PLANT & ANIMAL GENOME

Clfinder-Orthnet: creating comparative genomics frameworks for closely-related genomes using co-linearity networks

Dong-Ha Oh[*] and Maheshi Dassanayake

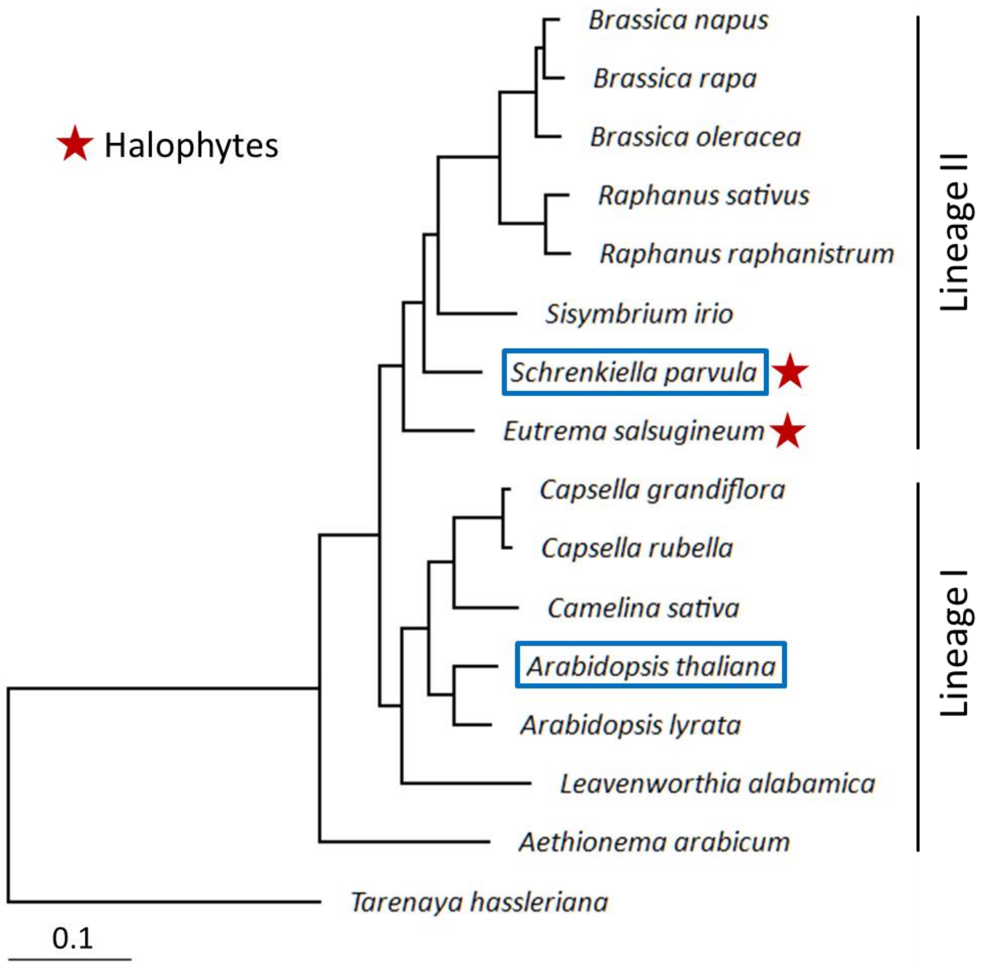
Department of Biological Sciences

Louisiana State University, Baton Rouge LA

*Presenter

Brassicaceae: one of the most sequenced and annotated plant family

Publicly available Brassicaceae genomes, 2017 Spring



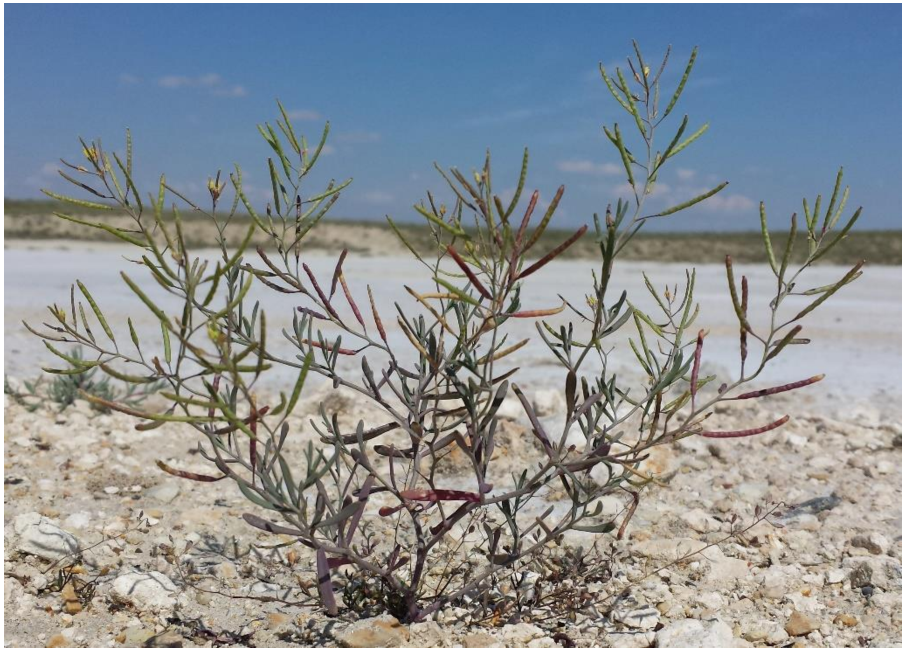
A comparative genomics framework including the two extremophyte/halophyte crucifers, *S. parvula* and *E. salsugineum*. The tree was based on 14,614 alignment of homologous gene clusters.

Schrenkiella parvula from Lake Tuz, Turkey

(NASA, 1994)

(Sun Jeon, c. 2010)

S. parvula in situ (Ozfidan-Konakci et al., Funct Plant Biol 2016)

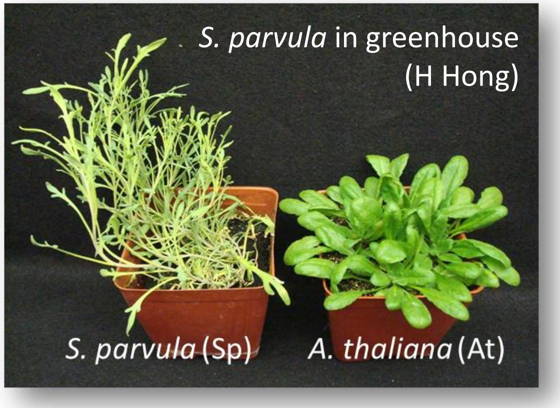


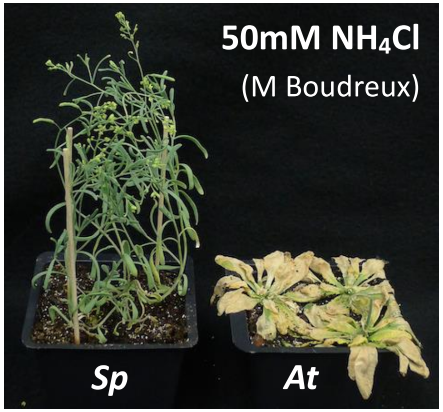
(Helvaci *et al.*, Int Geol Rev 2004)

	pH	Li^+	Na^+	K^+	Mg^{2+} (mg/L)
Burdur Lake	8.8	3	5,500	45	960
Tuz Lake	**6.0**	**325**	**61,000**	**12,000**	**37,500**
Seawater	7.2-8.4	0	10,800	392	1,290

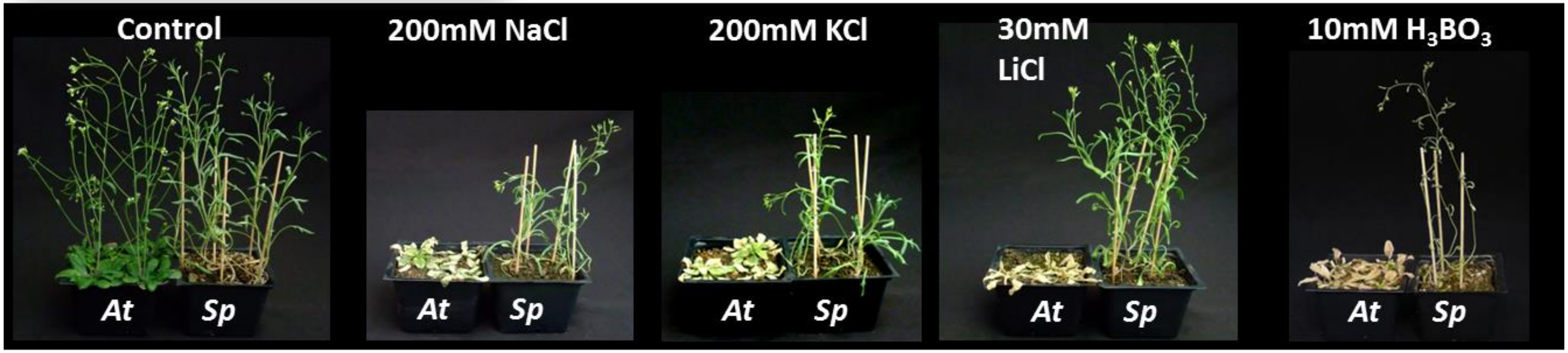
Schrenkiella parvula from Lake Tuz, Turkey

Adaptation to multi-ion salt stresses



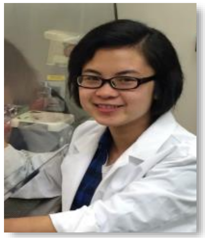


(Oh *et al.*, Plant Physiol 2014)



W489

P495

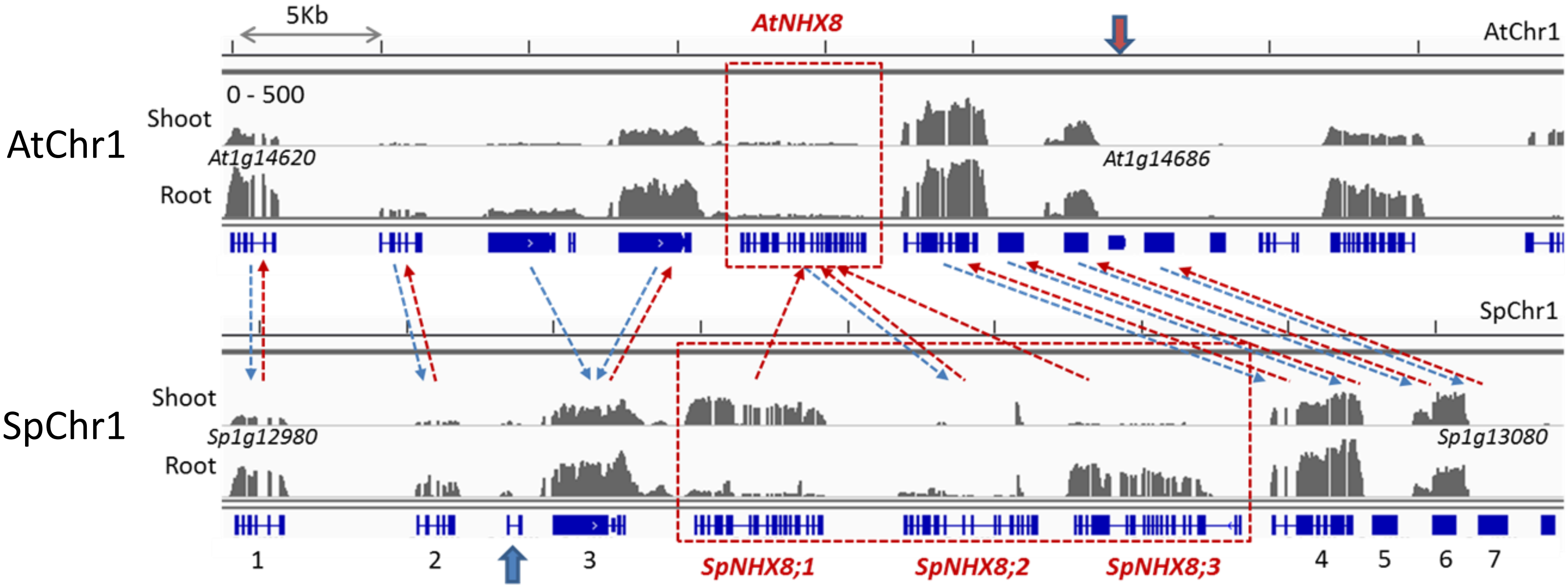


W562

W457

Comparison with a model species (i.e. *S. parvula* vs *A. thaliana*)

