



Jonathan Burns and Denys Kukushkin
University of South Florida

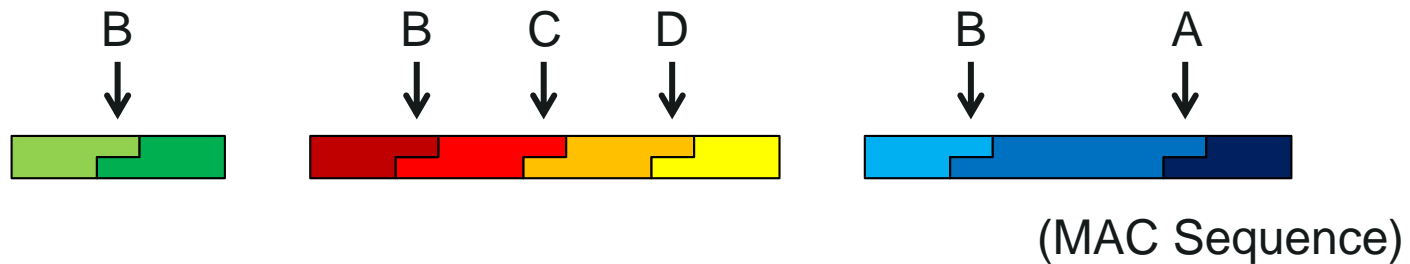
May 12, 2014

Visualizing Genetic Recombination

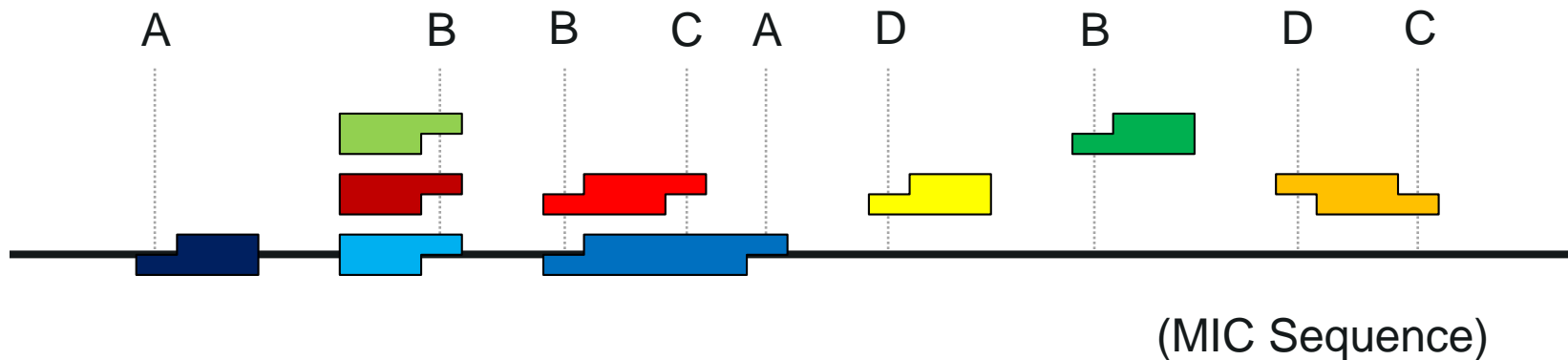
- Assembly Graphs
- Chord Diagrams

Assembly Graph Construction

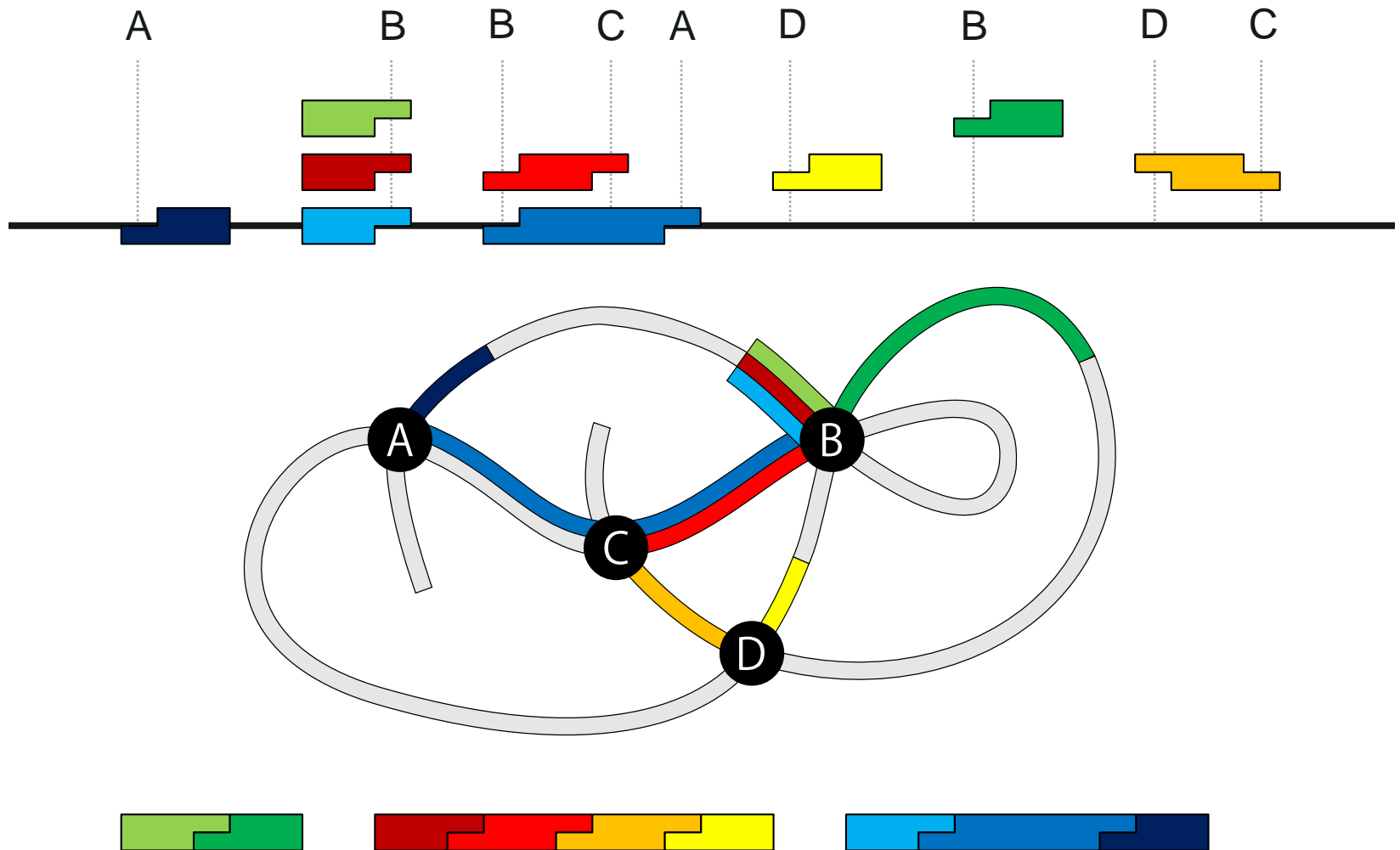
- Identify pointers by the overlaps in MAC sequence



- Assign two pointers the same symbol if they overlap in either MAC or MIC



Assembly Graph Construction

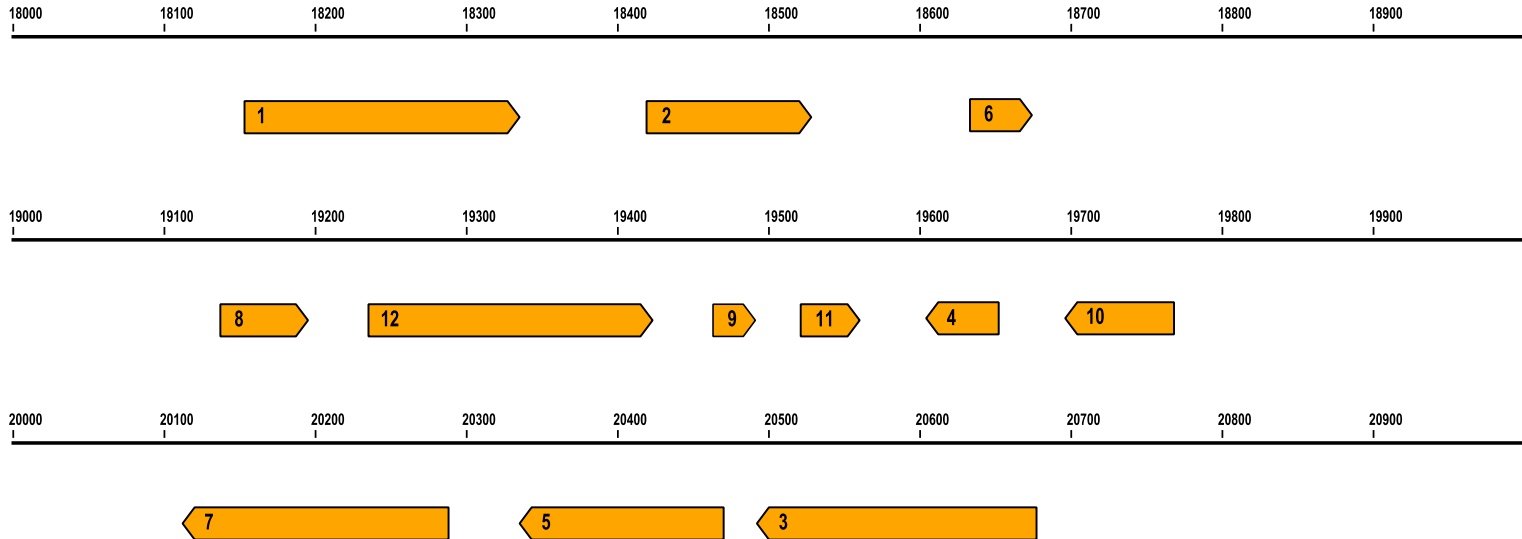


Assembly Graph Examples

- Single Contig
- Scaffold: Non-overlapping MDSs
- Scaffold: Fully-overlapping MDSs
- Scaffold: Partially-overlapping MDSs

Single Contig

(*Oxytricha trifallax*, MIC scaffold1572)

