome, and
- is not contained in a longer such region

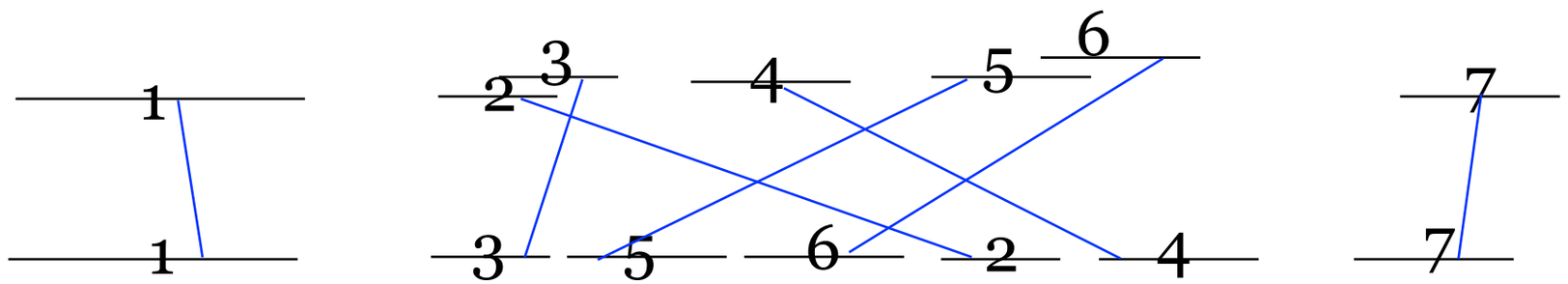
The idea is that these unique, exact matches should almost always be aligned in the true alignment.

(We'll see how to find these regions efficiently soon.)

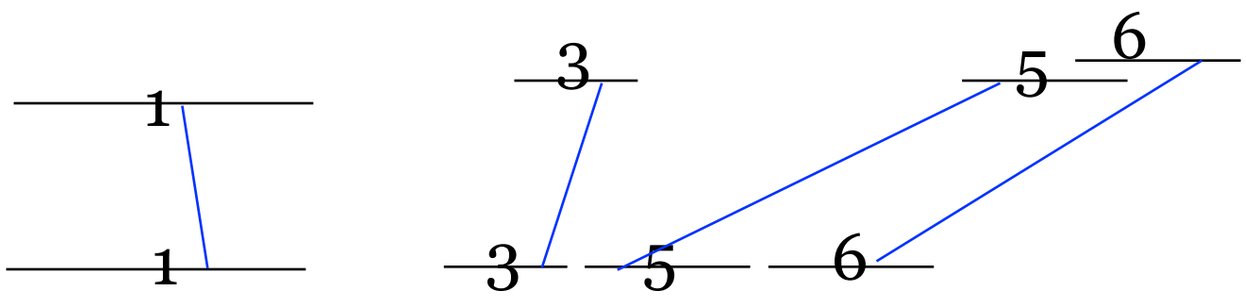
MUMmer

Delcher et al. "Alignment of whole genomes", *Nuc. Acids Res., 1999*

- Find all the MUMs between the two sequences
- Find the longest sequence of MUMs in a consistent order.

