Improving Genome Assemblies With Syntenic Alignments

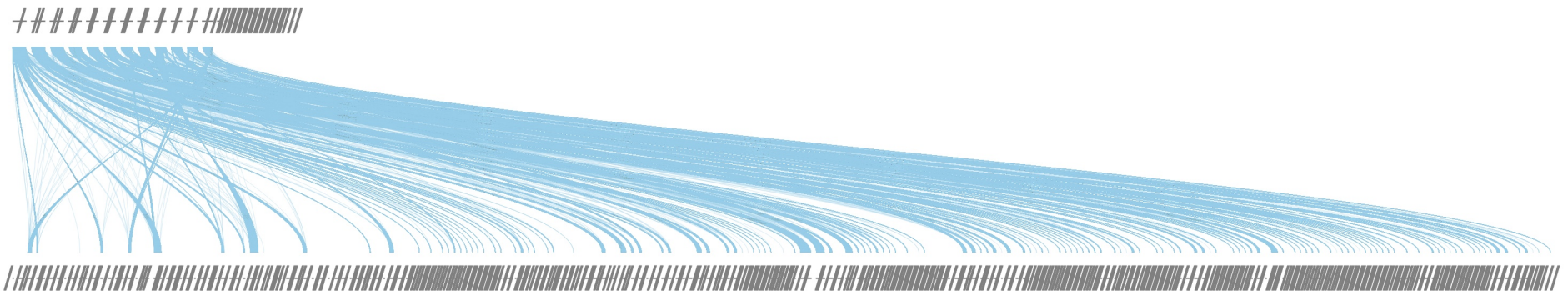
Karl Kaether

Abteilung Stadler, Leipzig

Bled 2025

Context

- Synteny alignments provide signal for reference-guided contig scaffolding



Context

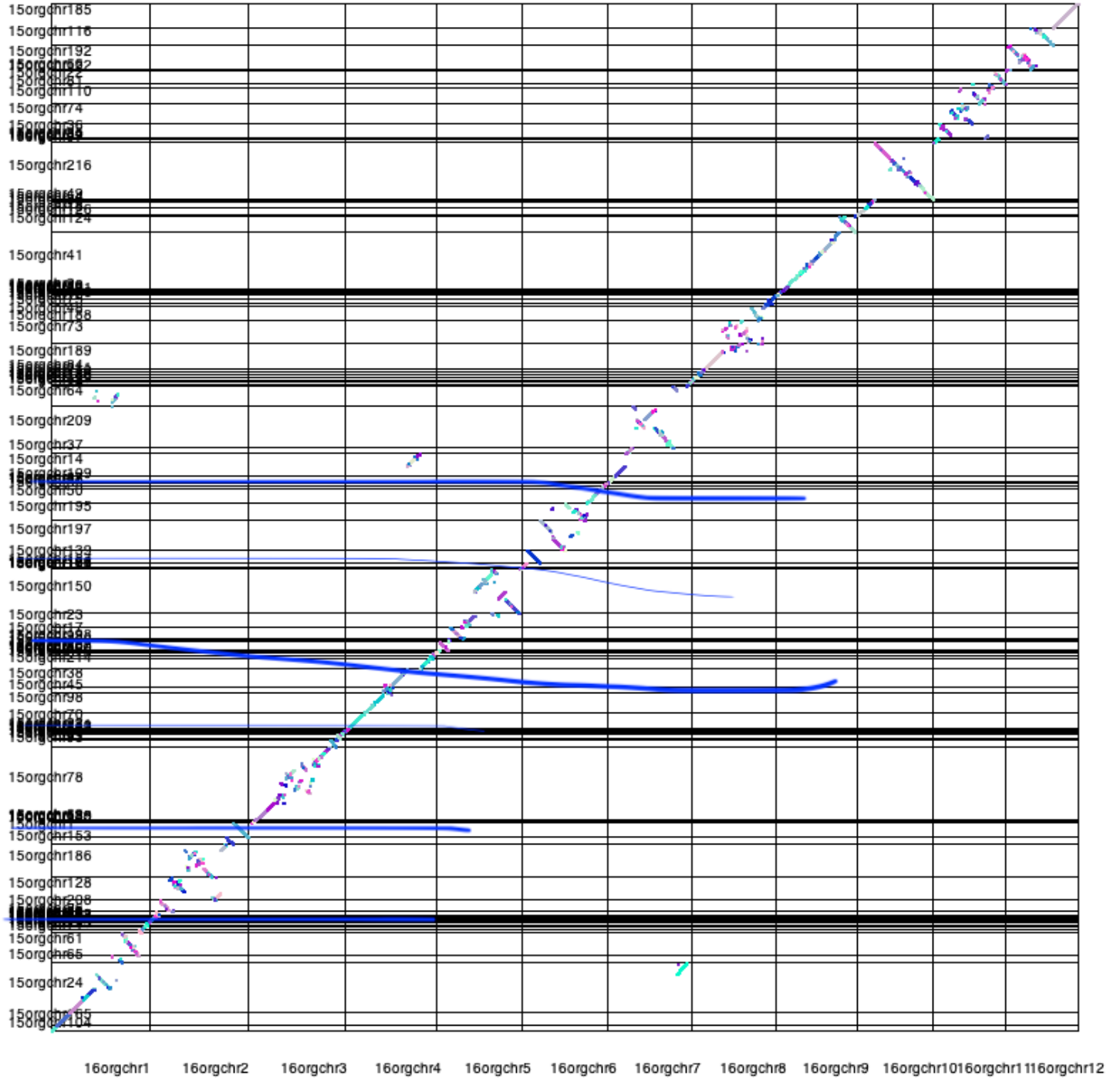
- Synteny alignments provide signal for reference-guided contig scaffolding
- note: not classical reference guided assembly but contig reordering
- tools exist: e.g. Ragout2 , MeDuSa, (Multi-)CSAR
- all use some kind of alignments + optimization to reduce number of rearrangements

