Annotation-free Identification of Synteny Anchors

Karl Kaether

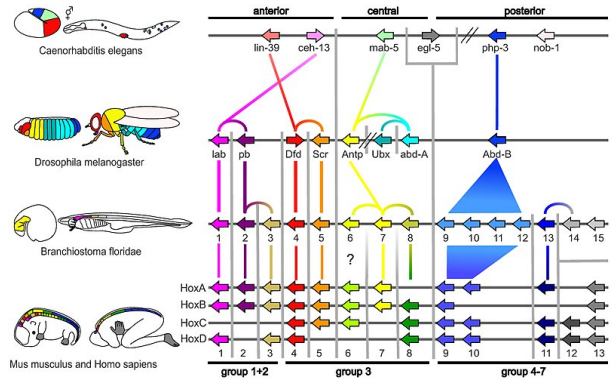
Abteilung Stadler, Leipzig

Bled 2024

Orthology Inference

- mostly based on sequence similarity
- e.g. (reciprocal) blast
- some problems[1,2,3]
  - sequence divergence
  - genome rearrangements

- **synteny** can help in such situations



https://en.wikipedia.org/wiki/Synteny

- in principle solved by global alignments
- realistic approaches use genome annotations

[1] Altenhoff et al., "The Quest for Orthologs benchmark service and consensus calls in 2020".

[2] Moyers and Zhang, "Further Simulations and Analyses Demonstrate Open Problems of Phylostratigraphy".

[3] Vakirlis, Carvunis, and McLysaght, "Synteny-based analyses indicate that sequence divergence is not the main source of orphan genes".