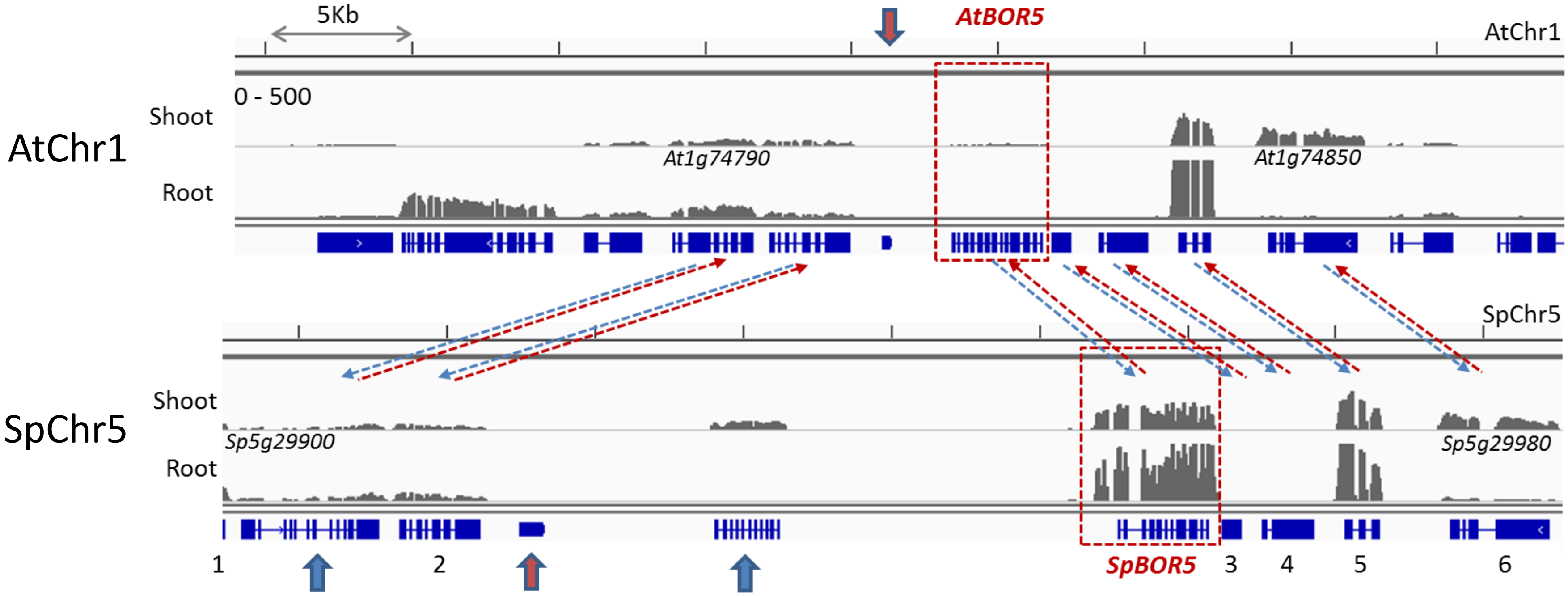
S. parvula-specific tandem duplication of *Na⁺/H⁺ EXCHANGER 8* (*NHX8*)?



(Oh *et al*., Plant Physiol 2014)

Comparison with a model species (i.e. *S. parvula* vs *A. thaliana*)

S. parvula-specific transposition near *BORON TRANSPORTER 5* (*BOR5*)?



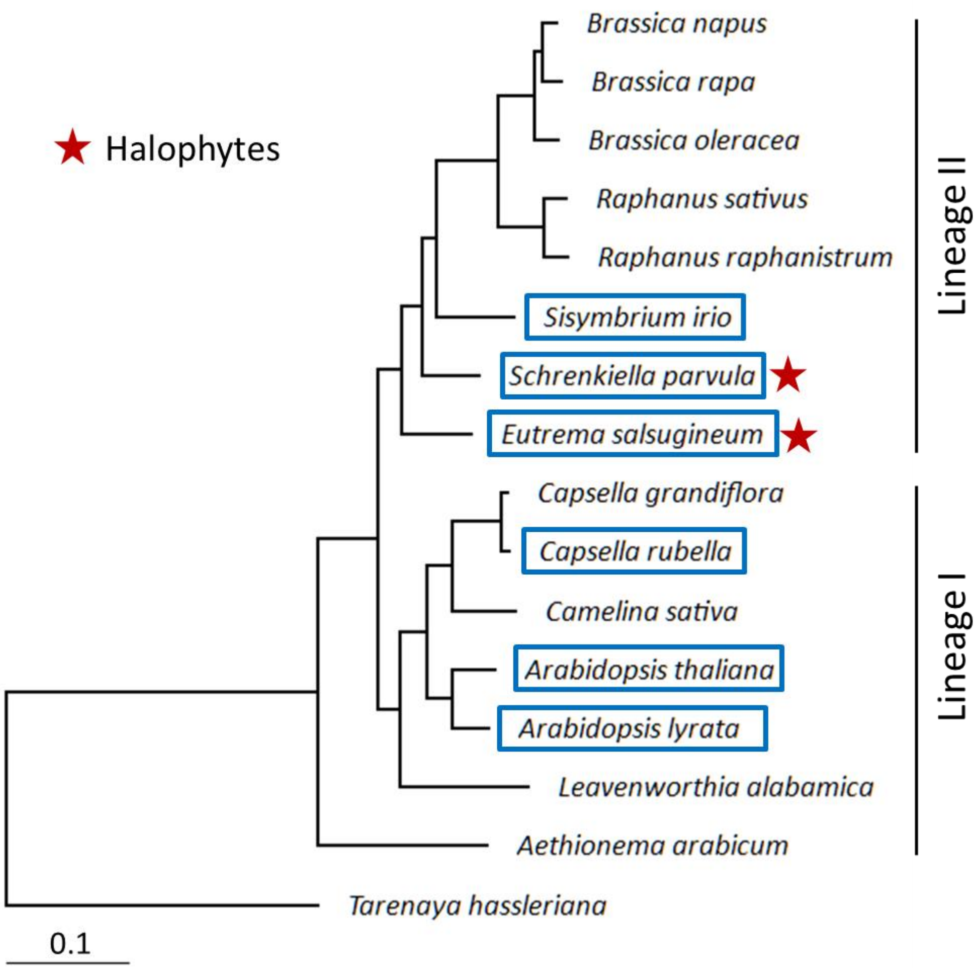
(Oh *et al.*, Plant Physiol 2014)

Criticisms for two-species comparison with a model plant:

Are these events really unique to *S. parvula*?

Expanding genome samples for comparative analyses

Publicly available Brassicaceae genomes, 2017 Spring



(Oh and Dassanayake, *DNA Res, 2018*)

Lineage I

🟦 Aly = *A. lyrata*

🟦 Ath = *A. thaliana*

🟩 Cru = *C. rubella*

Lineage II

🟨 Esa = *E. salsugineum*

🟧 Sir = *S. irio*

🟧 Spa = *S. parvula*

CLfinder-OrthNet

a pipeline to systematically compare multiple closely-related genomes

1. CLfinder (Co-Linearity finder)

   - summarizes annotation and detects tandem duplication (*td*) for each genome
   - detects co-linearity in gene orders in all pairs of genomes

Pair-wise comparisons by CLfinder – but this was not enough...

Summary of CLfinder results showing pairwise comparisons among 6 crucifer species

Query species	# protein-coding genes	CL type[a]	Target species						# TD[b] events (# TD genes)
			Aly	Ath	Cru	Esa	Sir	Spa	
Aly	32,657	*cl*		24,296	23,055	21,416	19,988	21,032	
		tr		4,881	5,375	6,668	8,104	6,478	2,163
		ls		2,876	3,611	3,954	3,902	4,530	(5,733)
		nd		604	616	619	663	617	
Ath	27,206	*cl*	23,436		22,683	21,187	19,821	20,851	
		tr	2,431		2,804	4,032	5,355	4,064	1,747
		ls	1,339		1,719	1,987	2,030	2,291	(4,770)
		nd	0		0	0	0	0	
Cru	26,521	*cl*	22,371	22,836		20,906	19,350	20,436	
		tr	3,036	2,836		4,338	5,817	4,267	1,752
		ls	950	666		1,112	1,154	1,646	(4,996)
		nd	164	183		165	200	172	
Esa	26,351	*cl*	20,384	20,884	20,460		19,699	20,612	
		tr	4,465	4,137	4,460		5,431	4,046	1,646
		ls	1,452	1,274	1,377		1,146	1,631	(4,461)
		nd	50	56	54		75	62	
Sir	32,524	*cl*	19,015	19,538	19,068	19,728		19,766	
		tr	3,062	2,860	2,998	2,722		2,697	1,795
		ls	5,520	5,148	5,496	5,054		4,225	(4,586)
		nd	4,927	4,978	4,962	5,020		5,836	
Spa	26,847	*cl*	19,849	20,358	19,934	20,380	19,546		
		tr	2,830	2,452	2,718	2,534	4,097		1,242
		ls	3,649	3,541	3,688	3,432	2,526		(3,049)
		nd	519	496	507	501	678		

[a] Co-linear (*cl*), transposed (*tr*), lineage-specific (*ls*), or not determined due to too small genome scaffold (*nd*)

[a] Using CLfinder parameters {window_size, num_CL_trshld, gap_CL_trshld} = { 20, 3, 20 }

[b] Using TDfinder parameter max_TD_loci_dist = 4

(Oh and Dassanayake, *DNA Res, 2018*)

CLfinder-OrthNet

a pipeline to systematically compare multiple closely-related genomes

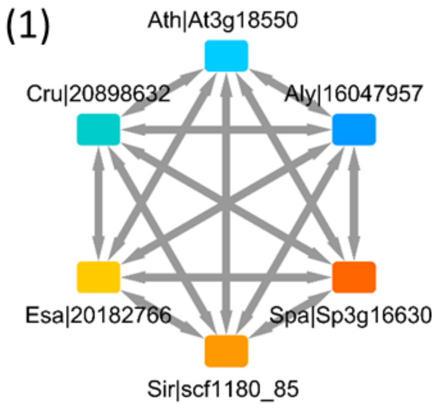
1. CLfinder (Co-Linearity finder)

2. ONfinder (OrthNet finder)

   - combines CLfinder results into networks of orthologs (OrthNet),
   - connects “best ortholog” and tandem duplicated (*td*) paralog nodes with the presence (*cl*) or absence (*tr*) of co-linearity as edge property,
   - *iterative Markov clustering (MCL) to “prune” and finalize OrthNets,
   - *searches and clusters OrthNets based on topology