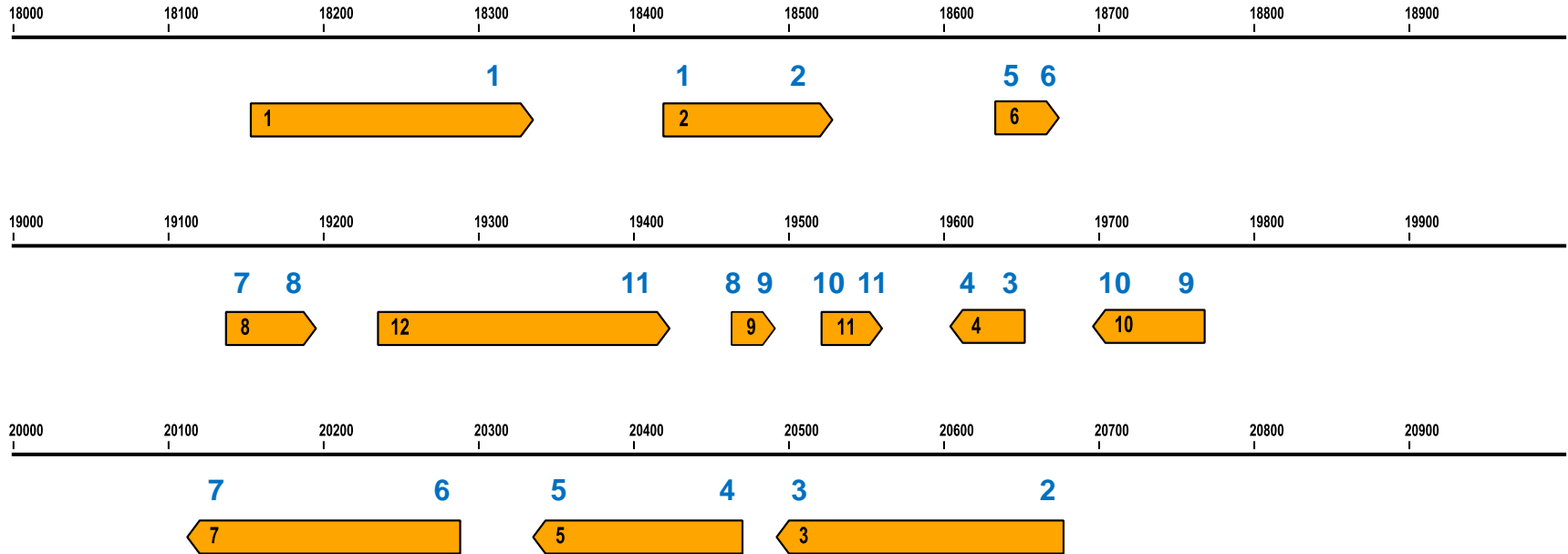


(Contig12928.0, MAC)



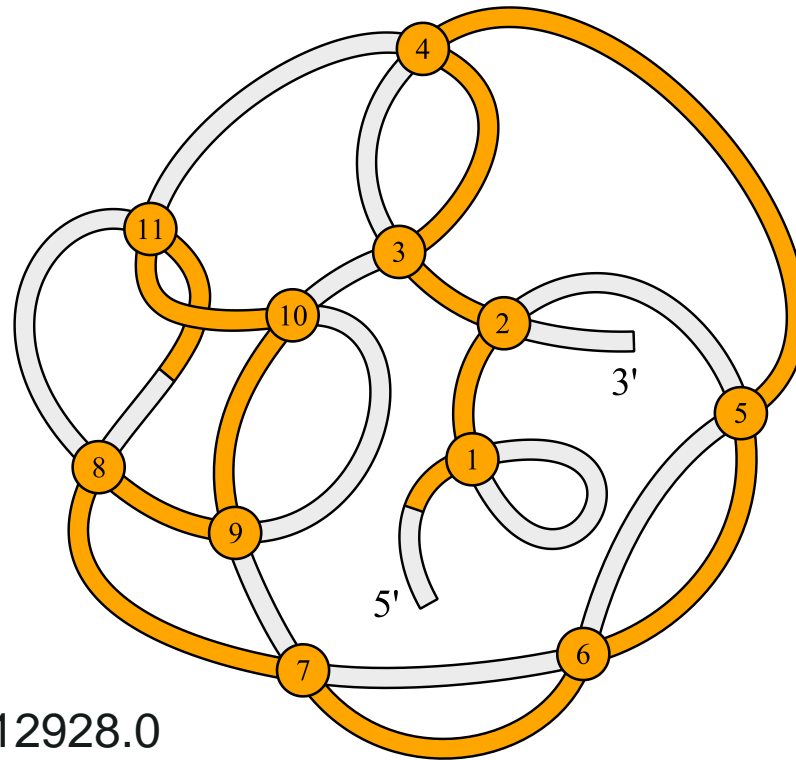
(Chen et. al., 2014)