A

B

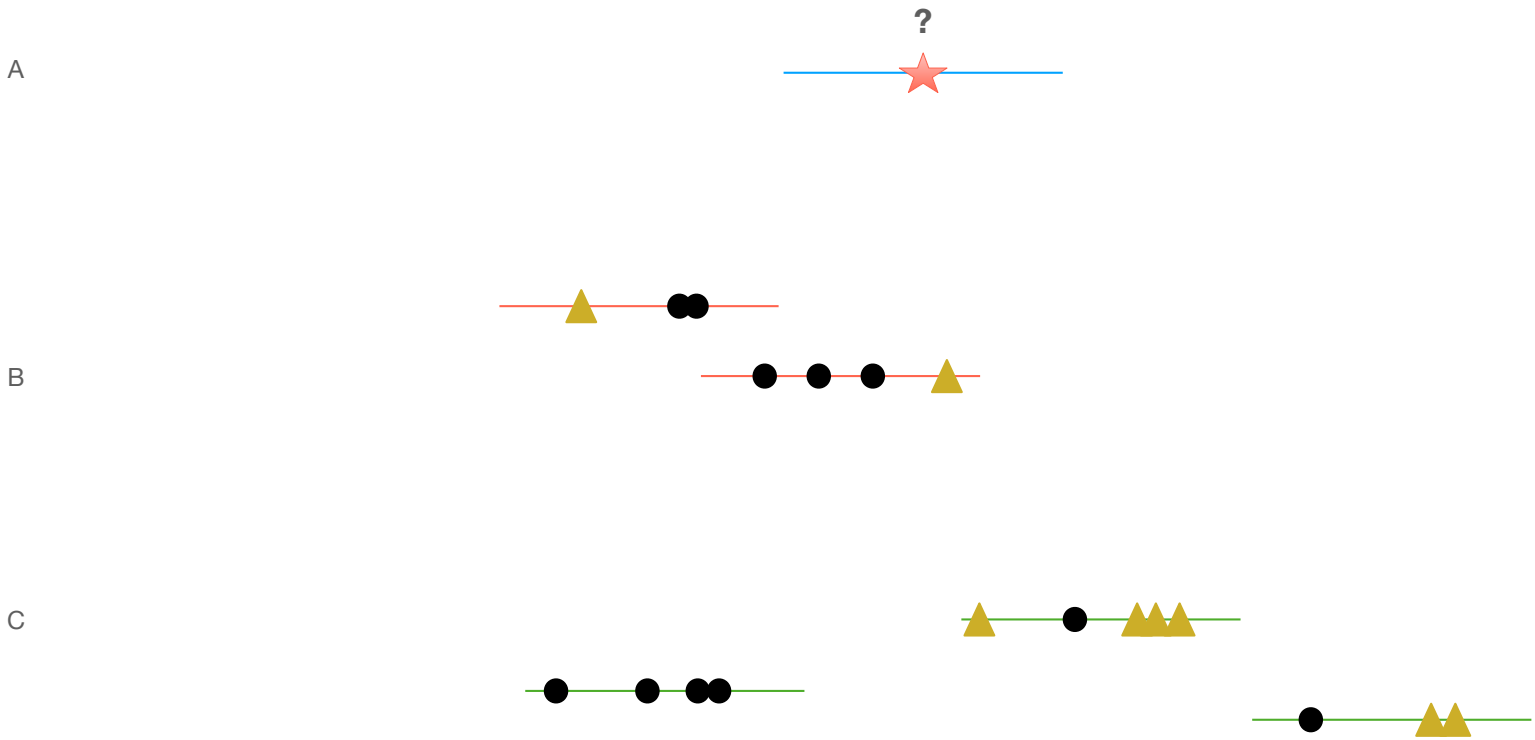
C

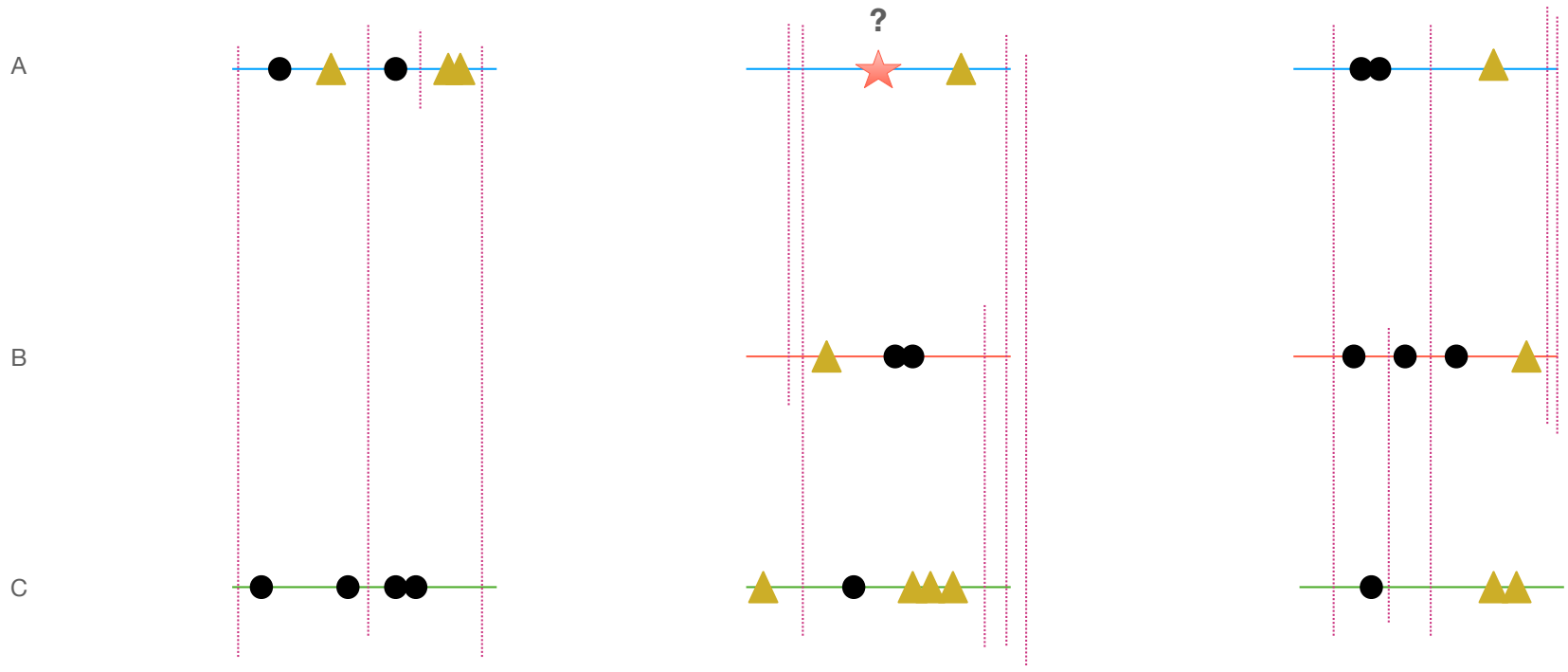
A

B

Drosophila_melanogaster_tRNA-Ala-AGC-1-1	OX371225.1	98.63	73	1	0	1	73	47815972	47816044	3e-29	130
Drosophila_melanogaster_tRNA-Ala-AGC-1-1	OX371225.1	98.63	73	1	0	1	73	47818702	47818630	3e-29	130
Drosophila_melanogaster_tRNA-Ala-AGC-1-1	OX371225.1	94.52	73	4	0	1	73	233858497	233858425	3e-24	113
Drosophila_melanogaster_tRNA-Ala-AGC-1-1	OX371225.1	94.52	73	4	0	1	73	233859067	233859139	3e-24	113
Drosophila_melanogaster_tRNA-Ala-AGC-1-1	OX371225.1	91.78	73	6	0	1	73	233863496	233863568	7e-21	102
Drosophila_melanogaster_tRNA-Ala-AGC-1-1	OX371225.1	85.14	74	9	2	1	73	303312418	303312490	1e-12	75.0
Drosophila_melanogaster_tRNA-Ala-AGC-1-1	OX371225.1	93.75	48	3	0	1	48	303310365	303310412	5e-12	73.1
Drosophila_melanogaster_tRNA-Ala-AGC-1-1	OX371225.1	91.67	48	4	0	1	48	303308851	303308898	2e-10	67.6
Drosophila_melanogaster_tRNA-Ala-AGC-1-1	OX371225.1	92.31	26	2	0	1	26	303311081	303311106	0.19	38.1
Drosophila_melanogaster_tRNA-Ala-AGC-1-1	OX371224.1	98.63	73	1	0	1	73	363661361	363661433	3e-29	130
Drosophila_melanogaster_tRNA-Ala-AGC-1-1	OX371224.1	97.26	73	2	0	1	73	319278696	319278624	1e-27	124
Drosophila_melanogaster_tRNA-Ala-AGC-1-1	OX371224.1	97.26	73	2	0	1	73	372231794	372231866	1e-27	124
Drosophila_melanogaster_tRNA-Ala-AGC-1-1	OX371224.1	94.52	73	4	0	1	73	320281395	320281467	3e-24	113
Drosophila_melanogaster_tRNA-Ala-AGC-1-1	OX371224.1	94.52	73	4	0	1	73	372219431	372219503	3e-24	113
Drosophila_melanogaster_tRNA-Ala-AGC-1-1	OX371224.1	94.52	73	4	0	1	73	372231012	372230940	3e-24	113
Drosophila_melanogaster_tRNA-Ala-AGC-1-1	OX371224.1	94.52	73	4	0	1	73	372274162	372274090	3e-24	113
Drosophila_melanogaster_tRNA-Ala-AGC-1-1	OX371224.1	93.15	73	5	0	1	73	306015974	306015902	1e-22	108
Drosophila_melanogaster_tRNA-Ala-AGC-1-1	OX371224.1	93.06	72	4	1	2	73	306016192	306016262	2e-21	104
Drosophila_melanogaster_tRNA-Ala-AGC-1-1	OX371224.1	87.67	73	8	1	1	73	6187391	6187462	2e-15	84.2
Drosophila_melanogaster_tRNA-Ala-AGC-1-1	OX371224.1	86.30	73	9	1	1	73	6102878	6102807	1e-13	78.7
Drosophila_melanogaster_tRNA-Ala-AGC-1-1	OX371224.1	86.30	73	9	1	1	73	178173075	178173146	1e-13	78.7
Drosophila_melanogaster_tRNA-Ala-AGC-1-1	OX371224.1	86.30	73	9	1	1	73	198903251	198903322	1e-13	78.7
Drosophila_melanogaster_tRNA-Ala-AGC-1-1	OX371224.1	86.30	73	9	1	1	73	372199550	372199479	1e-13	78.7
Drosophila_melanogaster_tRNA-Ala-AGC-1-1	OX371224.1	86.30	73	9	1	1	73	372231203	372231274	1e-13	78.7
Drosophila_melanogaster_tRNA-Ala-AGC-1-1	OX371224.1	84.93	73	10	1	1	73	281774135	281774064	5e-12	73.1
Drosophila_melanogaster_tRNA-Ala-AGC-1-1	OX371224.1	83.78	74	9	3	1	73	8114606	8114677	2e-10	67.6
Drosophila_melanogaster_tRNA-Ala-AGC-1-1	OX371224.1	87.10	31	4	0	8	38	319396726	319396696	0.67	36.2

C


?
A
B
C


A
B
C
?

