Whole Genome Alignment

Adam M. Phillippy

CMSC701: April 18, 2019

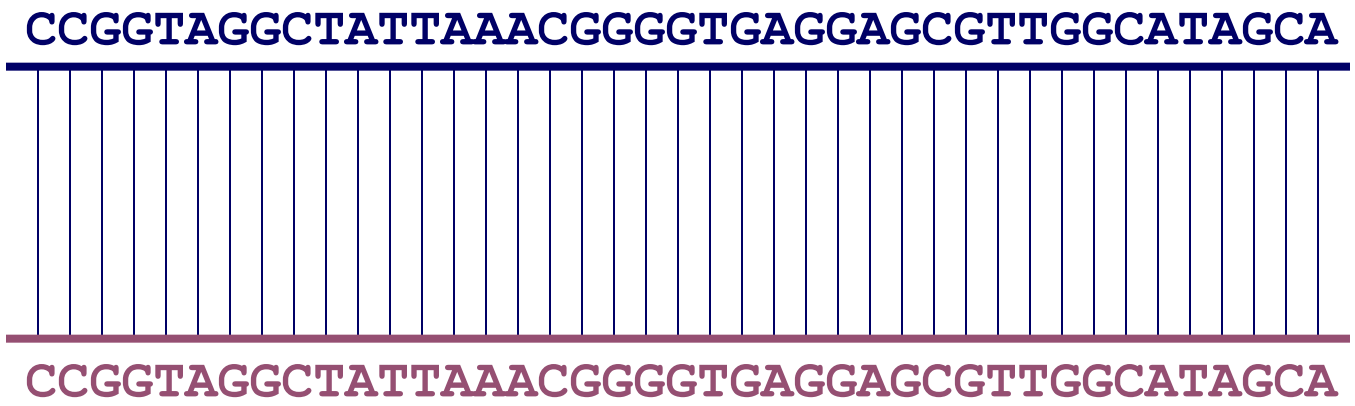
@aphillippy

National Human Genome Research Institute

The Forefront of Genomics®

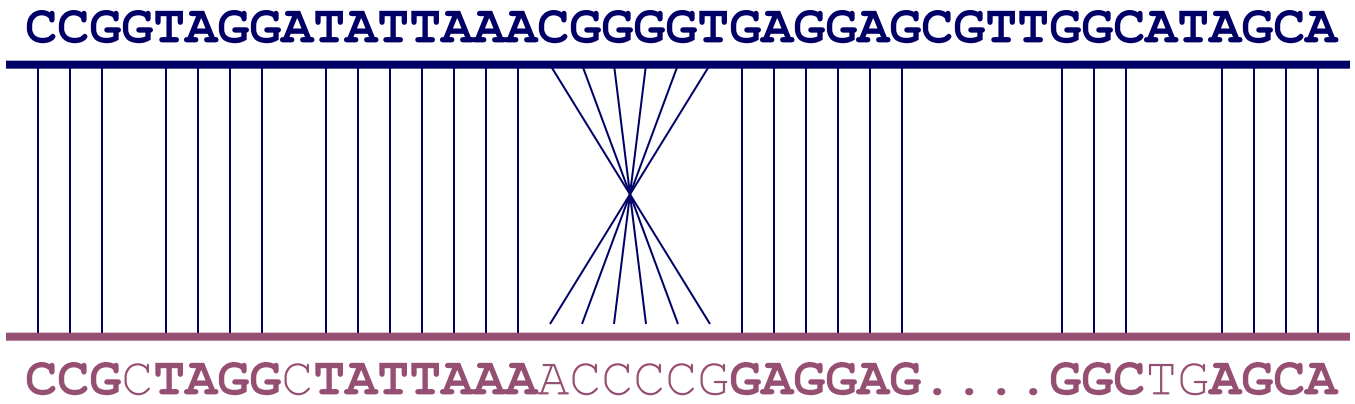
Goal of whole genome alignment

- For genomes *A* and *B*, find a mapping from each position in *A* to its corresponding position in *B*



Not so fast…

- Genome *A* may have insertions, deletions, translocations, inversions, or duplications with respect to *B*

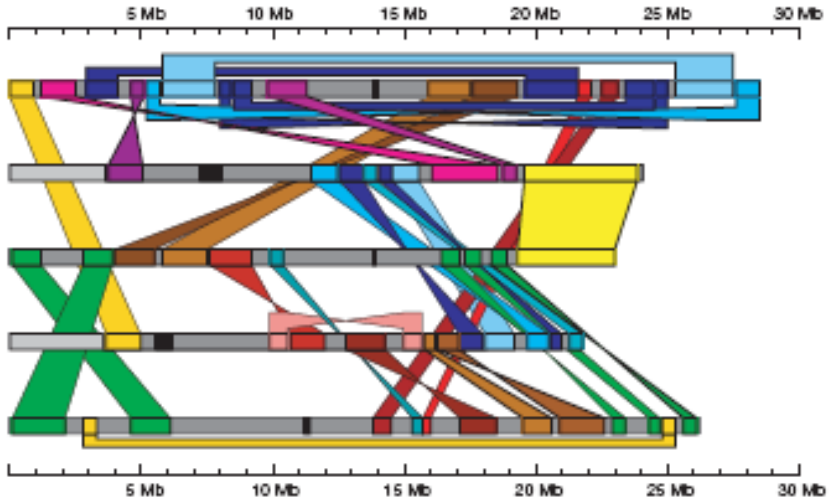


Global vs. local alignments

- Global pairwise alignment

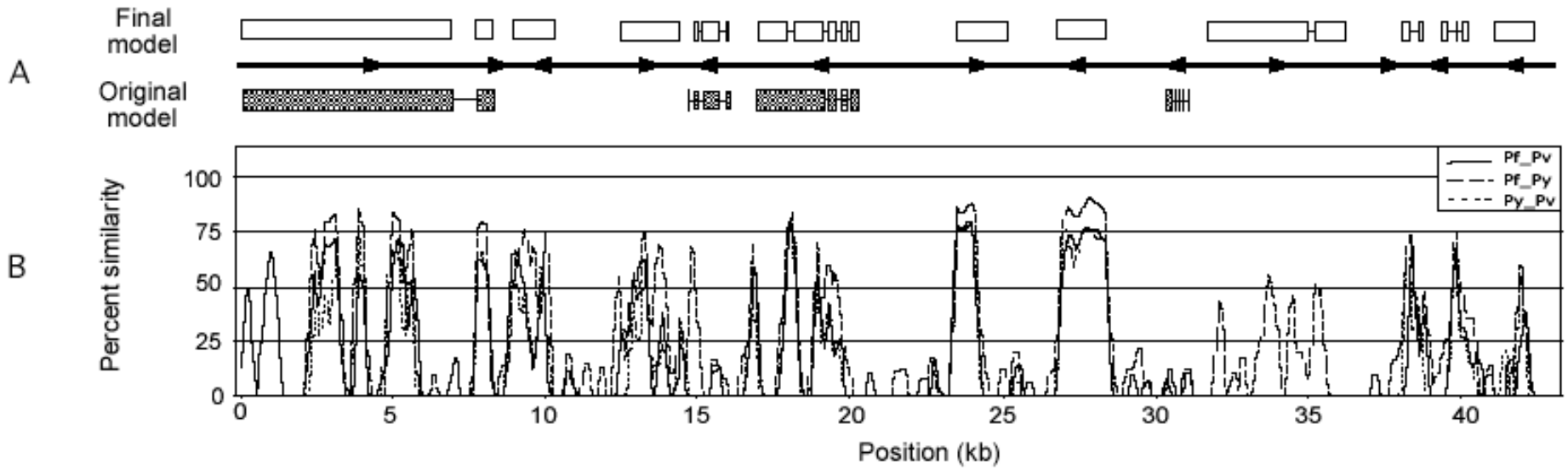
```
...AAGCTTGGCTTAGCTGCTAGGGTAGGCTTGGG...
...AAGCTGGGCTTAGTTGCTAG..TAGGCTTTGG...
           ^        ^       ^^        ^
```

- Whole genome alignment
  - A “collection” of local alignments



E.g. Segmental duplications in Arabidopsis (ca. 2000)

