



GENOME DUPLICATION AND GENE ANNOTATION: AN EXAMPLE FOR A REFERENCE PLANT SPECIES.

Alessandra Vigilante, Mara Sangiovanni, Chiara Colantuono, Luigi Frusciante and Maria Luisa Chiusano

Dept. of Soil, Plant, Environmental and Animal Production Sciences

CAB (Computer Aided Biosciences) group
Web: <http://cab.unina.it>

Arabidopsis thaliana WAS THE FIRST PLANT GENOME
TO BE COMPLETELY SEQUENCED



A REFERENCE GENOME SHOULD BE:

FULLY RELIABLE

SAFELY ANNOTATED

WELL UNDERSTOOD IN TERMS OF EVOLUTIONARY HISTORY

A REFERENCE GENOME SHOULD BE:

FULLY RELIABLE

SAFELY ANNOTATED

WELL UNDERSTOOD IN TERMS OF EVOLUTIONARY HISTORY

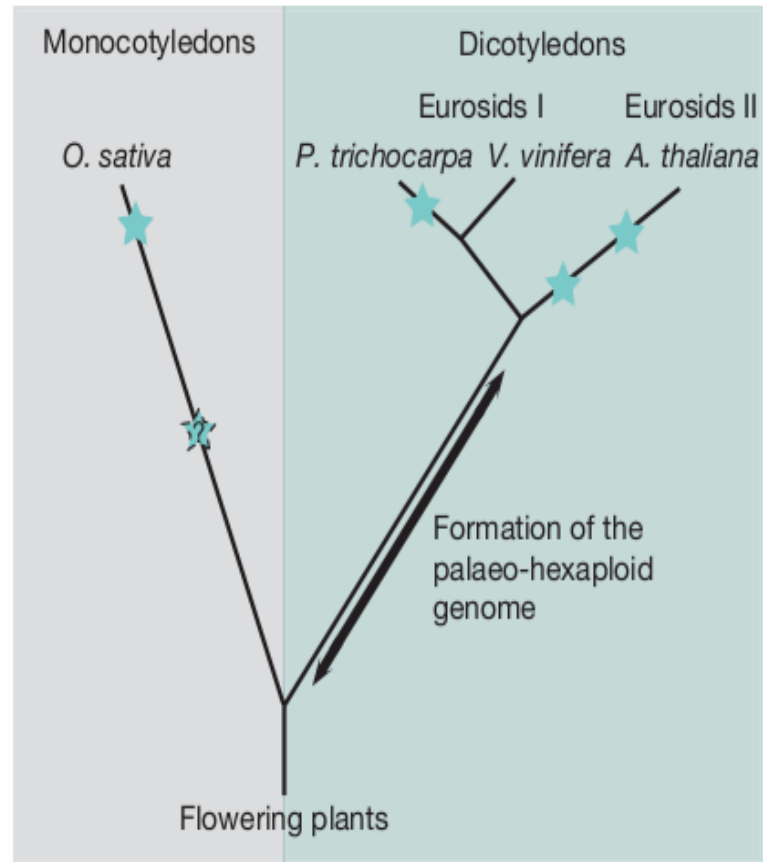
Arabidopsis thaliana GENOME IS:

GENE DENSE

COMPLEX BECAUSE
HIGHLY DUPLICATED AND CLAIMED TO BE ARCHEOPOLYPLOID

STILL NOT EXHAUSTIVELY ANNOTATED

BACKGROUND: *Arabidopsis thaliana* as a reference genome

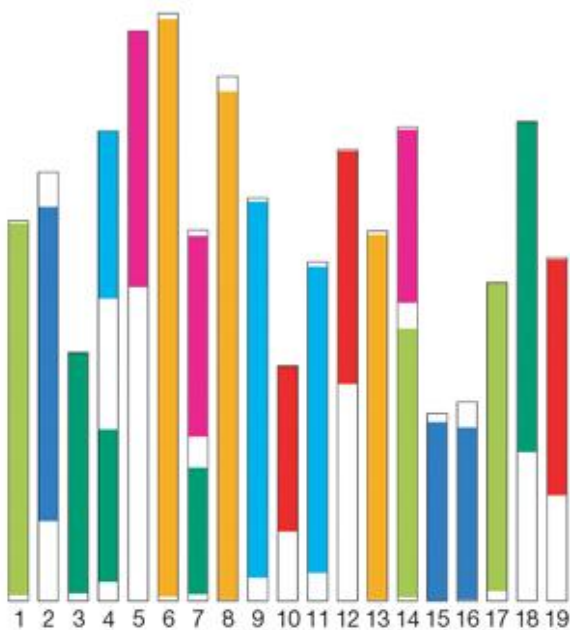


★ Whole genome duplication events

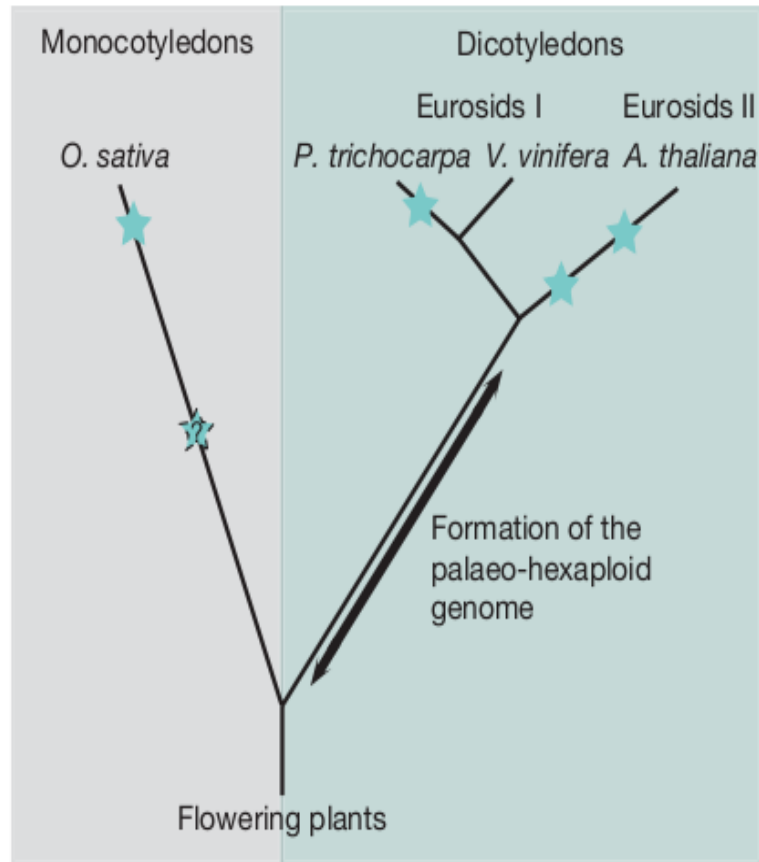
Nature (2007)

Jaillon O, et al. The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm phyla.

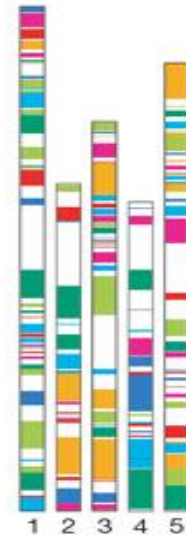
BACKGROUND: *Arabidopsis thaliana* as a reference genome



V. vinifera



★ Whole genome duplication events



A. thaliana

Nature (2007)

Jaillon O, et al. The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm phyla.