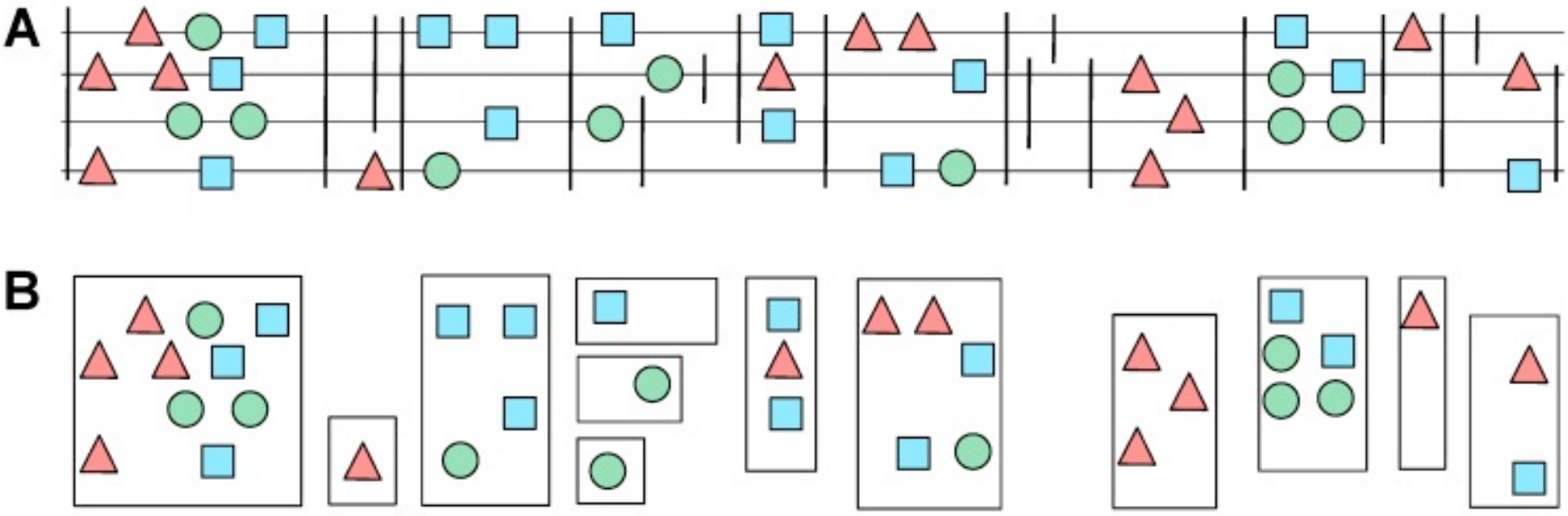
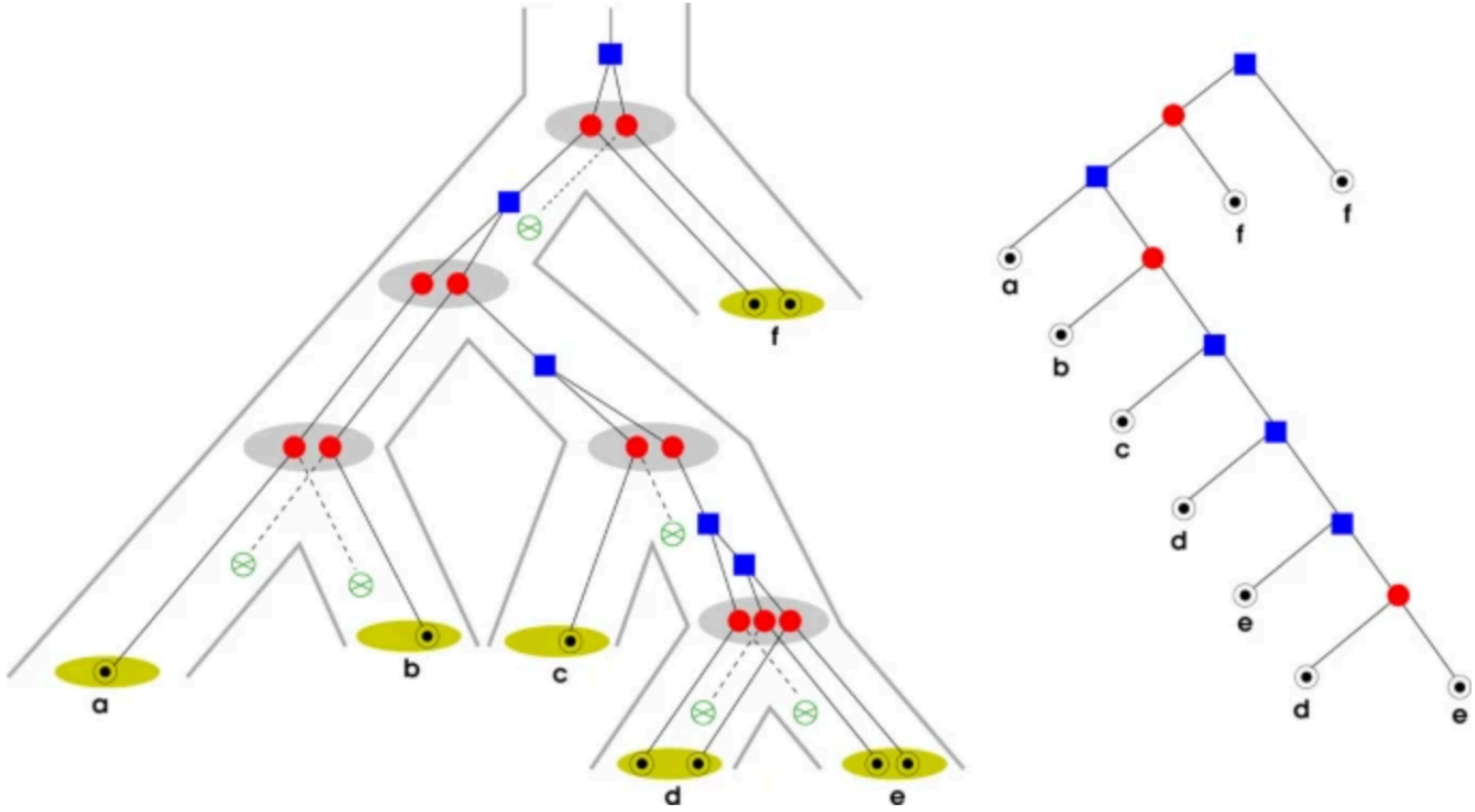
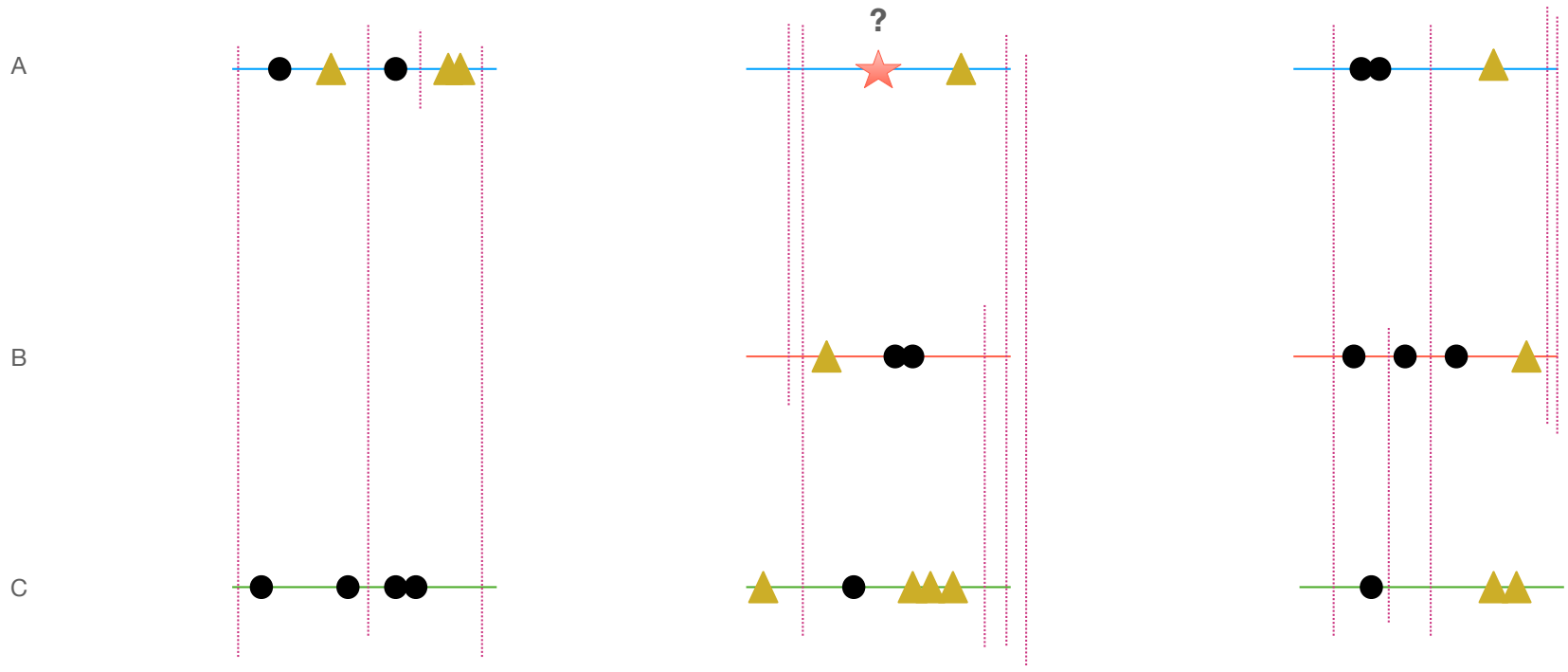

A
B

Figure 1




A
B
C
?

Finding Synteny Anchors

- usually based on genome annotations and possibly multiple sequence alignments[4,5]
  - MSAs too expensive for many genomes
  - annotation might be unavaliable
  - suffer from biases and errors[6]
    - ★ biased towards model species
    - ★ limited by assembly quality
    - ★ too sparse for synteny calculation
    - ★ contaminated by foreign species

[4] Wang et al., "MCScanX".

[5] Haas et al., "DAGchainer".

[6] Salzberg, "Next-generation genome annotation".

How to do this differently

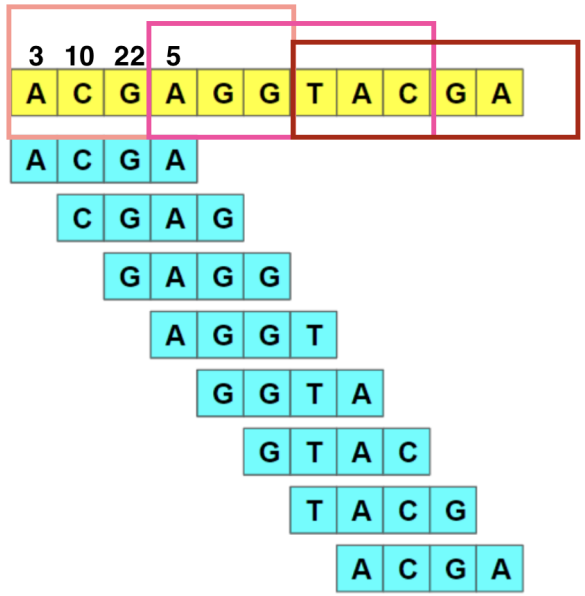
1. count k-mers

How to do this differently

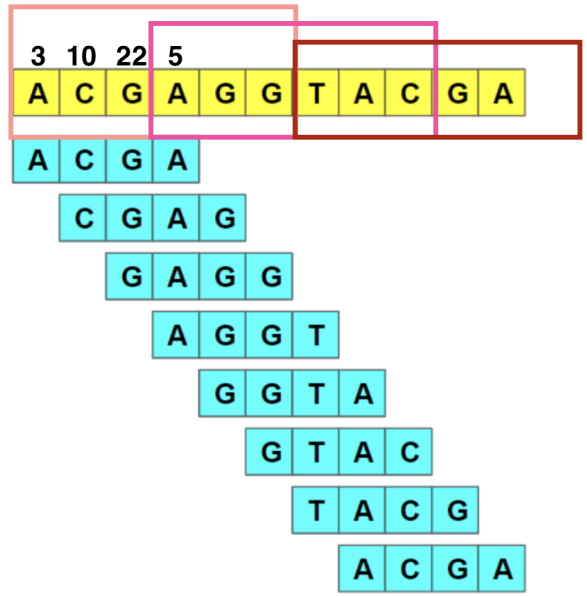
1. count k-mers
2. chop genome into windows

How to do this differently

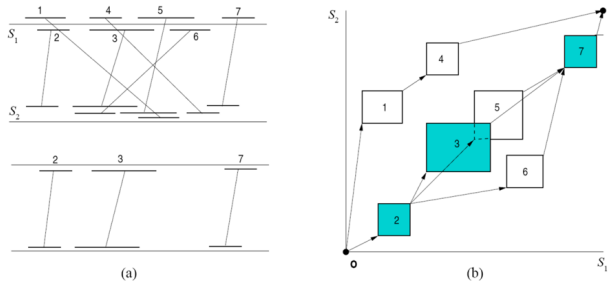
1. count k-mers
2. chop genome into windows
3. sum up counts per window and further use x best %