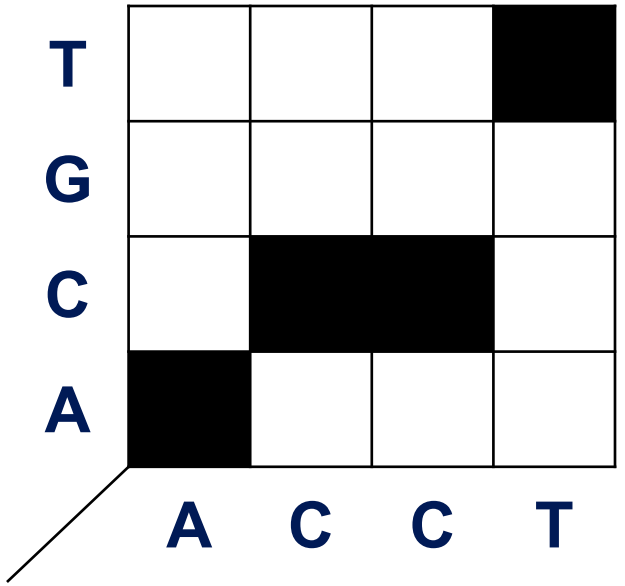
Global alignment visualization



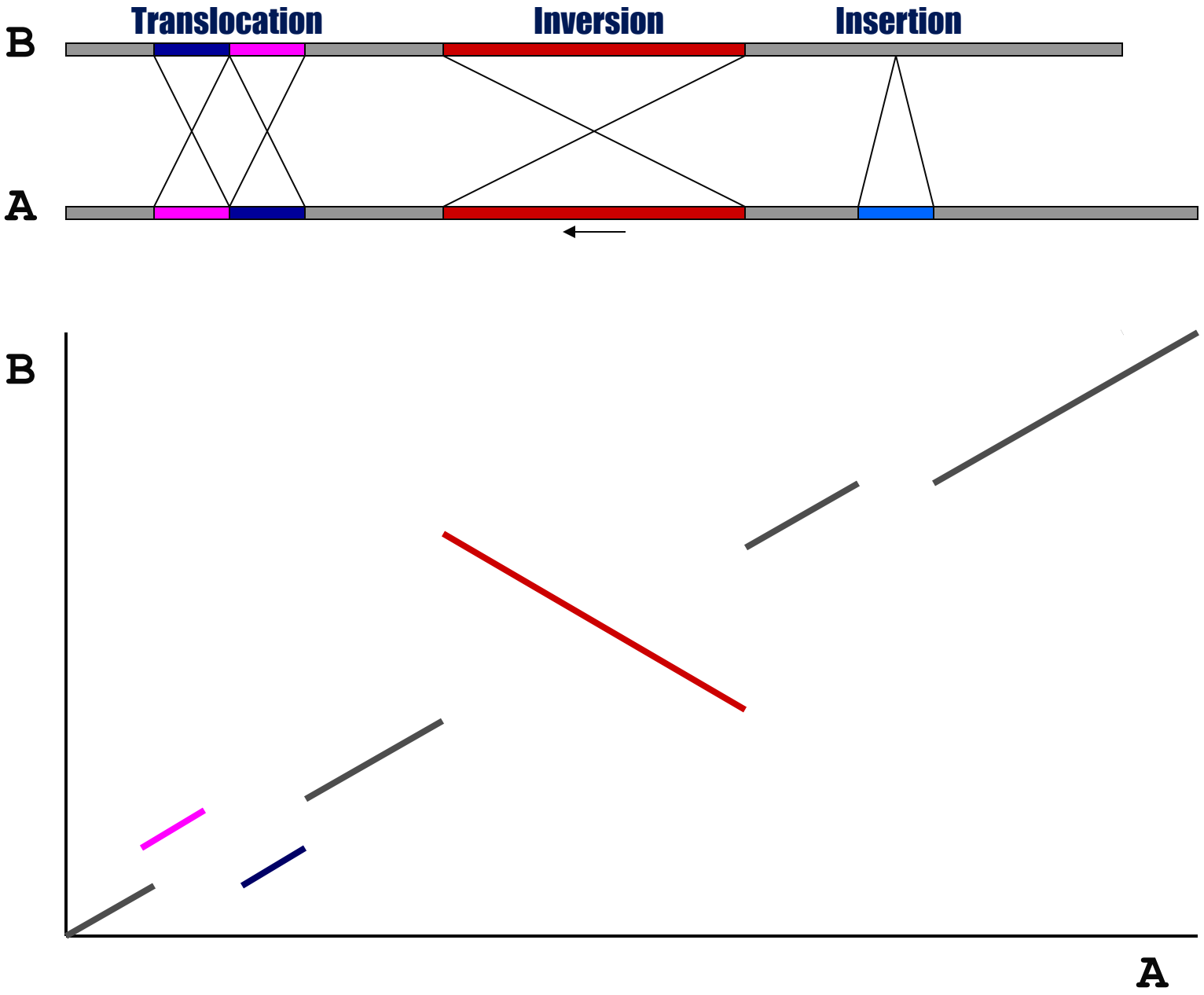
Whole genome alignment visualization

- How can we visualize *whole* genome alignments?

- With an alignment dot plot
  - *N* x *M* matrix
  - Let *i* = position in genome *A*
  - Let *j* = position in genome *B*
  - Fill cell *(i,j)* if A_i shows similarity to B_j

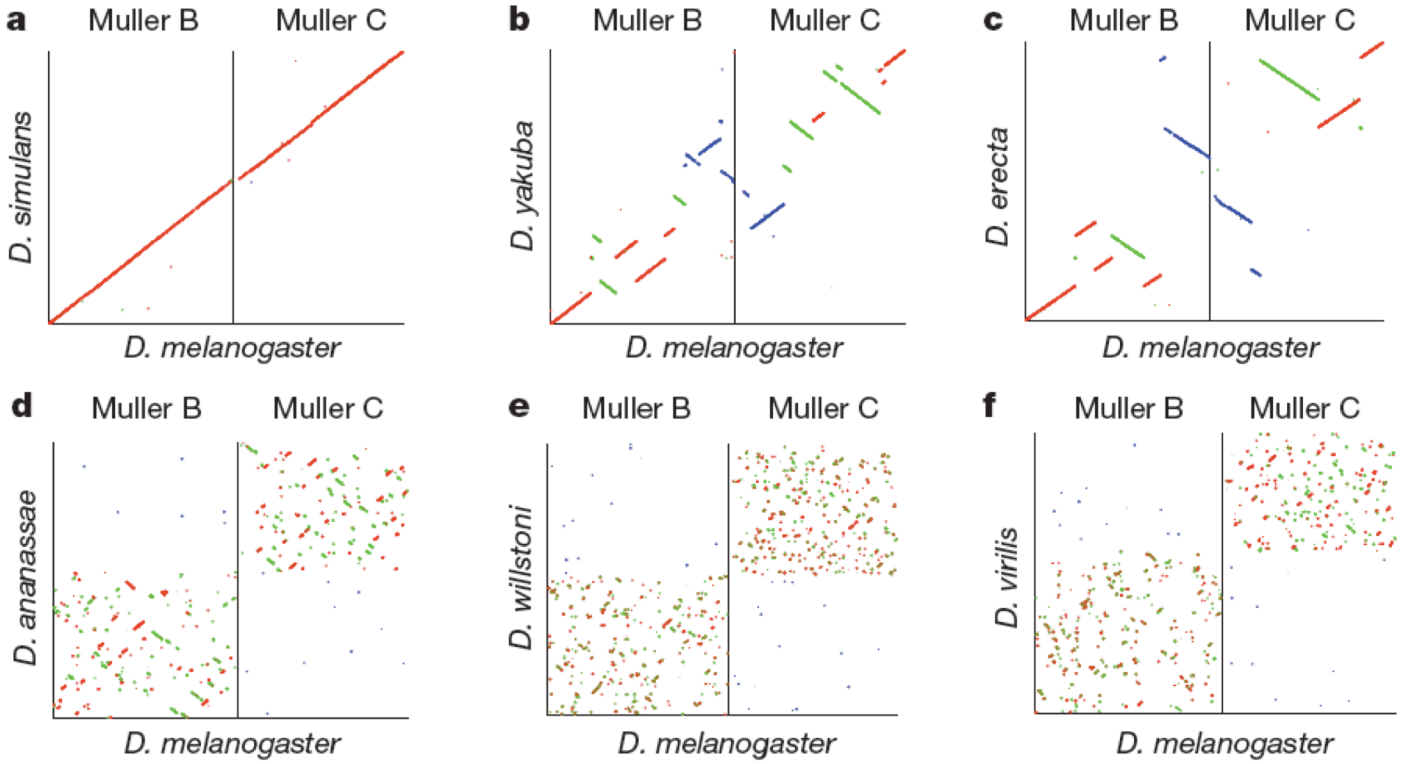


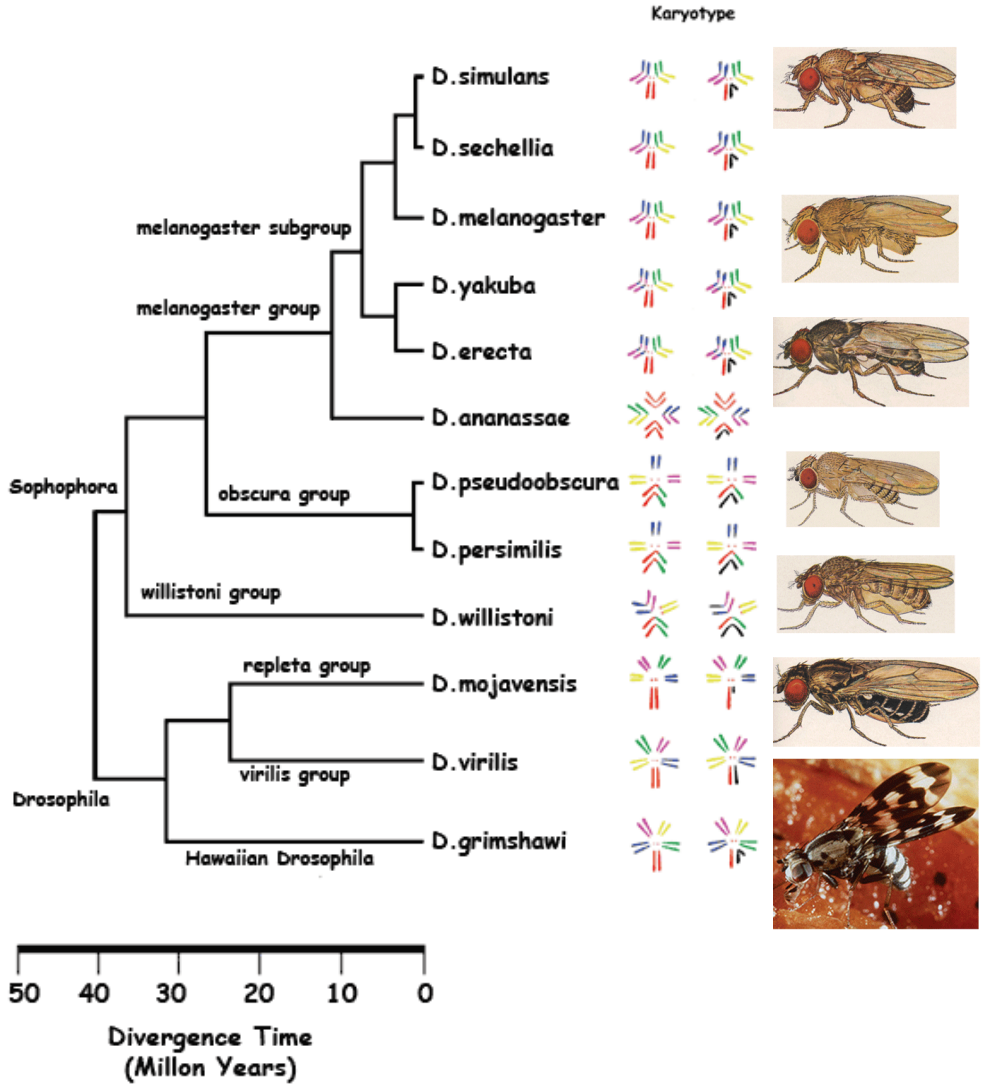
- A perfect alignment between *A* and *B* would completely fill the positive diagonal