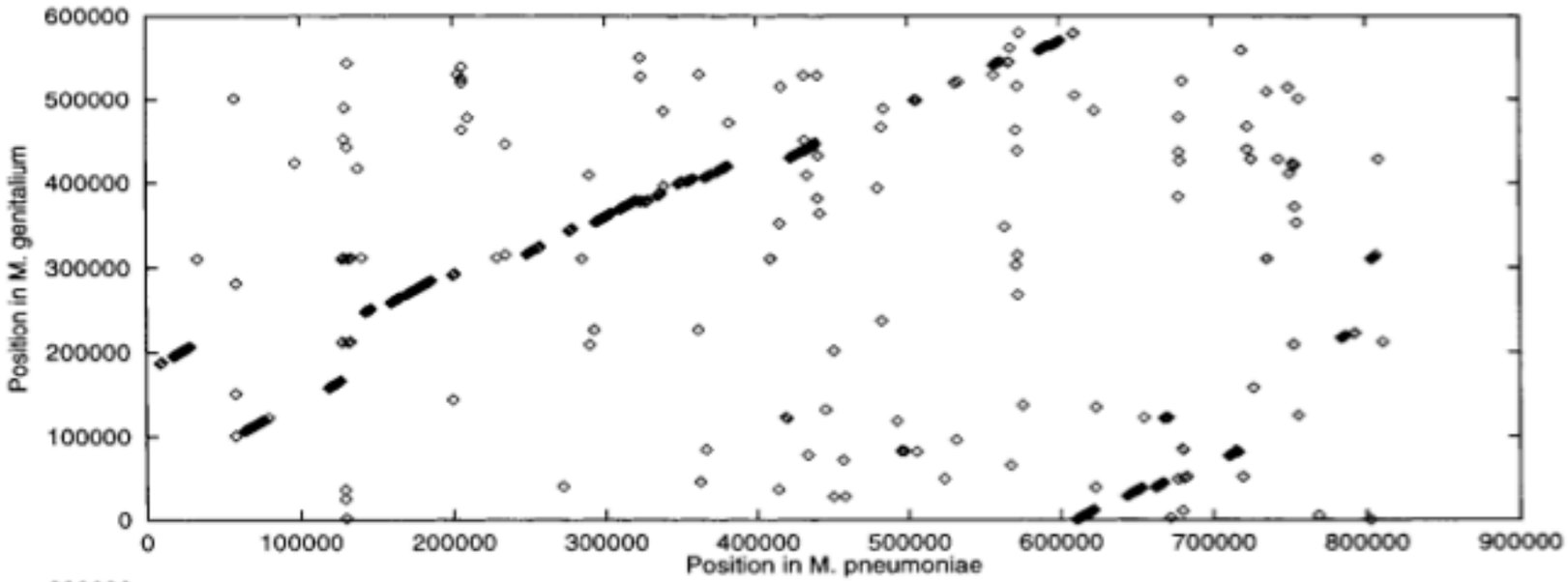


1. Label the MUMs in order in genome A.
2. Find the **longest sequence** of increasing MUM numbers in B.