How to do this differently

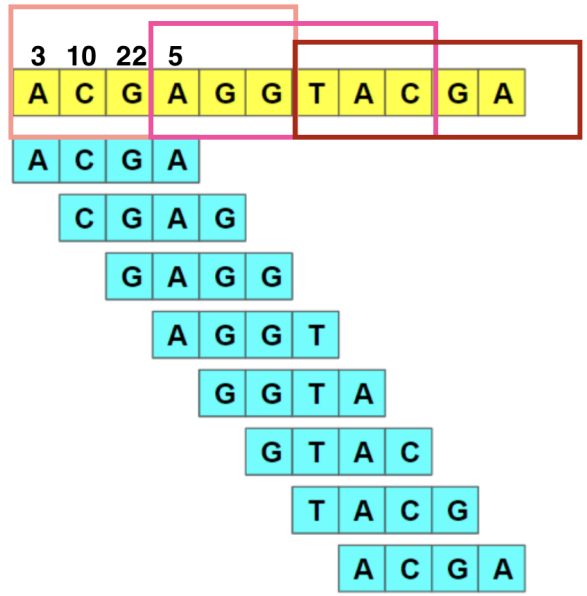


1. count k-mers
2. chop genome into windows
3. sum up counts per window and further use x best %

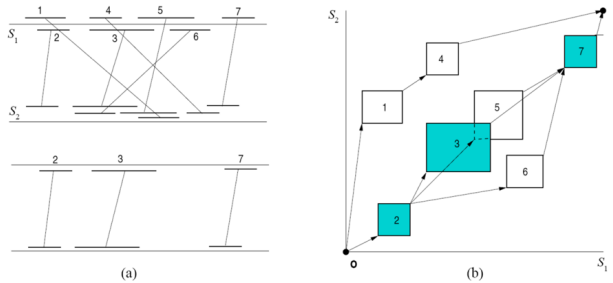


4. blast against own genome and chain hits $\rightarrow C(A, d^A)$ per genome

How to do this differently



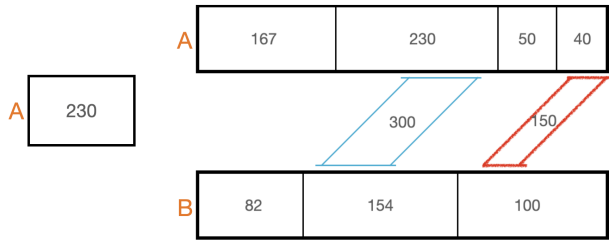
1. count k-mers
2. chop genome into windows
3. sum up counts per window and further use x best %

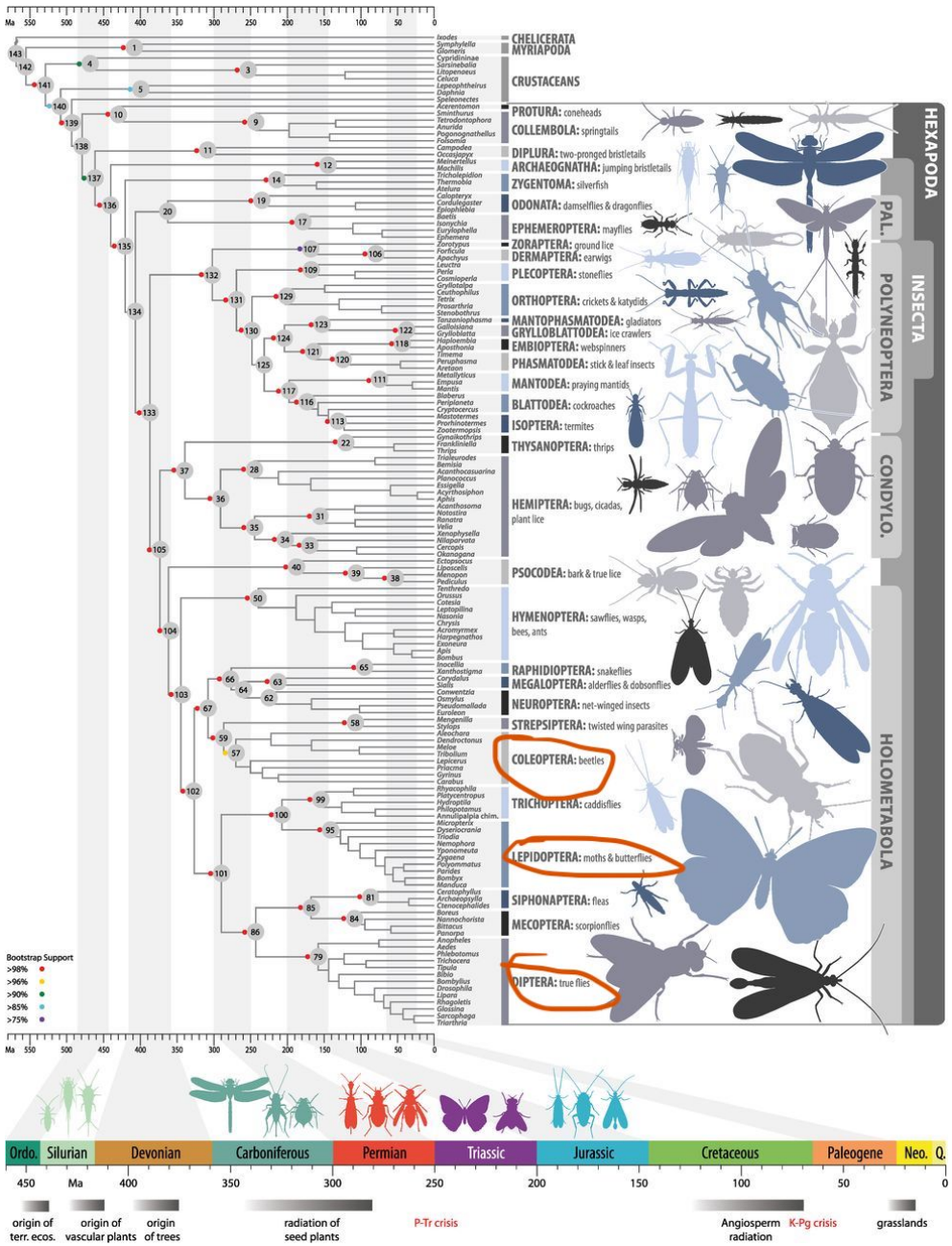


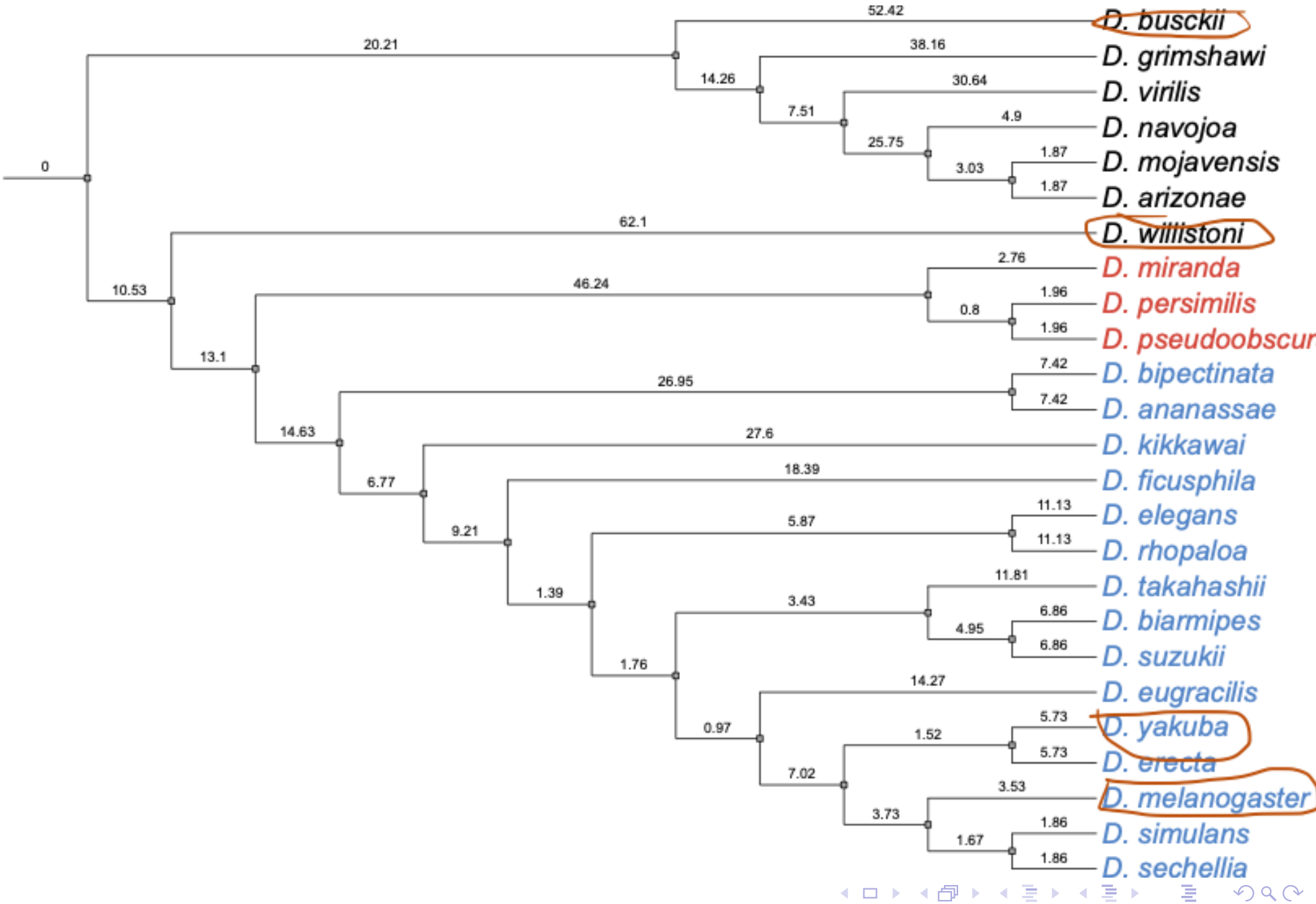
4. blast against own genome and chain hits $\rightarrow C(A, d^A)$ per genome
5. blast against $C(B, d^B)$ of other genomes and identify hits with distance d_1 satisfying $d_1 = \min(d_0^A, d_0^B) - tol \rightarrow$ synteny anchors