Single Contig



$$w = 1, 1, 2, 5, 6, 7, 8, 11, 8, 9, 10, 11, 4, 3, 10, 9, 7, 6, 5, 4, 3, 2$$

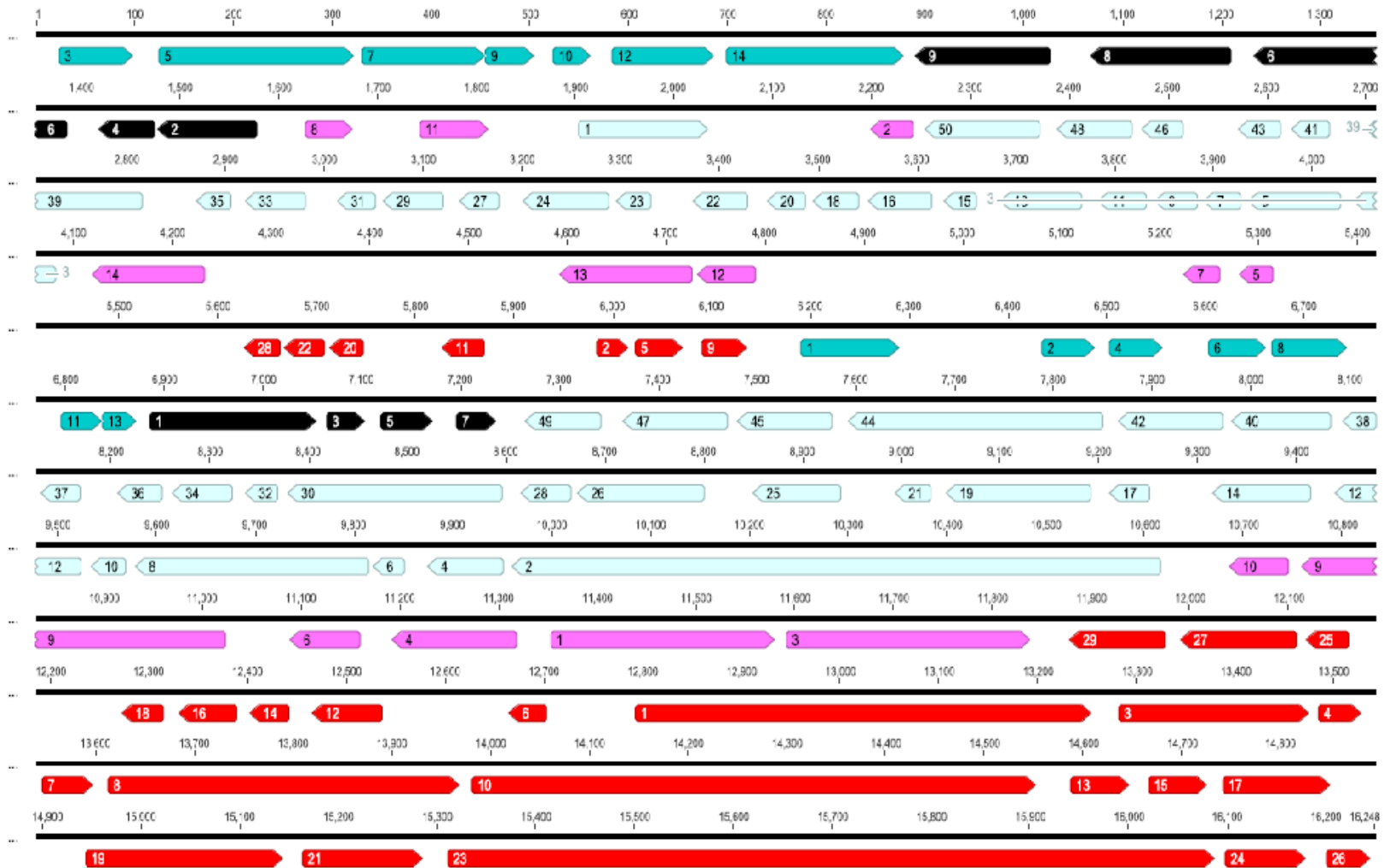
Single Contig



Contig12928.0

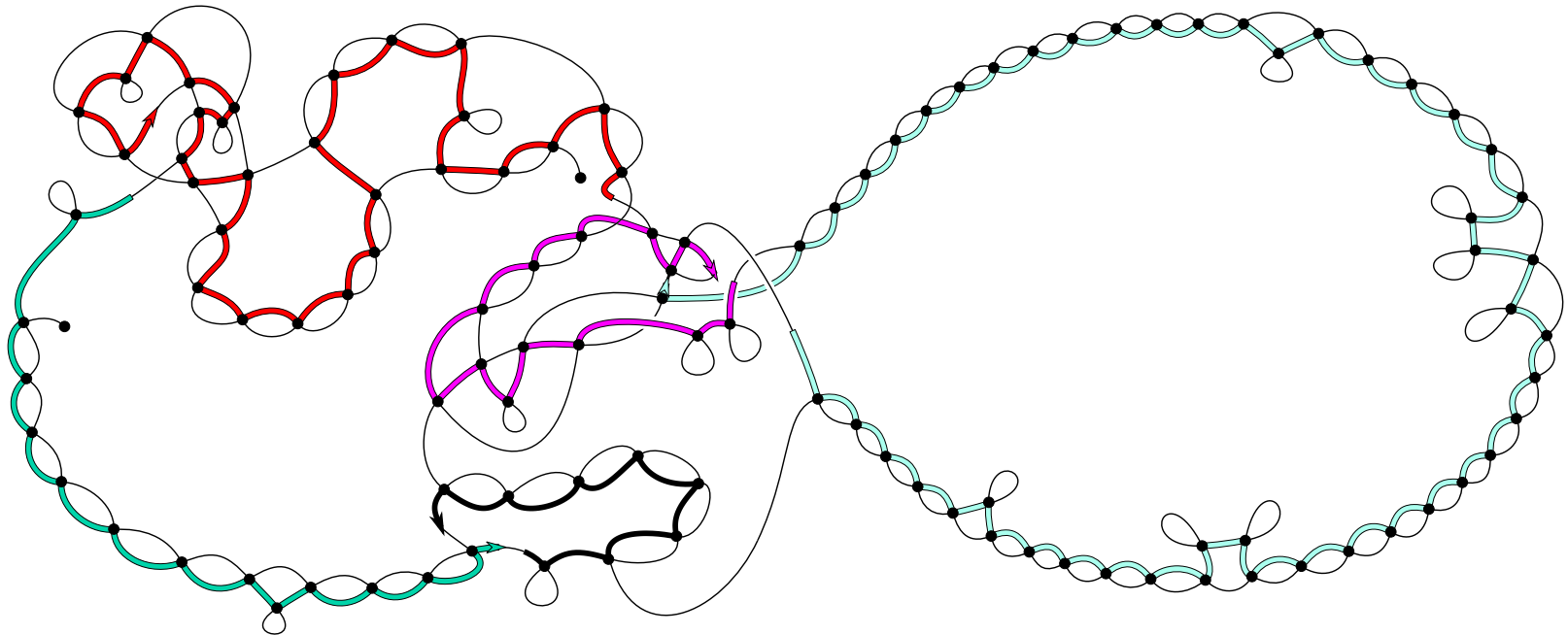
$w = 1, 1, 2, 5, 6, 7, 8, 11, 8, 9, 10, 11, 4, 3, 10, 9, 7, 6, 5, 4, 3, 2$

Scaffold: Non-overlapping MDSs

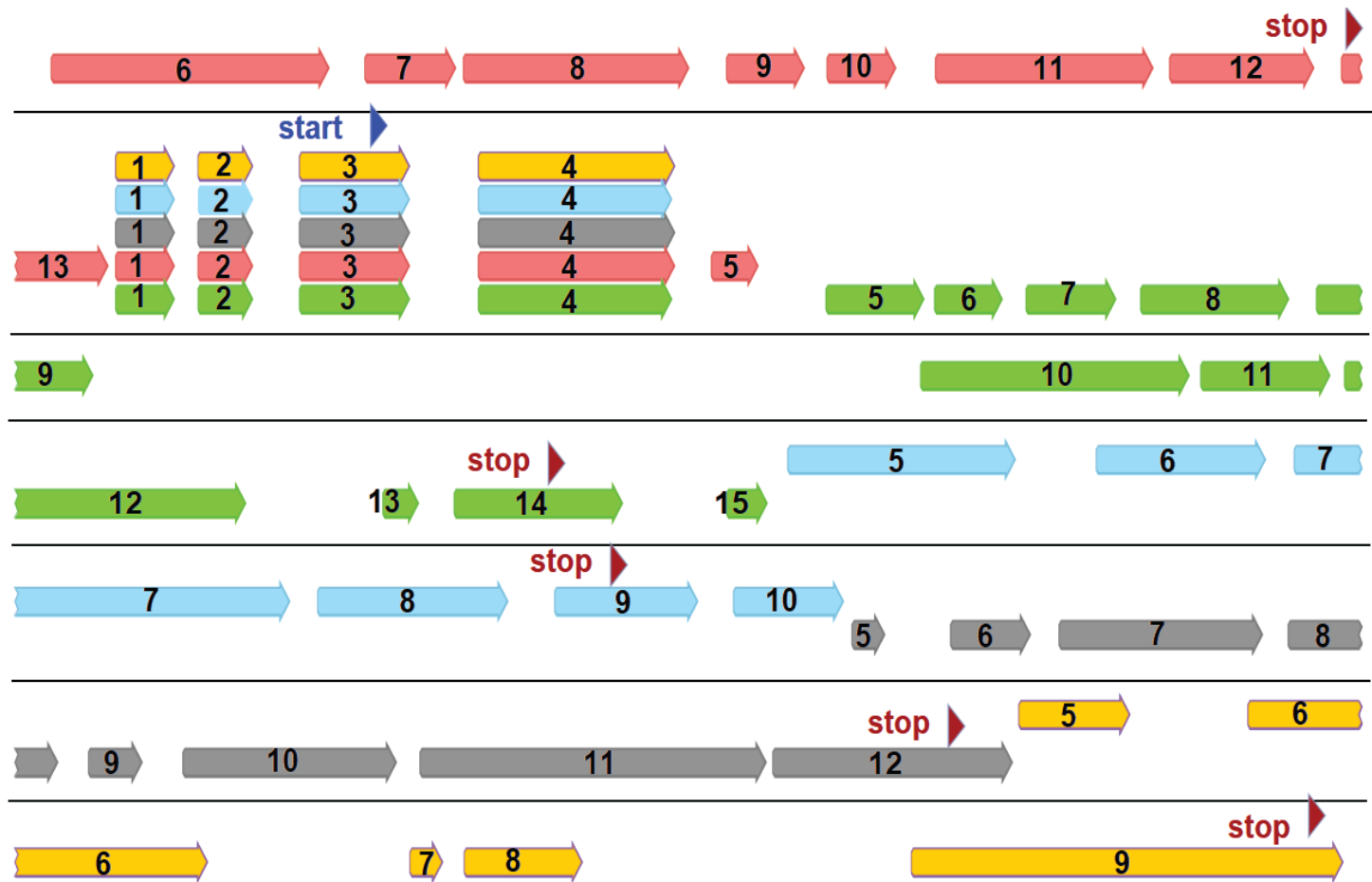


(Chen et. al., 2014)

Scaffold: Non-overlapping MDSs



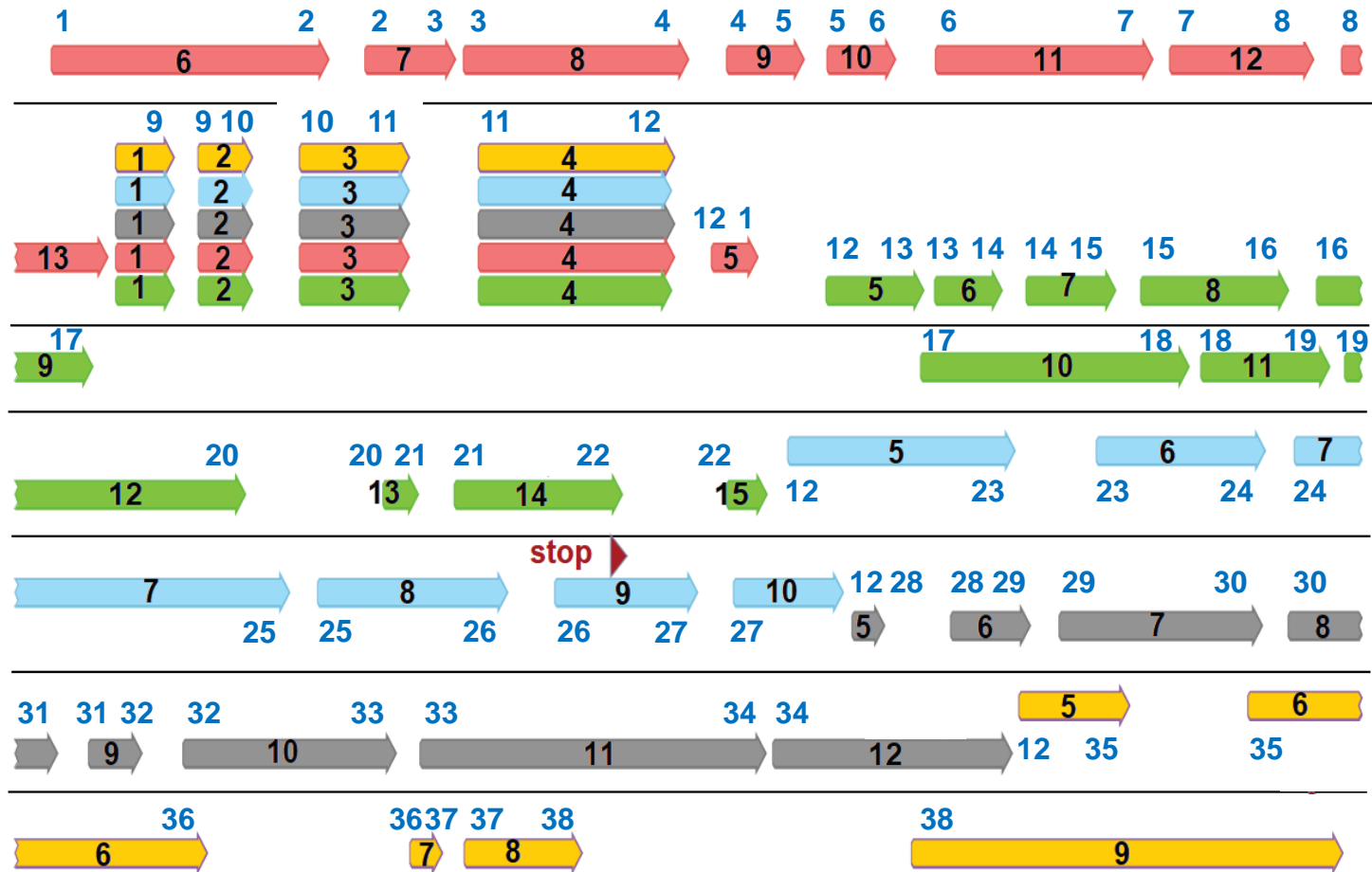
Scaffold: Fully-overlapping MDSs



(*Oxytricha trifallax*, MIC scaffold38094)