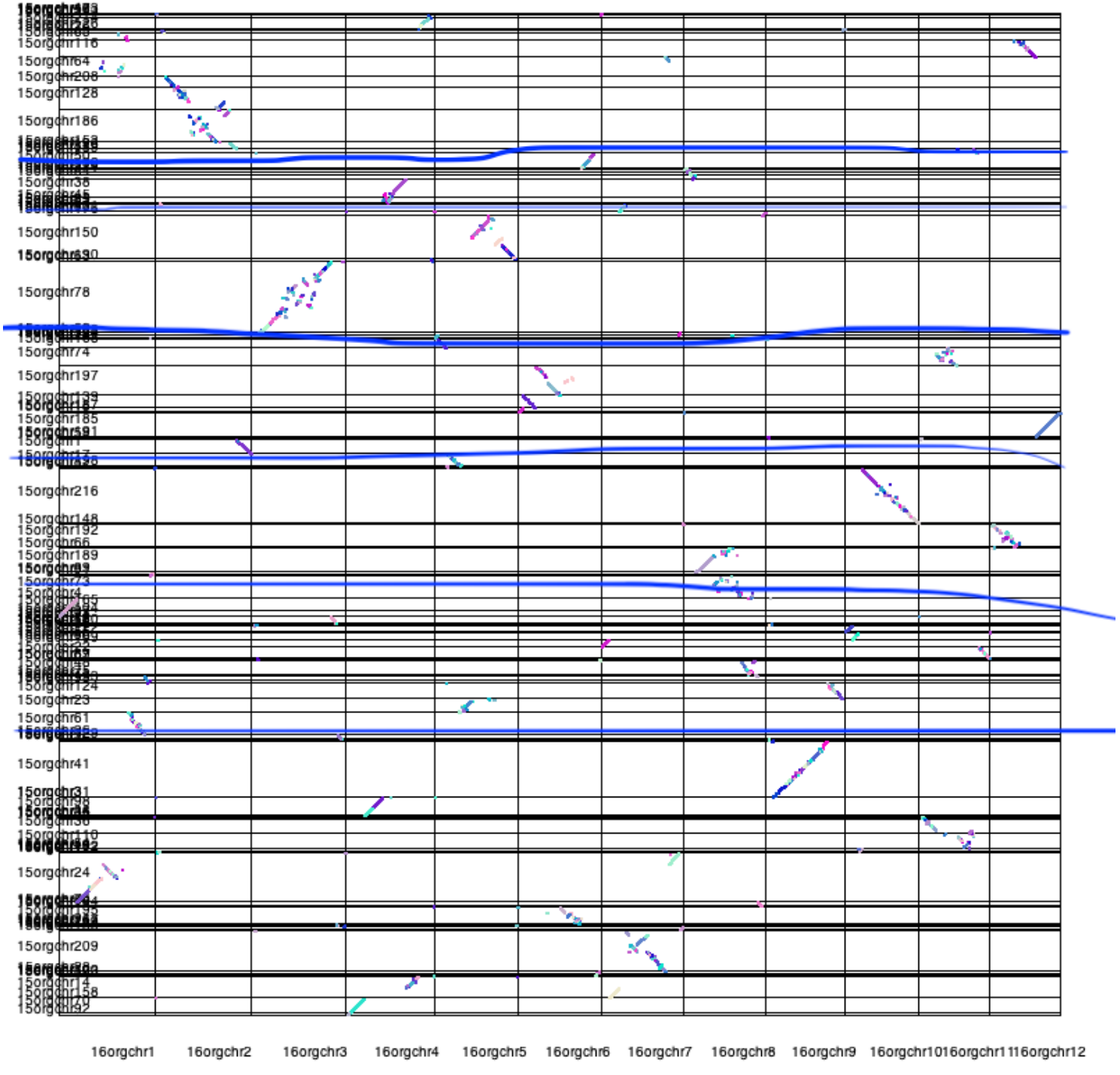
Lets give it a try

- need some model for evolutionary events
- would be nice to be able to include multiple references
- example: Solanum stellatiglandulosum - kindly assembled by Thomas

Lets give it a try

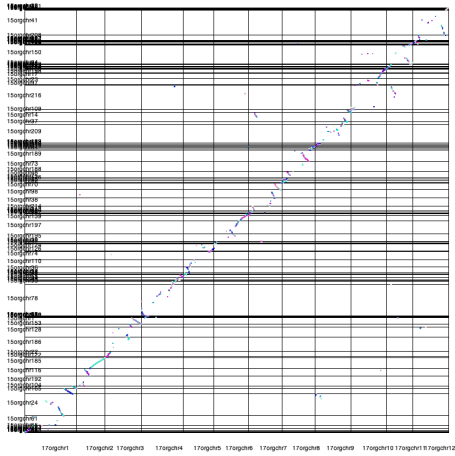
- need some model evolutionary events:
  - simple model - regard reference as fully assembled and contigs as reliable
  - enumerate where and which way anchors on target contigs align on reference chromosomes
  - orient contig according to the majority of alignments (and enumerate possible inversion events) - corresponds to minimizing inversion events
  - define the order of target contigs on reference chromsomes based on majority of alignments - corresponds to minimizing translocations