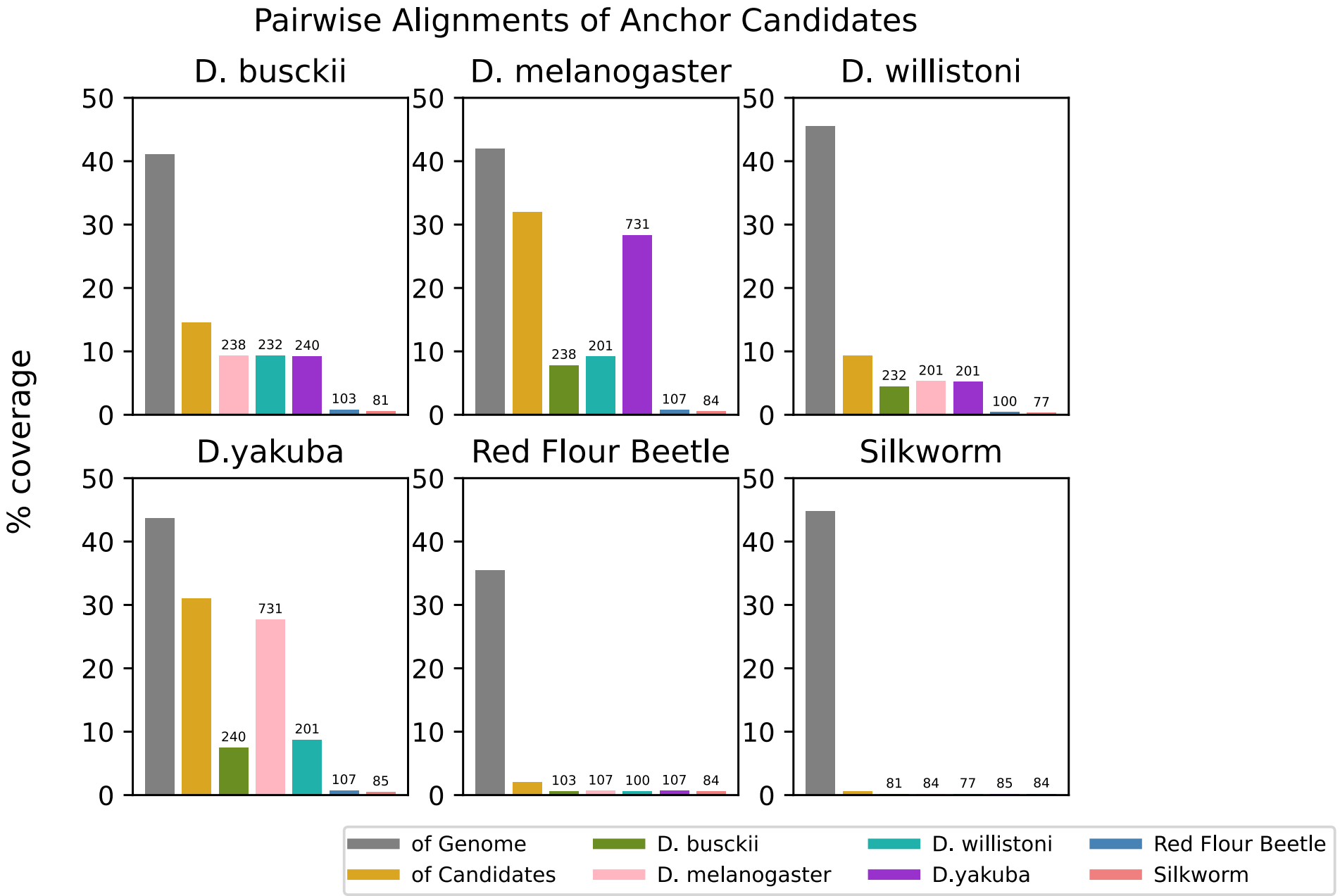
Anchor Candidate Mapping Across Genomes

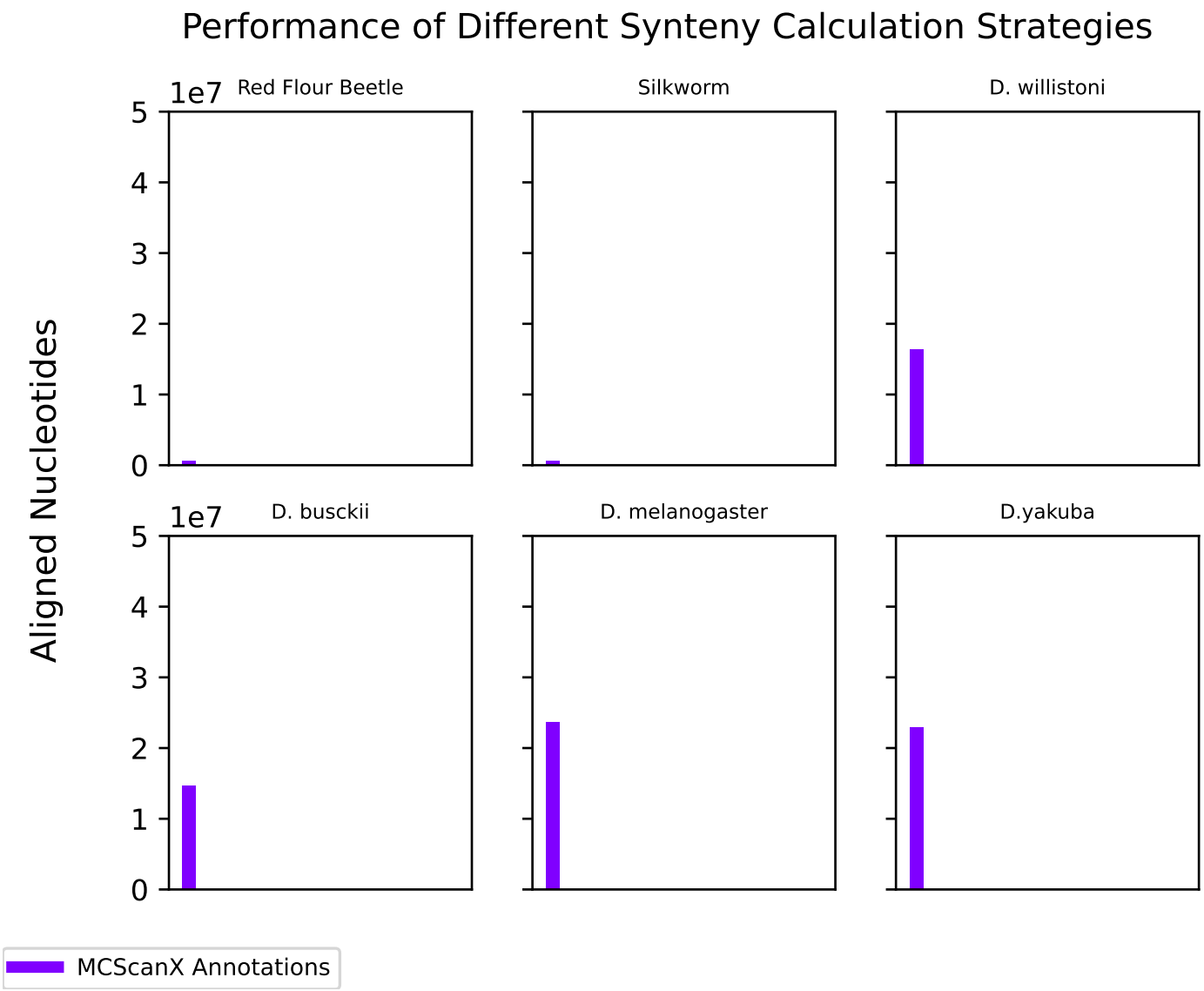
- triangle inequalitiy implies hits between genomes with $d_1 = \min(d_0^A, d_0^B)/2$ are best hit in other genome
- in reality hits with score $\geq$ score of region of candidate + tolerance taken