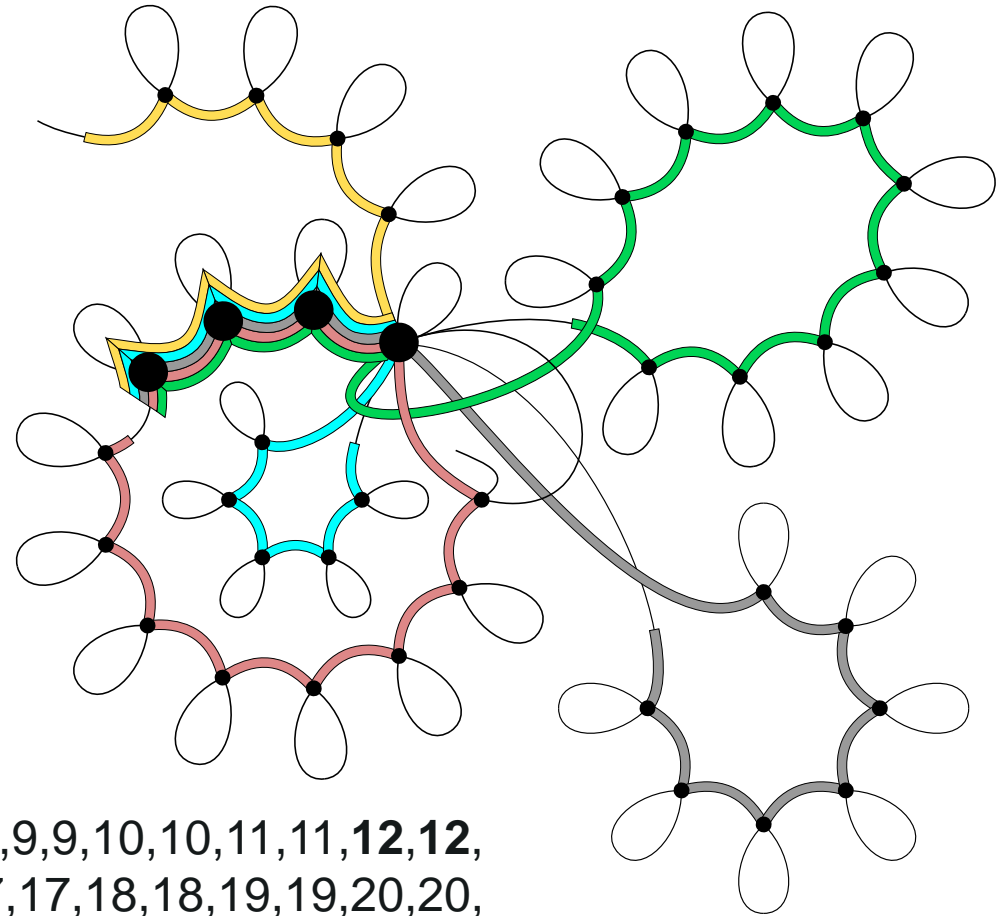
(Chen et. al., 2014) 100bp

Scaffold: Fully-overlapping MDSs



100bp

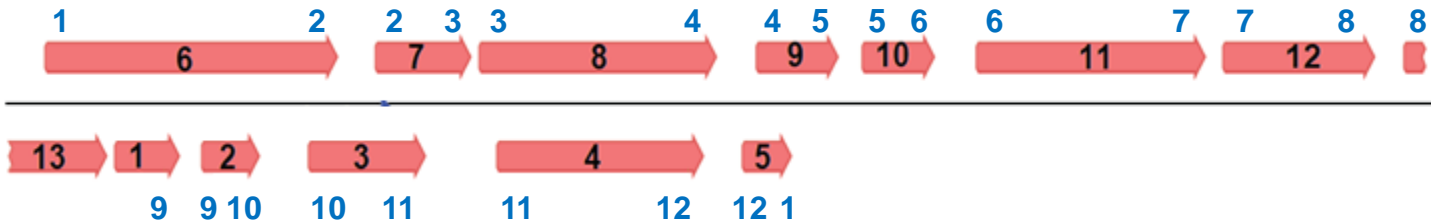
Scaffold: Fully-overlapping MDSs



$w = 1, 2, 2, 3, 3, 4, 4, 5, 5, 6, 6, 7, 7, 8, 8, 9, 9, 10, 10, 11, 11, \mathbf{12}, \mathbf{12},$
 $1, \mathbf{12}, 13, 13, 14, 14, 15, 15, 16, 16, 17, 17, 18, 18, 19, 19, 20, 20,$
 $21, 21, 22, 22, \mathbf{12}, 23, 23, 24, 24, 25, 25, 26, 26, 27, 27, \mathbf{12}, 28, 28, 29, 29,$
 $30, 30, 31, 31, 32, 32, 33, 33, 34, 34, \mathbf{12}, 35, 35, 36, 36, 37, 37, 38, 38$

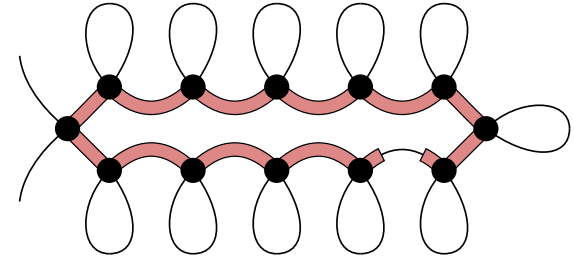
Individual Contigs within the Scaffold

$w = \underline{1,2,2,3,3,4,4,5,5,6,6,7,7,8,8,9,9,10,10,11,11,12,12,1,}$
 $12,13,13,14,14,15,15,16,16,17,17,18,18,19,19,20,20,$
 $21,21,22,22,12,23,23,24,24,25,25,26,26,27,27,12,28,$
 $28,29,29,30,30,31,31,32,32,33,33,34,34,12,35,35,36,$
 $36,37,37,38,38$

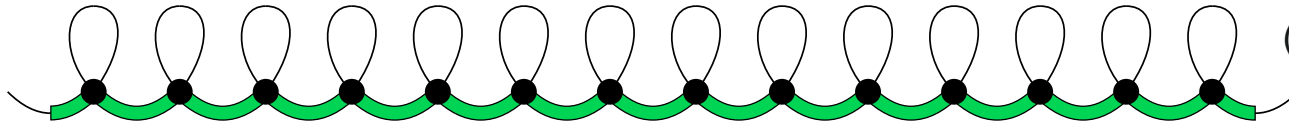


Individual Contigs within the Scaffold

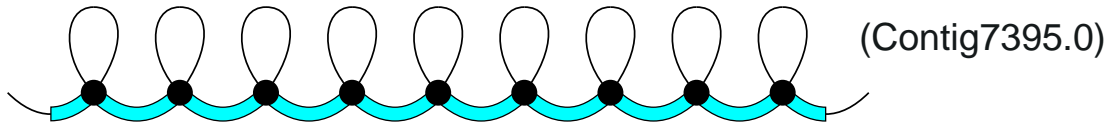
- $w_{\text{red}} = 1,2,2,3,3,4,4,5,5,6,6,7,7,8,8,9,9,10,10,11,11,12,12,1$
(Contig3325.0)



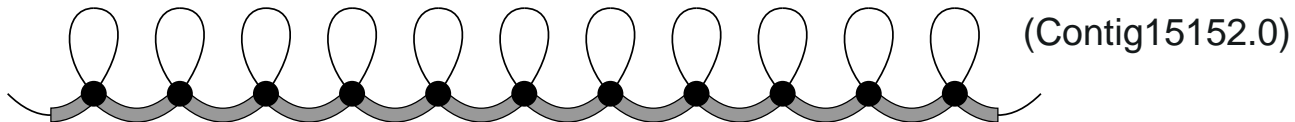
- $w_{\text{green}} = 9,9,10,10,11,11,12,12,13,13,14,14,15,15,16,16,17,17,18,18,19,19,20,20,21,21,22,22$
(Contig7507.0)



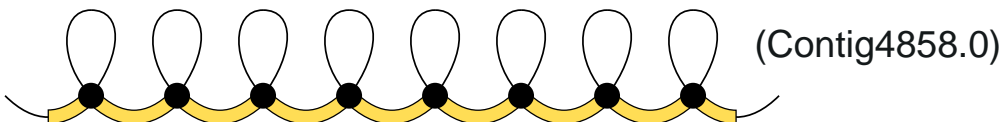
- $w_{\text{blue}} = 9,9,10,10,11,11,12,12,23,23,24,24,25,25,26,26,27,27$