NIH
NHGRI

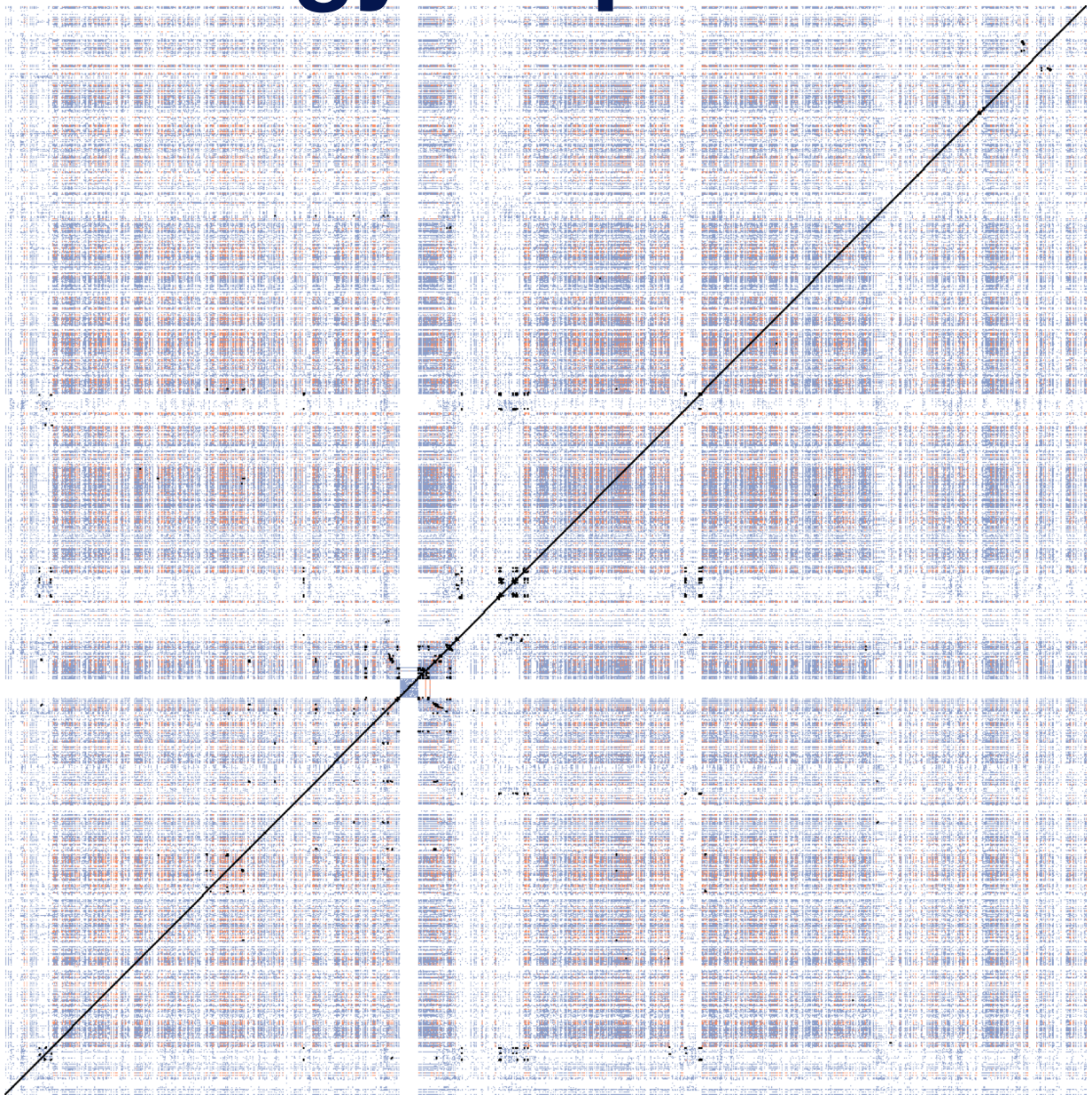
Drosophila genome evolution





NIH
NHGRI

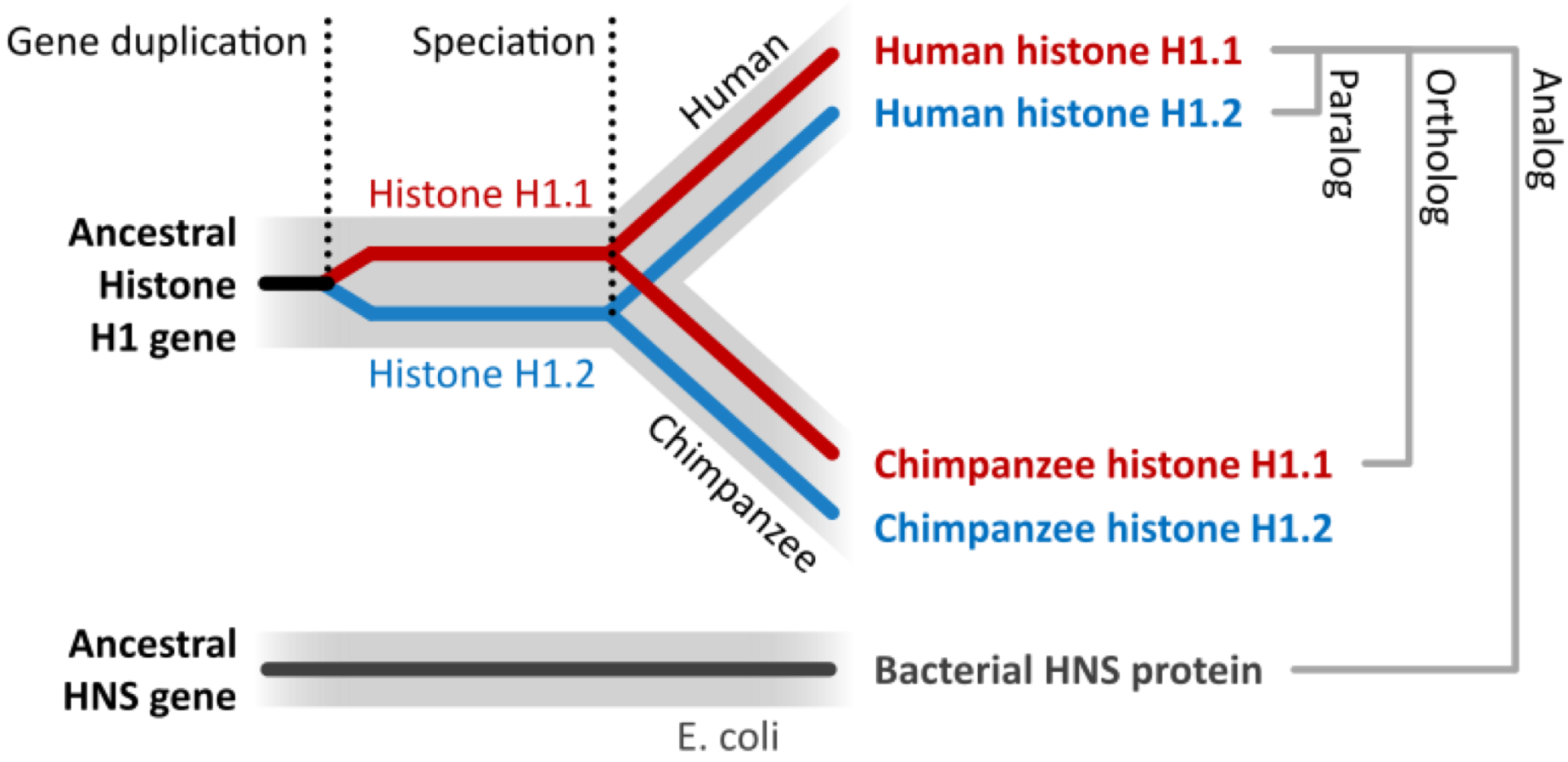
Homology map of human chr7



- Homology
  - Shared ancestry
- Showing:
  - >85% identity
  - Blue: 1–5 kbp
  - Red: 5–10 kbp
  - Black: >10 kbp