15orgchr185
15orgchr116
15orgchr192
15orgchr110
15orgchr74
15orgchr216
15orgchr124
15orgchr41
15orgchr73
15orgchr189
15orgchr64
15orgchr209
15orgchr37
15orgchr14
15orgchr50
15orgchr195
15orgchr197
15orgchr150
15orgchr23
15orgchr38
15orgchr45
15orgchr98
15orgchr78
15orgchr153
15orgchr186
15orgchr128
15orgchr61
15orgchr65
15orgchr24
16orgchr1
16orgchr2
16orgchr3
16orgchr4
16orgchr5
16orgchr6
16orgchr7
16orgchr8
16orgchr9
16orgchr10
16orgchr11
16orgchr12

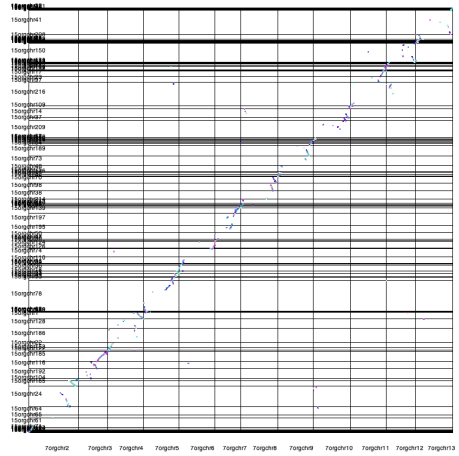

15orgchr116
15orgchr64
15orgchr208
15orgchr128
15orgchr186
15orgchr38
15orgchr150
15orgchr78
15orgchr74
15orgchr197
15orgchr185
15orgchr216
15orgchr148
15orgchr192
15orgchr66
15orgchr189
15orgchr23
15orgchr4
15orgchr124
15orgchr23
15orgchr61
15orgchr41
15orgchr24
15orgchr209
15orgchr14
15orgchr158
16orgchr1
16orgchr2
16orgchr3
16orgchr4
16orgchr5
16orgchr6
16orgchr7
16orgchr8
16orgchr9
16orgchr10
16orgchr11
16orgchr12

Lets give it a try

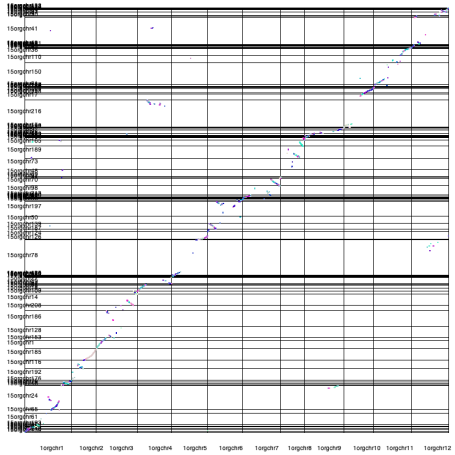
- would be nice to be able to include multiple references:
  - followed Multi-CSAR algorithm



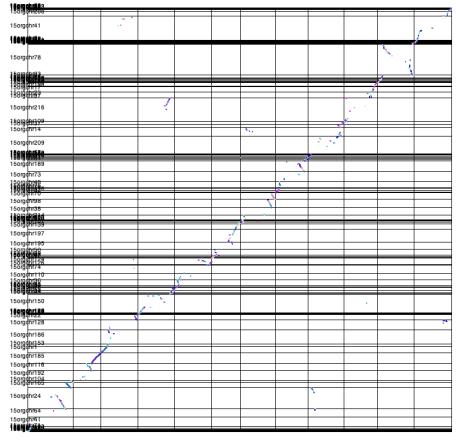
(a) Solanum tuberosum



(b) Solanum lycopersicum



(c) Capsicum annuum



(d) Solanum americanum

(a)

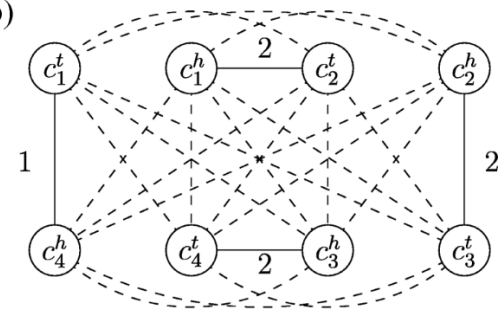
$T = \{c_1, c_2, c_3, c_4\}$

$S_1 = (+c_1, +c_2, +c_3)$

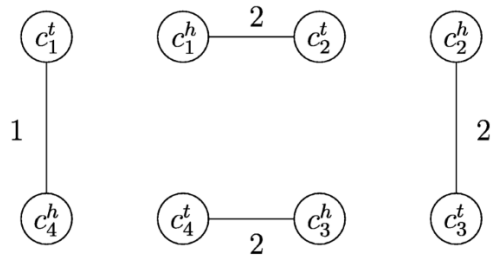
$S_2 = (+c_2, +c_3, +c_4)$

$S_3 = (-c_2, -c_1, -c_4, -c_3)$

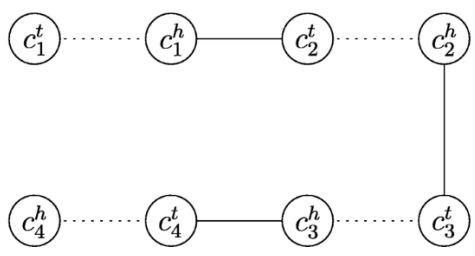
(b)