* More on these: W1069 (Systems Genomics)

Encoding the evolutionary history of a locus as a network (OrthNet)

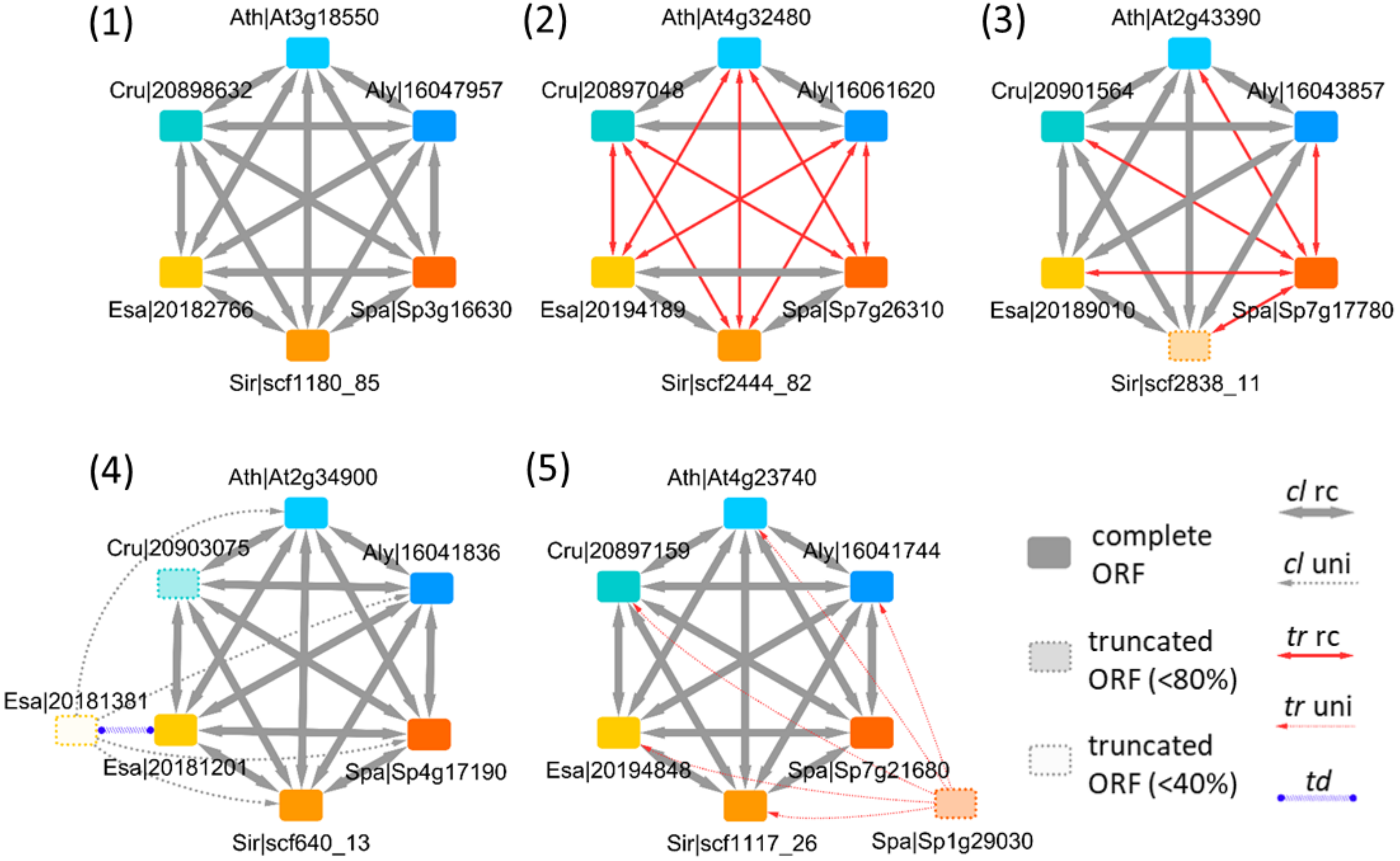


- ■ complete ORF
- ▨ truncated ORF (<80%)
- □ truncated ORF (<40%)

- *cl* rc ↔
- *cl* uni ←
- *tr* rc ↔
- *tr* uni ←
- *td* ●—●

(Oh and Dassanayake,
DNA Res, 2018)

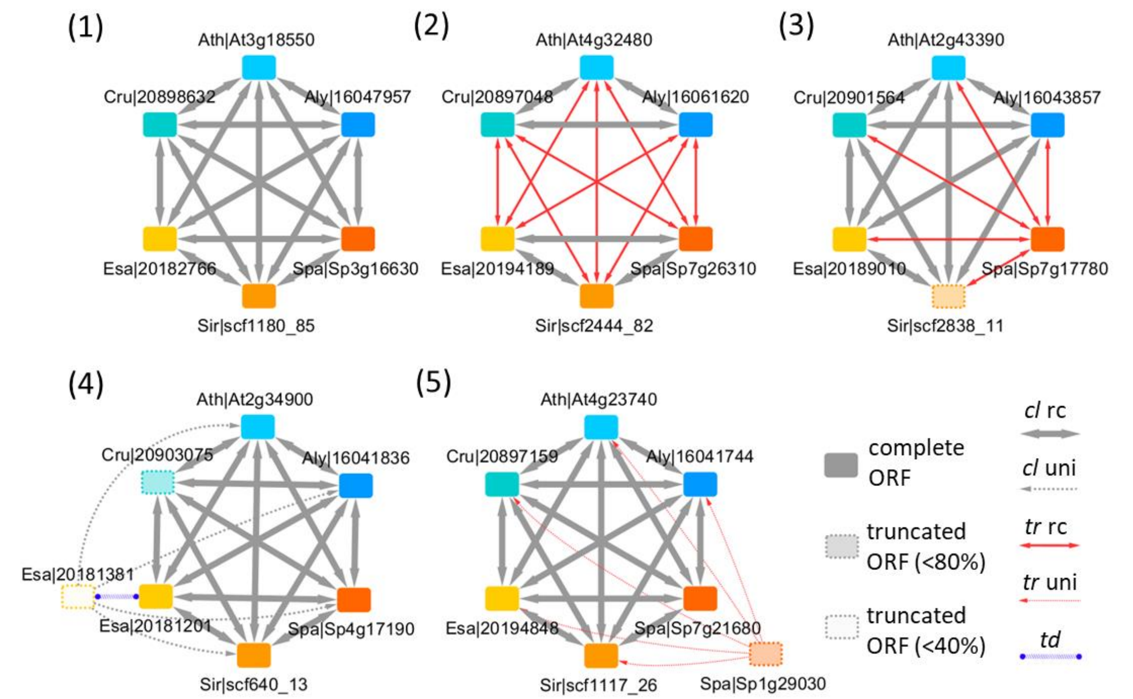
Encoding the evolutionary history of a locus as a network (OrthNet)



(Oh and Dassanayake, *DNA Res, 2018*)

Encoding the evolutionary history of a locus as a network (OrthNet)

enables detection of all orthologs showing a shared "evolutionary context/history"

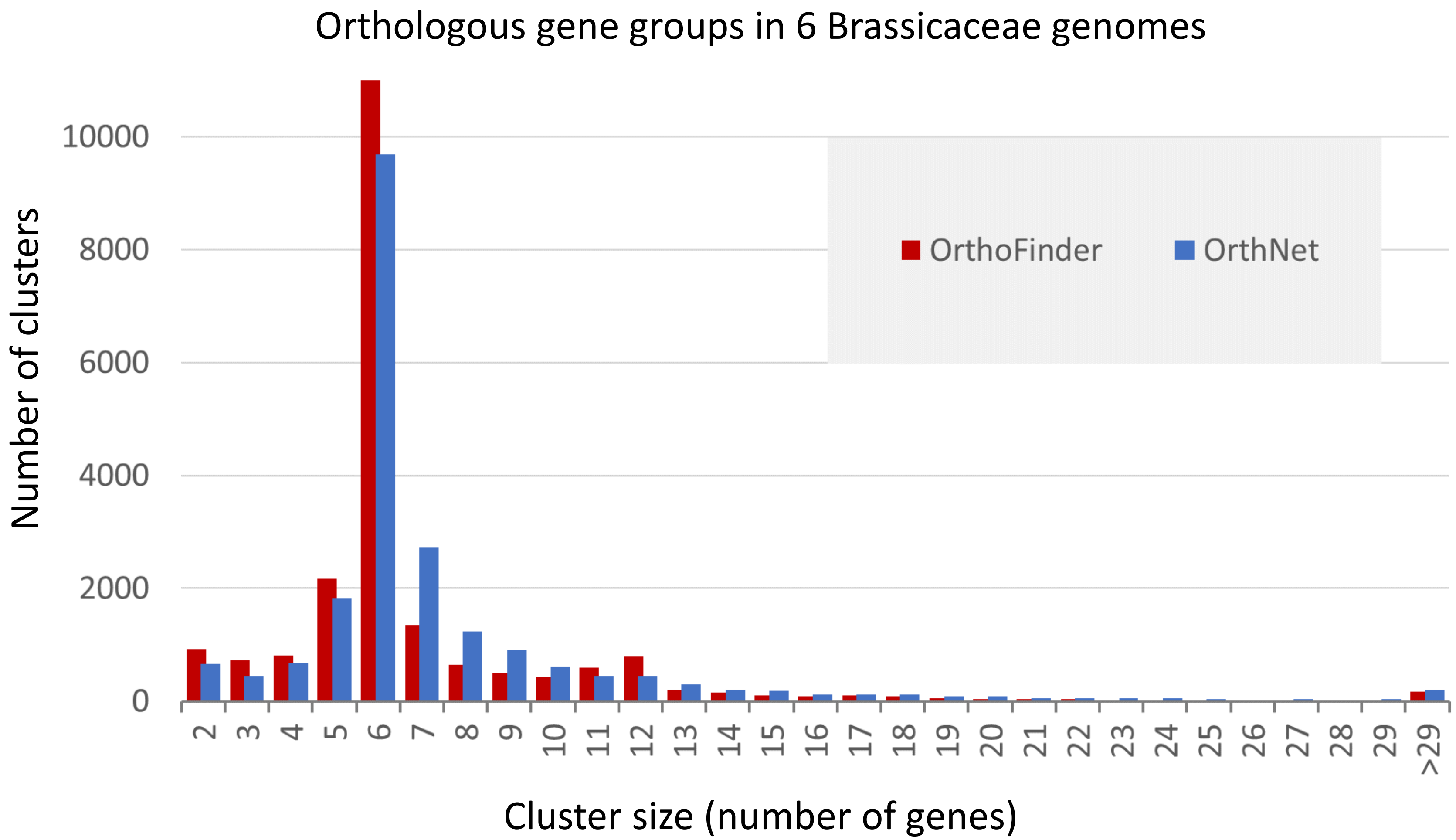


Evolutionary history	Number of OrthNets	Example OrthNet ID (Annotation)
(1) All co-linear, single copy	7,034	On_6361 (*BRANCHED 1*)
(2) Transposition (*tr*) between Lineage I and II	50	On_8867 (unknown function)
(3) Spa-specific transposition	12	On_12904 (unknown function)
(4) Esa-specific tandem duplication (*td*)	44	On_4974 (*IMBIBITION-INDUCIBLE 1*)
(5) Spa-specific transposition-duplication (*tr-d*)	70	On_4629 (LRR kinase family)

(Oh and Dassanayake, *DNA Res, 2018*)