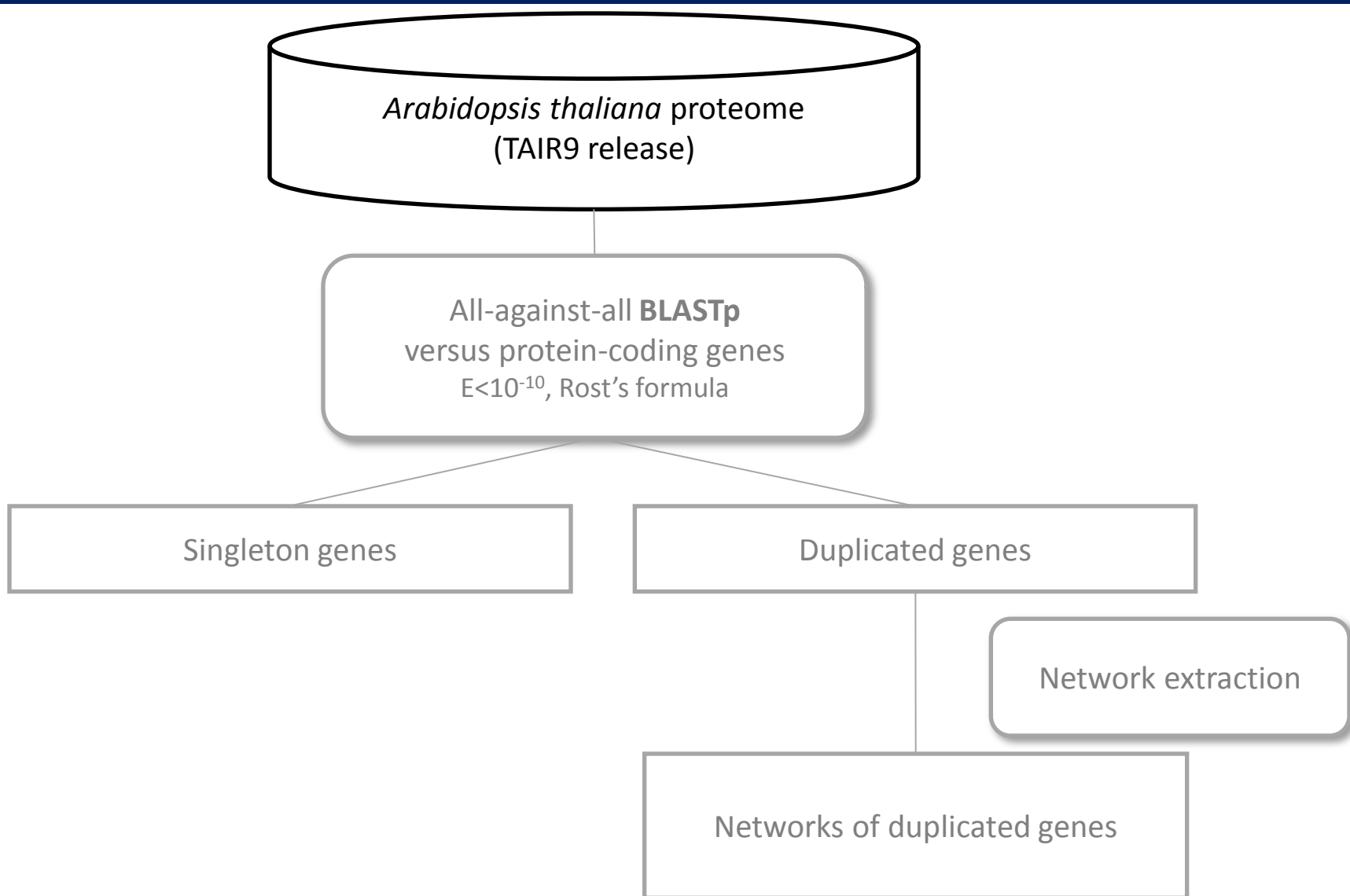
IT COULD BE USEFUL TO REVIEW THE GENOME IN TERMS OF RELATIONSHIPS BETWEEN
DUPLICATED GENES/PARALOGS