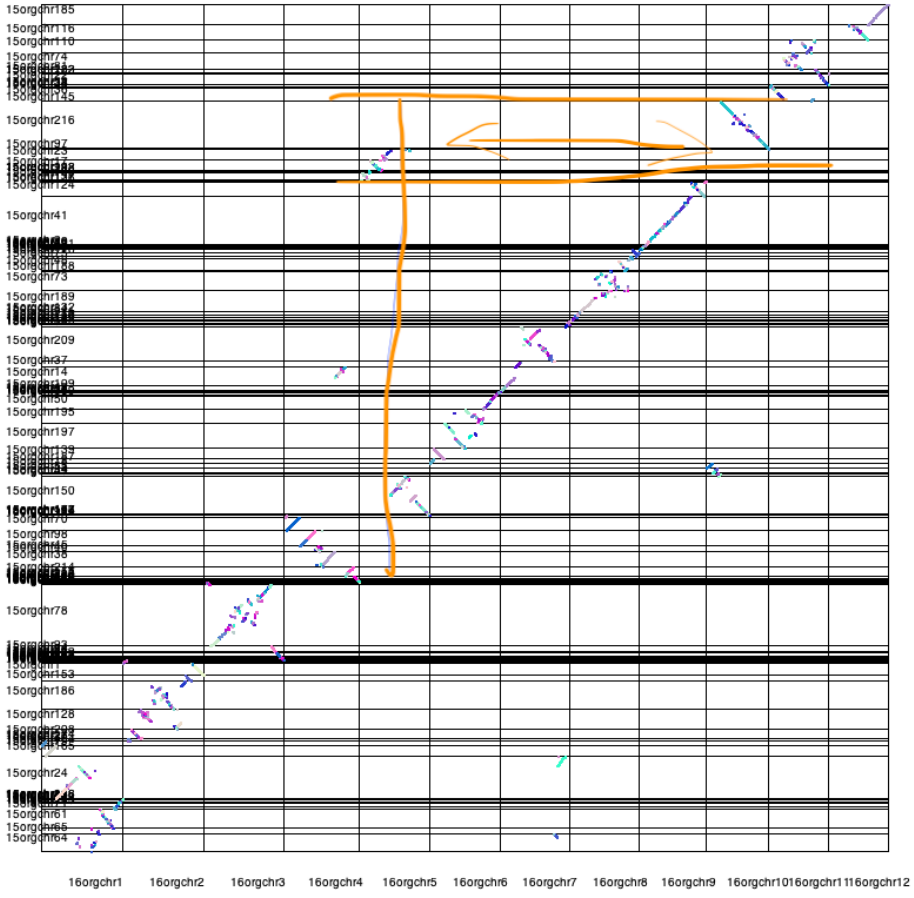
(c)



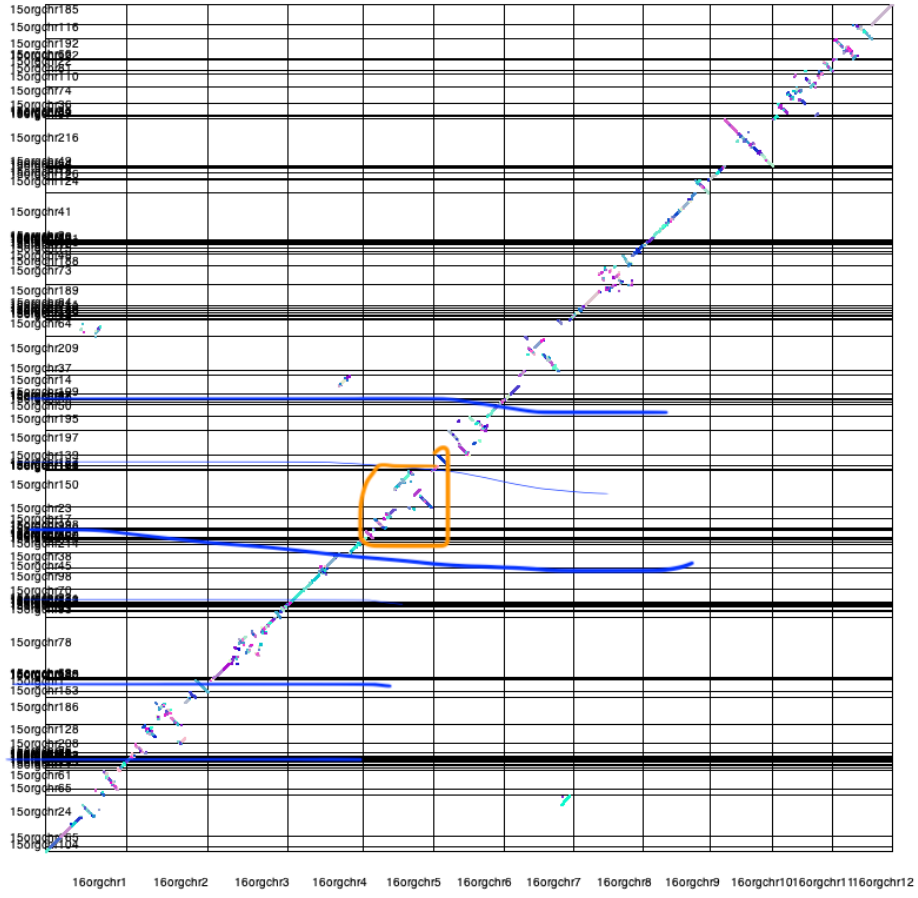
(d)