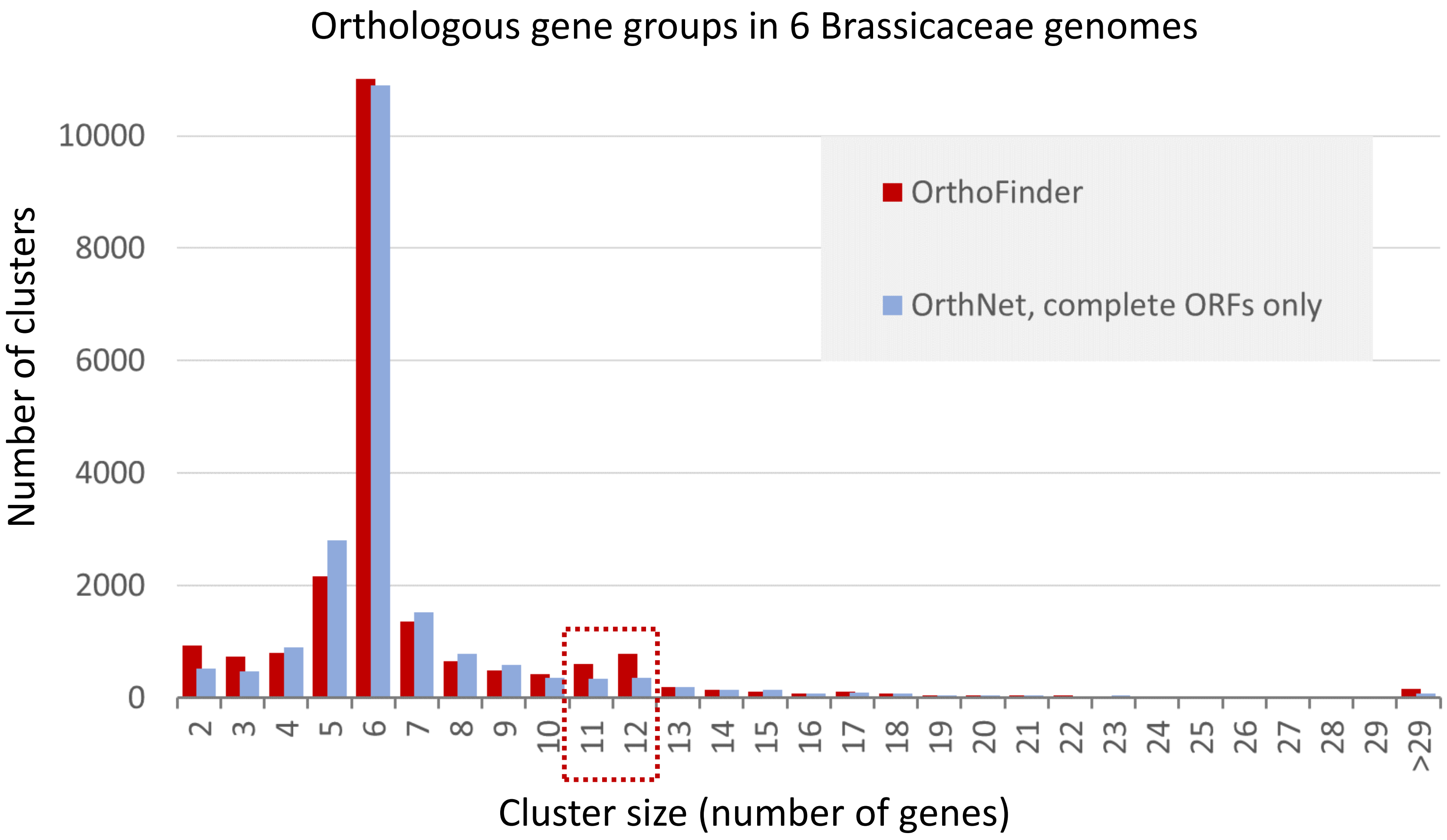
Orthology inference using OrthNets (sequence similarity + co-linearity)

Comparison of OrthNet and OrthoFinder



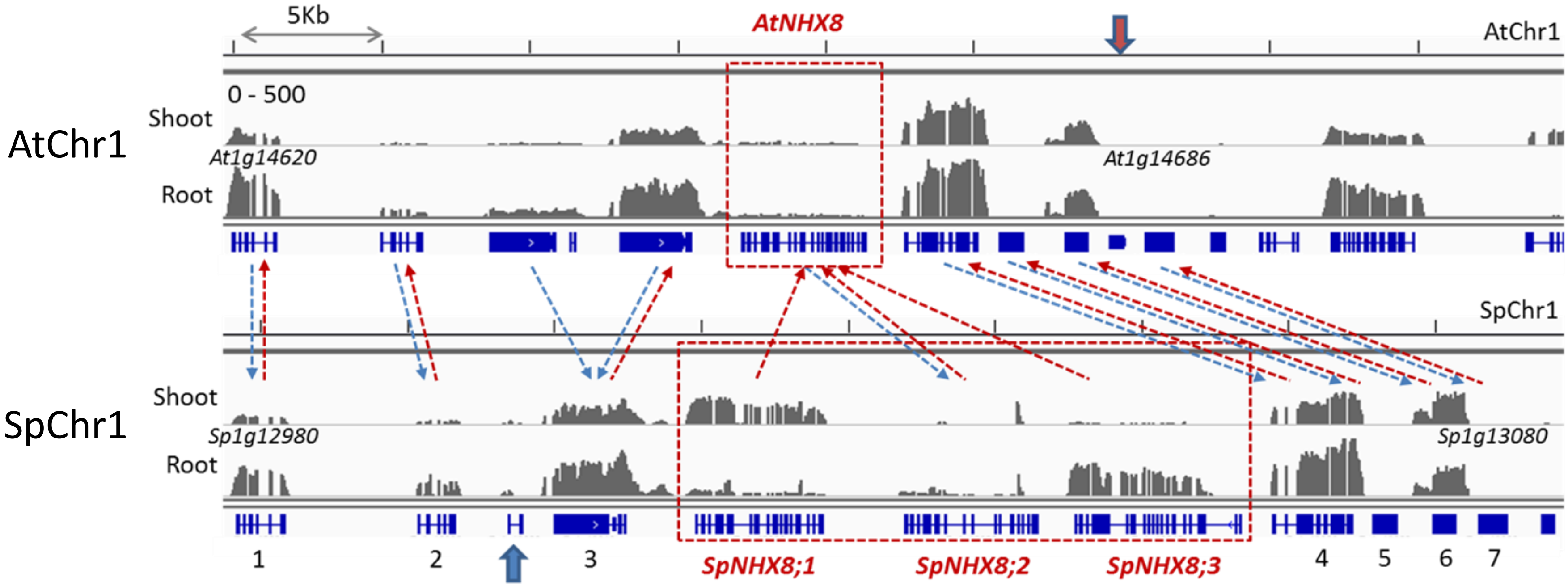
Orthology inference using OrthNets (sequence similarity + co-linearity)

Comparison of OrthNet and OrthoFinder



OrthNet reveals the evolutionary history of orthologous gene groups

How many of "*S. parvula*-specific" events really unique to *S. parvula*?

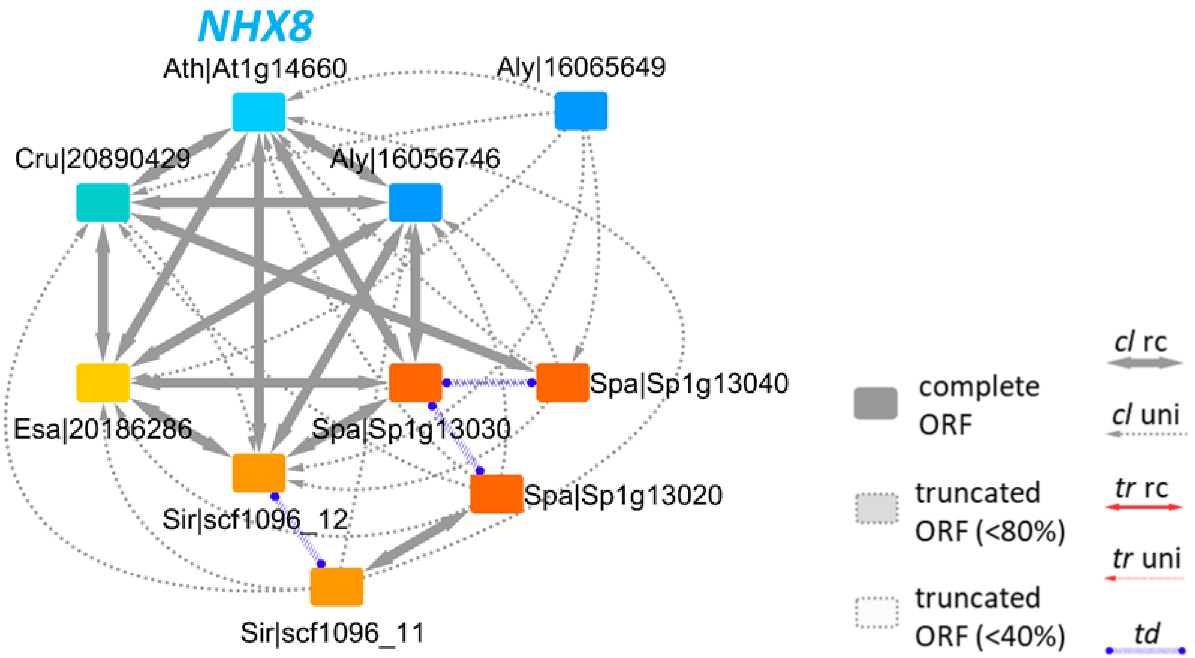


(Oh *et al*., Plant Physiol 2014)

Is this tandem duplication unique to *S. parvula*?

OrthNet reveals the evolutionary history of orthologous gene groups

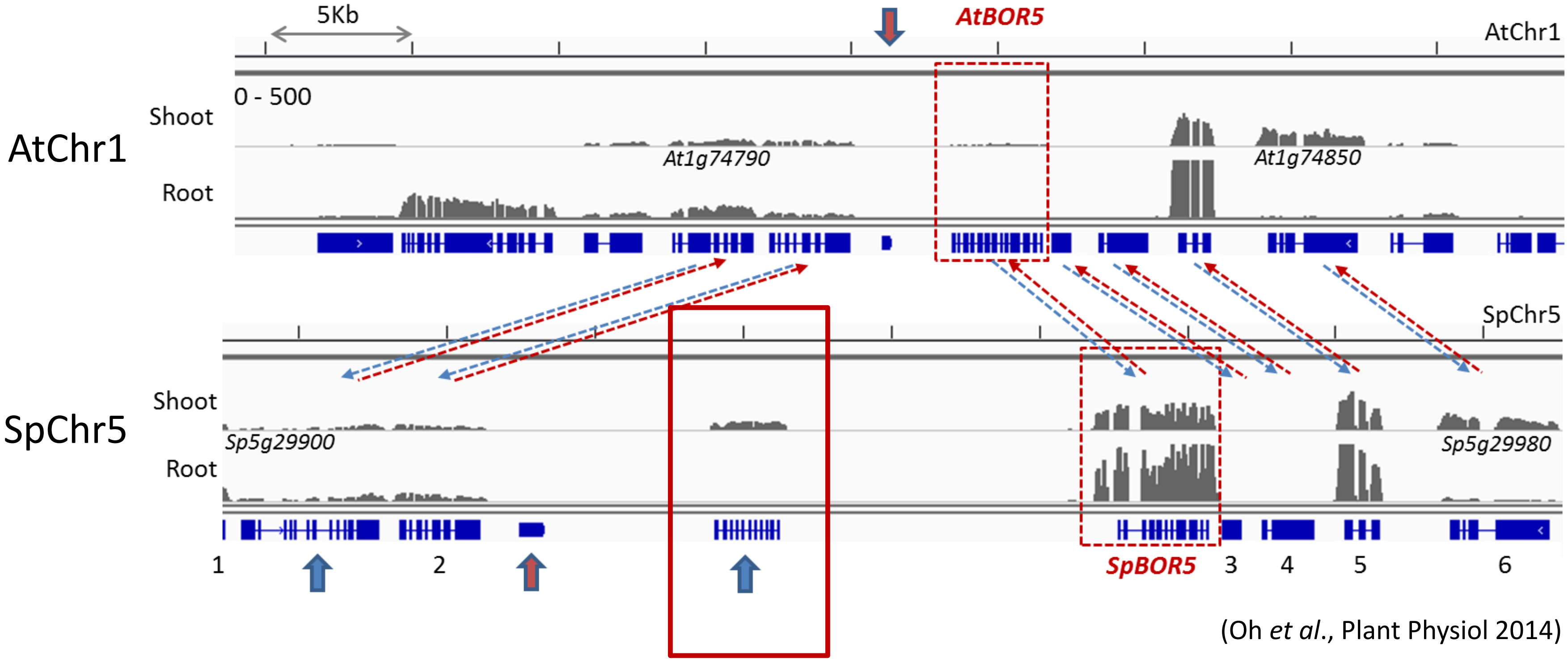
How many of “*S. parvula*-specific” events really unique to *S. parvula*?