IT COULD BE USEFUL TO REVIEW THE GENOME IN TERMS OF RELATIONSHIPS BETWEEN
DUPLICATED GENES/PARALOGS

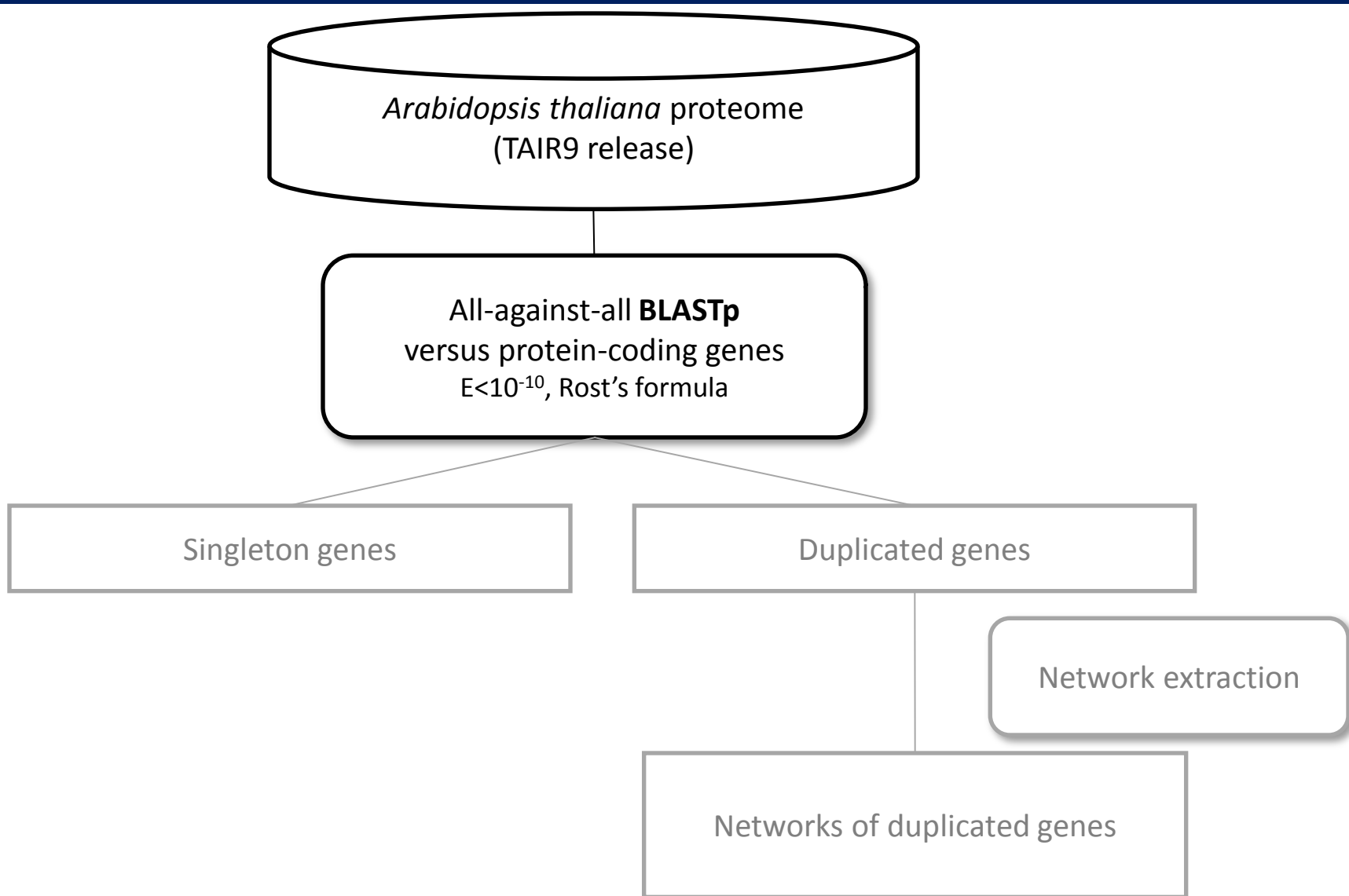


NETWORKS OF PARALOG GENES

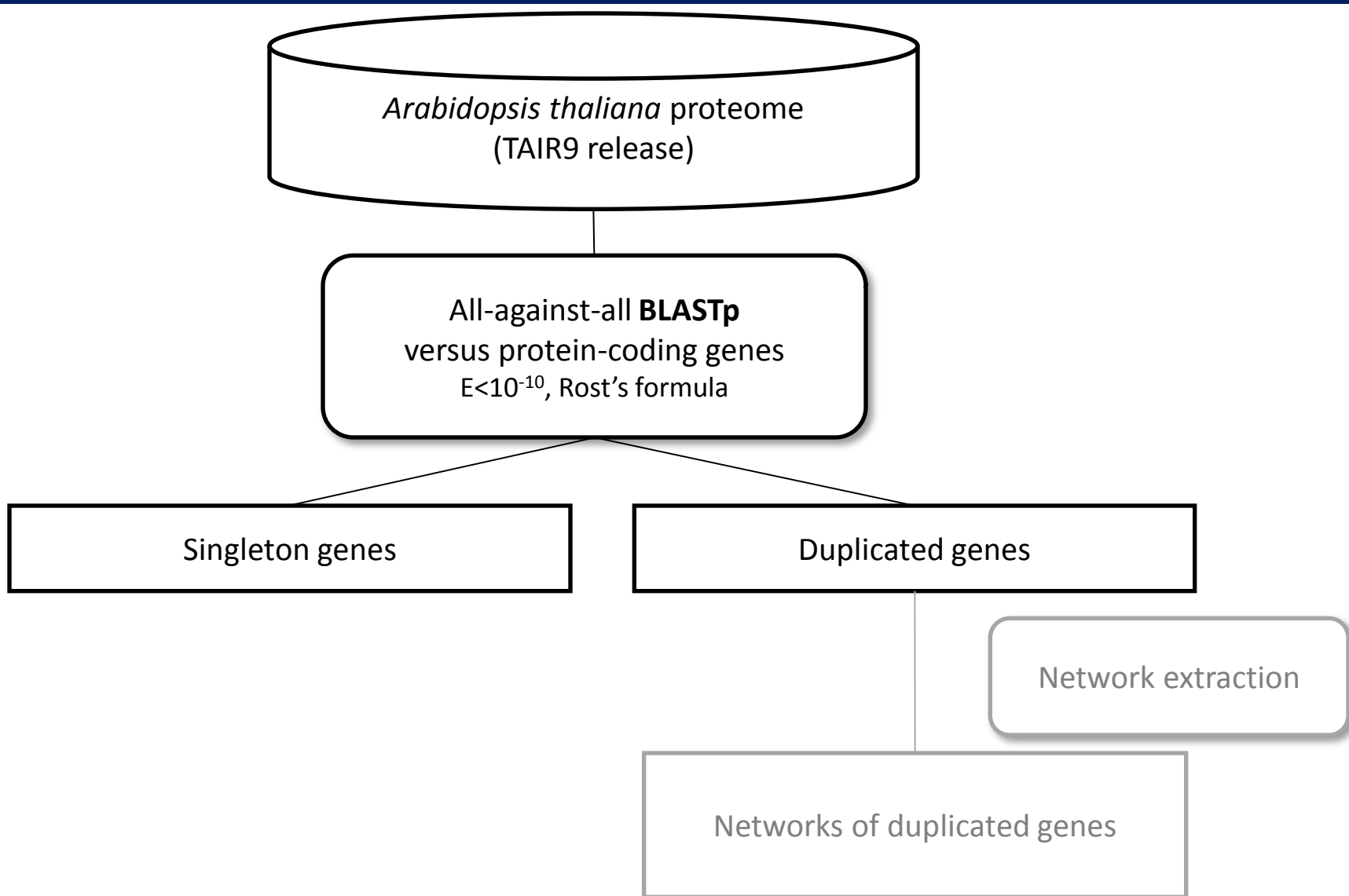
Gene duplication analysis: pipeline



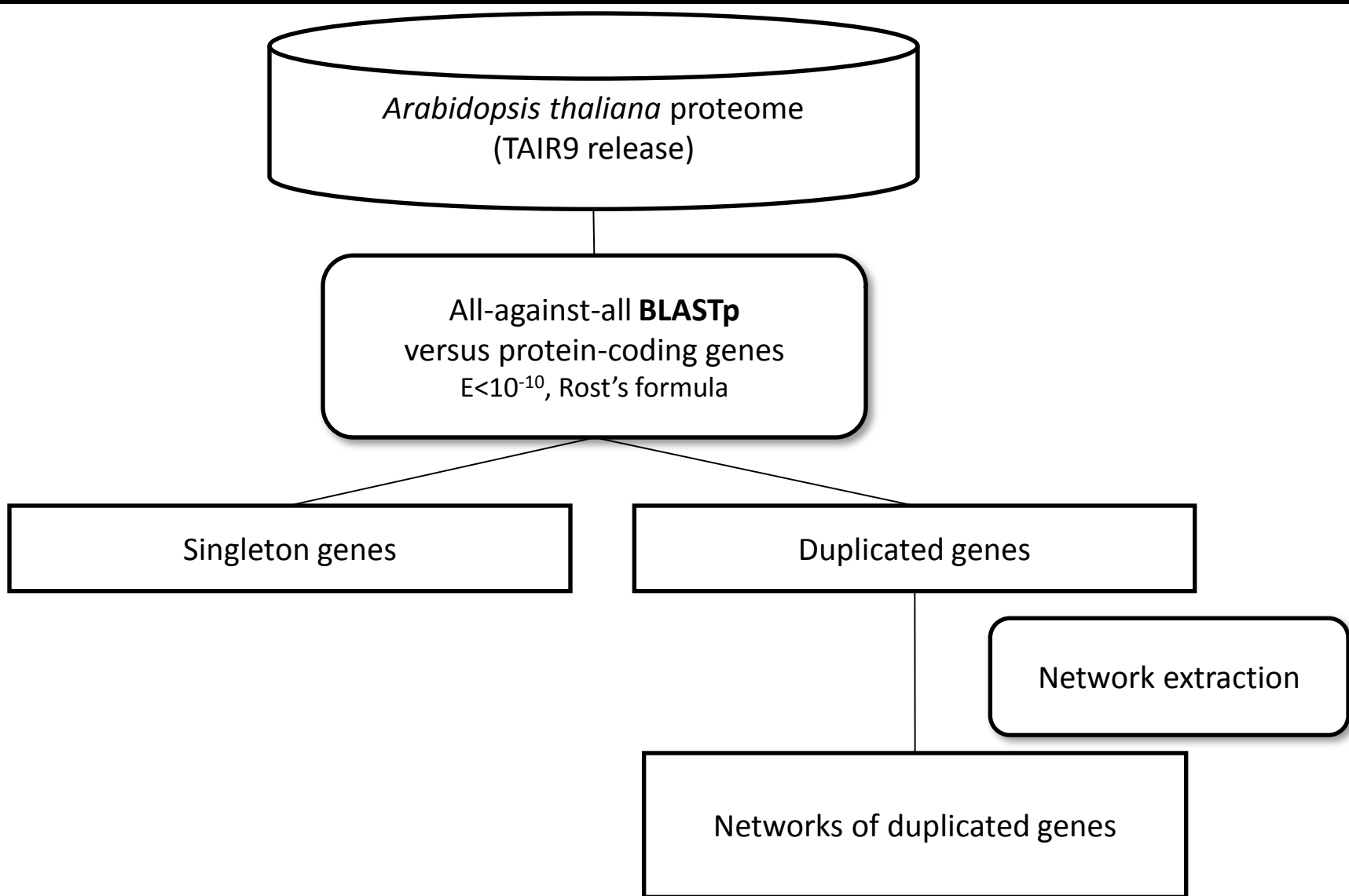
Gene duplication analysis: pipeline



Gene duplication analysis: pipeline



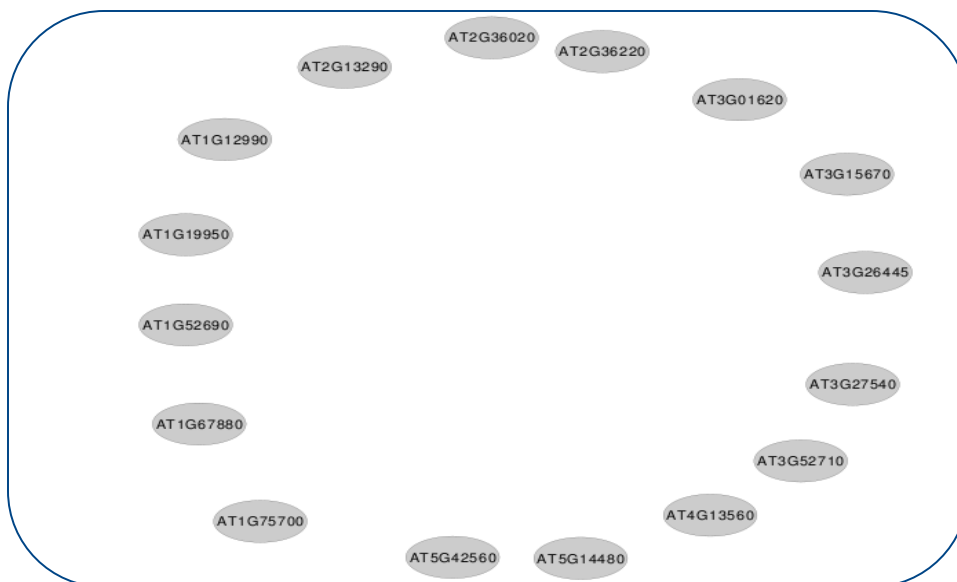
Gene duplication analysis: pipeline



A. thaliana GENES AND PARALOGIES ARE REPRESENTED AS AN (UNDIRECTED) **GRAPH**
 $G(V,E)$

WHERE:

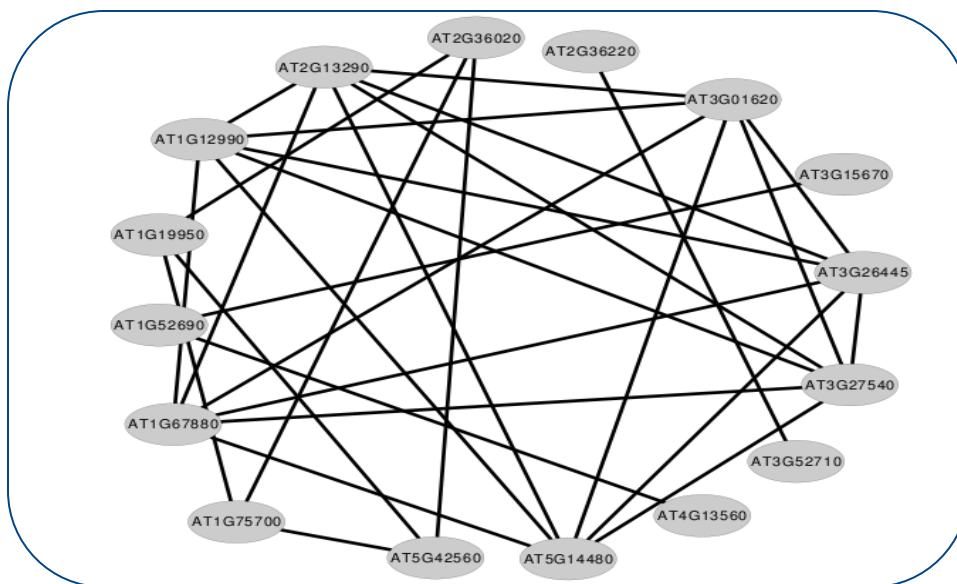
- $V = \{v_1, \dots, v_N\}$ = genes
- $E = \{e_1, \dots, e_M\}$ = paralogies



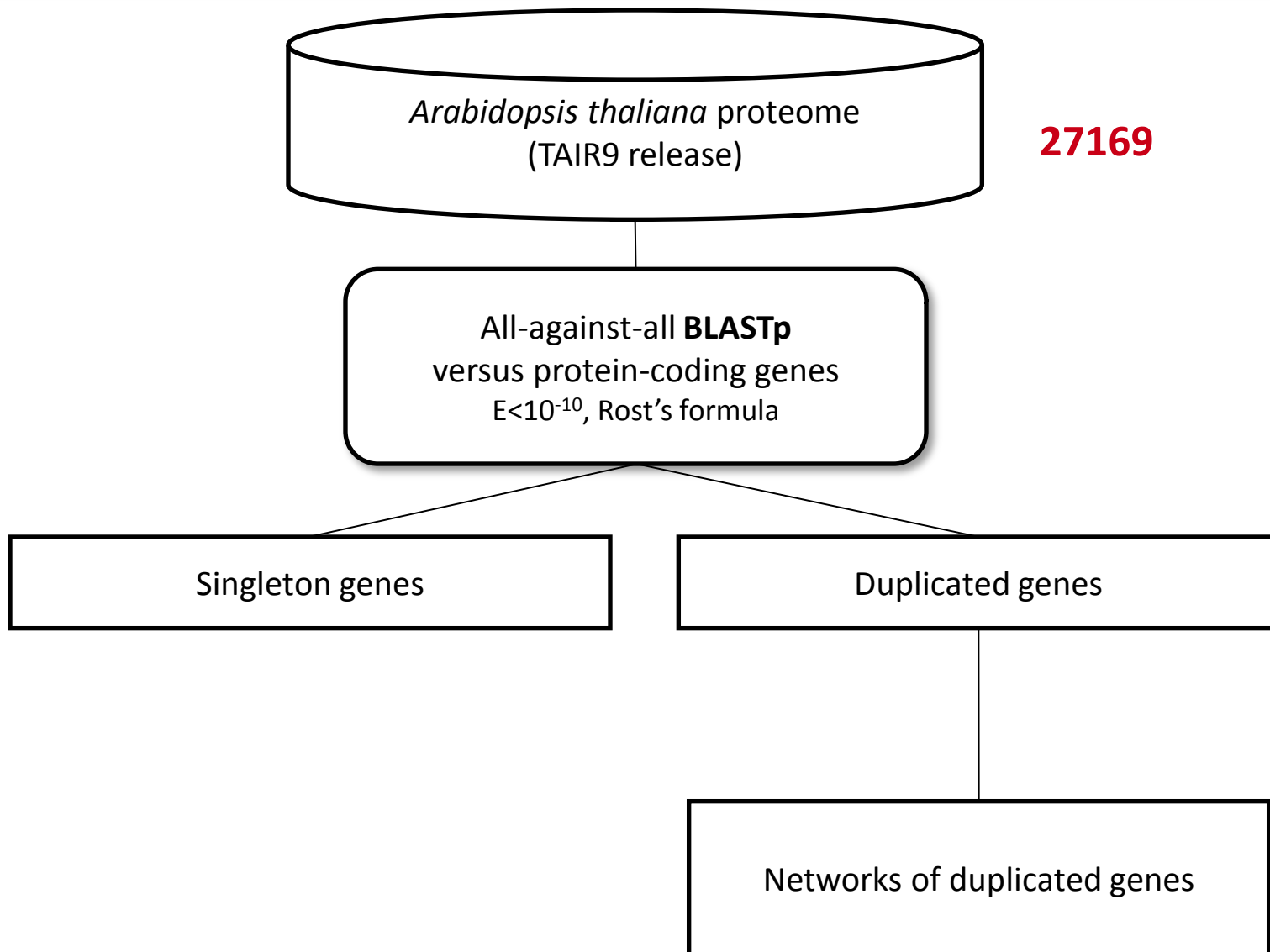
A. thaliana GENES AND PARALOGIES ARE REPRESENTED AS AN (UNDIRECTED) **GRAPH**
 $G(V,E)$

WHERE:

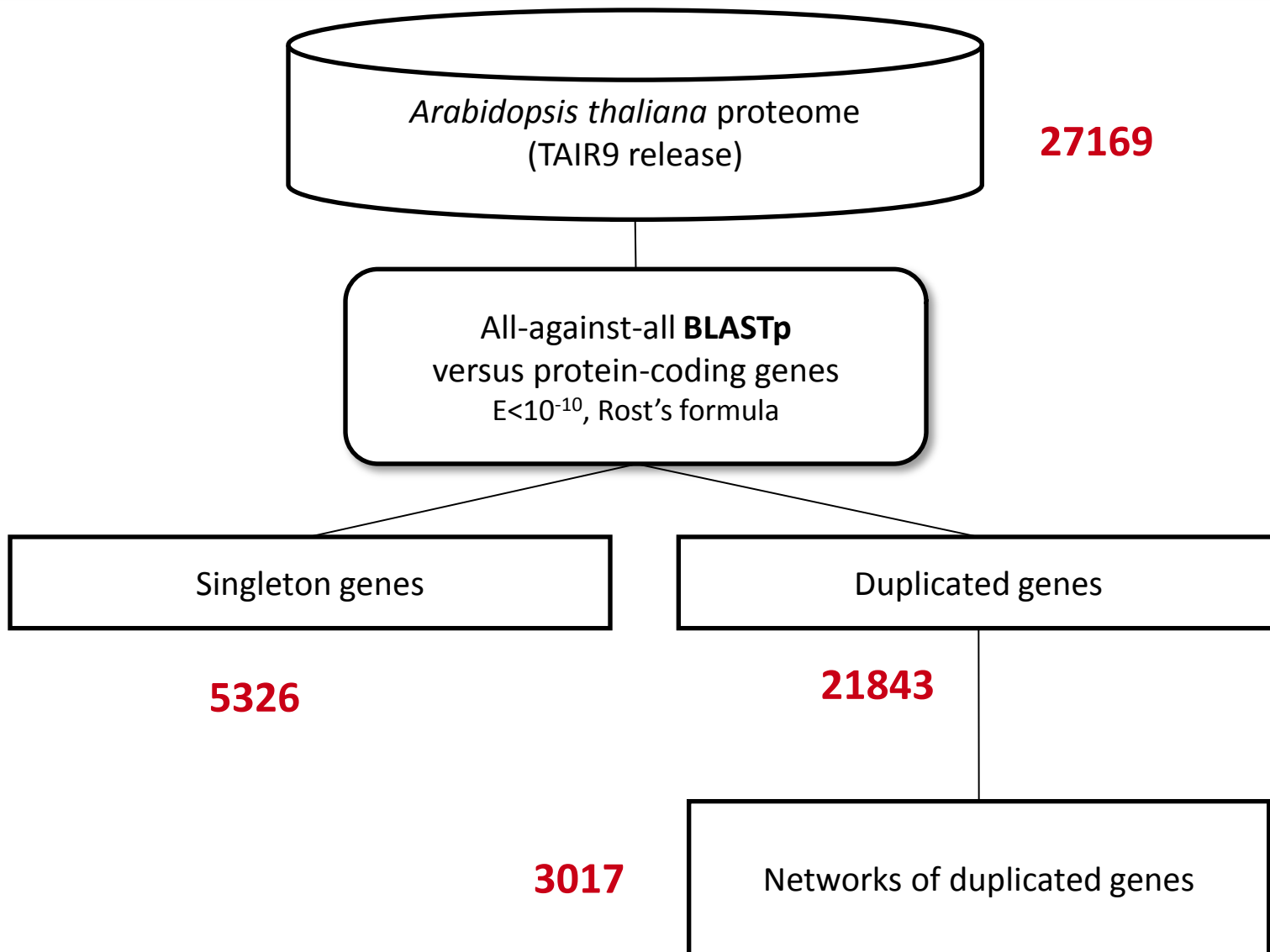
- $V = \{v_1, \dots, v_N\}$ = genes
- $E = \{e_1, \dots, e_M\}$ = paralogies