(e) Scaffold of $T = (+c_1, +c_2, +c_3, +c_4)$



(a) Consensus vs. S.torvum



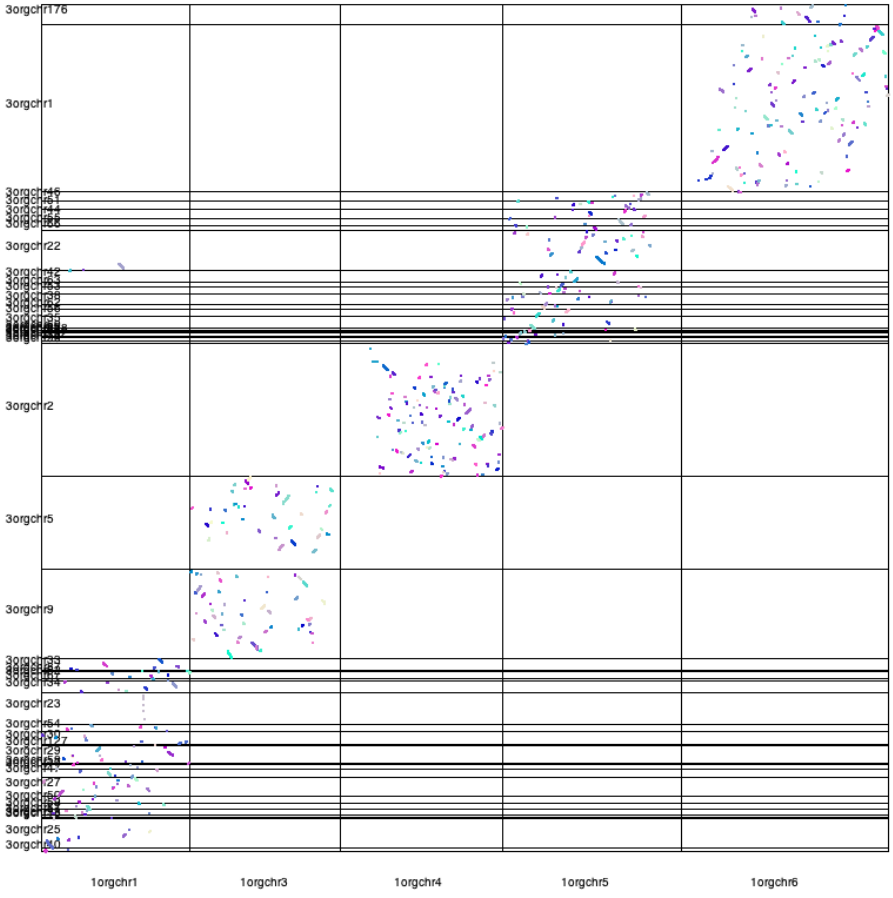
(b) Solanum Torvum

Table: Improved N50 from original 28 MB

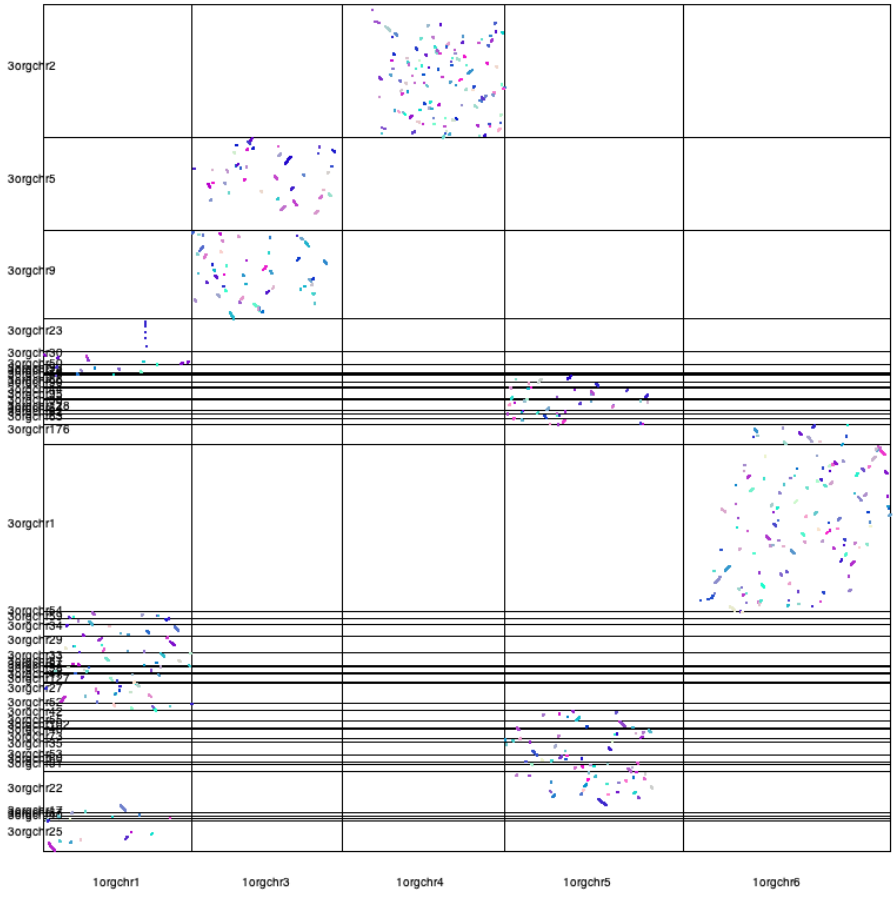
Reference Genome	scaffolded N50 all	scaffolded N50 limited
C.annuum	119 MB	92 MB
S.americanum	107 MB	95 MB
S.candolleanum	83 MB	57 MB
S. dulcamara	104 MB	87 MB
S. galapagense	107 MB	100 MB
S. laciniatum	51 MB	35 MB
S. lycopersicum	107 MB	107 MB
S. lyratum	108 MB	104 MB
S. melongena	107 MB	100 MB
S. pimpinellifolium	107 MB	100 MB