- *NHX8* tandem duplication was found in *S. irio* (Sir), too.
- However, only *S. parvula* contained three copies.

OrthNet reveals the evolutionary history of orthologous gene groups

How many of "*S. parvula*-specific" events really unique to *S. parvula*?



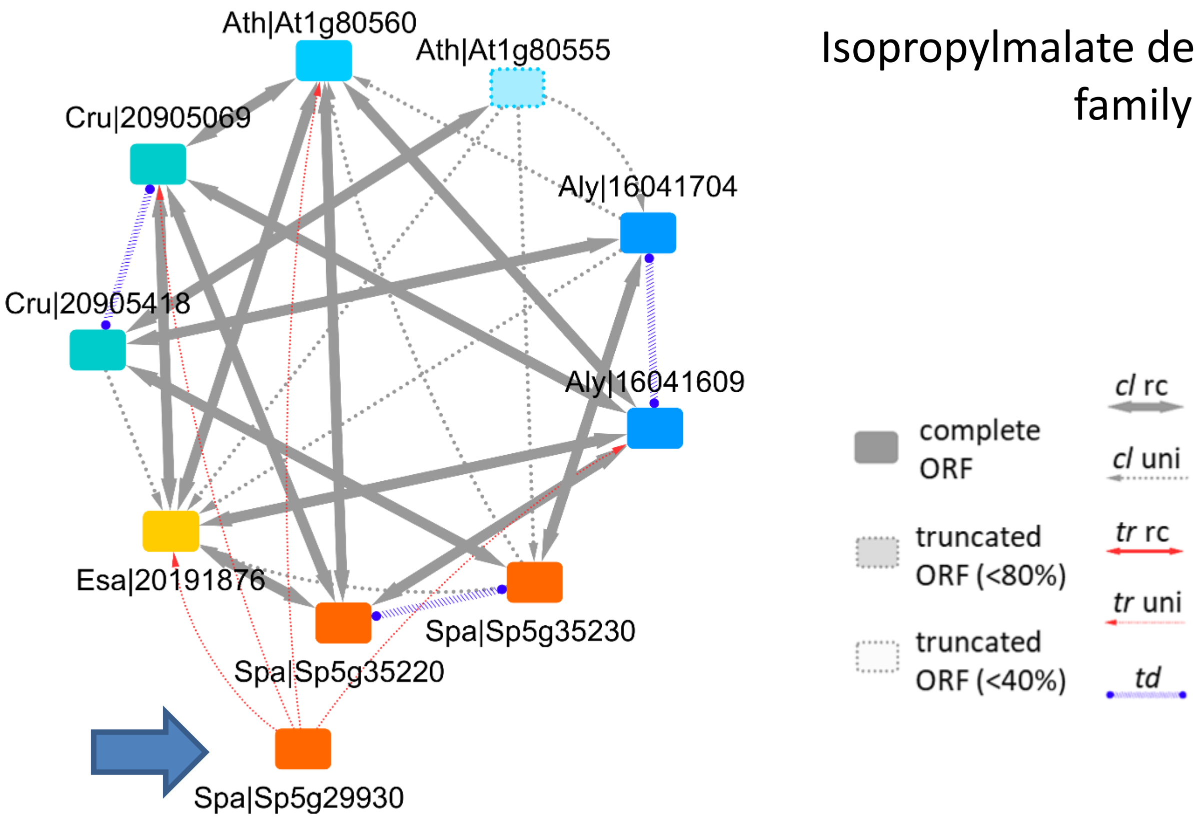
(Oh *et al.*, Plant Physiol 2014)

Is this transposition unique to *S. parvula*?

OrthNet reveals the evolutionary history of orthologous gene groups

How many of "*S. parvula*-specific" events really unique to *S. parvula*?

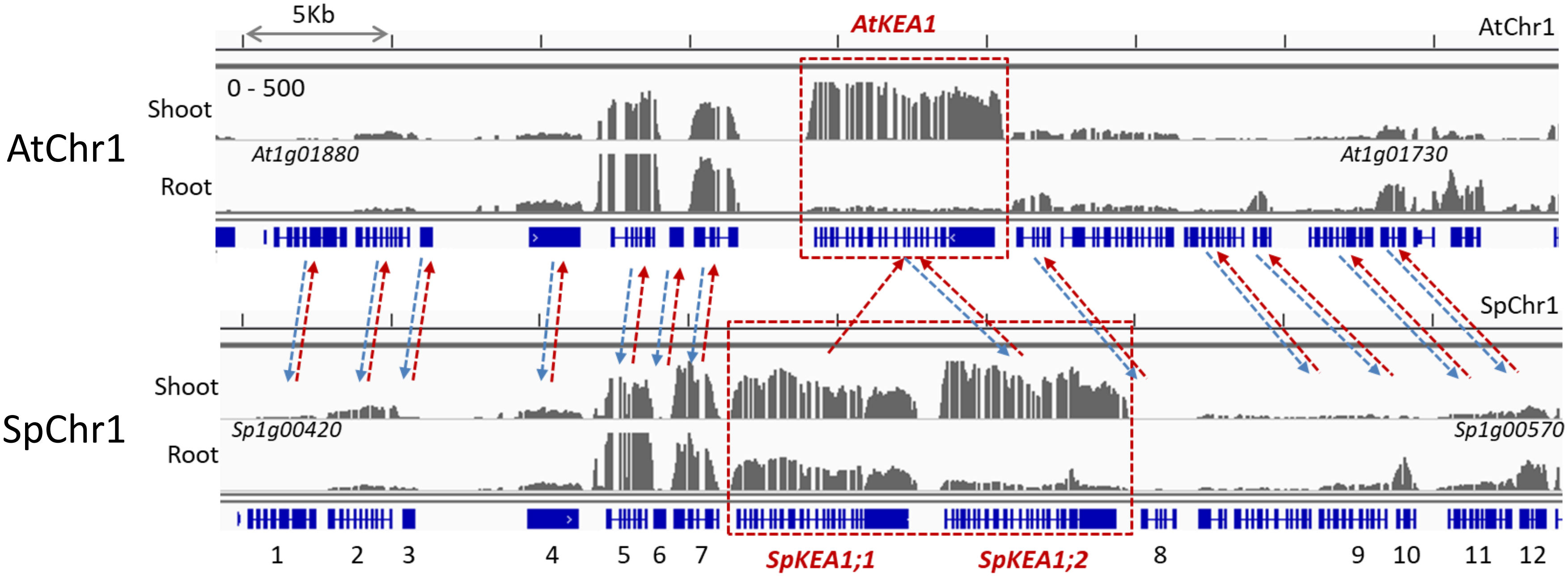
Isopropylmalate dehydrogenase family



Yes, the transposition-duplication (*tr-d*) of Sp5g29930 (the gene inserted at the 5' of SpBOR5) was unique to *S. parvula*.

OrthNet reveals the evolutionary history of orthologous gene groups

How many of “*S. parvula*-specific” events really unique to *S. parvula*?

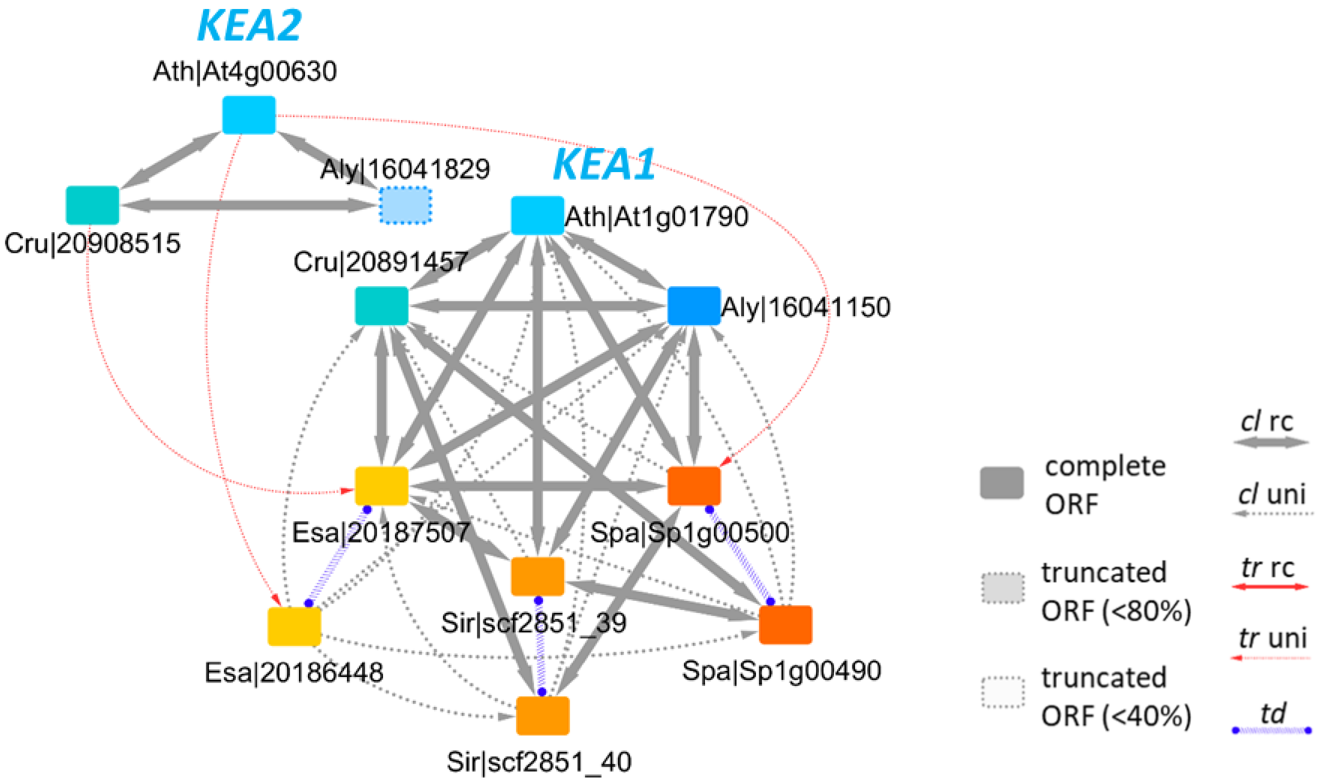


(Oh *et al.*, Plant Physiol 2014)

Is this tandem duplication unique to *S. parvula*?