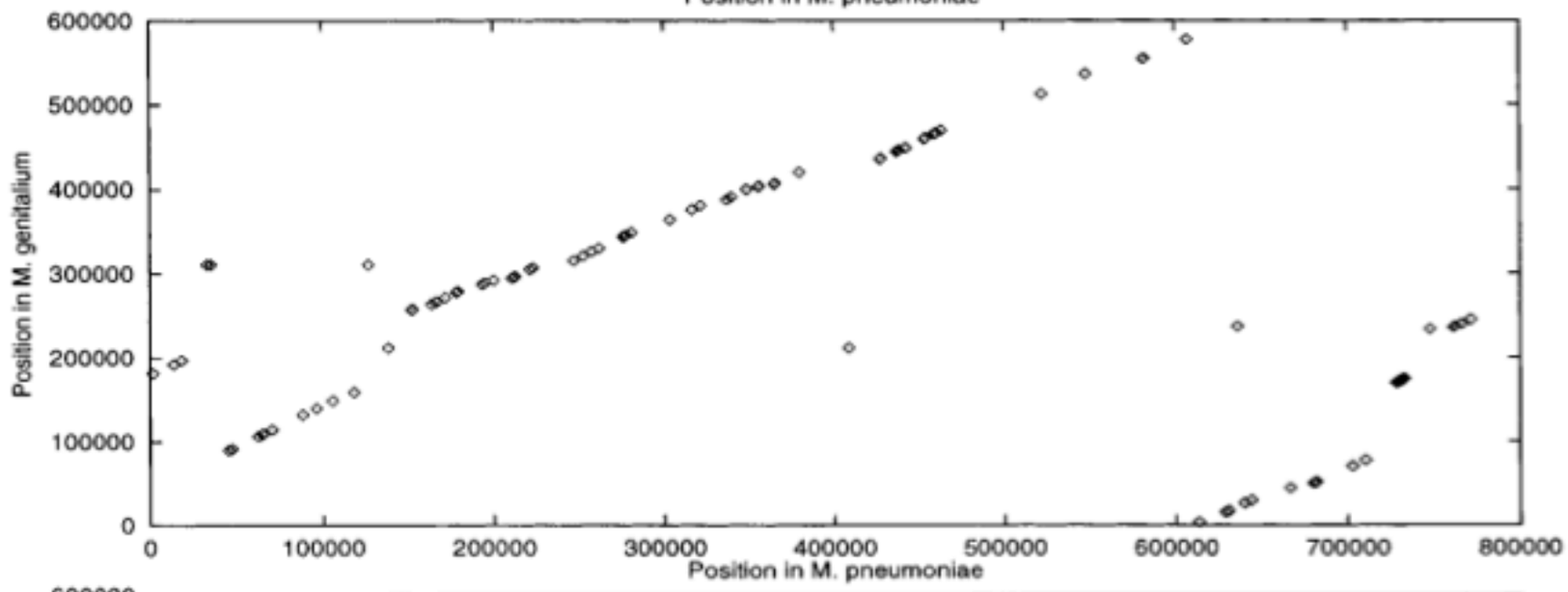


1, 3, 5, 6, 7

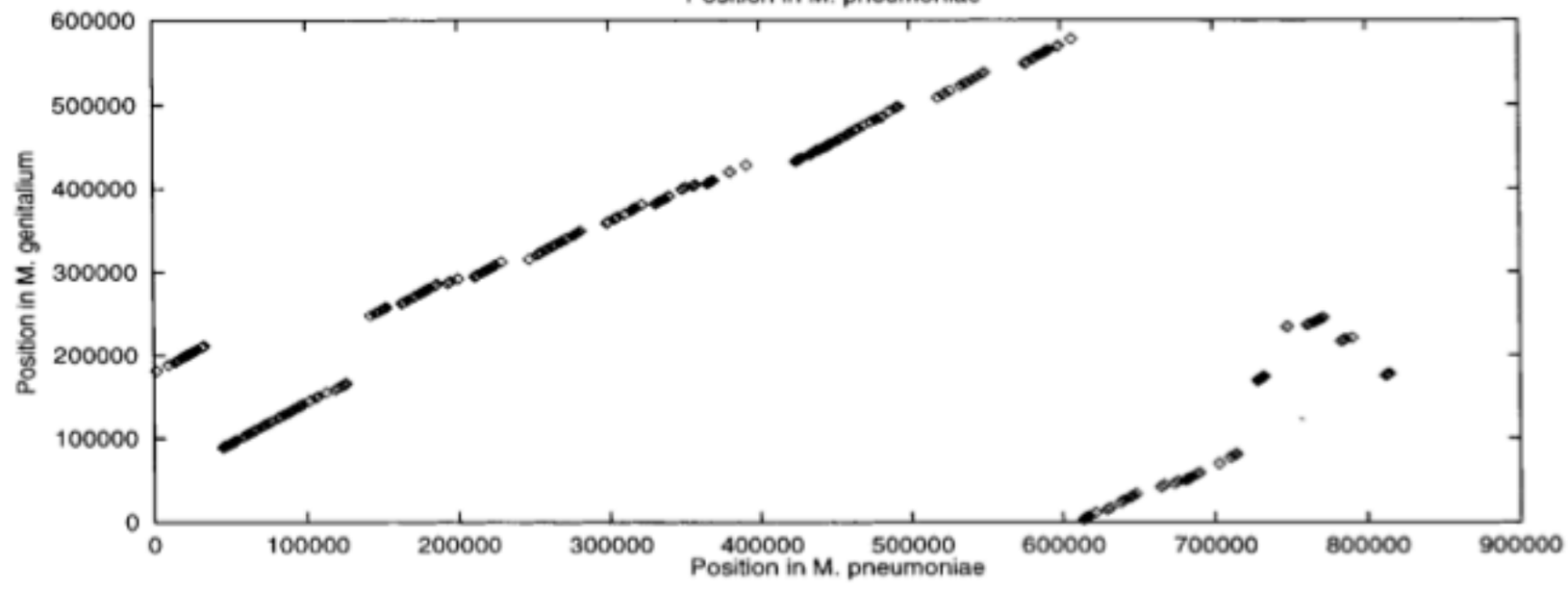
- Compute pairwise alignment only between adjacent MUMs.