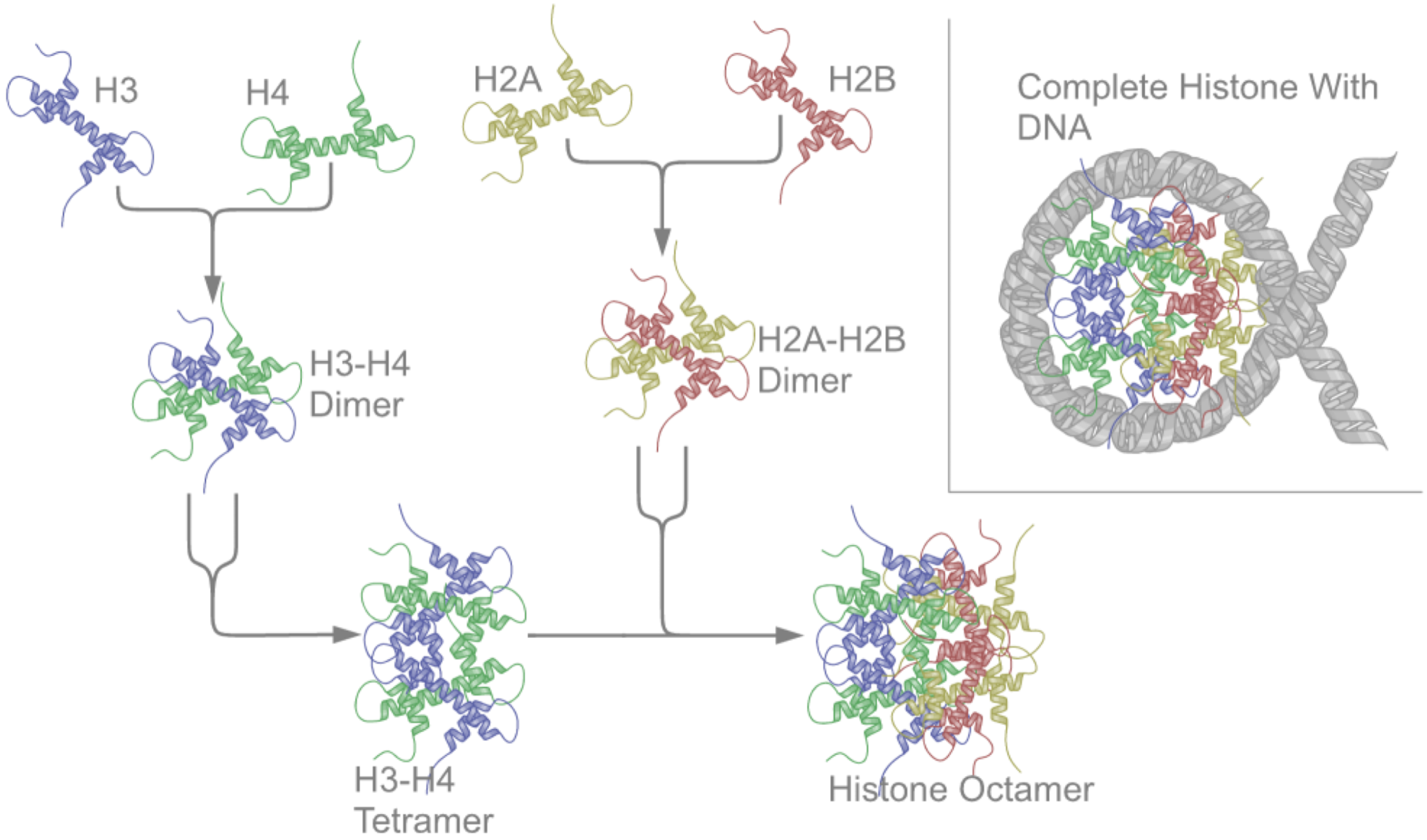
NIH
NHGRI

Homologs: orthologs vs. paralogs



NIH
NHGRI

Paralogous genes



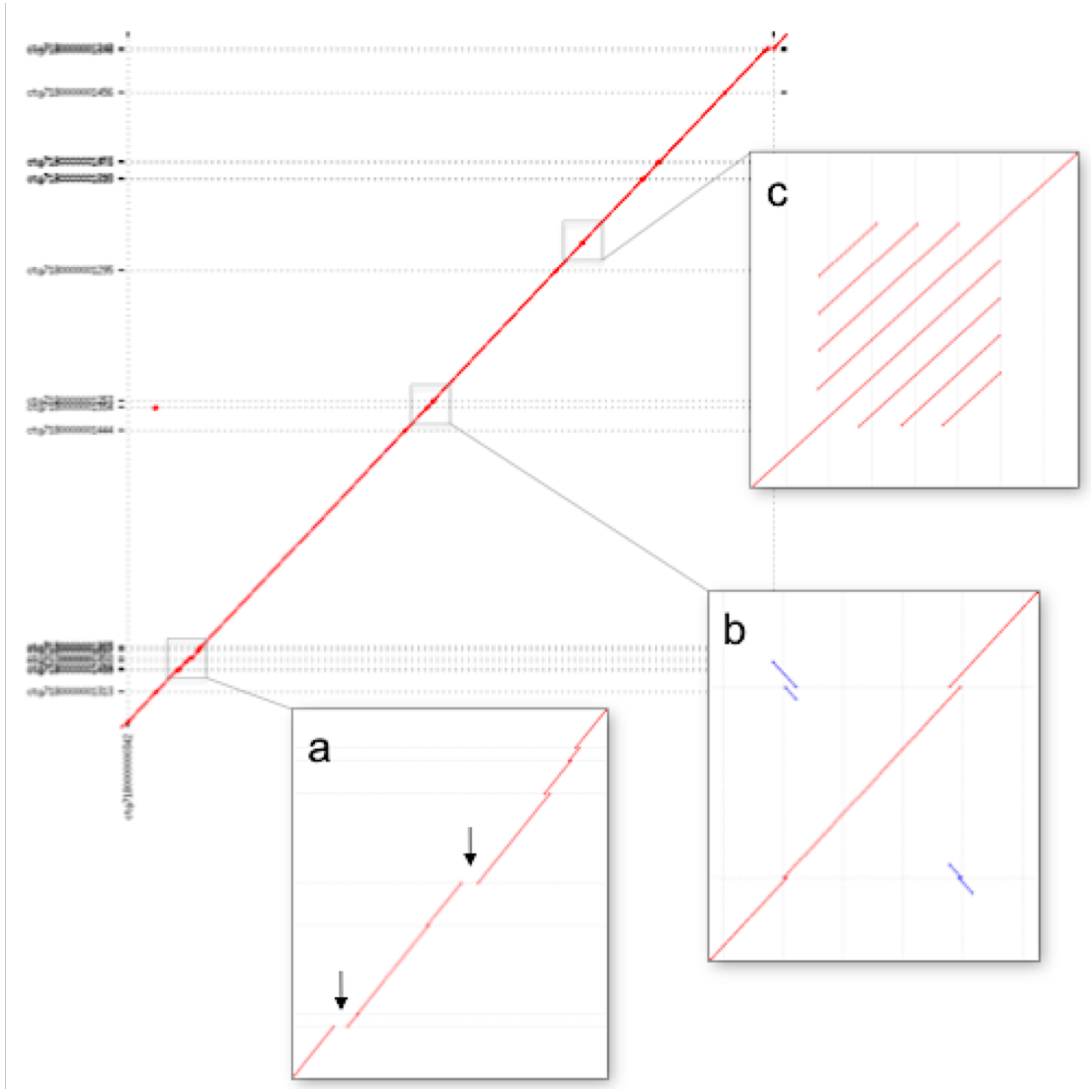
NIH
NHGRI

Goal of whole genome alignment

- For genomes *A* and *B*, find a mapping from each position in *A* to its orthologous position in *B*

- Requires an evolutionary model
  - Compared to nucleotide substitution models genome rearrangement models are not well defined
  - Typically based on “synteny” arguments
  - Genomic regions tend to be inherited in blocks

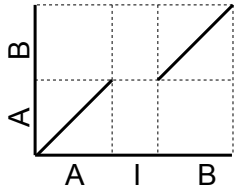
Dotplot quiz



http://mummer.sf.net/manual/AlignmentTypes.pdf

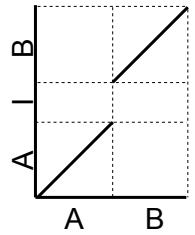
Insertion into Reference

R: AIB
Q: AB



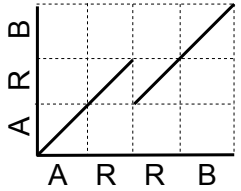
Insertion into Query

R: AB
Q: AIB



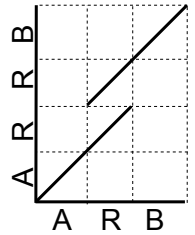
Collapse Query

R: ARRB
Q: ARB



Collapse Reference

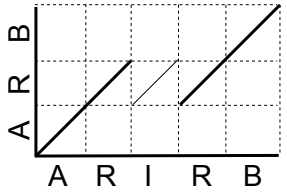
R: ARB
Q: ARRB



Collapse Query
w/ Insertion

R: ARIRB
Q: ARB

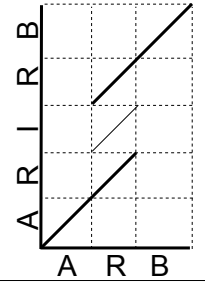
Exact tandem alignment if I=R



Collapse Reference
w/Insertion

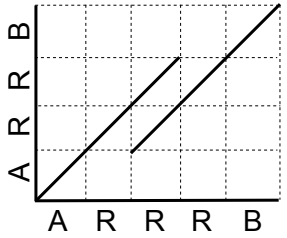
R: ARB
Q: ARIRB

Exact tandem alignment if I=R



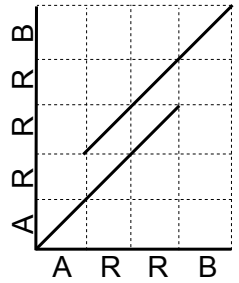
Collapse Query

R: ARRRB
Q: ARRB



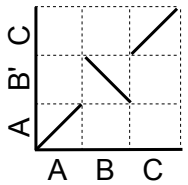
Collapse Reference

R: ARRB
Q: ARRRB



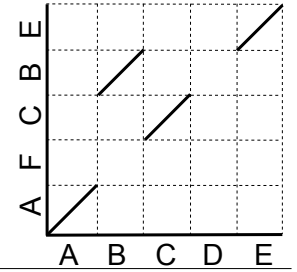
Inversion

R: ABC
Q: AB'C



Rearrangement
w/ Disagreement

R: ABCDE
Q: AFCBE



MUMmer

- Maximal Unique Matcher (MUM)
  - match
    - exact match of a minimum length
  - maximal
    - cannot be extended in either direction without a mismatch
  - unique
    - occurs only once in both sequences (MUM)
    - occurs only once in the reference sequence (MAM)
    - occurs one or more times in either sequence (MEM)

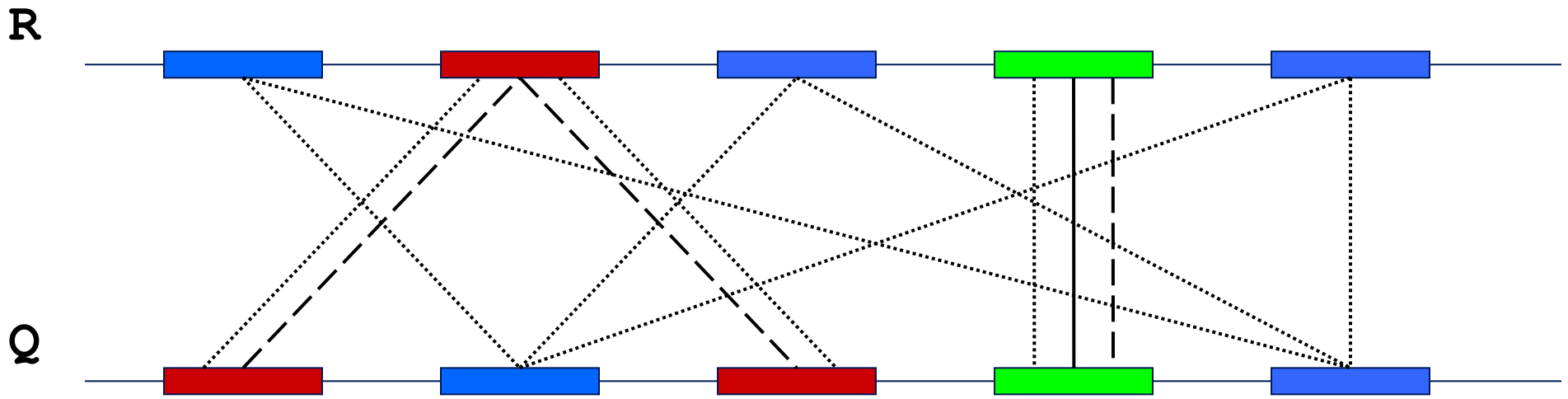
Alignment of whole genomes. Delcher *et al.* 1999