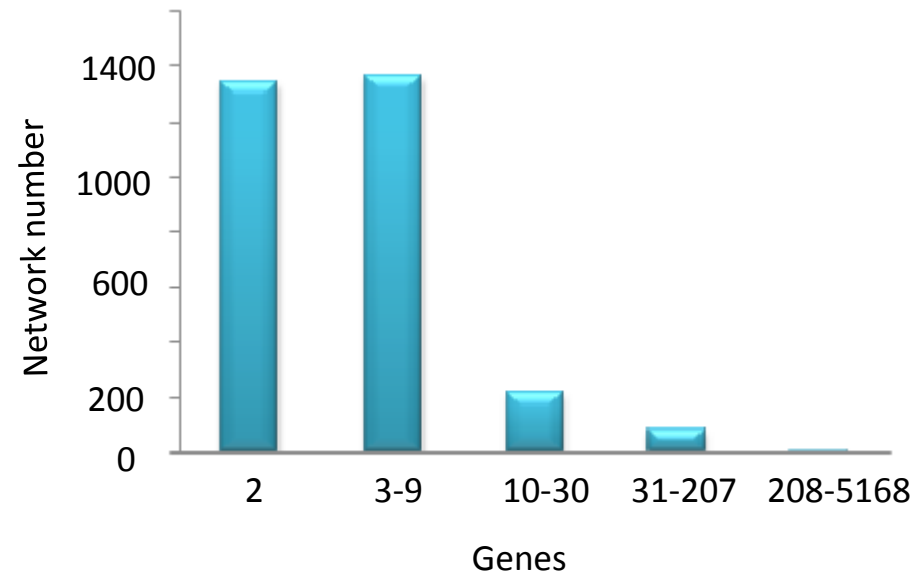
RESULTS: Gene duplication analysis



RESULTS: Gene duplication analysis



RESULTS: Gene duplication analysis



A network contains all and only the genes that share at least one paralogy relationship.

Each gene belongs to one and only one network.

Arabidopsis chromosomes



NETWORKS ARE A USEFUL TOOL TO DEEPLY INVESTIGATE RELATIONSHIPS BETWEEN SUBSETS OF GENES SHARING DUPLICATION RELATIONSHIPS

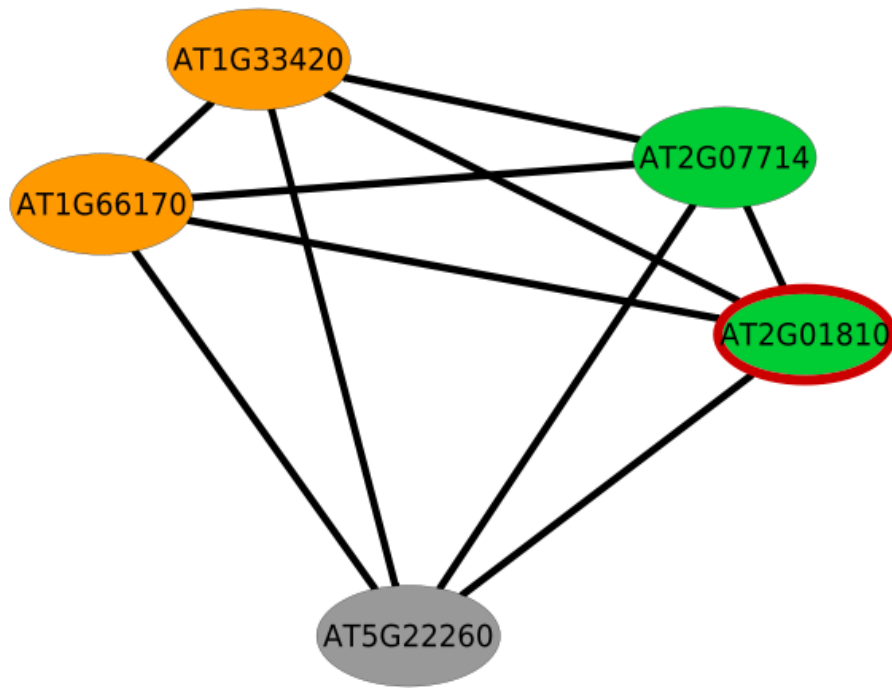
NETWORKS CAN BE A USEFUL TOOL TO REFINE GENE ANNOTATION

RESULTS: Networks applications for the annotation of unknown information

The 19% of the proteome is still annotated as “unknown protein”

UNKNOWN PROTEIN GENES (TAIR9)