OrthNet reveals the evolutionary history of orthologous gene groups

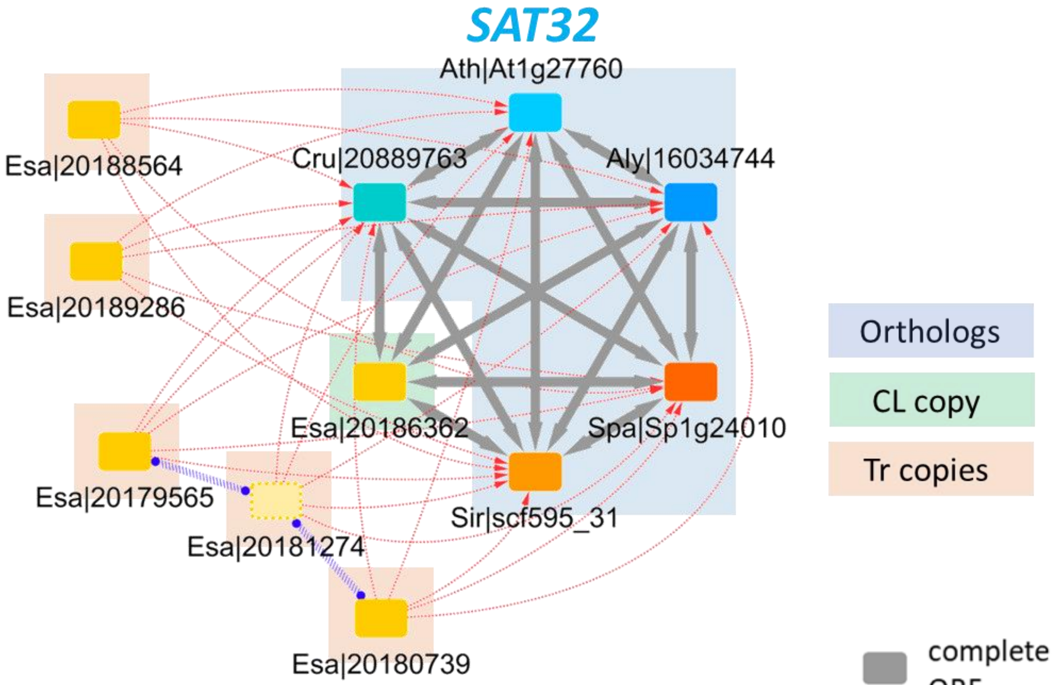
How many of "*S. parvula*-specific" events really unique to *S. parvula*?

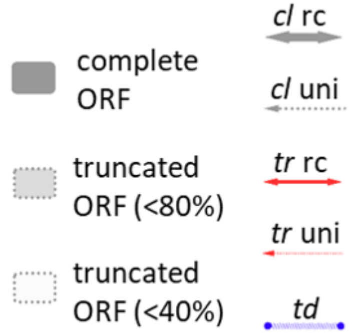


- *SpKEA1* tandem duplication was found in all Lineage II species.
- Lineage I species shared a transposition-duplication (*tr-d*) of *KEA1* (annotated as *KEA2*).

OrthNet reveals the evolutionary history of orthologous gene groups

SALT TOLERANCE 32 (*SAT32*): the largest *E. salsugineum*-specific *tr-d* event





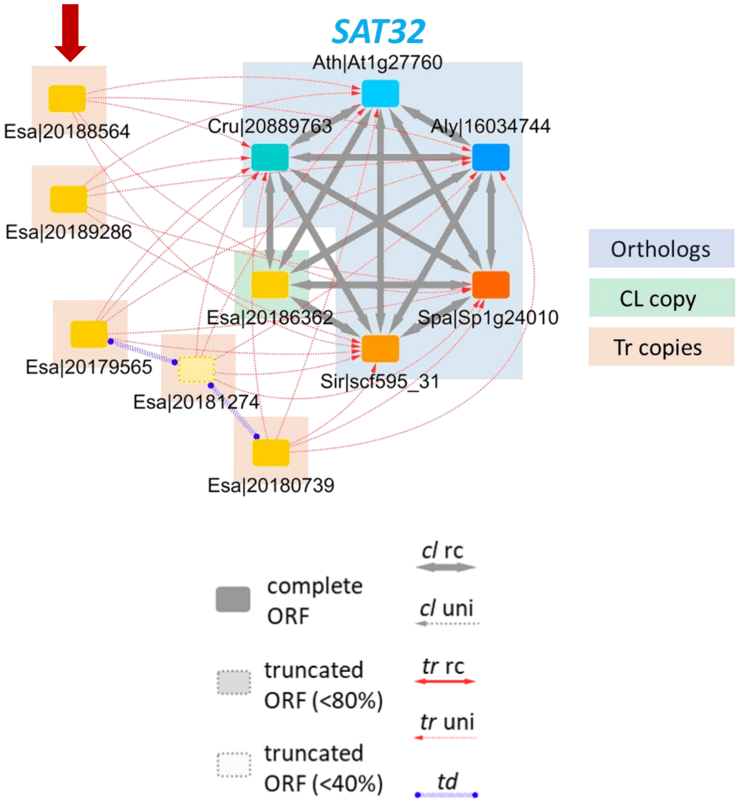
More detail on *tr-d* events in Brassicaceae:

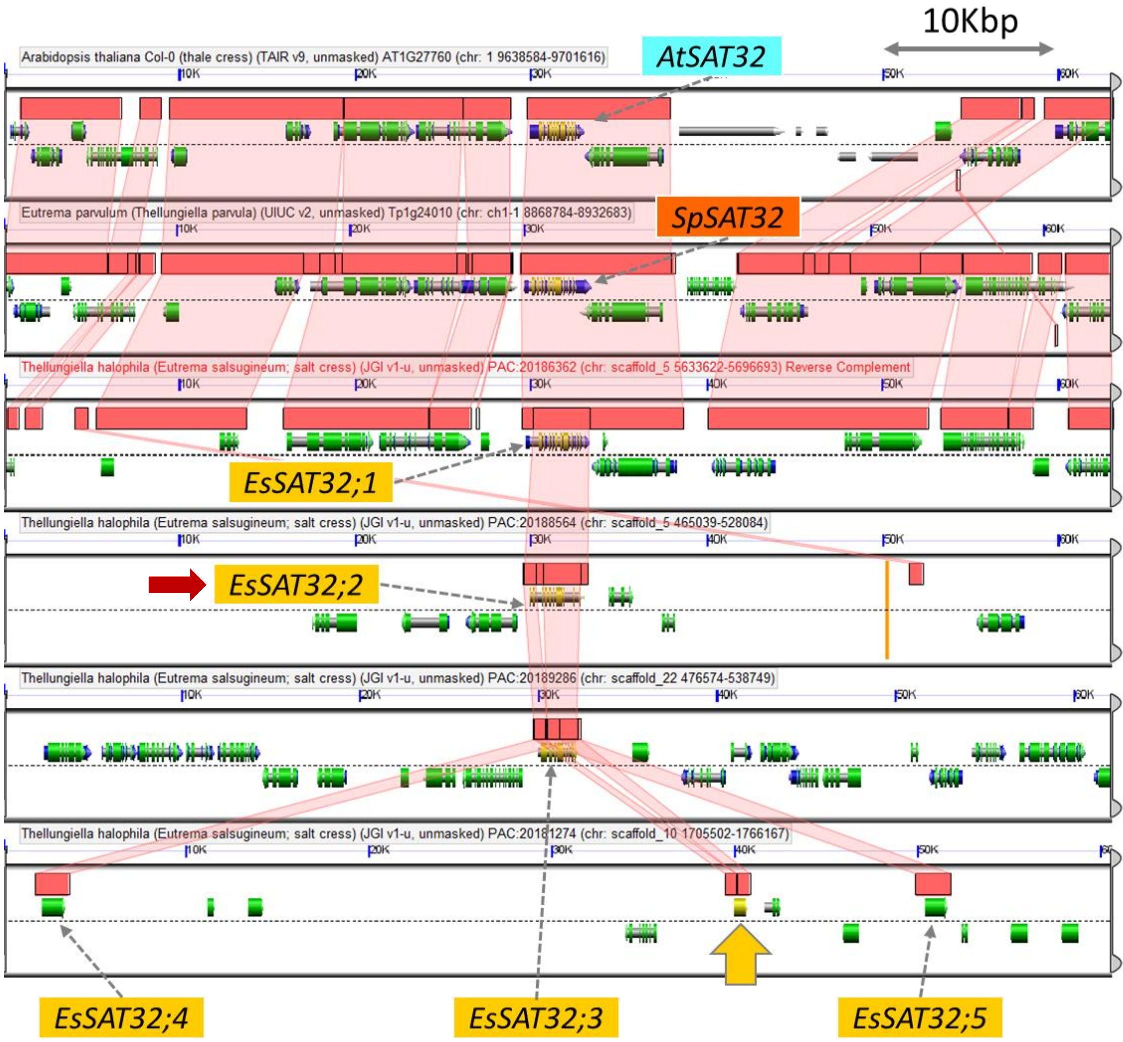
W1069 (Systems Genomics)

(Oh and Dassanayake, *DNA Res, 2018*)