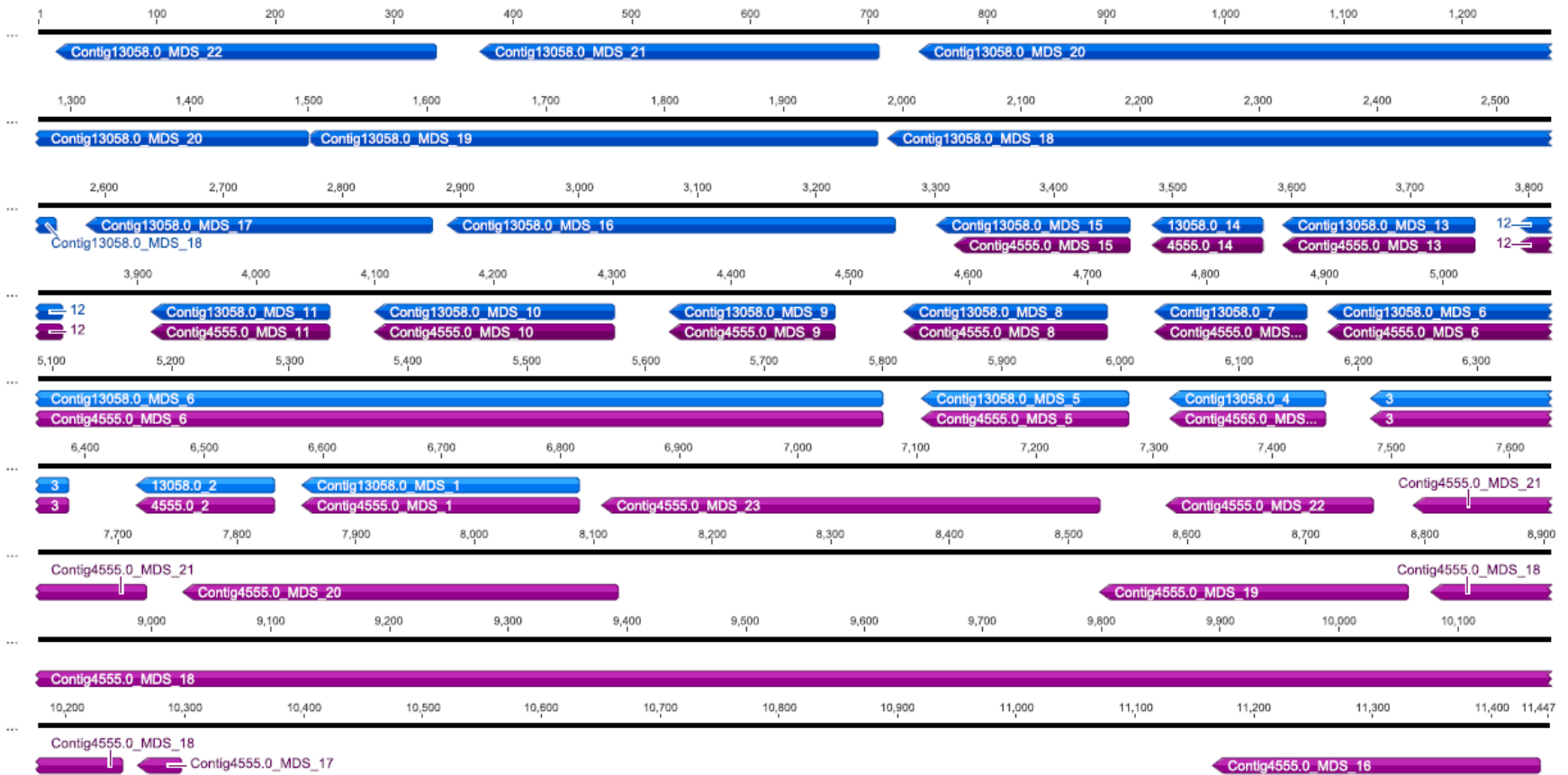
- $w_{\text{gray}} = 9,9,10,10,11,11,12,12,28,28,29,29,30,30,31,31,32,32,33,33,34,34$



- $w_{\text{gold}} = 9,9,10,10,11,11,12,12,35,35,36,36,37,37,38,38$



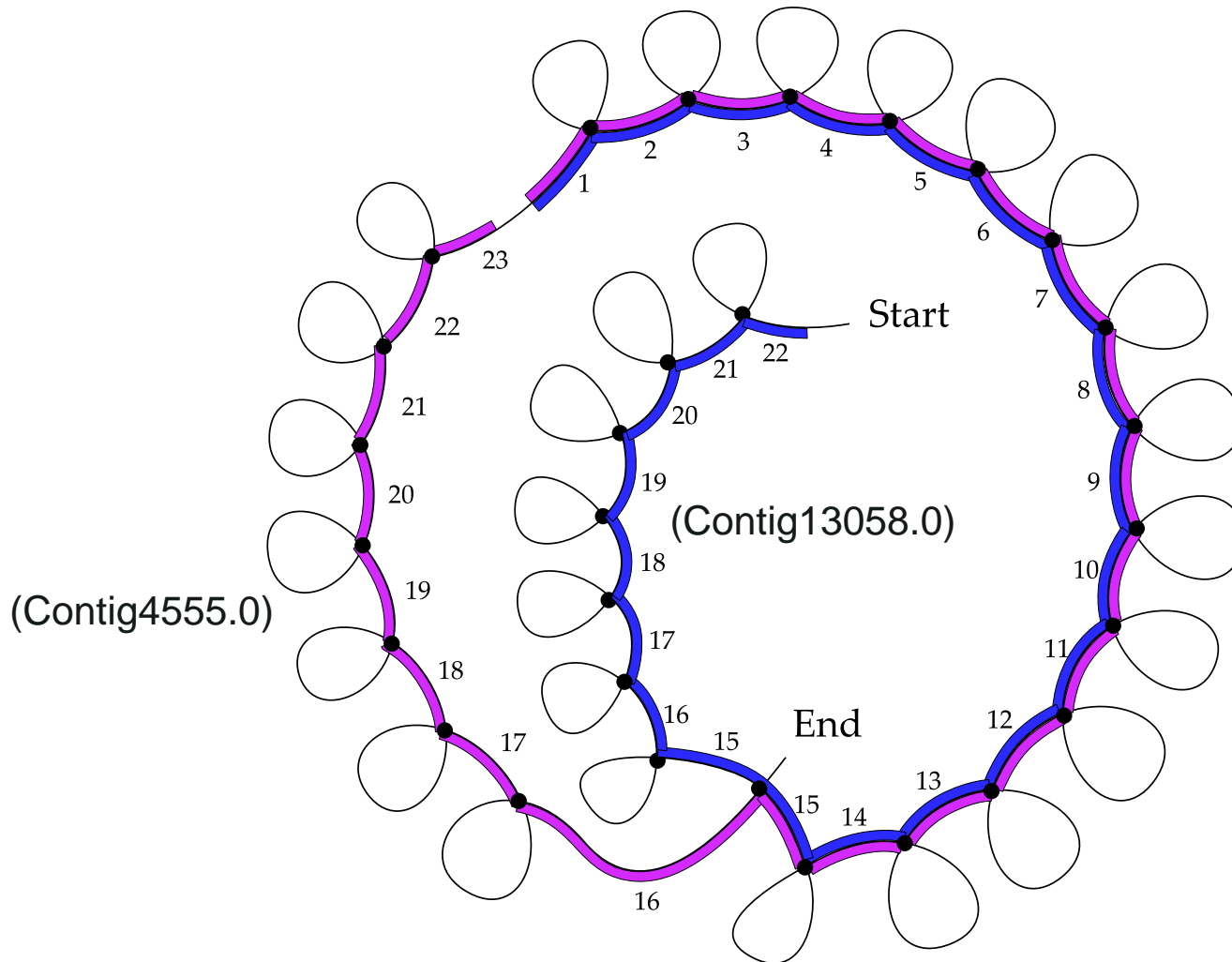
Scaffold: Partially-overlapping MDSs



(*Oxytricha trifallax*, MIC scaffold4620)

(Chen et. al., 2014)

Scaffold: Partially-overlapping MDSs



Chord Diagrams

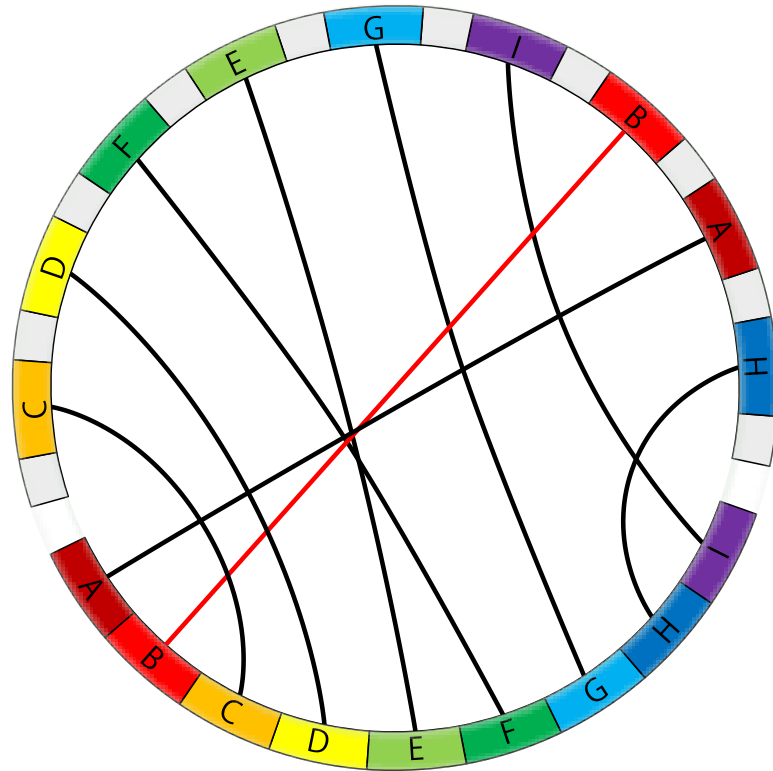
Precursor

Oxytricha nova, Actin I



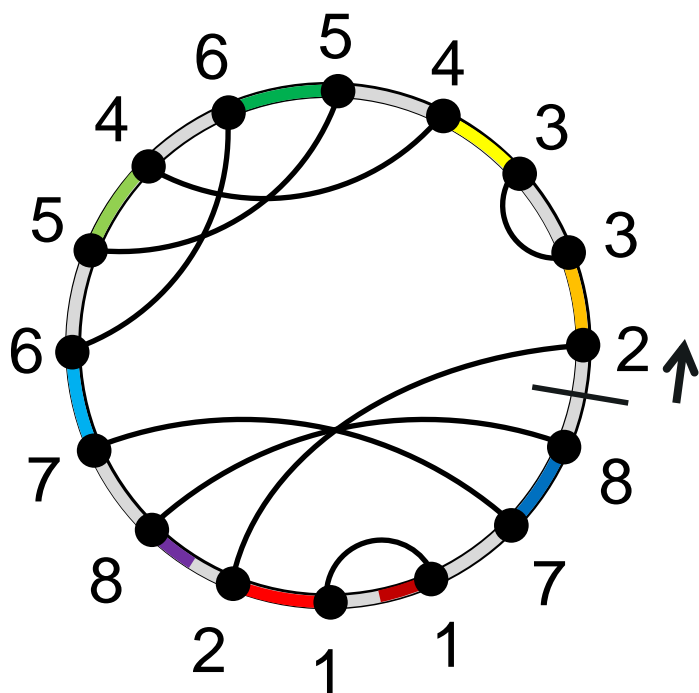
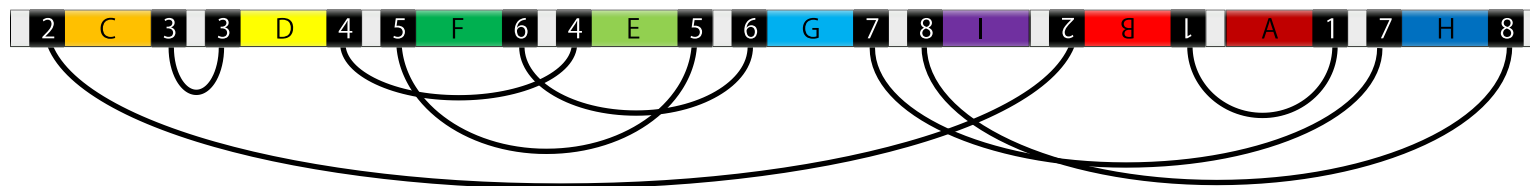
(Prescott 1992)