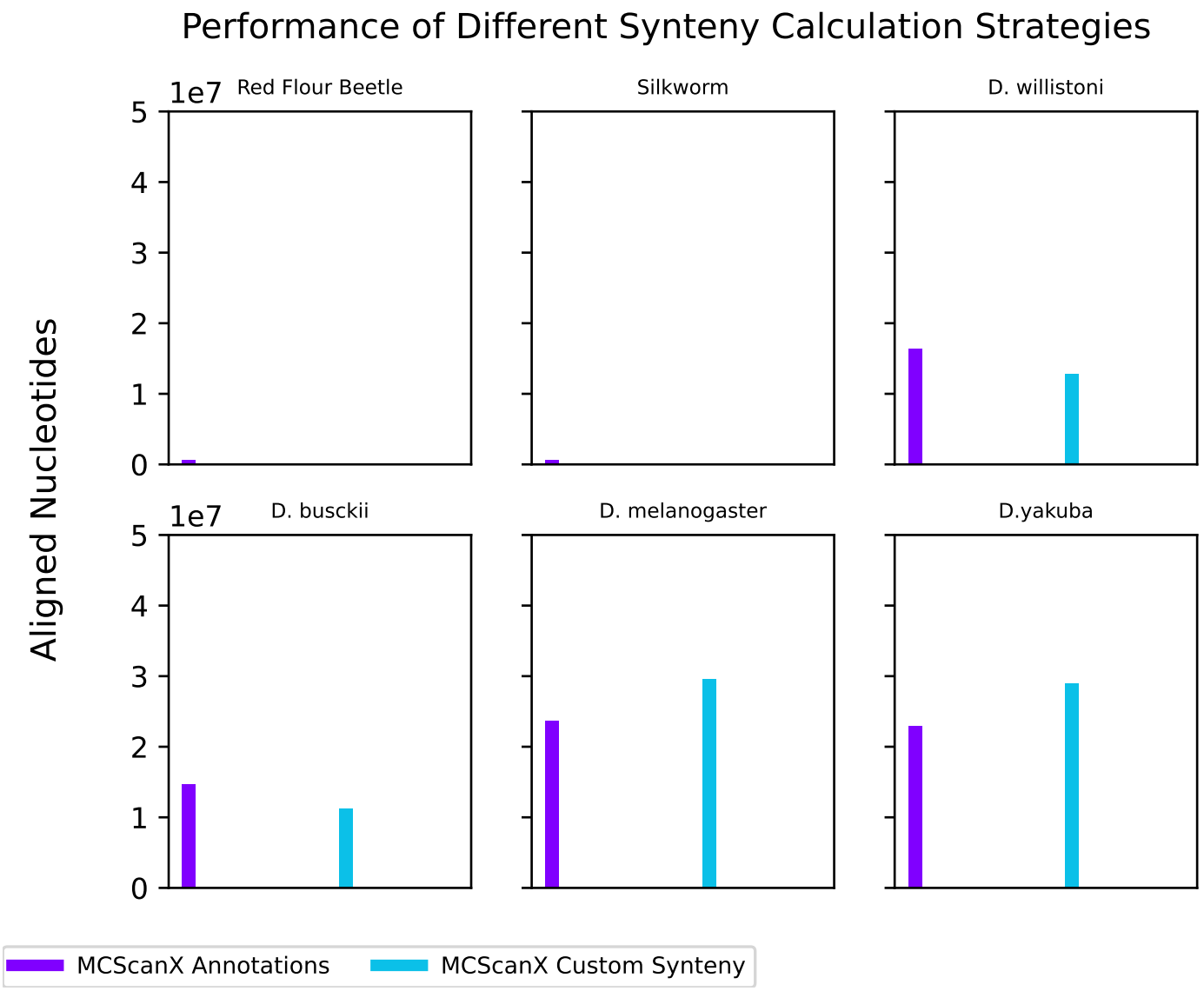


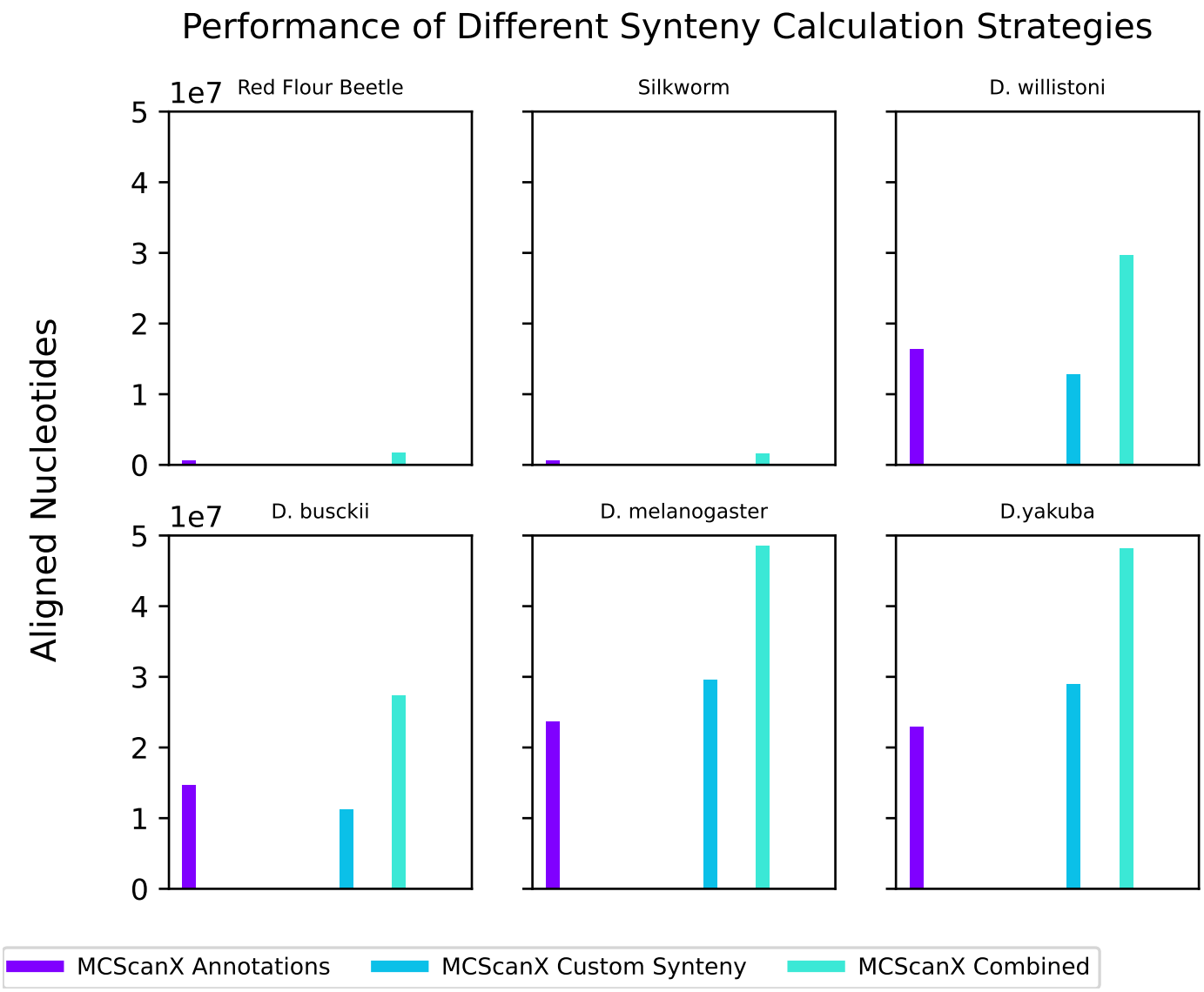

HEXAPODA
INSECTA
PAL.
POLYNEOPTERA
CONDYLO.
HOLOMETABOLA
CHELICERATA
MYRIAPODA
CRUSTACEANS
PROTURA: coneheads
COLLEMBOLA: springtails
DIPLURA: two-pronged bristletails
ARCHAEOGNATHA: jumping bristletails
ZYGENTOMA: silverfish
ODONATA: damselflies & dragonflies
EPHEMEROPTERA: mayflies
ZORAPTERA: ground lice
DERMAPTERA: earwigs
PLECOPTERA: stoneflies
ORTHOPTERA: crickets & katydids
MANTOPHASMATODEA: gladiators
GRYLLOBLATTODEA: ice crawlers
EMBIOPTERA: webspinners
PHASMATODEA: stick & leaf insects
MANTODEA: praying mantids
BLATTODEA: cockroaches
ISOPTERA: termites
THYSANOPTERA: thrips
HEMIPTERA: bugs, cicadas, plant lice
PSOCODEA: bark & true lice
HYMENOPTERA: sawflies, wasps, bees, ants
RAPHIDIOPTERA: snakeflies
MEGALOPTERA: alderflies & dobsonflies
NEUROPTERA: net-winged insects
STREPSIPTERA: twisted wing parasites
COLEOPTERA: beetles
TRICHOPTERA: caddisflies
LEPIDOPTERA: moths & butterflies
SIPHONAPTERA: fleas
MECOPTERA: scorpionflies
DIPTERA: true flies
Bootstrap Support
>98%
>96%
>90%
>85%
>75%
Ma 550 500 450 400 350 300 250 200 150 100 50 0
Ordo. Silurian Devonian Carboniferous Permian Triassic Jurassic Cretaceous Paleogene Neo. Q.
450 Ma 400 350 300 250 200 150 100 50 0
origin of terr. ecos.
origin of vascular plants
origin of trees
radiation of seed plants
P-Tr crisis
Angiosperm radiation
K-Pg crisis
grasslands