TOTAL	SINGLETON	DUPLICATED
5274	2445	2829

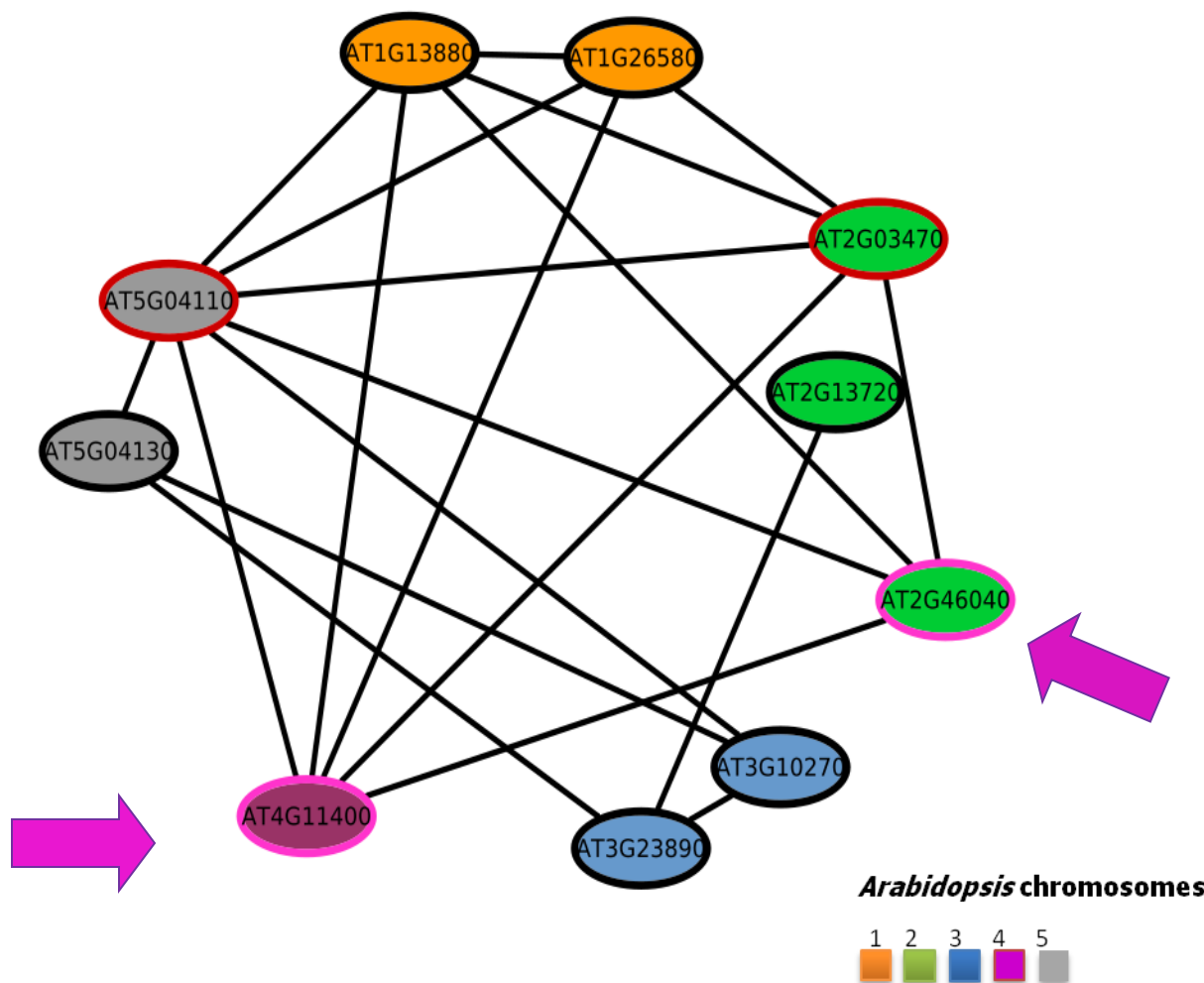


NETWORKS CAN HELP IN REFINING GENE ANNOTATION

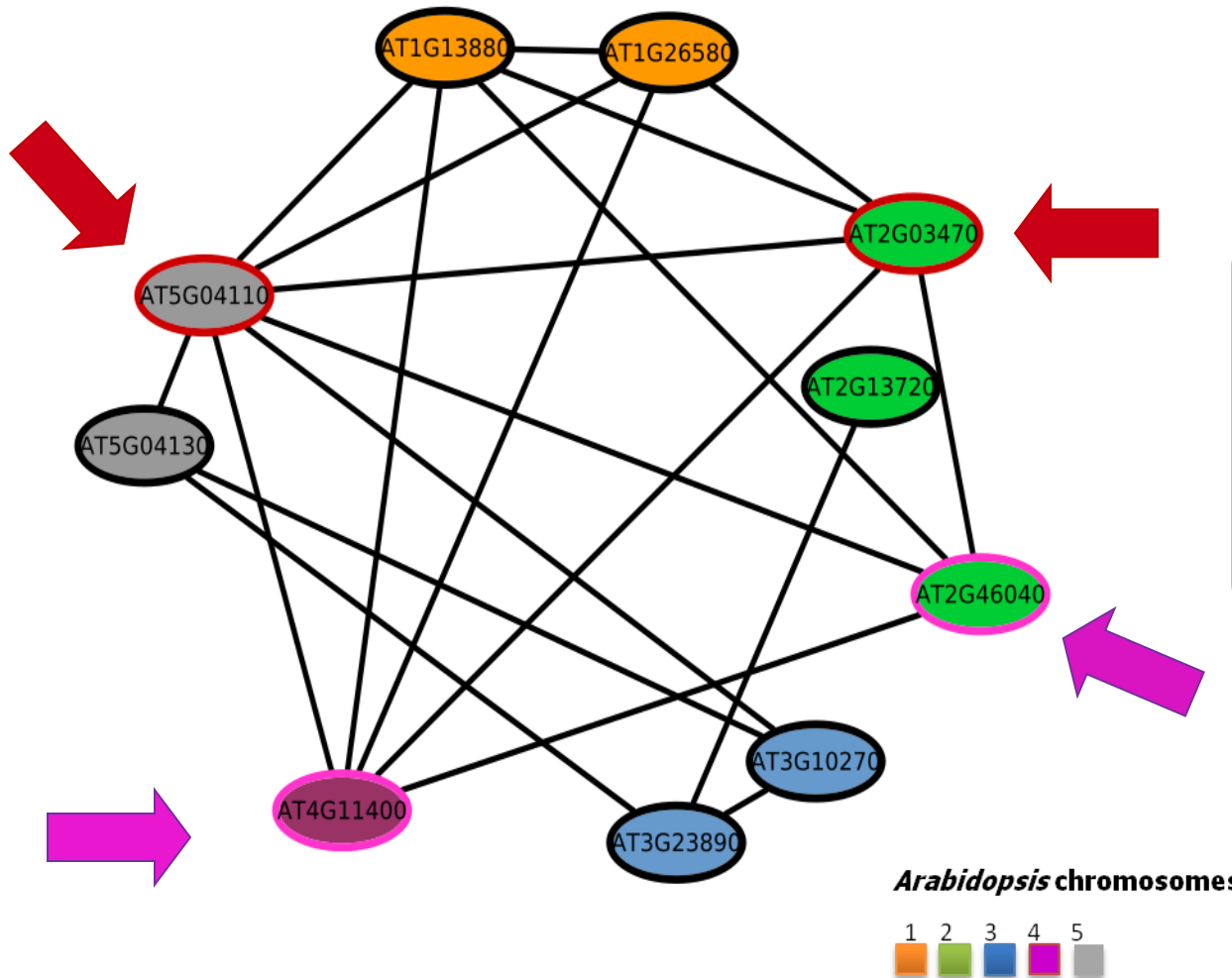
Arabidopsis chromosomes



RESULTS: Networks applications for study of relationships between gene families



RESULTS: Networks applications for study of relationships between gene families



Networks can be a useful tool for highlighting evolutionary relationships between different gene families

Arabidopsis chromosomes



RESULTS: Networks of duplicated genes

	GENE NUMBER	NET NUMBER	% NETS	% PROTEIN CODING GENES
	1	0	0	20
	2 - 5168	3017	100	80
TOTALS	27169	3017	100	100

RESULTS: Networks of duplicated genes

	GENE NUMBER	NET NUMBER	% NETS	% PROTEIN CODING GENES
	1	0	0	20
	2 - 5168	3017	100	80
TOTALS	27169	3017	100	100

	GENE NUMBER	NET NUMBER	% NETS	% PROTEIN CODING GENES
	2	1347	44.65	10
	3 - 9	1370	45.41	23
	10 - 30	216	7.16	13
	31 - 207	83	2.75	15
→	208 - 5168	1	0.03	19
TOTALS	21843	3017	100	80