MEMs and MUMs

MUM : maximal unique match

MAM : maximal almost-unique match

MEM : maximal exact match

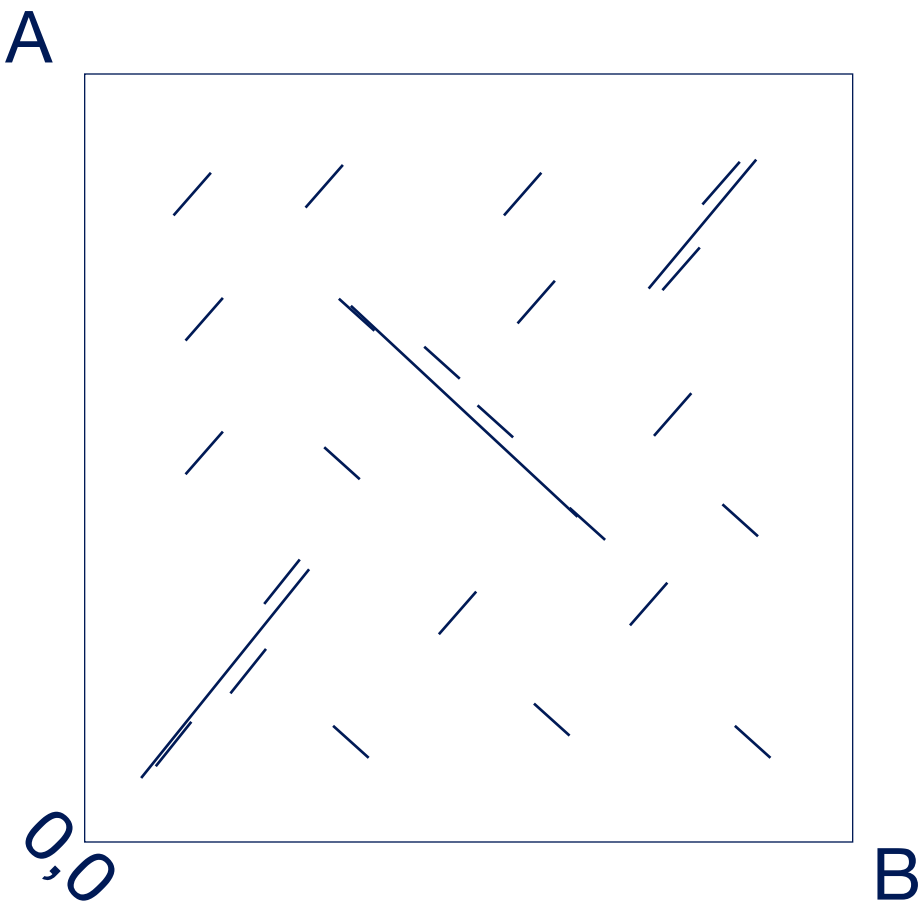


MEMs scale quadratically w/ repeats

How do we turn MUMs into alignments?

- Nucmer approach
  - Find all exact matches
  - Chain exact matches
  - Gapped alignment

- With which algorithms?



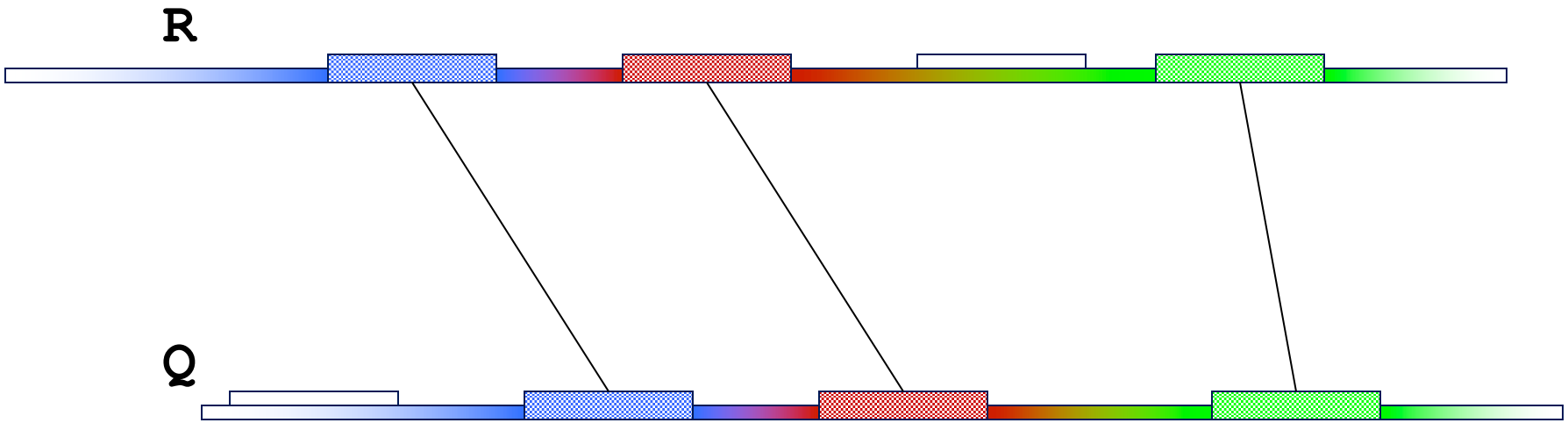
Fast algorithms for large-scale genome alignment and comparison. Delcher *et al.* 2002

Seed, chain, and extend

FIND all MUMs

CHAIN consistent MUMs

EXTEND alignments



Seed

NIH
NHGRI

Suffix indexes for MUM finding

	Suffix tree	Suffix array	FM Index
Time: Does P occur?	O(n)	O(n log m)	O(n)
Time: Count k occurrences of P	O(n + k)	O(n log m)	O(n)
Time: Report k locations of P	O(n + k)	O(n log m + k)	O(n + k)
Space	O(m)	O(m)	O(m)
Needs T?	yes	yes	no
Bytes per input character	>15	~4	~0.5

$m = |T|, n = |P|, k =$ # occurrences of P in T

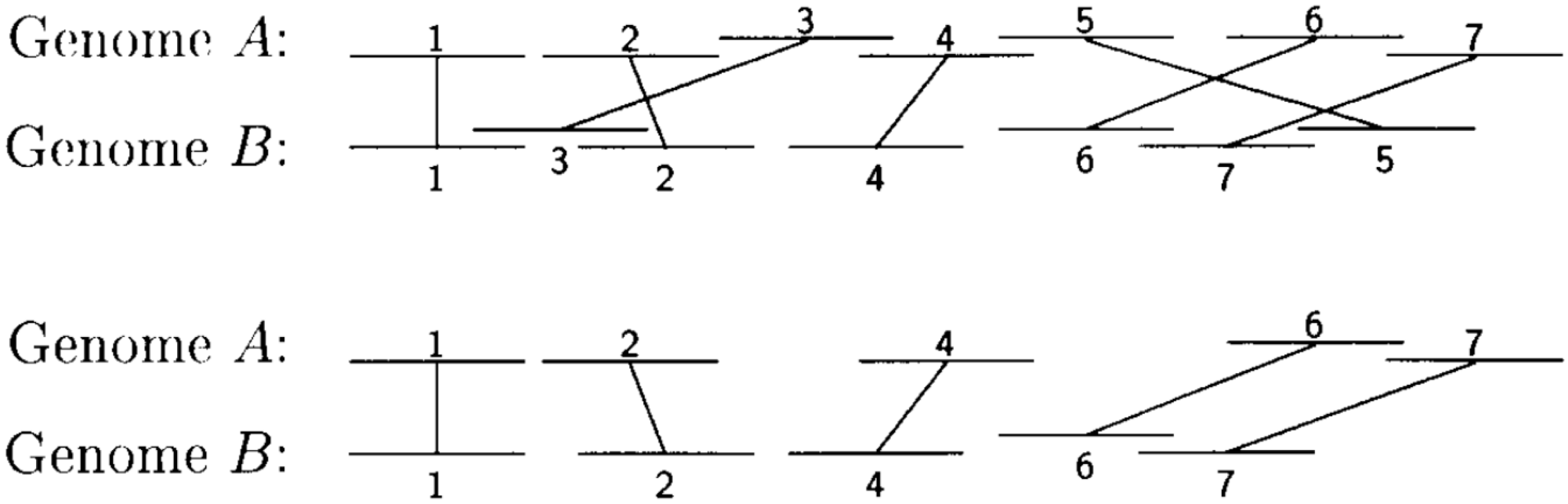
Chain

NIH
NHGRI

Global: Longest Increasing Subsequence

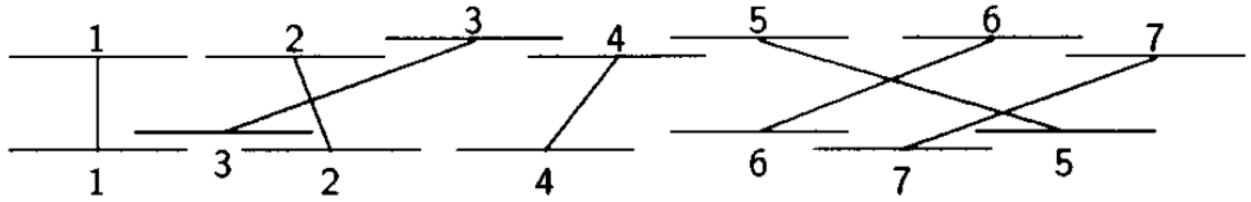
- A **subsequence** of a **permutation** is a collection of elements of the permutation in the order that they appear. For example, (5, 3, 4) is a subsequence of (5, 1, 3, 4, 2).
- A subsequence is **increasing** if the elements of the subsequence increase. For example, given the permutation (8, 2, 1, 6, 5, 7, 4, 3, 9), an increasing subsequence is (2, 6, 7, 9).
  - **Given:** A a permutation π of length n.
  - **Return:** A longest increasing subsequence of π.