OrthNet reveals the evolutionary history of orthologous gene groups

SALT TOLERANCE 32 (*SAT32*): the largest *E. salsugineum*-specific *tr-d* event

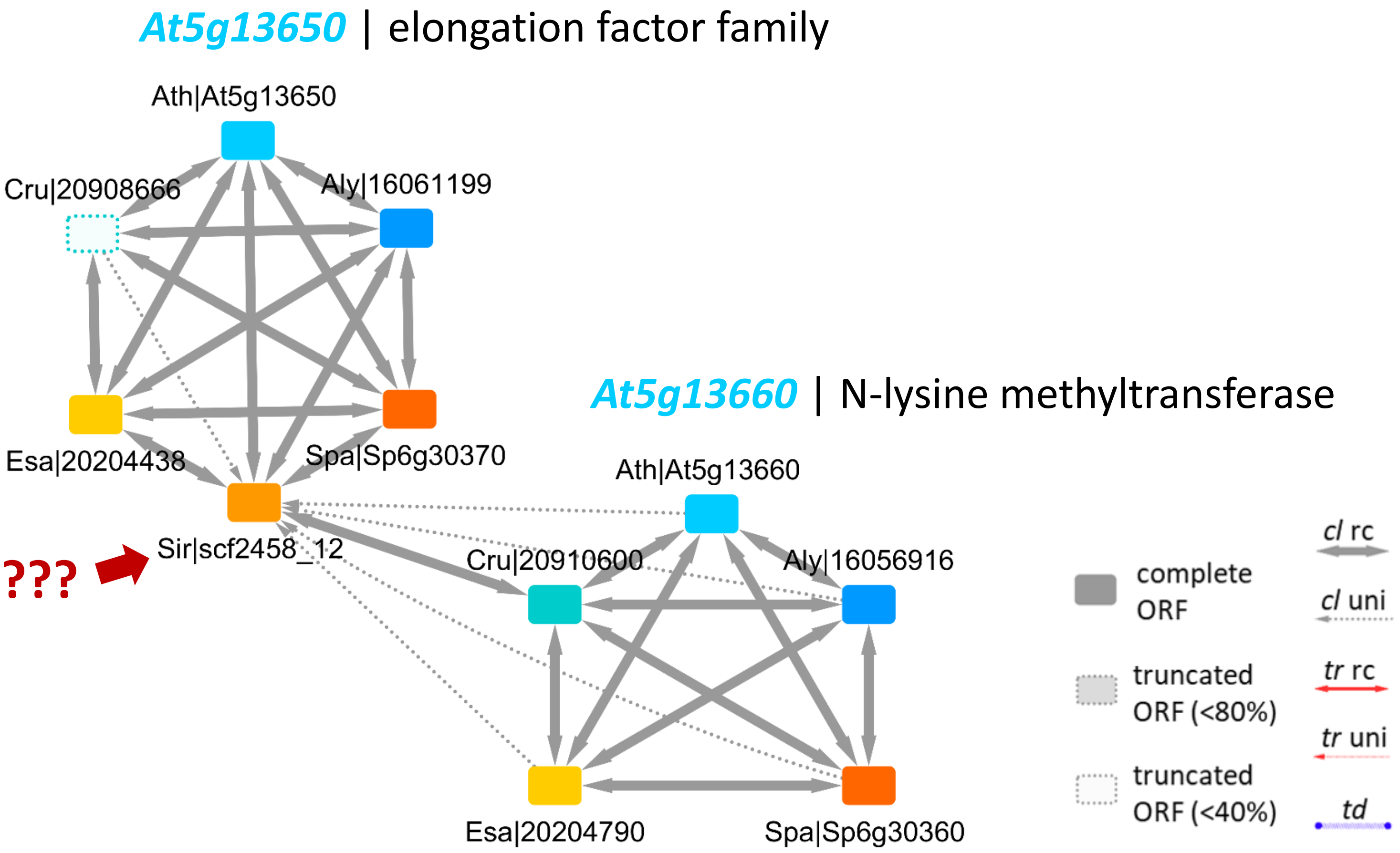



10Kbp
Arabidopsis thaliana Col-0 (thale cress) (TAIR v9, unmasked) AT1G27760 (chr: 1 9638584-9701616)
AtSAT32
Eutrema parvulum (Thellungiella parvula) (UIUC v2, unmasked) Tp1g24010 (chr: ch1-1 8868784-8932683)
SpSAT32
Thellungiella halophila (Eutrema salsugineum; salt cress) (JGI v1-u, unmasked) PAC:20186362 (chr: scaffold_5 5633622-5696693) Reverse Complement
EsSAT32;1
Thellungiella halophila (Eutrema salsugineum; salt cress) (JGI v1-u, unmasked) PAC:20188564 (chr: scaffold_5 465039-528084)
EsSAT32;2
Thellungiella halophila (Eutrema salsugineum; salt cress) (JGI v1-u, unmasked) PAC:20189286 (chr: scaffold_22 476574-538749)
Thellungiella halophila (Eutrema salsugineum; salt cress) (JGI v1-u, unmasked) PAC:20181274 (chr: scaffold_10 1705502-1766167)
10K
20K
30K
40K
50K
60K
EsSAT32;4
EsSAT32;3
EsSAT32;5

(Oh and Dassanayake, *DNA Res, 2018*)

An (unexpected) application: detecting “chimeric” gene models

A characteristic OrthNet topology caused by fused genes

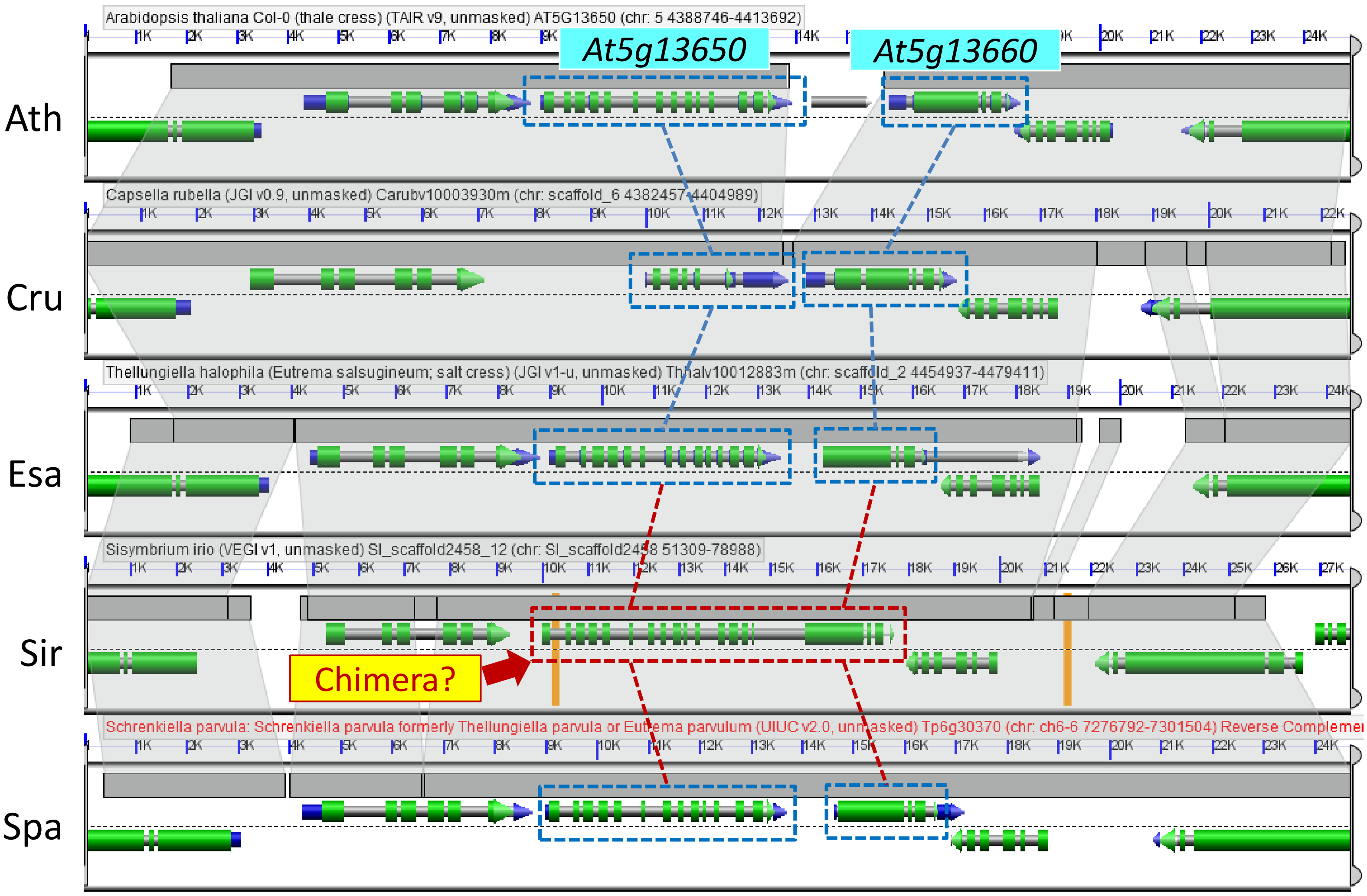


At5g13650 | elongation factor family

At5g13660 | N-lysine methyltransferase

An (unexpected) application: detecting “chimeric” gene models

A characteristic OrthNet topology caused by fused genes



Arabidopsis thaliana Col-0 (thale cress) (TAIR v9, unmasked) AT5G13650 (chr: 5 4388746-4413692)

Capsella rubella (JGI v0.9, unmasked) Carubv10003930m (chr: scaffold_6 4382457-4404989)

Thellungiella halophila (Eutrema salsugineum; salt cress) (JGI v1-u, unmasked) Thhalv10012883m (chr: scaffold_2 4454937-4479411)

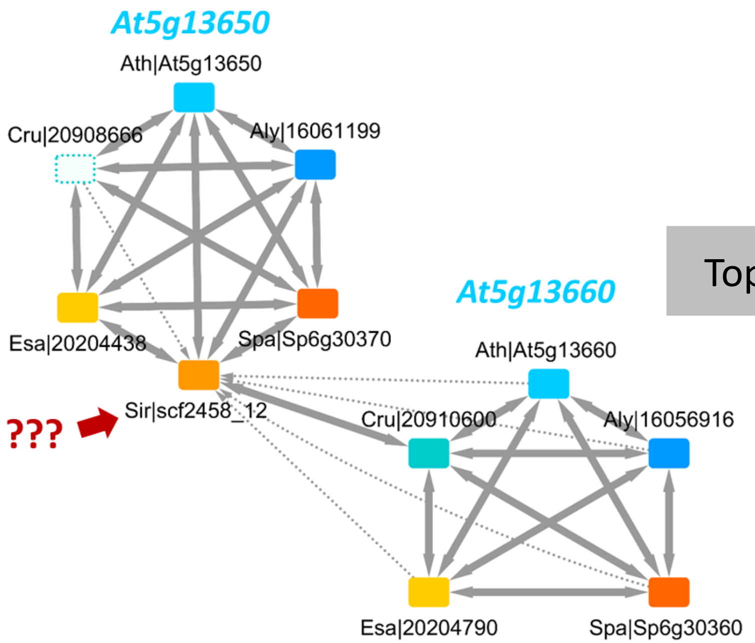
Sisymbrium irio (VEGI v1, unmasked) SI_scaffold2458_12 (chr: SI_scaffold2458 51309-78988)

Schrenkiella parvula: Schrenkiella parvula formerly Thellungiella parvula or Eutrema parvulum (UIUC v2.0, unmasked) Tp6g30370 (chr: ch6-6 7276792-7301504) Reverse Complemen

https://genomevolution.org/r/15k8l

An (unexpected) application: detecting “chimeric” gene models

A characteristic OrthNet topology caused by fused genes



“Chimera” OrthNet topology =

Multiple functionally unrelated, often adjacent, genes captured in the same OrthNet, due to a node in a single (or a subset of) species.