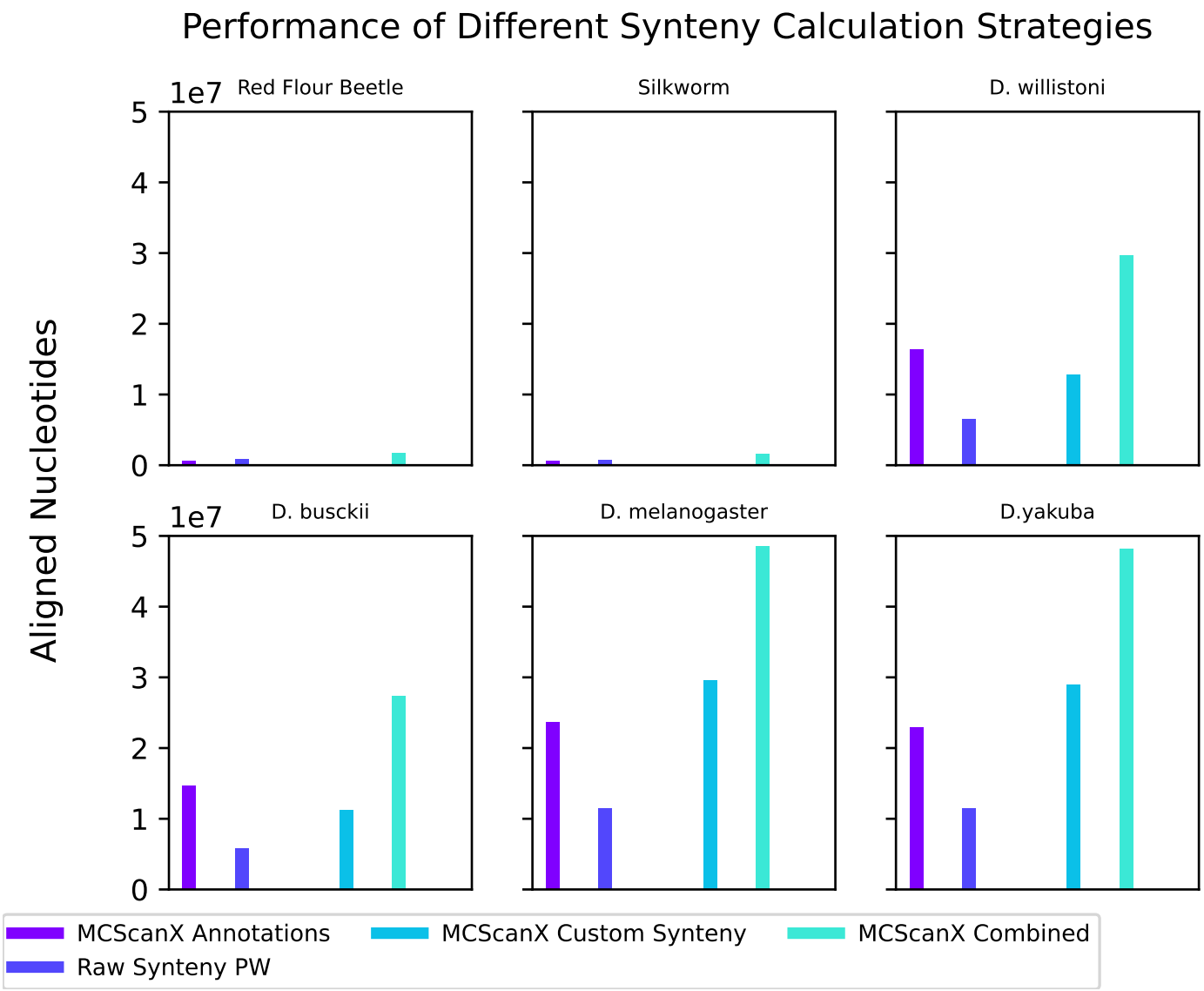

0
20.21
52.42
38.16
14.26
30.64
7.51
4.9
25.75
1.87
3.03
1.87
D. busckii
D. grimshawi
D. virilis
D. navojoa
D. mojavensis
D. arizonae
62.1
D. willistoni
10.53
2.76
46.24
1.96
0.8
1.96
D. miranda
D. persimilis
D. pseudoobscur
13.1
7.42
26.95
7.42
D. bipectinata
D. ananassae
14.63
27.6
D. kikkawai
6.77
18.39
D. ficusphila
11.13
9.21
5.87
11.13
D. elegans
D. rhopaloa
11.81
1.39
3.43
6.86
4.95
6.86
D. takahashii
D. biarmipes
D. suzukii
1.76
14.27
D. eugracilis
5.73
0.97
1.52
5.73
D. yakuba
D. erecta
7.02
3.53
D. melanogaster
3.73
1.86
1.67
1.86
D. simulans
D. sechellia

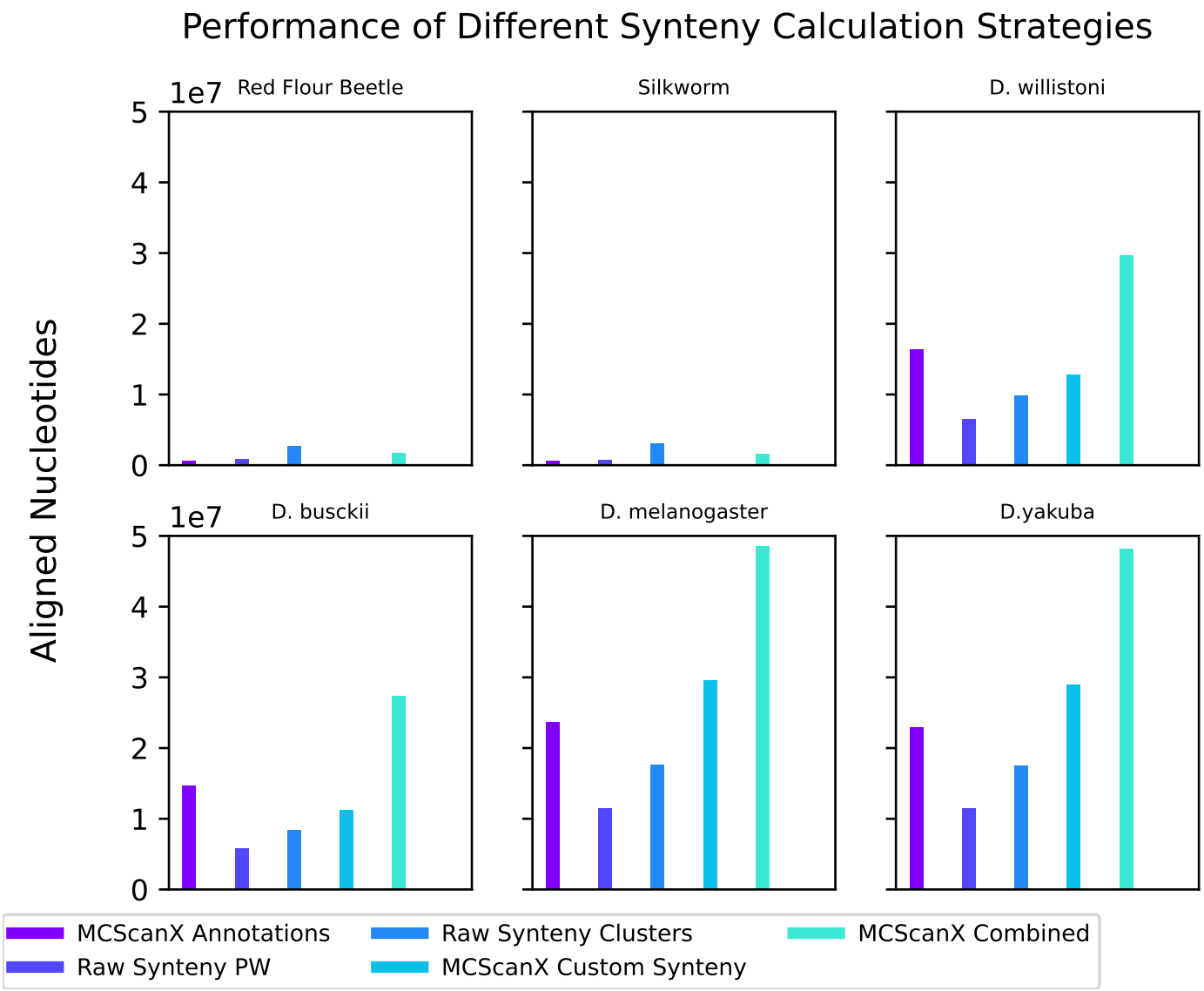

Pairwise Alignments of Anchor Candidates
D. busckii
D. melanogaster
D. willistoni
D.yakuba
Red Flour Beetle
Silkworm
% coverage
0
10
20
30
40
50
238 232 240 103 81
238 201 731 107 84
232 201 201 100 77
240 731 201 107 85
103 107 100 107 84
81 84 77 85 84
of Genome
of Candidates
D. busckii
D. melanogaster
D. willistoni
D.yakuba
Red Flour Beetle
Silkworm