S. pinnatisectum	107 MB	95 MB
S. pseudocapsicum	79 MB	63 MB
S. retroflexum	77 MB	31 MB
S. spirale	73 MB	28 MB
S. torvum	108 MB	108 MB
S. tuberosum	107MB	100 MB
S. viarum	107 MB	99 MB
Consensus neutral weights	100 MB	102 MB
Consensus informed weights	101 MB	100 MB

Take home

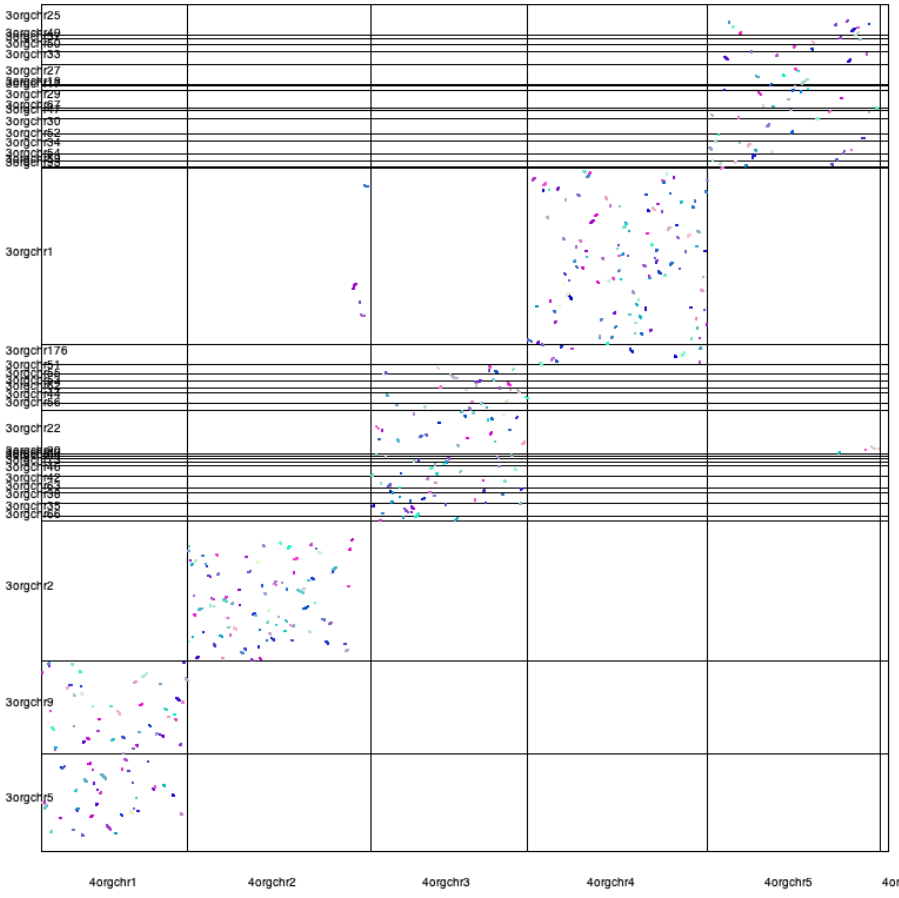
- reference-guided contig scaffolding possible without MSA or gene homology data
- way more benchmarking needed to see within what boundaries it can work
- potentially faster and more scalable
- really interesting bit are the potential breakpoints