Applied LIS example



NIH
NHGRI

LIS pseudo code




```
for i in [1..n]
    for j in [1..i)
        score = max[j] + len(i) – olap(j,i)
        if score > max[i]
            max[i] = score
            pre[i] = j
```

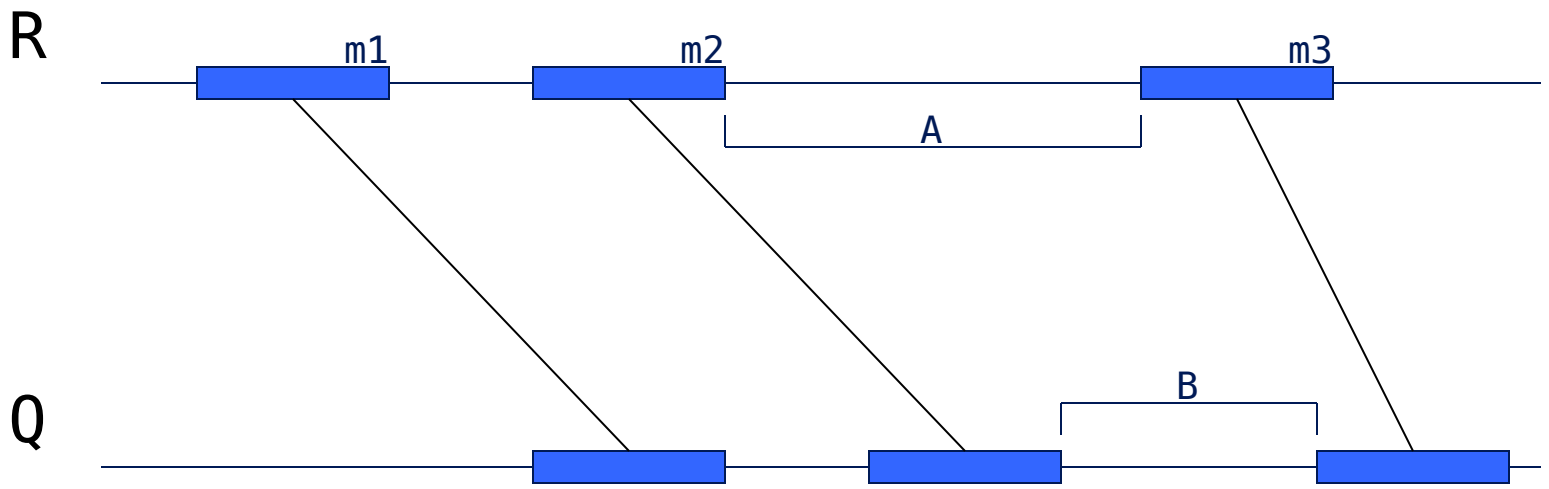

Can also define arbitrary gap penalties, and speed up with early termination heuristics: see minimap2

Local: Agglomerative Clustering

cluster length = $\sum m_i$

gap distance = A

diagonal difference = |B – A| or |B – A| / B



Agglomerative pseudo code

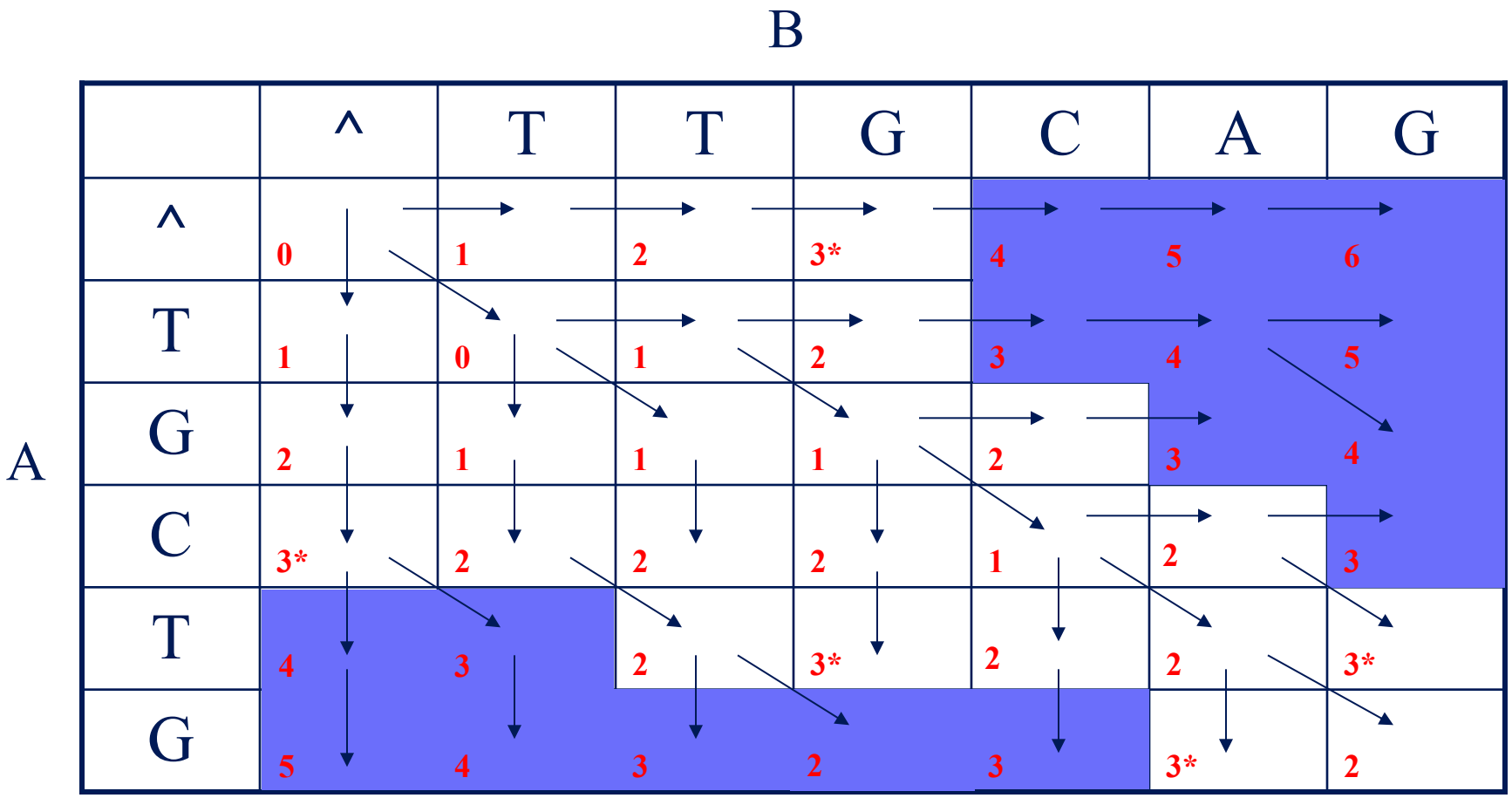
```
for i in [1..n)
   for j in [i+1..n]
      if gap(i,j) < g and diagdiff(i,j) < d
         union(find(i),find(j))
```