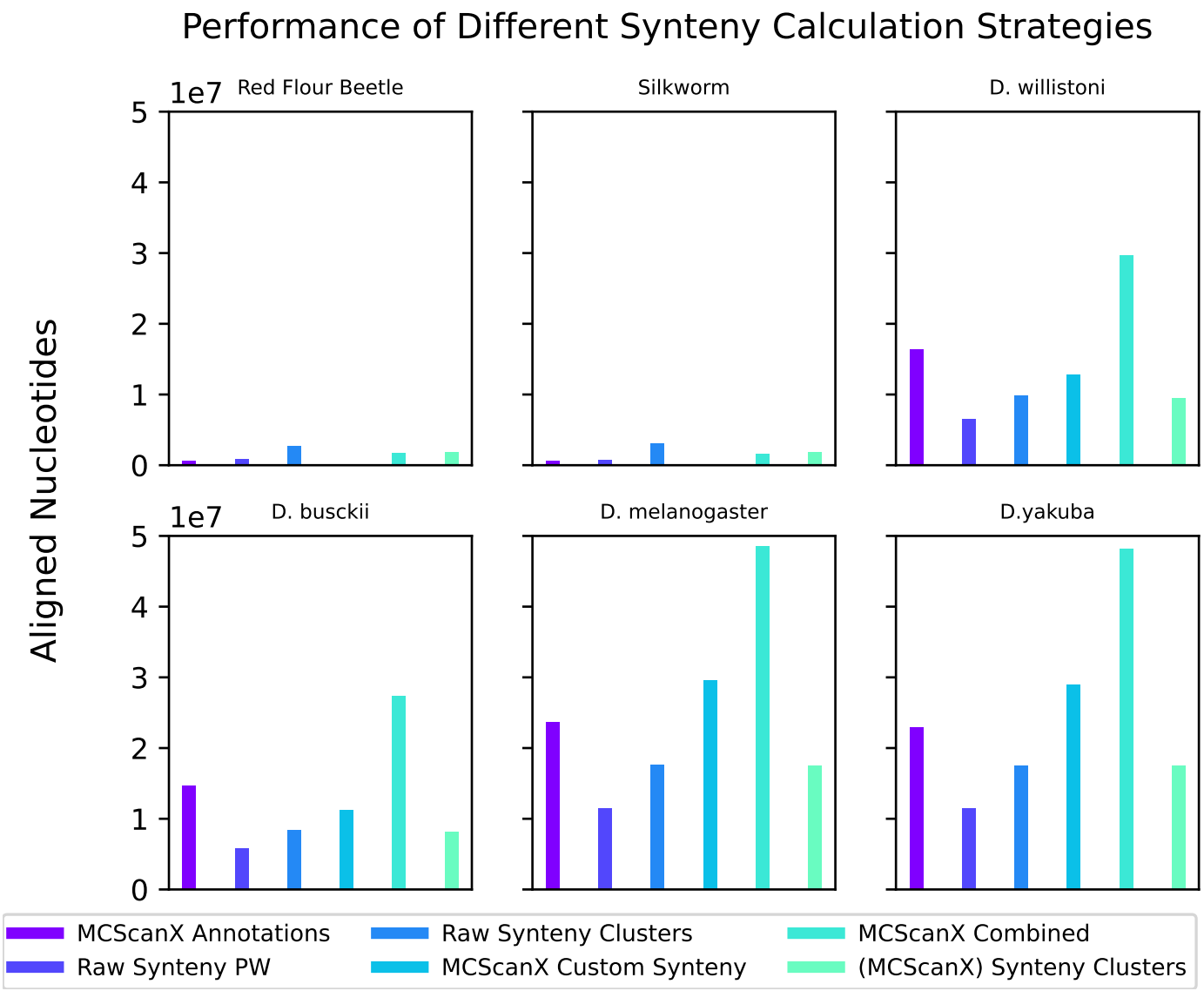

Performance of Different Synteny Calculation Strategies
Aligned Nucleotides
1e7
Red Flour Beetle
Silkworm
D. willistoni
D. busckii
D. melanogaster
D.yakuba
0
1
2
3
4
5
MCScanX Annotations


Performance of Different Synteny Calculation Strategies
Aligned Nucleotides
1e7
Red Flour Beetle
Silkworm
D. willistoni
D. busckii
D. melanogaster
D.yakuba
0
1
2
3
4
5
MCScanX Annotations
MCScanX Custom Synteny


Performance of Different Synteny Calculation Strategies
Aligned Nucleotides
1e7
Red Flour Beetle
Silkworm
D. willistoni
D. busckii
D. melanogaster
D.yakuba
0
1
2
3
4
5
MCScanX Annotations
MCScanX Custom Synteny
MCScanX Combined


Performance of Different Synteny Calculation Strategies
Aligned Nucleotides
1e7
Red Flour Beetle
Silkworm
D. willistoni
D. busckii
D. melanogaster
D.yakuba
0
1
2
3
4
5
MCScanX Annotations
Raw Synteny PW
MCScanX Custom Synteny
MCScanX Combined


Performance of Different Synteny Calculation Strategies
Aligned Nucleotides
1e7
Red Flour Beetle
Silkworm
D. willistoni
D. busckii
D. melanogaster
D.yakuba
0
1
2
3
4
5
MCScanX Annotations
Raw Synteny PW
Raw Synteny Clusters
MCScanX Custom Synteny
MCScanX Combined


Performance of Different Synteny Calculation Strategies
Aligned Nucleotides
1e7
Red Flour Beetle
Silkworm
D. willistoni
D. busckii
D. melanogaster
D.yakuba
0
1
2
3
4
5
MCScanX Annotations
Raw Synteny PW
Raw Synteny Clusters
MCScanX Custom Synteny
MCScanX Combined
(MCScanX) Synteny Clusters

Computational Resources

- set of 70 diptera (true flies) genomes
  - avg. size $\sim$ 415 MB
  - from very poor to very good assemblies
  - I am told they cover the phylogeny of true flies well
- results
  - takes around 2 weeks on Leipzig Bioinf cluster
  - core results around 15 GB in python dict(s)
  - coverage of genome with candidates overall: 32.58 % $\pm$ 4.05
  - length of $\sim$ 471 $\pm$ 663 and spacing in between of $\sim$ 918 $\pm$ 190
  - for candidates with $\geq$ 5 matches length is around 1000 and spacing 16000
  - how much of candidates are aligned somewhere (to any other species): 20.53 % $\pm$ 18.23
  - how much of candidates are aligned counting the mean of all alignments per candidate: 3.20 % $\pm$ 2.29

- Altenhoff, Adrian M et al. "The Quest for Orthologs benchmark service and consensus calls in 2020". In: *Nucleic Acids Research* 48 (W1 July 2, 2020), W538–W545. ISSN: 0305-1048. DOI: 10.1093/nar/gkaa308. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7319555/ (visited on 09/08/2023).
- Haas, Brian J. et al. "DAGchainer: a tool for mining segmental genome duplications and synteny". In: *Bioinformatics* 20.18 (Dec. 12, 2004), pp. 3643–3646. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/bth397. URL: https://doi.org/10.1093/bioinformatics/bth397 (visited on 09/08/2023).
- Moyers, Bryan A. and Jianzhi Zhang. "Further Simulations and Analyses Demonstrate Open Problems of Phylostratigraphy". In: *Genome Biology and Evolution* 9.6 (June 1, 2017), pp. 1519–1527. ISSN: 1759-6653. DOI: 10.1093/gbe/evx109. URL: https://doi.org/10.1093/gbe/evx109 (visited on 09/08/2023).
- Salzberg, Steven L. "Next-generation genome annotation: we still struggle to get it right". In: *Genome Biology* 20.1 (May 16, 2019),

Thanks for the attention :)