Product



— Indicates MDS is inverted

Different Chord Diagrams



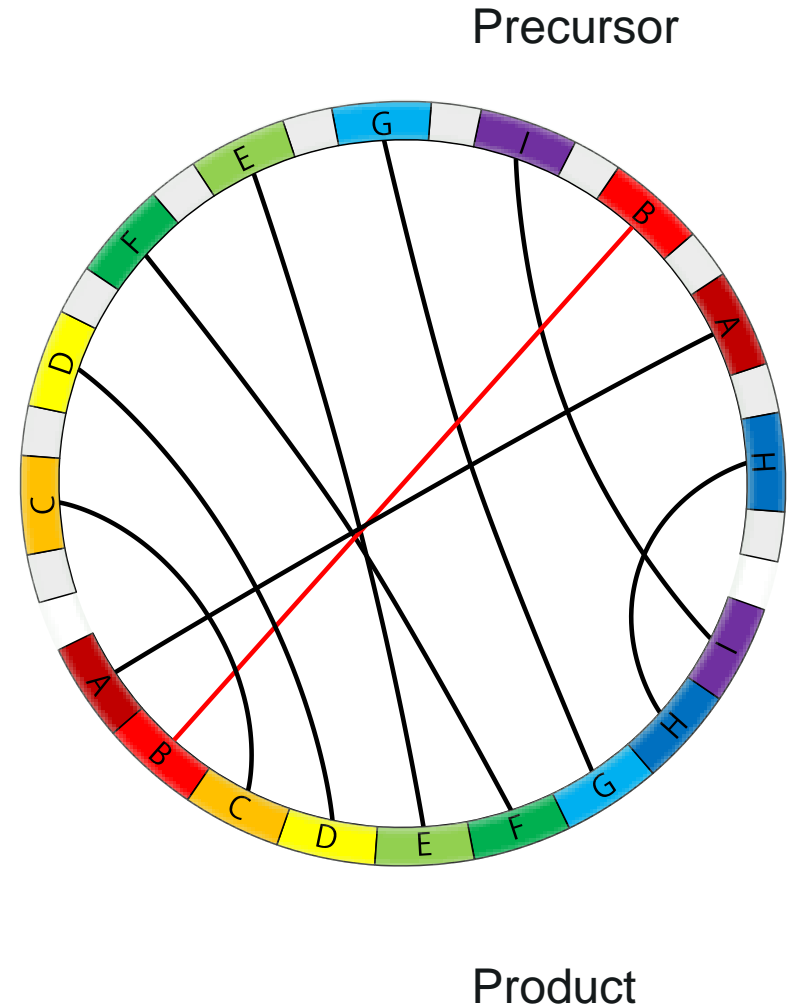
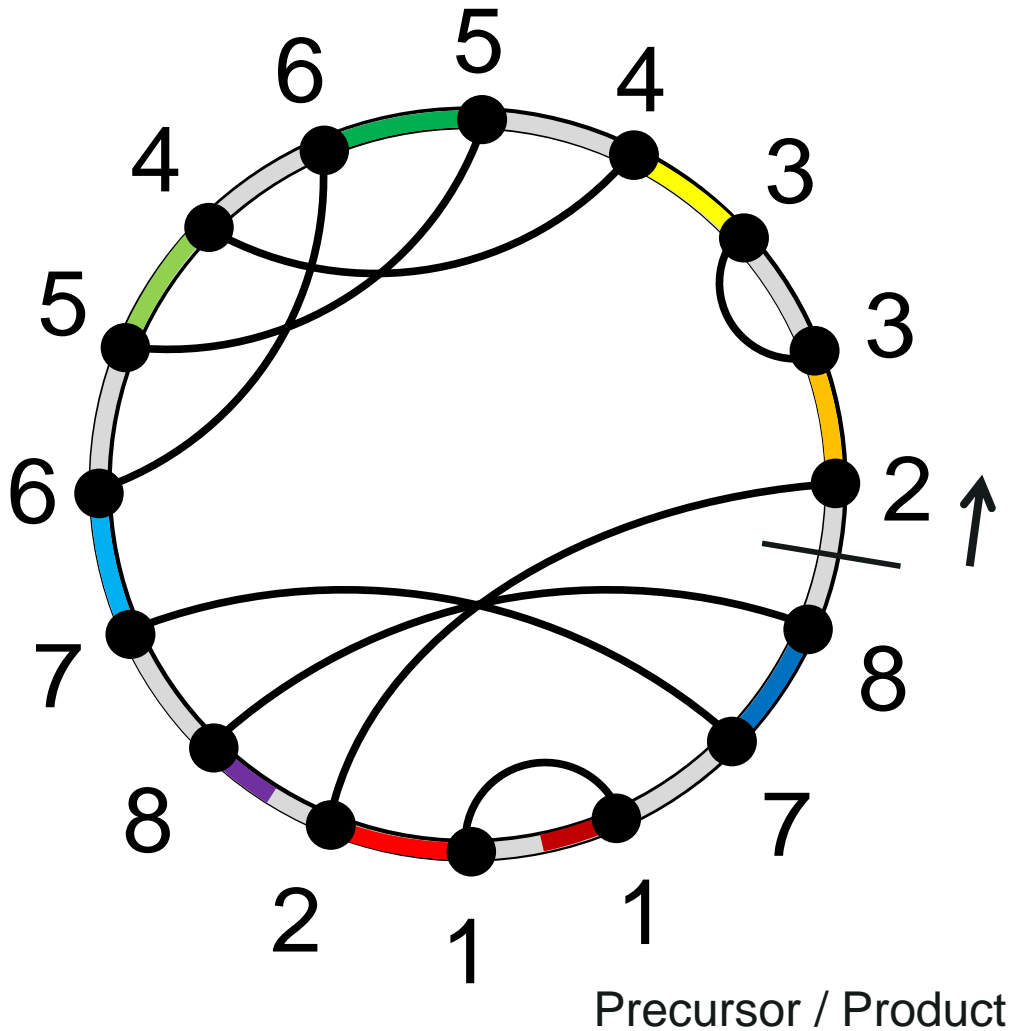
Precursor / Product

Precursor corresponds to the outside of the circle

Product corresponds to the path starting at MDS 1 and moving along chords and arcs sequentially until the last MDS



Comparing Chord Diagrams



Data Analysis

Data Processing

Scaffold Data

scaffold7244	oxytricha	Contig4067.0_MDS	2026	2172	.	+	.	Name=Contig4067.0_MDS_2
scaffold7244	oxytricha	Contig1019.1_MDS	2026	2173	.	+	.	Name=Contig1019.1_MDS_2
scaffold7244	oxytricha	Contig4067.0_MDS	2219	2532	.	+	.	Name=Contig4067.0_MDS_3
scaffold7244	oxytricha	Contig1019.1_MDS	2219	2532	.	+	.	Name=Contig1019.1_MDS_3
scaffold7244	oxytricha	Contig5190.0_MDS	2219	2532	.	+	.	Name=Contig5190.0_MDS_3
scaffold7244	oxytricha	Contig1019.1_MDS	2555	3462	.	+	.	Name=Contig1019.1_MDS_4
scaffold7244	oxytricha	Contig4067.0_MDS	2555	3463	.	+	.	Name=Contig4067.0_MDS_4
scaffold7244	oxytricha	Contig5190.0_MDS	2555	3463	.	+	.	Name=Contig5190.0_MDS_4
scaffold7244	oxytricha	Contig1019.1_MDS	3551	3928	.	+	.	Name=Contig1019.1_MDS_5
scaffold7244	oxytricha	Contig4067.0_MDS	3640	3928	.	+	.	Name=Contig4067.0_MDS_5
scaffold7244	oxytricha	Contig5190.0_MDS	3640	3928	.	+	.	Name=Contig5190.0_MDS_5
scaffold7244	oxytricha	Contig4067.0_MDS	3956	4308	.	+	.	Name=Contig4067.0_MDS_6
scaffold7244	oxytricha	Contig5190.0_MDS	3956	4308	.	+	.	Name=Contig5190.0_MDS_6
scaffold7244	oxytricha	Contig1019.1_MDS	3956	4309	.	+	.	Name=Contig1019.1_MDS_6
scaffold7244	oxytricha	Contig1019.1_MDS	4328	4864	.	+	.	Name=Contig1019.1_MDS_7
scaffold7244	oxytricha	Contig4067.0_MDS	4328	4836	.	+	.	Name=Contig4067.0_MDS_7
scaffold7244	oxytricha	Contig5190.0_MDS	4328	4864	.	+	.	Name=Contig5190.0_MDS_7
scaffold7244	oxytricha	Contig1019.1_MDS	4890	4953	.	+	.	Name=Contig1019.1_MDS_8
scaffold7244	oxytricha	Contig5190.0_MDS	4890	4953	.	+	.	Name=Contig5190.0_MDS_8
scaffold7244	oxytricha	Contig4067.0_MDS	4988	5017	.	+	.	Name=Contig4067.0_MDS_8