Create connected components based on small separation and similar diagonals

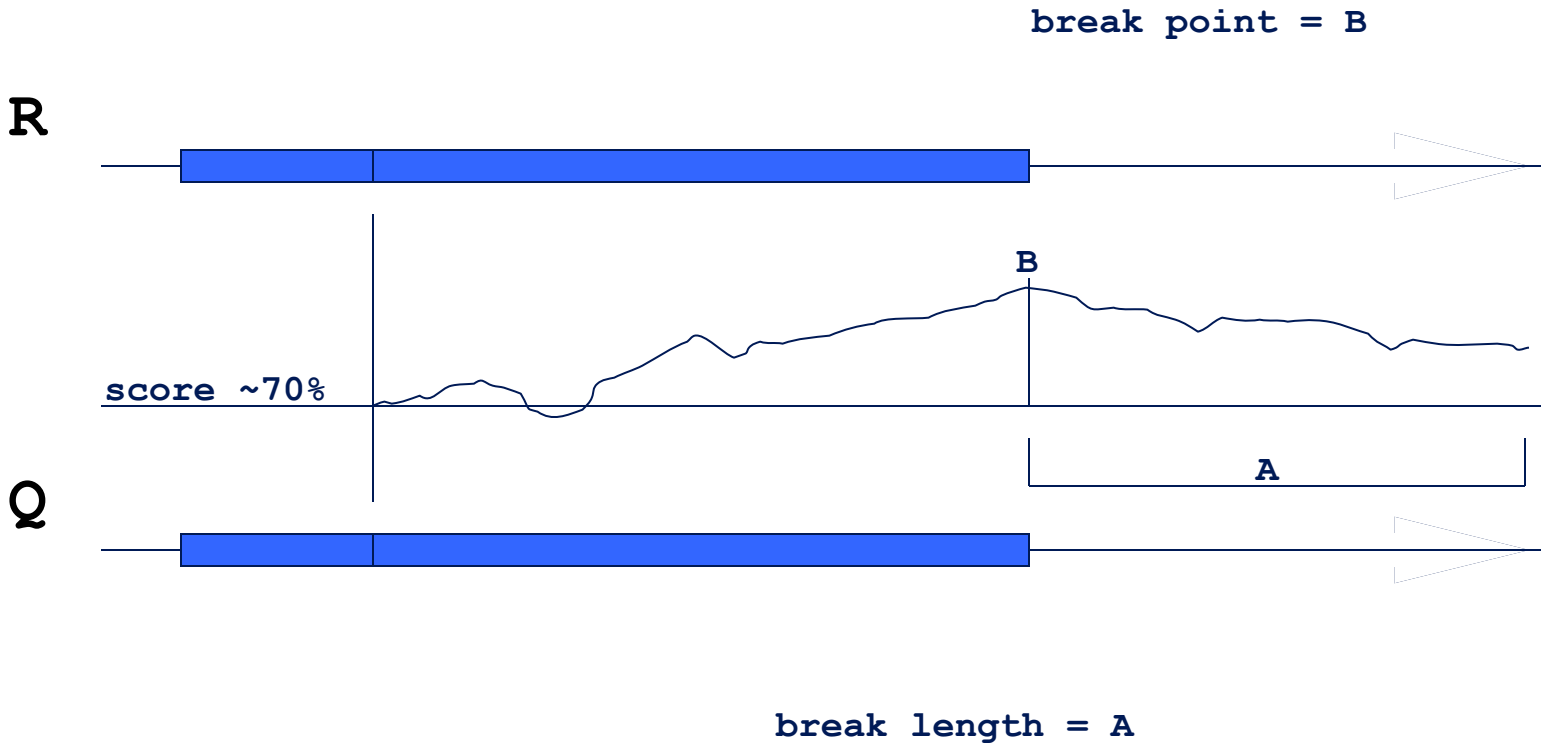
Extend

NIH
NHGRI

Banded alignment extension



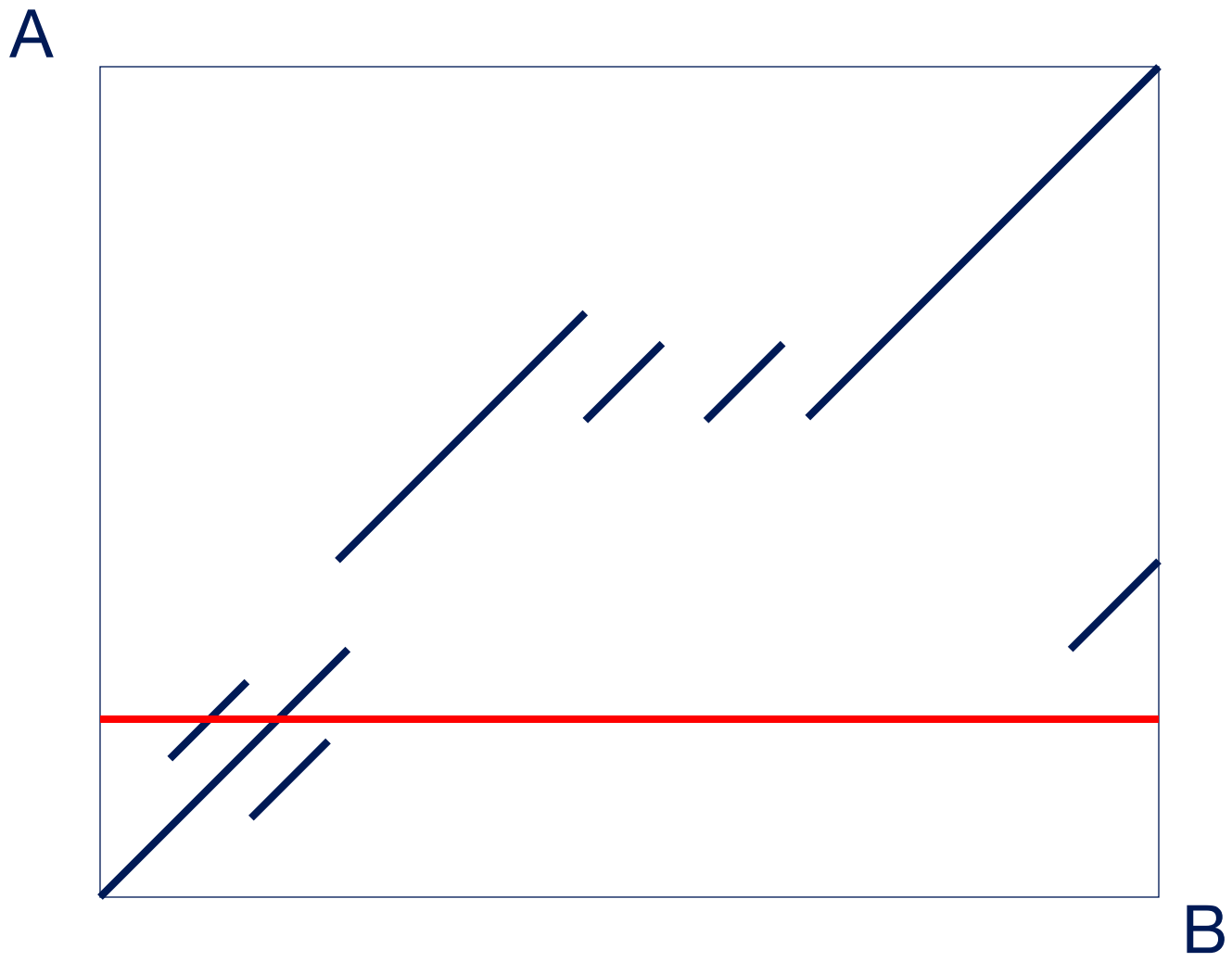
Alignment extension



Filter

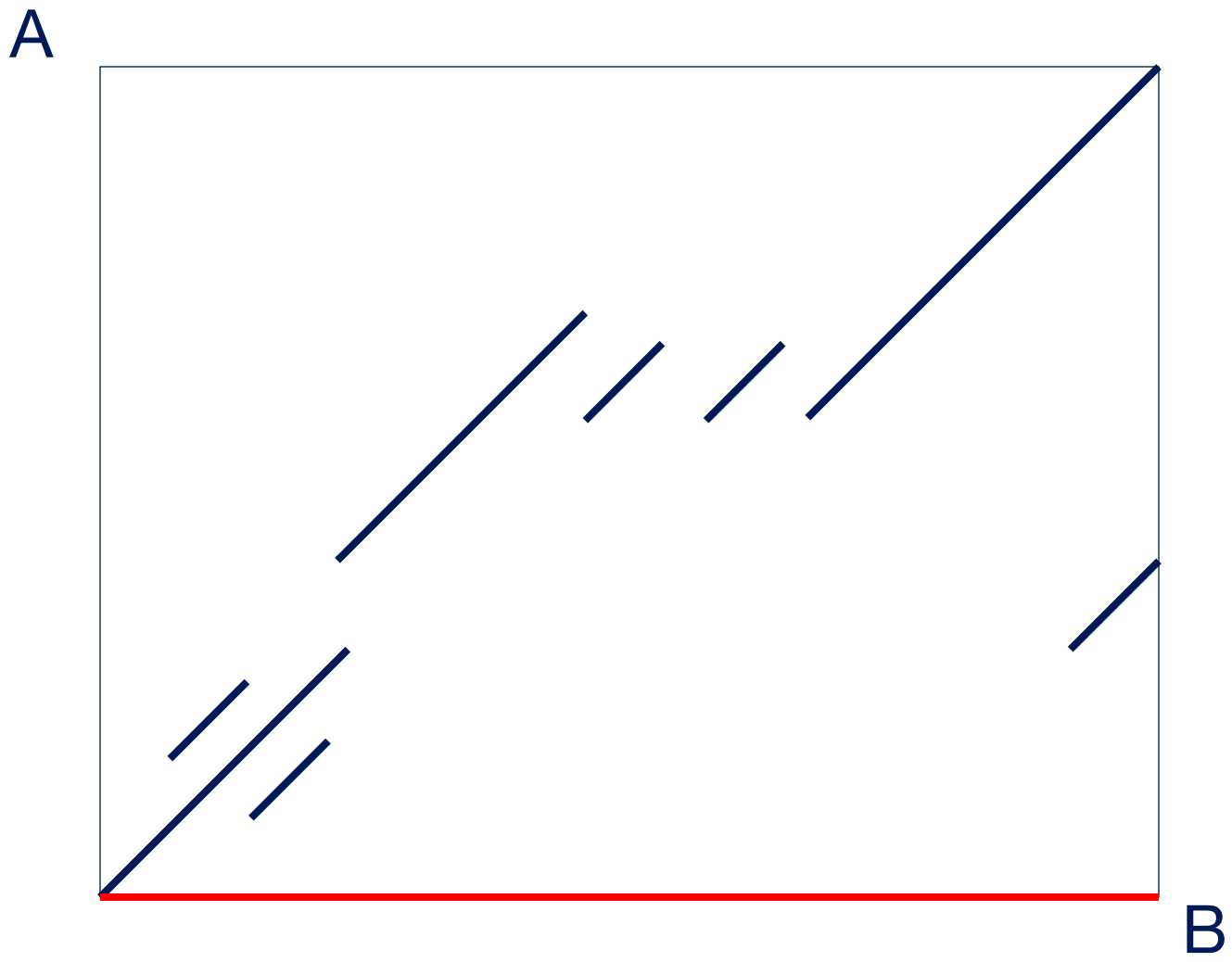
NIH
NHGRI

Which are orthologs?



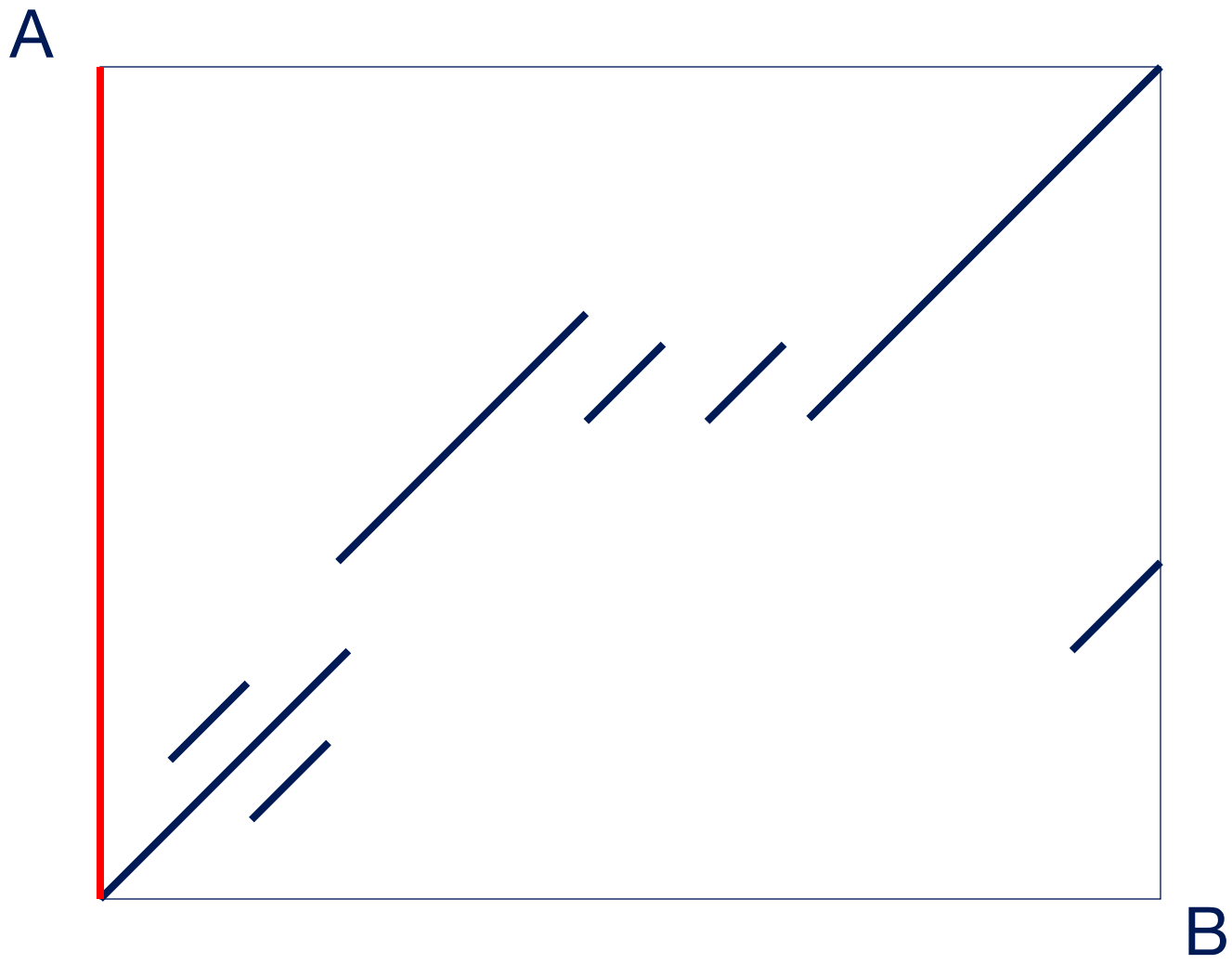
Mark a segment as “best” if it has the highest score of all segments intersecting the sweep line