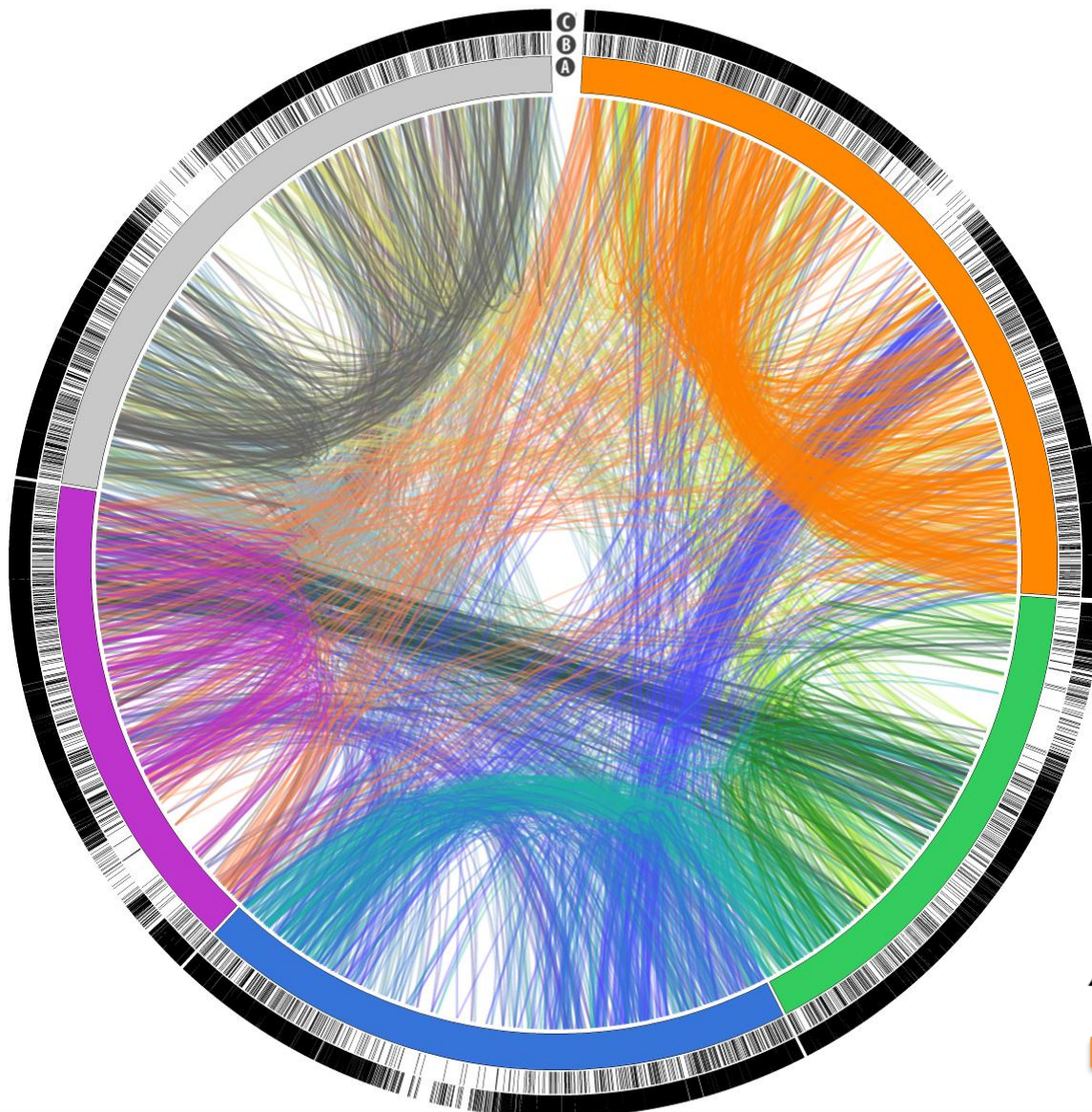
RESULTS: Networks of duplicated genes

	GENE NUMBER	NET NUMBER	% NETS	% PROTEIN CODING GENES
	1	0	0	20
	2 - 5168	3017	100	80
TOTALS	27169	3017	100	100

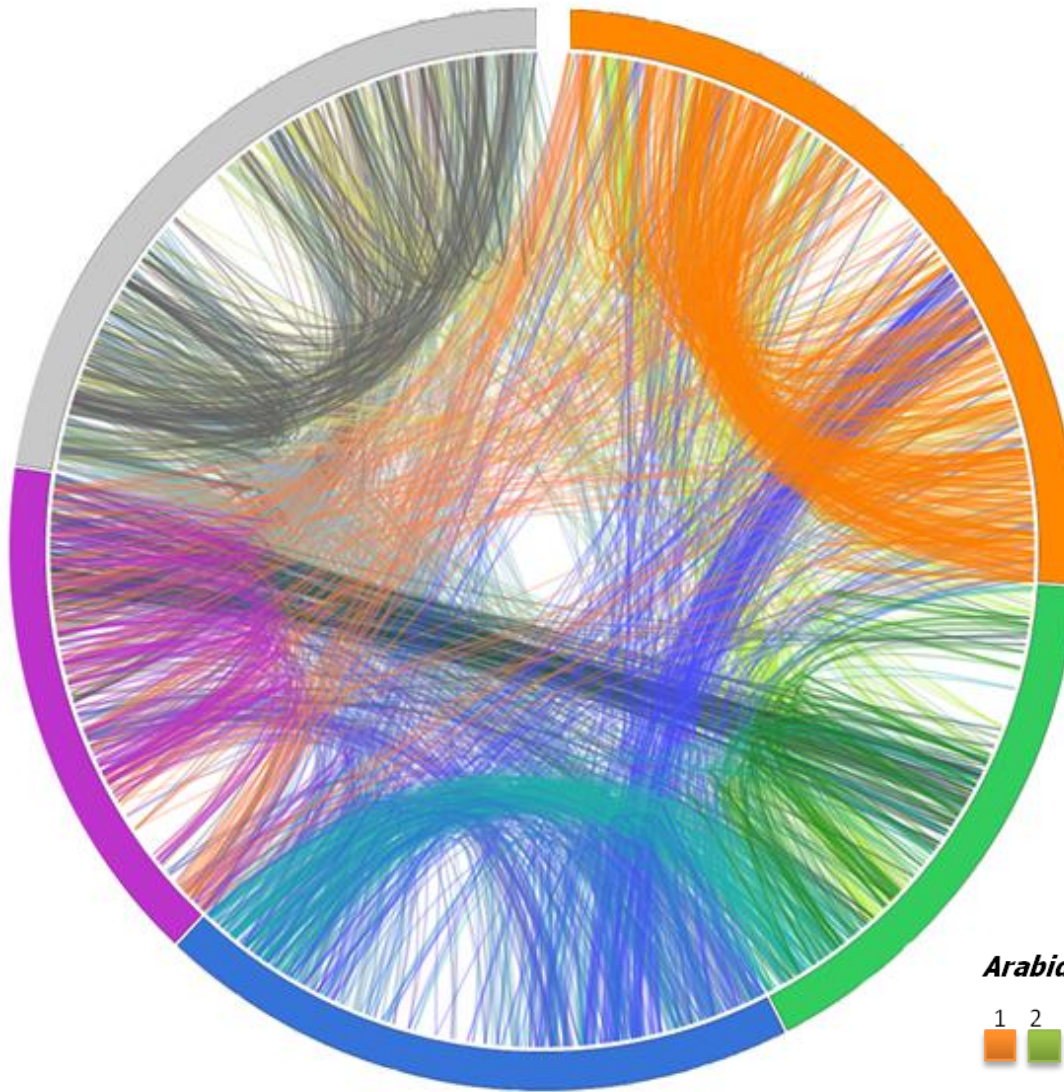


	GENE NUMBER	NET NUMBER	% NETS	% PROTEIN CODING GENES
	2	1347	44.65	10
	3 - 9	1370	45.41	23
	10 - 30	216	7.16	13
	31 - 207	83	2.75	15
	208 - 5168	1	0.03	19
TOTALS	21843	3017	100	80