Filters

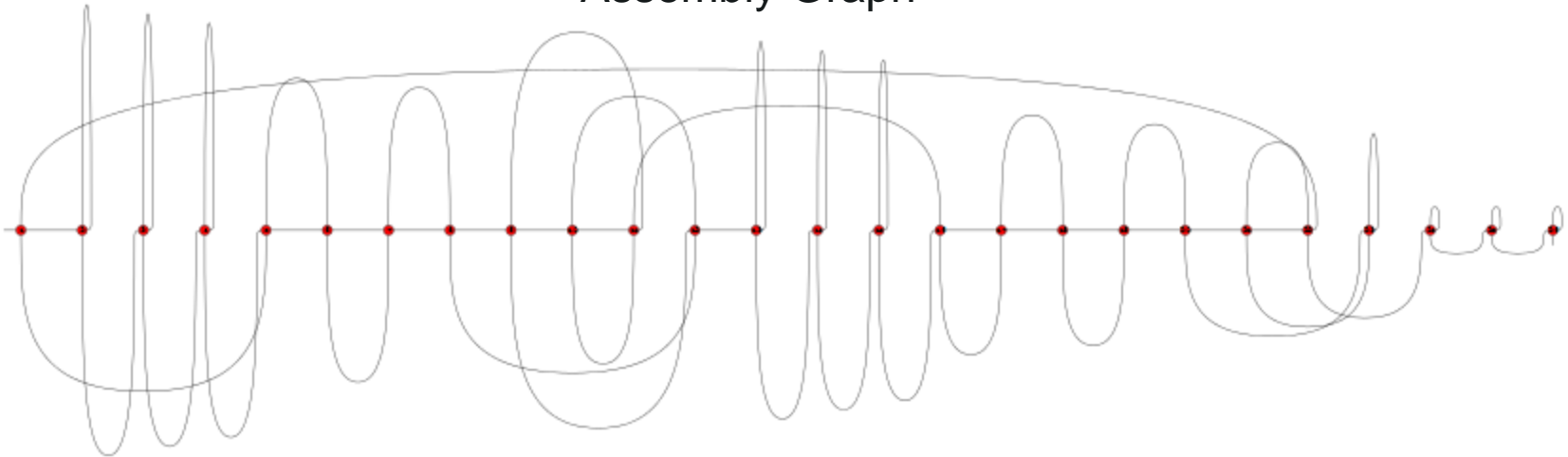


Assembly Words

```
scaffold1001, "1,1,2,2,3,3,4,4,5,5,6,6,7,8,8",
scaffold10186, "1,2,2,3,3,4,4,5,5,6,6,7,7,1"
scaffold10401, "1,1,2,2,3,3,4,4,5,5,6,6,7,7,ε
scaffold10573, "1,1,2,2,3,4,4,3,5,5,6,6,7,7,ε
scaffold10593, "1,1,2,2,3,3,4,4,5,5,6,6,7,7,ε
scaffold1107, "1,1,2,2,3,3,4,4,5,5,6,6,7,7,8,
scaffold1278, "1,1,2,2,3,3,4,4,5,5,6,6"
```



Assembly Graph

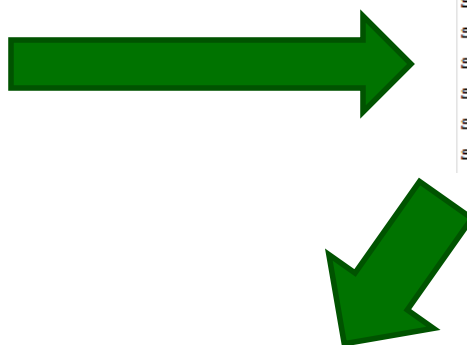


Data Processing

Scaffold Data

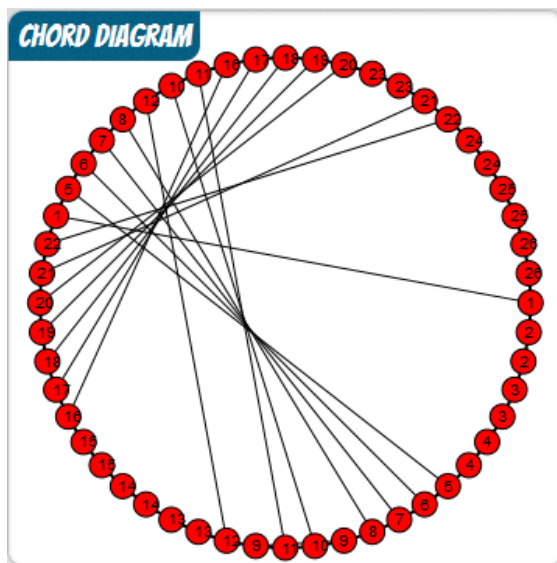
scaffold7244	oxytricha	Contig4067.0_MDS	2026	2172	.	+	.	Name=Contig4067.0_MDS_2
scaffold7244	oxytricha	Contig1019.1_MDS	2026	2173	.	+	.	Name=Contig1019.1_MDS_2
scaffold7244	oxytricha	Contig4067.0_MDS	2219	2532	.	+	.	Name=Contig4067.0_MDS_3
scaffold7244	oxytricha	Contig1019.1_MDS	2219	2532	.	+	.	Name=Contig1019.1_MDS_3
scaffold7244	oxytricha	Contig5190.0_MDS	2219	2532	.	+	.	Name=Contig5190.0_MDS_3
scaffold7244	oxytricha	Contig1019.1_MDS	2555	3462	.	+	.	Name=Contig1019.1_MDS_4
scaffold7244	oxytricha	Contig4067.0_MDS	2555	3463	.	+	.	Name=Contig4067.0_MDS_4
scaffold7244	oxytricha	Contig5190.0_MDS	2555	3463	.	+	.	Name=Contig5190.0_MDS_4
scaffold7244	oxytricha	Contig1019.1_MDS	3551	3928	.	+	.	Name=Contig1019.1_MDS_5
scaffold7244	oxytricha	Contig4067.0_MDS	3640	3928	.	+	.	Name=Contig4067.0_MDS_5
scaffold7244	oxytricha	Contig5190.0_MDS	3640	3928	.	+	.	Name=Contig5190.0_MDS_5
scaffold7244	oxytricha	Contig4067.0_MDS	3956	4308	.	+	.	Name=Contig4067.0_MDS_6
scaffold7244	oxytricha	Contig5190.0_MDS	3956	4308	.	+	.	Name=Contig5190.0_MDS_6
scaffold7244	oxytricha	Contig1019.1_MDS	3956	4309	.	+	.	Name=Contig1019.1_MDS_6
scaffold7244	oxytricha	Contig1019.1_MDS	4328	4864	.	+	.	Name=Contig1019.1_MDS_7
scaffold7244	oxytricha	Contig4067.0_MDS	4328	4836	.	+	.	Name=Contig4067.0_MDS_7
scaffold7244	oxytricha	Contig5190.0_MDS	4328	4864	.	+	.	Name=Contig5190.0_MDS_7
scaffold7244	oxytricha	Contig1019.1_MDS	4890	4953	.	+	.	Name=Contig1019.1_MDS_8
scaffold7244	oxytricha	Contig5190.0_MDS	4890	4953	.	+	.	Name=Contig5190.0_MDS_8
scaffold7244	oxytricha	Contig4067.0_MDS	4988	5017	.	+	.	Name=Contig4067.0_MDS_8

Filters



Assembly Words

```
scaffold1001, "1, 1, 2, 2, 3, 3, 4, 4, 5, 5, 6, 6, 7, 8, 8,"  
scaffold10186, "1, 2, 2, 3, 3, 4, 4, 5, 5, 6, 6, 7, 7, 1,"  
scaffold10401, "1, 1, 2, 2, 3, 3, 4, 4, 5, 5, 6, 6, 7, 7, 8,"  
scaffold10573, "1, 1, 2, 2, 3, 4, 4, 3, 5, 5, 6, 6, 7, 7, 8,"  
scaffold10593, "1, 1, 2, 2, 3, 3, 4, 4, 5, 5, 6, 6, 7, 7, 8,"  
scaffold1107, "1, 1, 2, 2, 3, 3, 4, 4, 5, 5, 6, 6, 7, 7, 8,"  
scaffold1278, "1, 1, 2, 2, 3, 3, 4, 4, 5, 5, 6, 6"
```



Data Filters

(Chen et. al., 2014)

All Contigs (17283)

Contigs on single scaffold (10778)

Contigs on
multiple
scaffolds (6505)

Overlapping
contigs (3998)

Non-overlapping contigs (6780)

Interleaving
(1152)

Non-interleaving (5628)

Singe contig on
scaffold (2480)

Multiple contigs
on scaffold
(3148)

Properties of single contig assembly graphs

- Nesting Index
- Genus range
- Genus spectrum

Properties for multi contig assembly graphs

- Nesting Index
- Crossing number
- Switching number
- Assembly Number
- Genus range
- Genus spectrum