Best for *A*



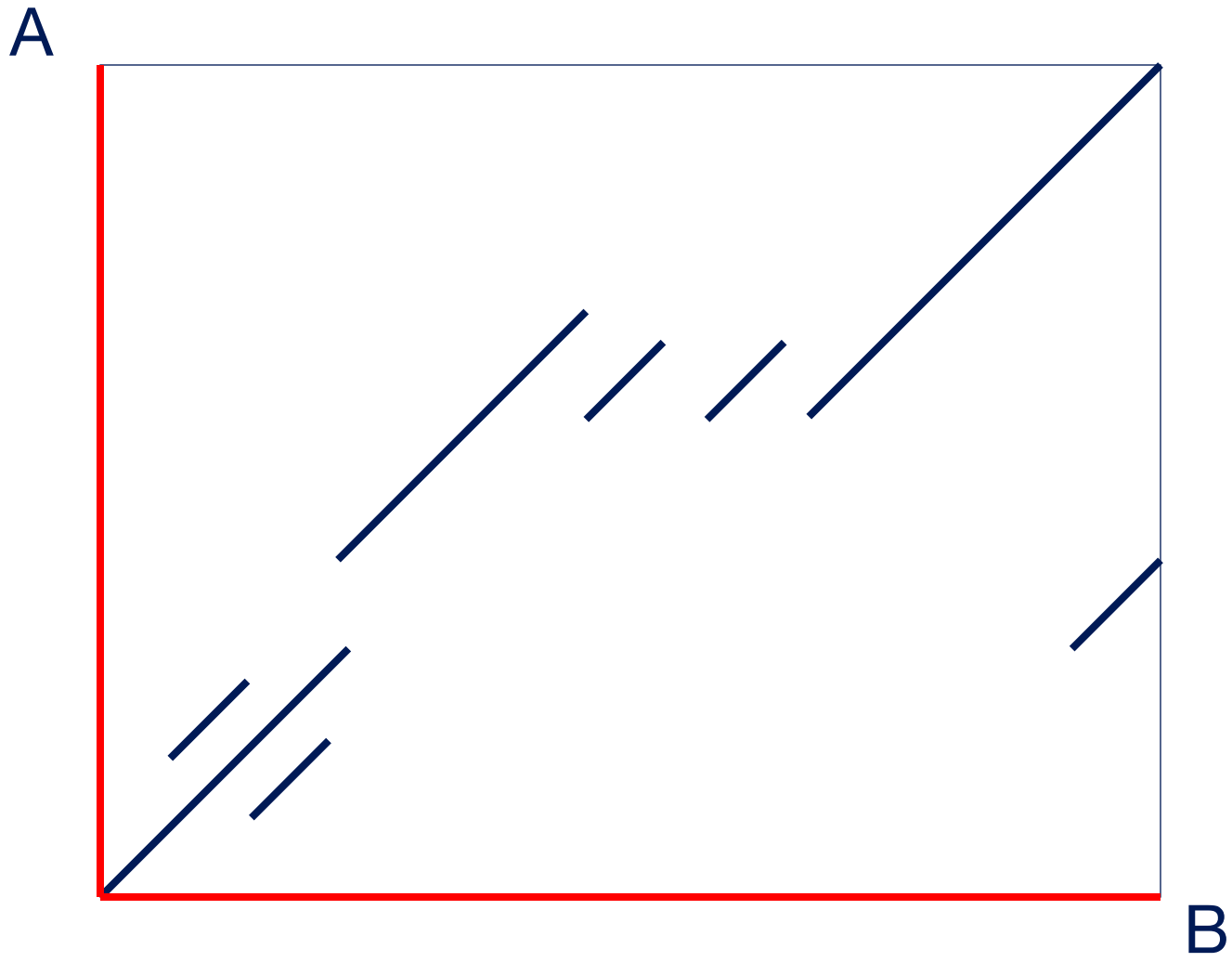
NIH
NHGRI

Best for *B*



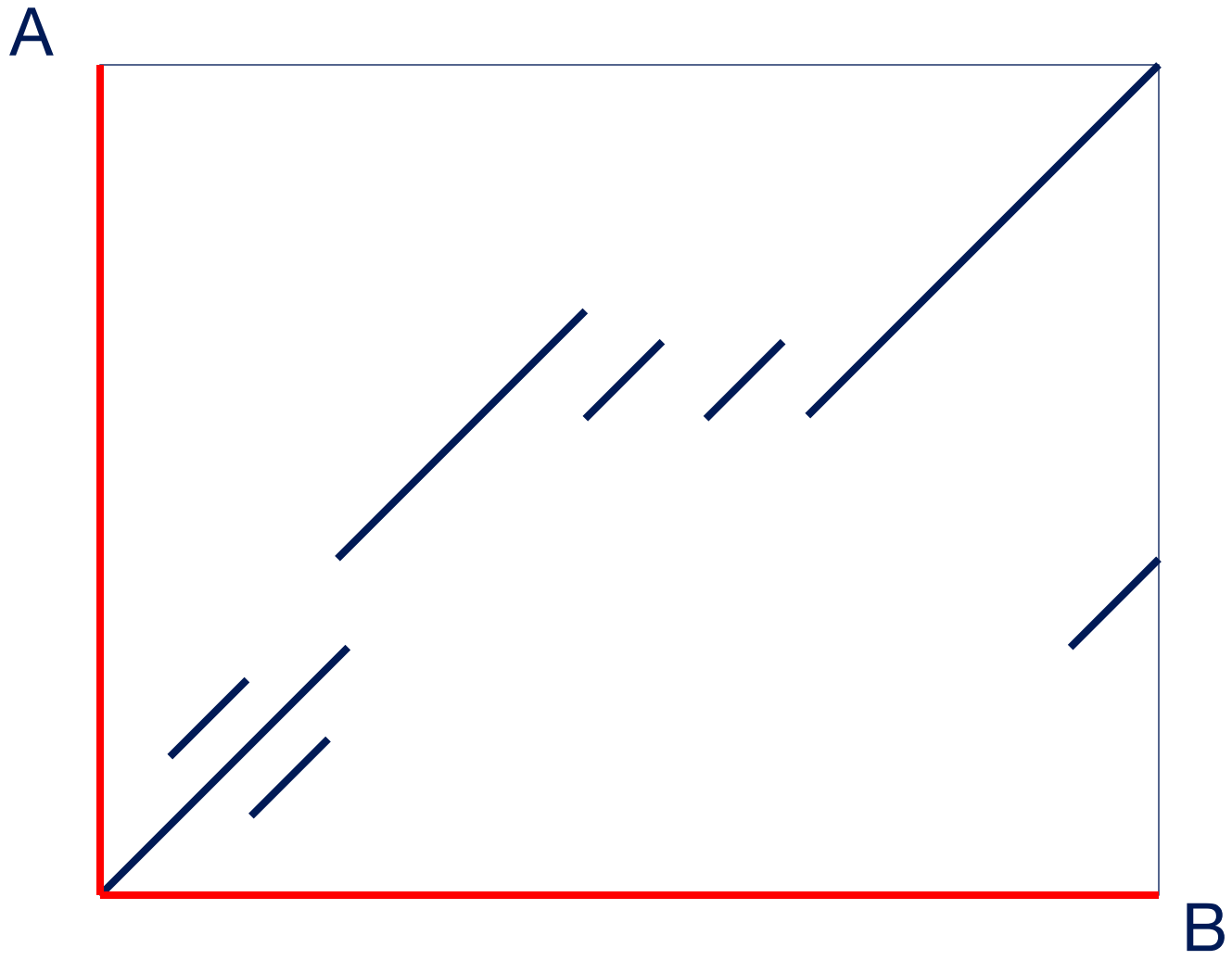
NIH
NHGRI

Best for *A and B* (orthologs only)



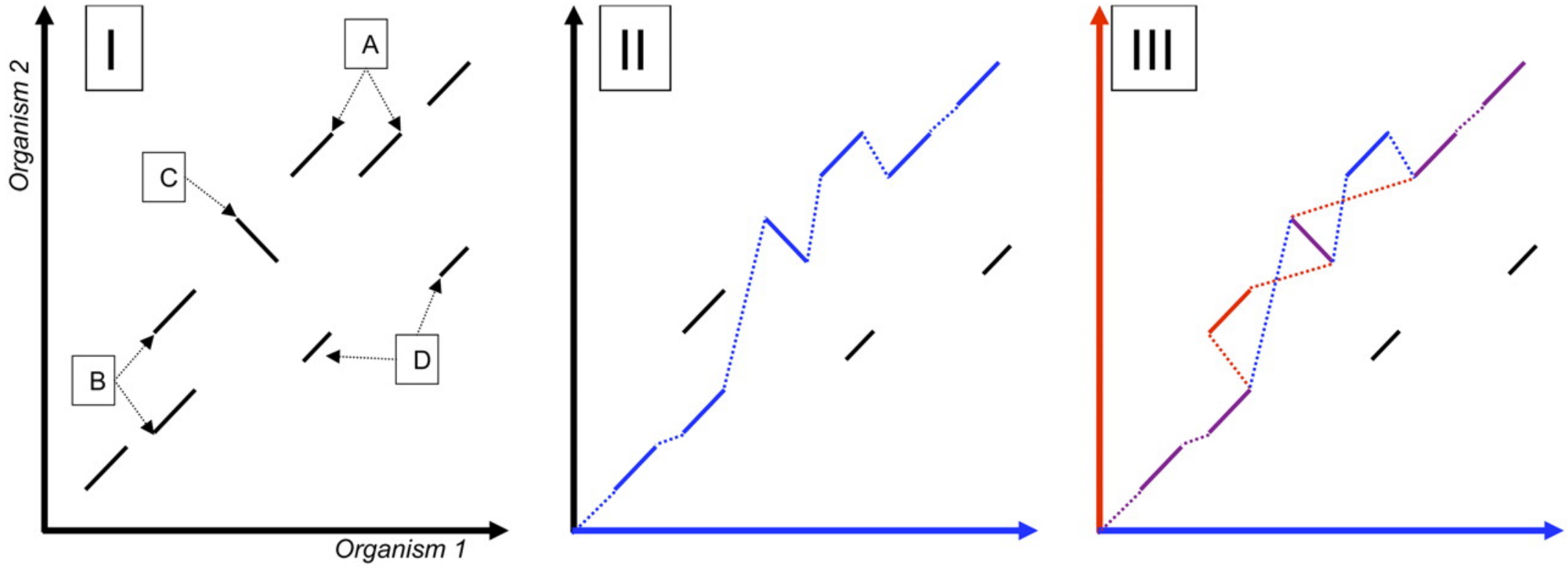
NIH
NHGRI

Best for *A or B* (paralogs too)



NIH
NHGRI

SuperMap



Multiple whole-genome alignments without a reference organism. Dubchak *et al.* 2009