All pairwise alignment between 1000bp regions

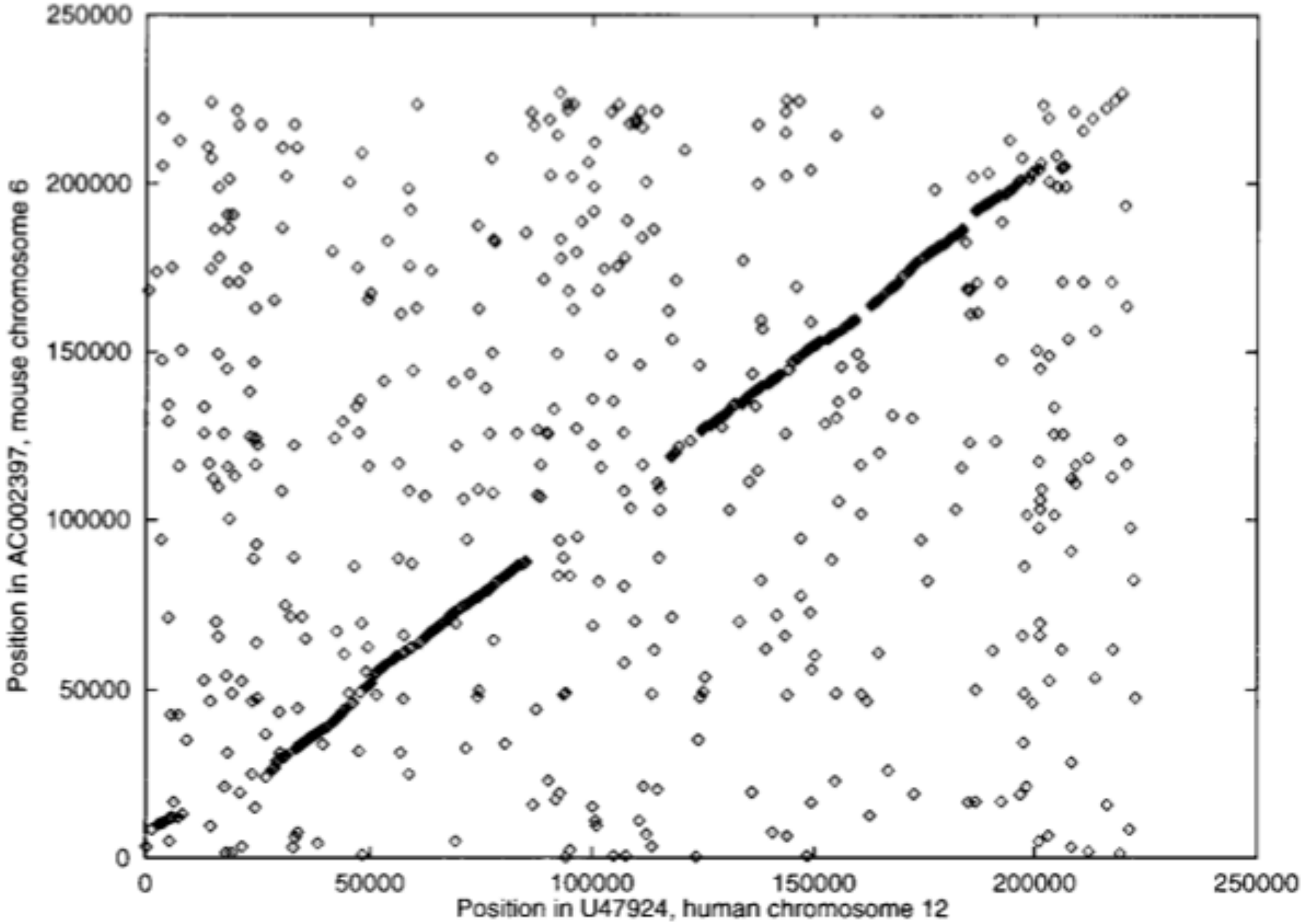


Exactly matching kmers



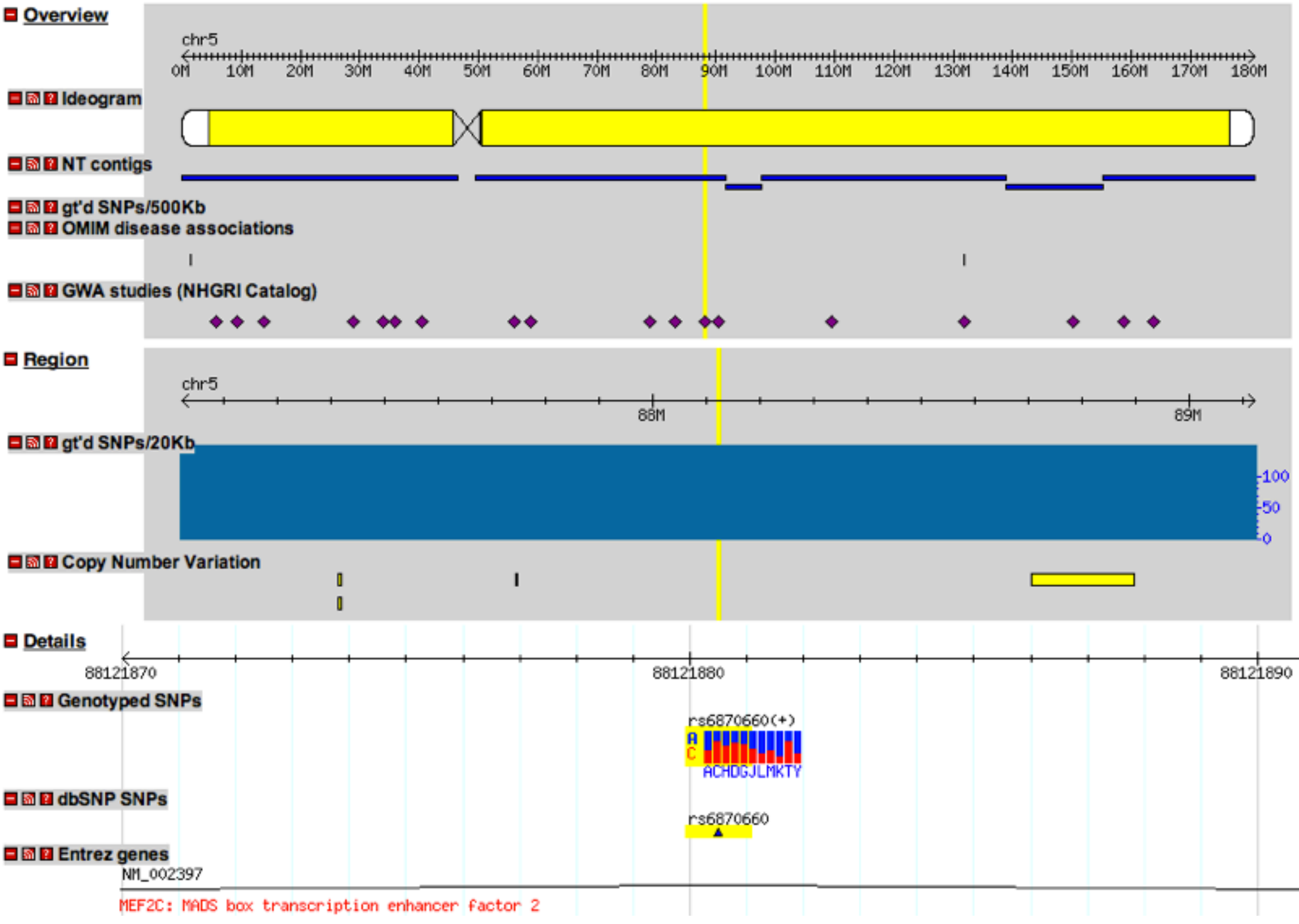
MUMmer

Human vs. Mouse



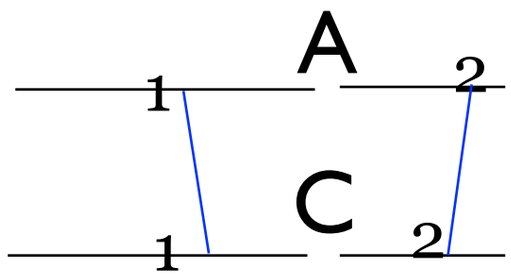
SNPs - Single Nucleotide Polymorphisms

Population descriptors:ASW: African ancestry in Southwest USA, **CEU**: Utah residents with Northern and Western European ancestry from the CEPH collection, **CHB**: Han Chinese in Beijing, China, **CHD**: Chinese in Metropolitan Denver, Colorado, **GIH**: Gujarati Indians in Houston, Texas, **JPT**: Japanese in Tokyo, Japan, **LWK**: Luhya in Webuye, Kenya, **MEX**: Mexican ancestry in Los Angeles, California, **MKK**: Maasai in Kinyawa, Kenya, **TSI**: Tuscan in Italy, **YRI**: Yoruban in Ibadan, Nigeria.



Finding SNPs in MUMmer alignments

SNPs will usually appear as 2 MUMs separated by a single base:



Summary

- Aligning whole genomes requires algorithms even faster than $O(n^2)$.
- It also requires being able to handle inversions, rearrangements, and transpositions.
- MUMmer does this by anchoring the alignment using maximal unique matches (MUMs) as anchors.
- **Finding** the MUMs is an interesting problem in its own right.
  - The solution uses a data structure called *suffix trees*, which we study next.