Topology search

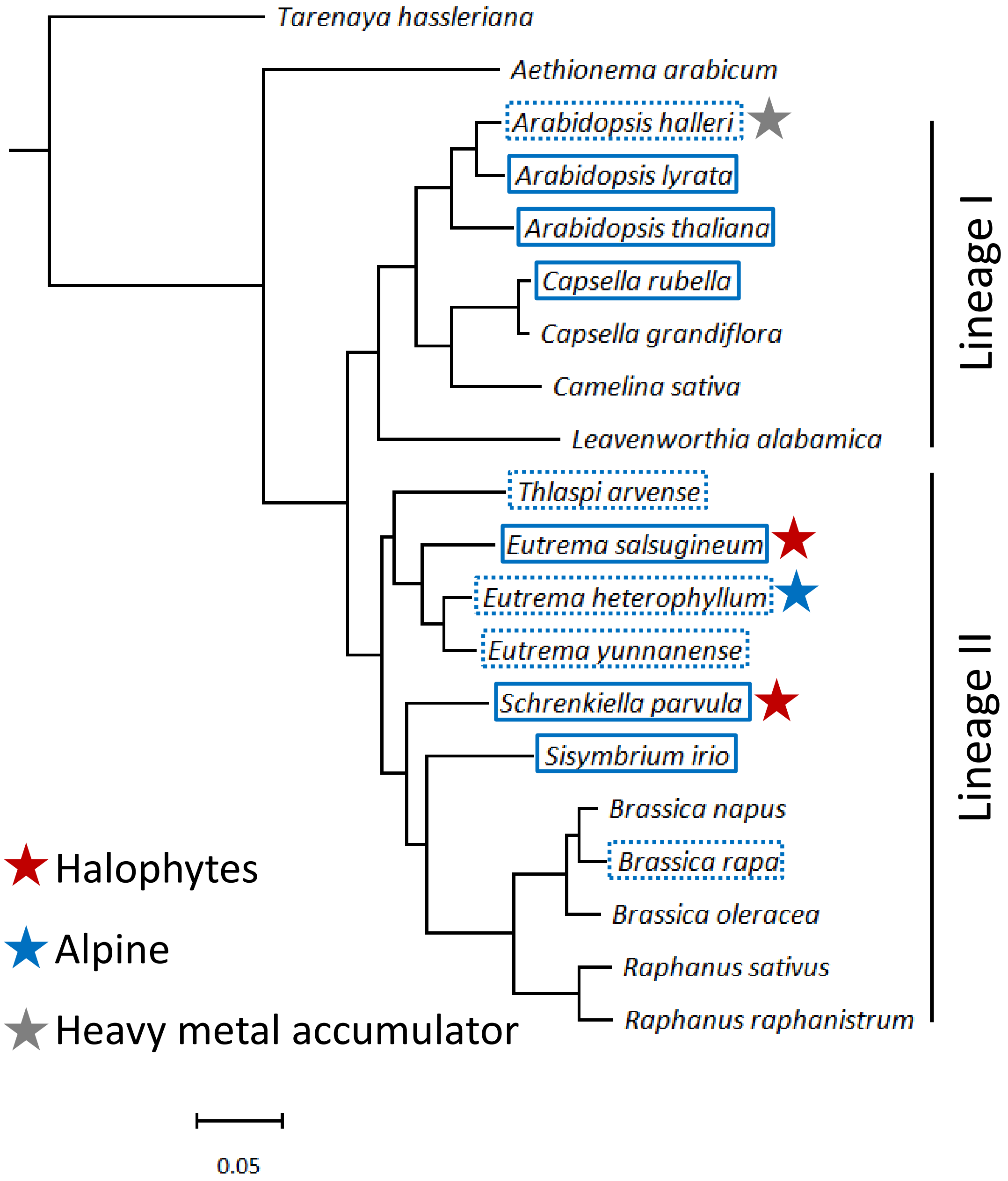
Species	Genome annotation	# of chimeric gene models
Aly	Phytozome 107	47
Ath	Araport11	0
Cru	Phytozome 183	11
Esa	Phytozome 173	8
Sir	ver. 0.2	**285**
Spa	ver. 2.1	10

RNA-Seq or Iso-Seq

Improved gene models

(or confirming a gene fusion event)

Brassicaceae genomes 2018 Fall (part)



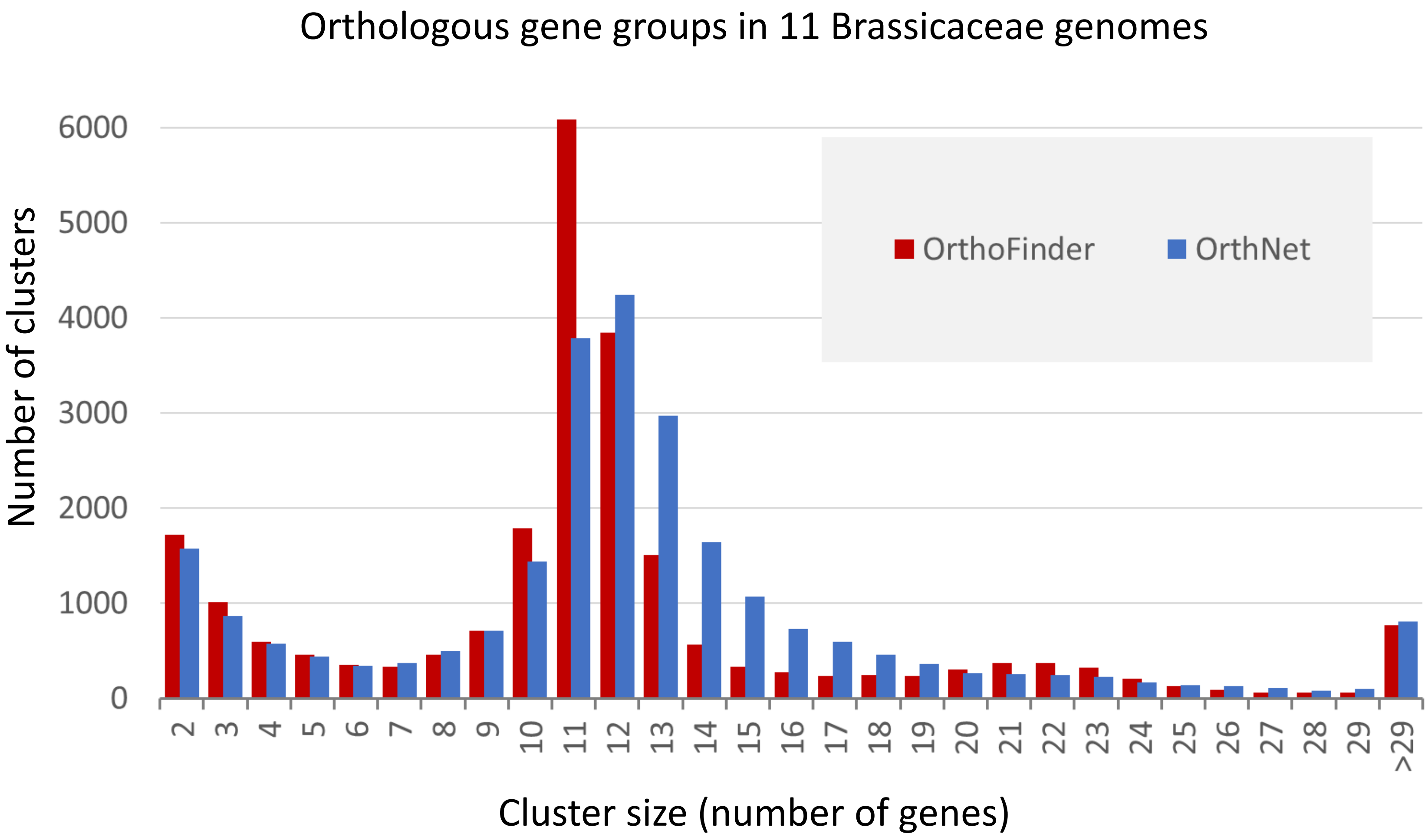
(Guo et al., *DNA Res*, 2018)

Ehe = *E. heterophyllum*

Ehe = *E. yunnanense*

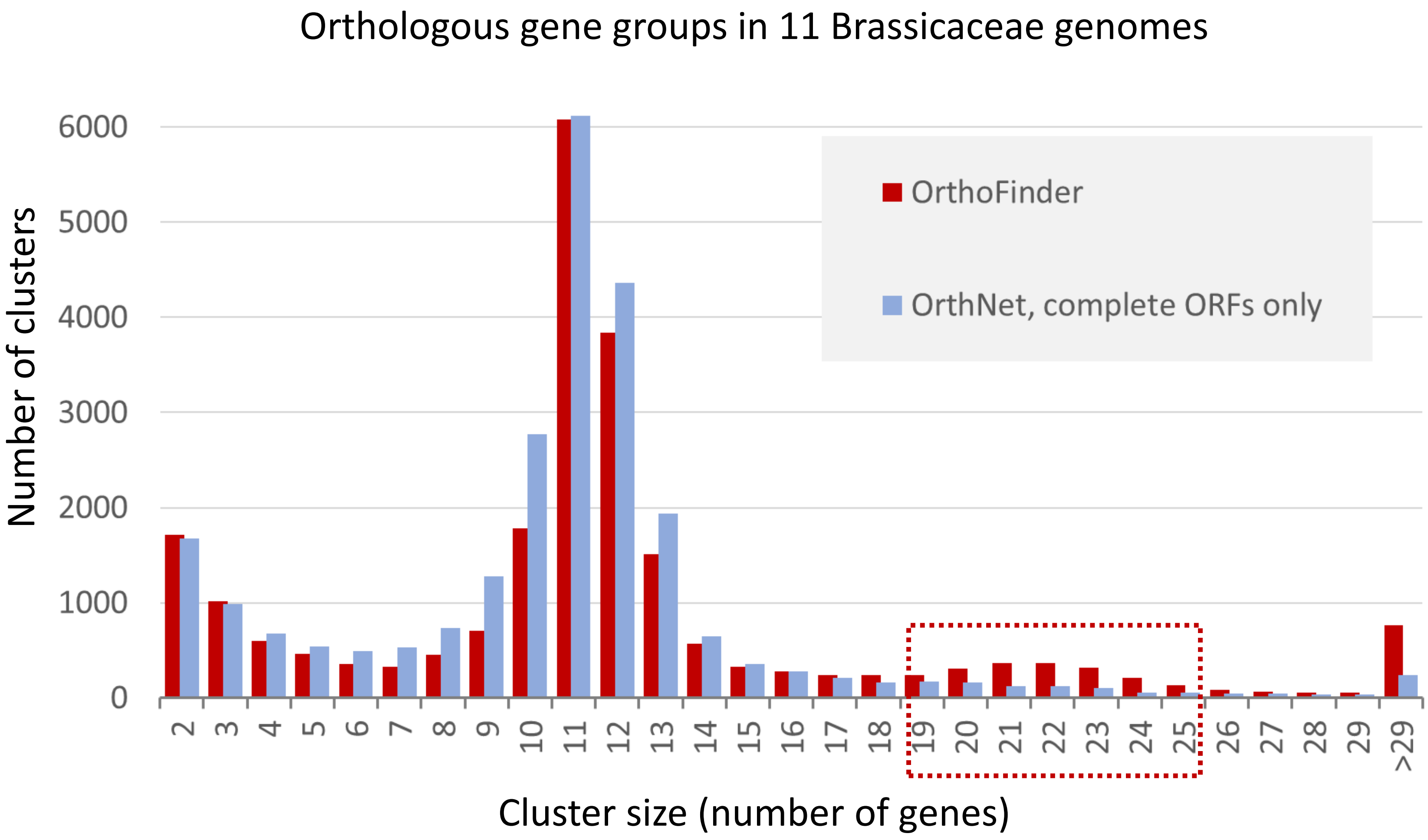
Orthology inference using OrthNets (sequence similarity + co-linearity)

Comparison of OrthNet and OrthoFinder



Orthology inference using OrthNets (sequence similarity + co-linearity)

Comparison of OrthNet and OrthoFinder



Conclusions

- OrthNet encodes evolutionary histories of each gene locus
- OrthNet complements existing tools (e.g. OrthoFinder) in inferring orthology.*
- OrthNet detects truncated and chimeric gene models.
- CLfinder-OrthNet prepares a newly annotated genome (or a set of genomes) into a comparative genomics framework.

*More in W1069 (Systems Genomics)
C20 (Digital Tools and Resources)
Poster #56

Thank you!

Dassanayake Lab.
Maheshi Dassanayake
Guannan Wang
Pramod Pantha
Keiu-Nga Tran
Chathura Wijesinghage

LSU expanded plant biology group
John Larkin, Aaron Smith, Jim Moroney, David Longstreth, and everyone in the group.

Visitors and collaborators (partial list)
Jose Dinneny
Jeongmoo Park
David Medoza-Cozatl
Simon Barak
Michell Arland
Mariana Vargas
Ying Sun
Tai-ping Sun
Sunghoon Lee
Gil Eshel
Patrick Finnegan
...

Funding by

NSF

USDA
United States Department of Agriculture

LSU

RDA
RURAL DEVELOPMENT ADMINISTRATION
Next-Generation BioGreen 21 Program
SSAC Systems & Synthetic Biology Agrobiotech Center