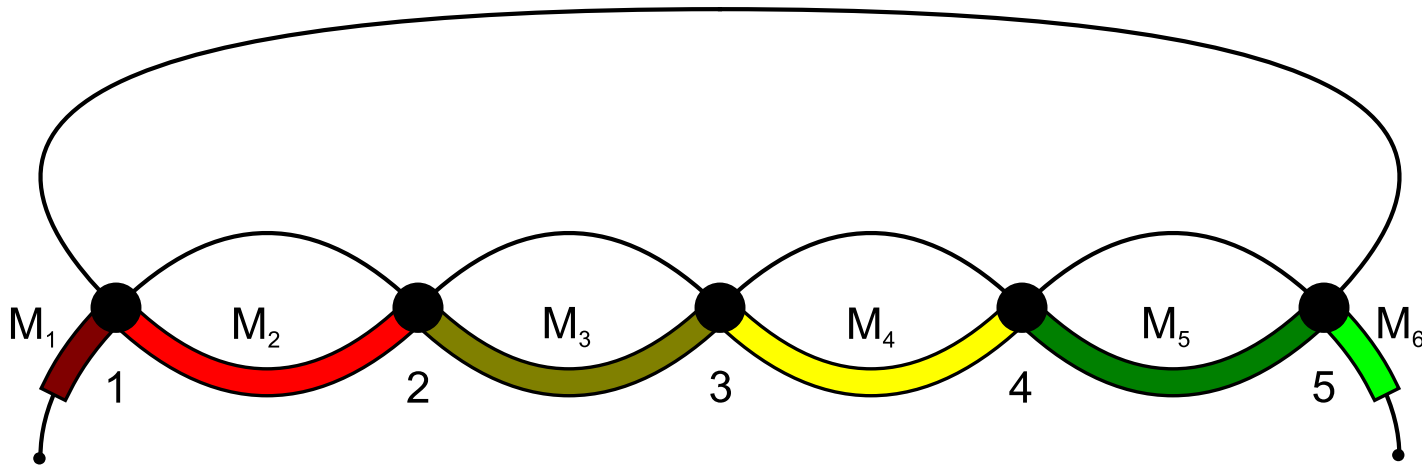
Properties of chord diagrams

- Braid Index
- Genus range
- Genus spectrum

Nesting Indices of *Oxytricha trifallax*

Nesting Index

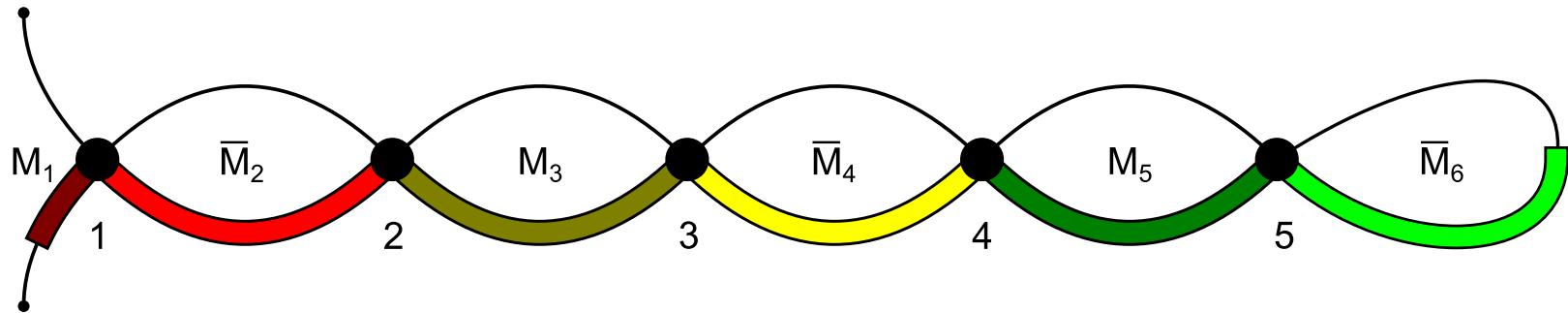
- Repeat Word: $a_1 a_2 \dots a_n a_1 a_2 \dots a_n$
- Ex: 1234512345



- $M_1 M_3 M_5 M_2 M_4 M_6$

Nesting Index

- Return Word: $a_1 a_2 \dots a_n a_n \dots a_2 a_1$
- Ex: 1234554321



- $M_1 M_3 M_5 \bar{M}_6 \bar{M}_4 \bar{M}_2$

Nesting Index

- Define two reduction operations:
 1. Removal of repeat and return words
123321456456 → empty
 2. Removal of a single letter
132321 → 1221
- Nesting Index is the minimum number of the above reduction operations that produces an empty word

Nesting Index

- 123324156456
- 123324156456
- 123324156456
- Nesting index is 2

Single Contig on Scaffold

Nesting Index	# of Genes
Single MDS	420
1	1906
2	130
3	3
4	7
5	0
6	4
7	3

Nesting Index	# of Genes
8	3
9	1
10	0
11	2
12	0
13	0
14	0
15	1

Scaffolds with a single Contig having high Nesting Index

Nesting Index	Scaffold	Contig
15	scaffold3034	Contig16661.0
11	scaffold201	Contig15305.0
11	scaffold736	Contig9968.0
9	scaffold310	Contig12119.0
8	scaffold1766	Contig18598.0
8	scaffold32745	Contig9447.0
8	scaffold33105	Contig1089.0.5
7	scaffold33218	Contig9029.0.0
7	scaffold33988	Contig13839.0
7	scaffold36972	Contig20991.0.0

Multiple Contigs on Scaffold (Filtered)

Nesting Index	# of Genes
1	20
2	93
3	3
4	2
5	2

Nesting Index	# of Genes
6	0
7	0
8	1
≥ 8	2

Scaffold with High Nesting Index

Scaffold	Contigs	Nesting Index
scaffold10573 (Nesting Index 8)	Contig11249.0	1
	Contig12411.0	1
	Contig14890.0	1
	Contig17133.0	1
	Contig17135.0	1
	Contig17623.0	1
	Contig18782.0	1
	Contig19908.0	1
	Contig20362.0	1
	Contig20860.0	1
	Contig29.0	8
	Contig374.0	1
	Contig6086.0	1
Contig6468.0	1	

Scaffold with High Nesting Index

Scaffold	Contigs	Nesting Index
scaffold35844 (Nesting Index ≥ 8)	Contig11030.0	0
	Contig12771.0	1
	Contig17508.0	1
	Contig17528.0	1
	Contig17650.0	≥ 8
	Contig18803.0	1
	Contig9101.0.0	2

Scaffold with High Nesting Index

Scaffold	Contigs	Nesting Index
Scaffold35966 (Nesting Index ≥ 8)	Contig11026.0	2
	Contig12848.0	0
	Contig14040.0	1
	Contig1423.1	1
	Contig14353.0	1
	Contig14941.0	1
	Contig15469.0	1
	Contig18103.0	1
	Contig18706.0	1
	Contig20747.0	1
Contig9655.0	12	

Data Cleaning

Data Cleaning

MDS merging

- Merge two MDSs into one if:
 - Two consecutive MDSs of the same contig are less than 10 bp apart
 - There is no IES in between them
- What if two MDSs are more than 10bp apart and there is no IES in between?

Data Cleaning

Redundancy elimination

- Remove redundant contig if:
 - Contig is a subset of another contig
 - Contigs do not share any MDS
- What if two contigs intersect each other at some (but not all) MDSs and are not listed in the file?

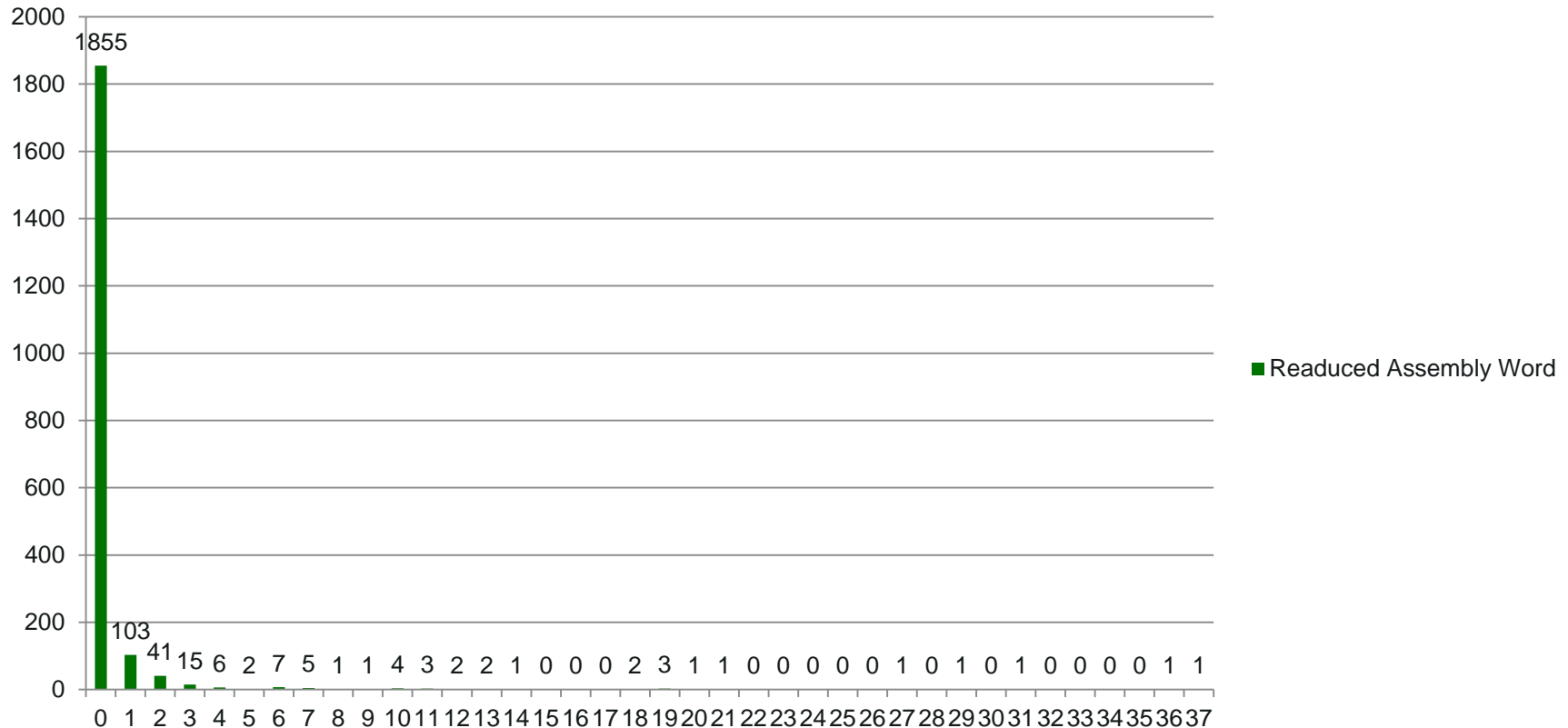
Removing conventional IESs

Reduced Assembly Word

- Given a word(Ex. 12334421) delete consecutive pairs of letters (Ex. 1221)
- Corresponds to the removal of conventional IESs
- We extracted Reduced Assembly Words out of scaffolds that contain only one contig

Statistics on single contig scaffolds

Reduced Assembly Words



X-axis – # of letters left after consecutive pairs of letters were removed

Y-axis – # of words

Total number of scaffolds: 2060

References

- A. Angeleska, N. Jonoska, M. Saito, DNA Rearrangement through assembly graphs. *Discrete and Applied Math*, **157** (2009) pp. 3020-3037.
- Arredondo, Ryan, Reductions On Double Occurrence Words. Proceedings of the Forty-fourth Southeastern International Conference on Combinatorics, Graph Theory and Computing. *Congr. Numer.* 218 (2013), pp. 43-56.
- Chen, Xiao, et al., The architecture of a scrambled genome reveals a massive level of DNA rearrangement, 2014 (preprint)
- Prescott, David M., Greslin, Arthur F., Scrambled actin I gene in the micronucleus of *Oxytricha nova*, *Developmental Genetics*, 13:1 (1992) pp. 66-74

Thank You