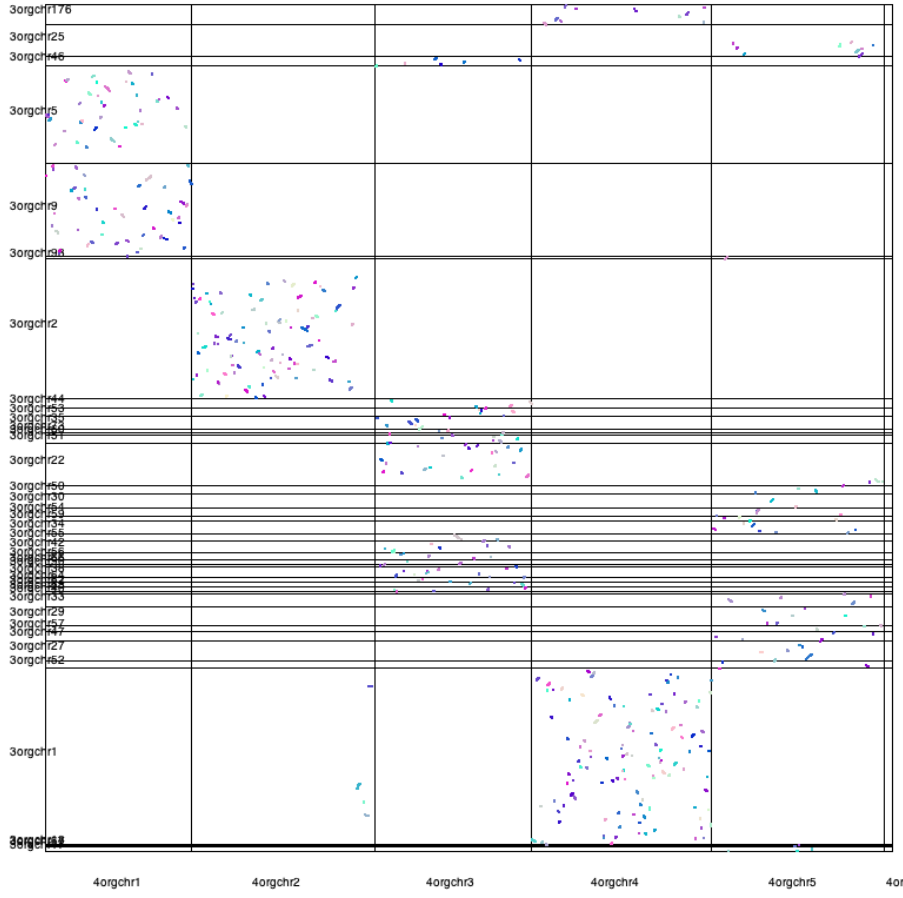
(a) D.mel



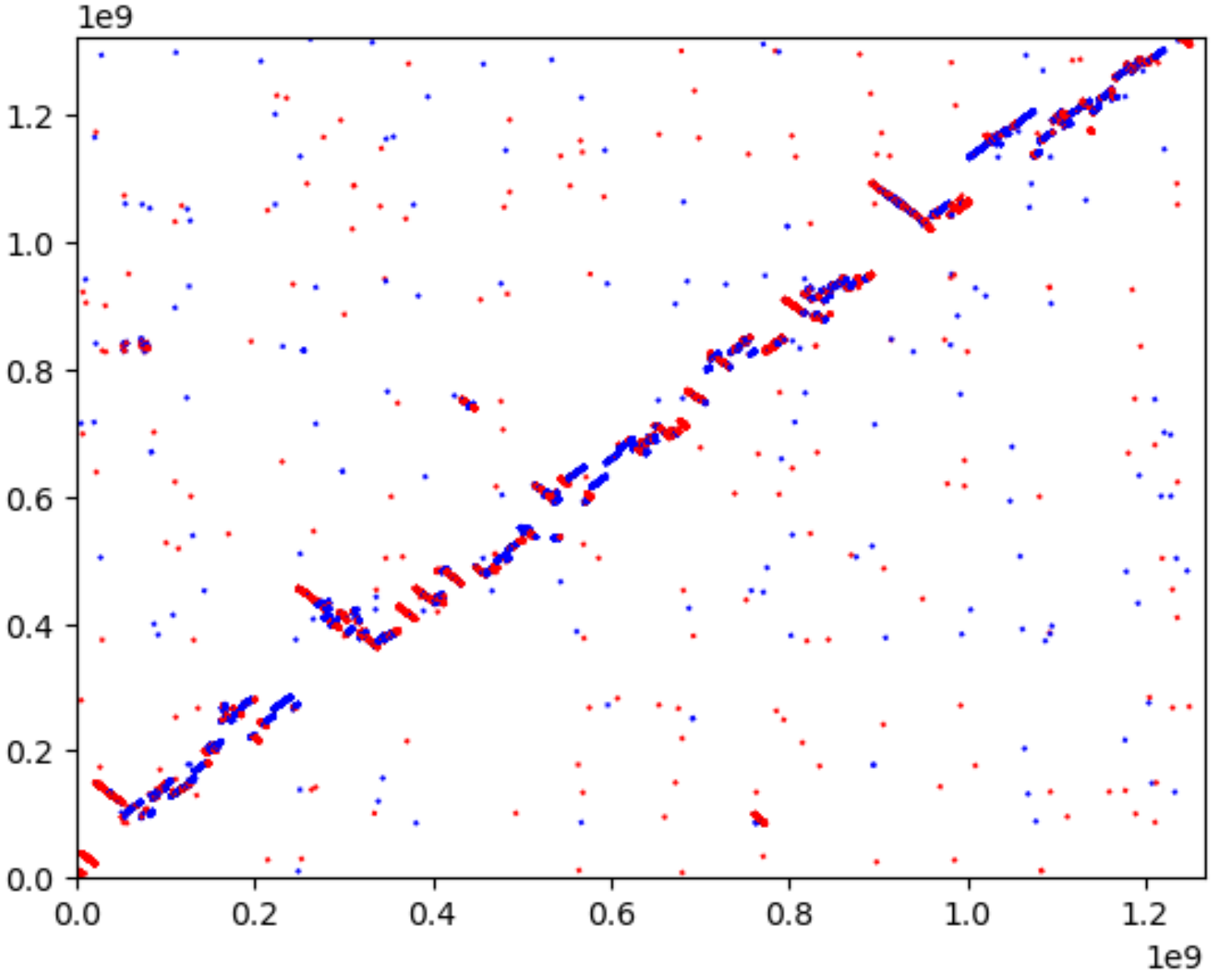
(b) D.mel CSAR

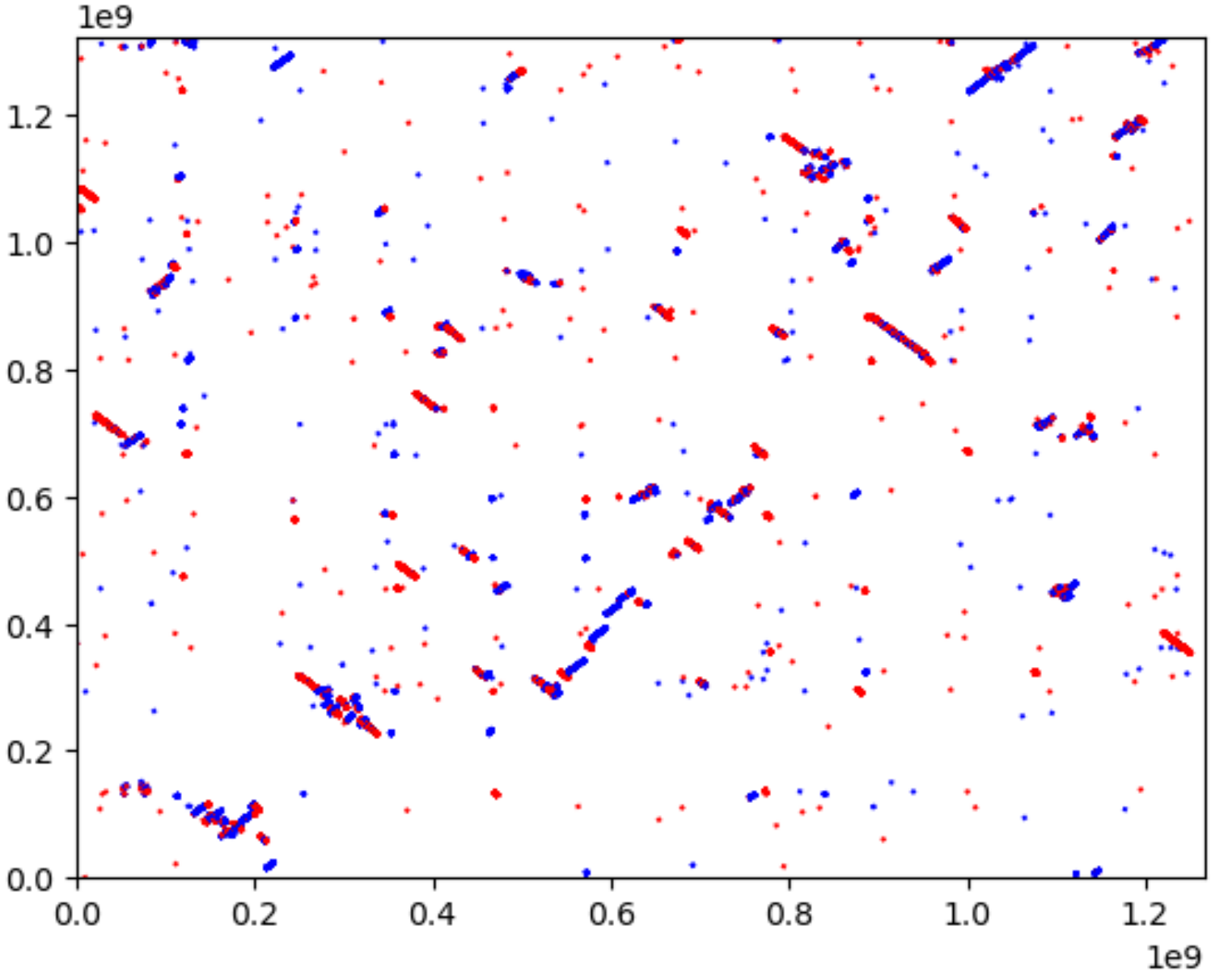


(a) D.busckii



(b) D.busckii CSAR


1e9
1.2
1.0
0.8
0.6
0.4
0.2
0.0
0.0
0.2
0.4
0.6
0.8
1.0
1.2
1e9


1e9
1.2
1.0
0.8
0.6
0.4
0.2
0.0
0.0
0.2
0.4
0.6
0.8
1.0
1.2
1e9