RESULTS: 2-gene networks



RESULTS: 2-gene networks

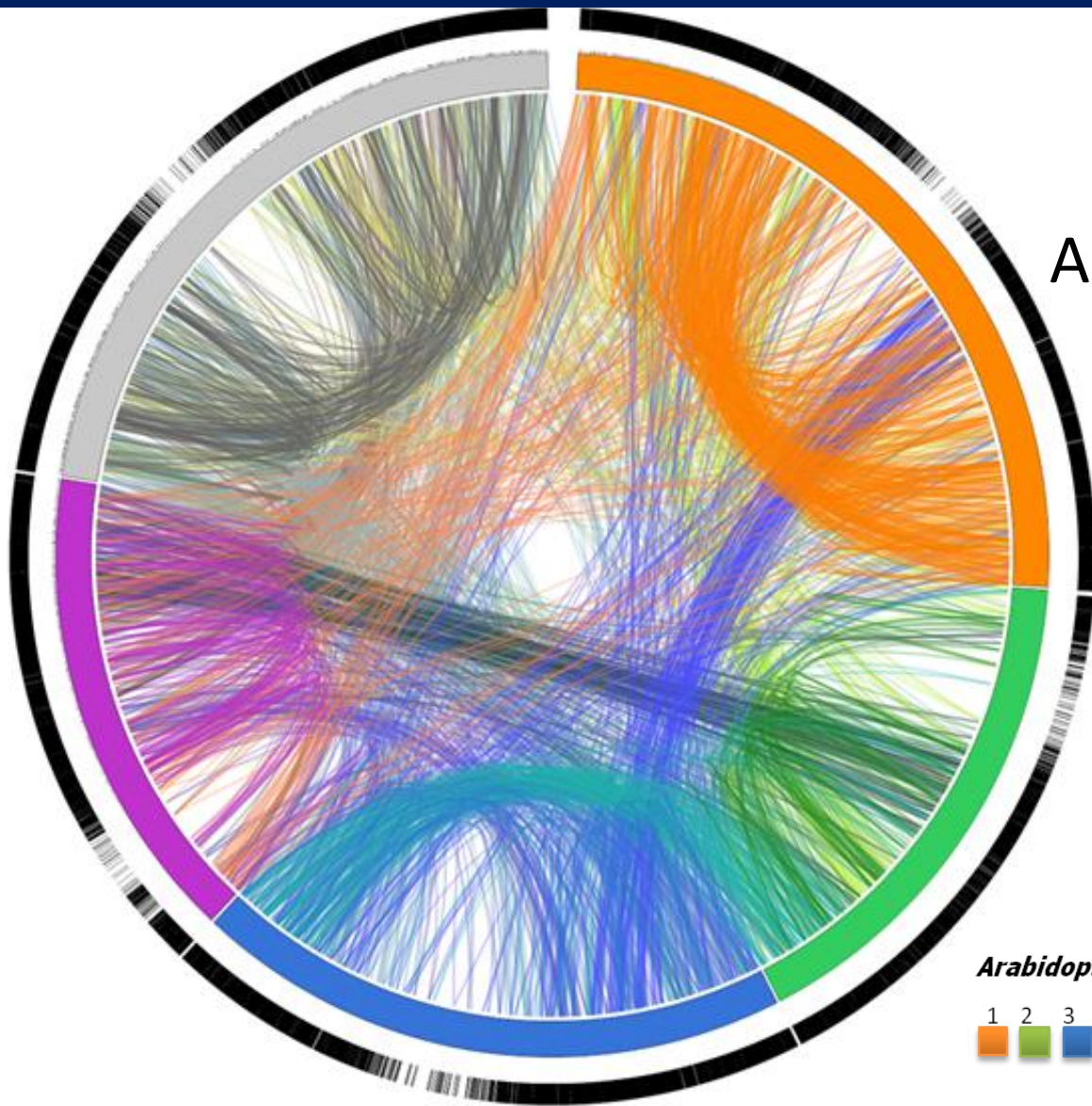


Arabidopsis thaliana
chromosomes

Arabidopsis chromosomes



RESULTS: 2-gene networks

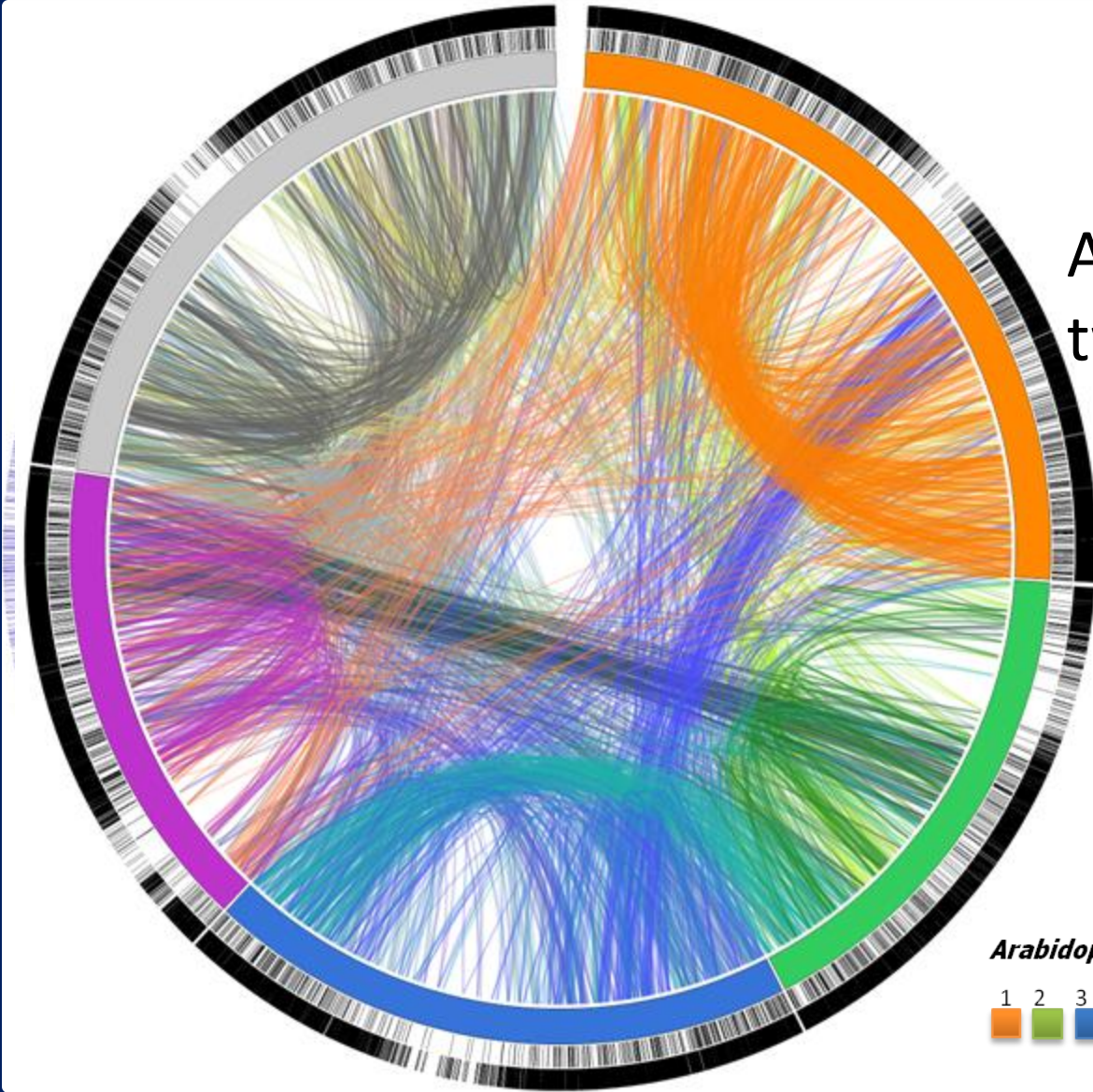


All protein-coding genes

Arabidopsis chromosomes



RESULTS: 2-gene networks



All genes involved in
two-gene networks

Arabidopsis chromosomes





RESULTS: 2-gene networks and singleton genes

AT LEAST 5% OF THE ENTIRE PROTEOME **HAS A SINGLE PARALOGY** RELATIONSHIP

AT LEAST 20% OF THE ENTIRE PROTEOME **IS A SINGLETON**



ABOUT **ONE QUARTER** OF THE ENTIRE PROTEOME HAS **ZERO OR AT MOST ONE** PARALOGY RELATIONSHIP

WHAT ABOUT SINGLETON GENES?

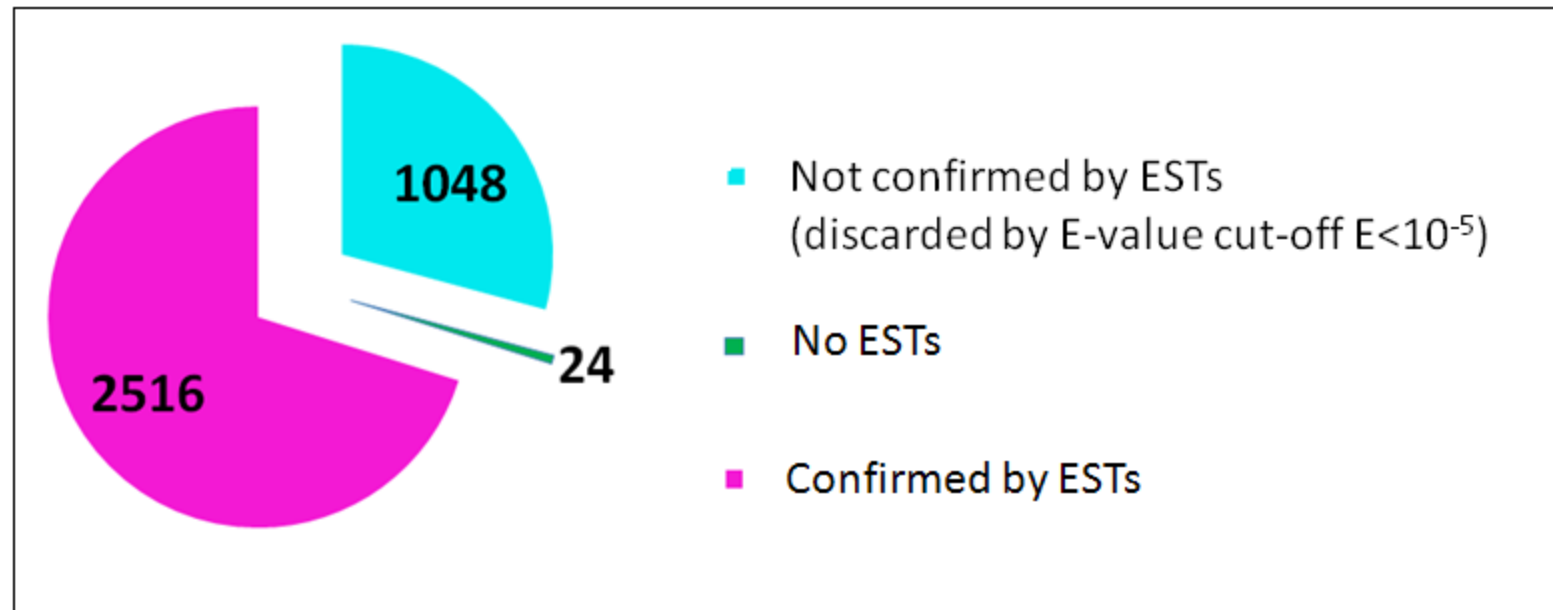
RESULTS: Singleton genes analysis

Singleton genes
not having protein-coding
paralogs

3588

Blastn mRNA
sequences against
ESTs

Singleton genes validated or
not validated by ESTs



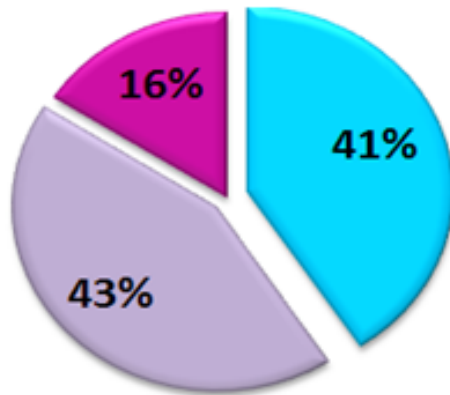
RESULTS: Singleton genes analysis

Singleton genes
not having protein-coding
paralogs

3588

**Comparative
analysis**
(Biomart Ensembl)

Singleton's orthologs in
different species (*O.sativa*,
V.vinifera, *S.bicolor*,
P.trichocarpa)

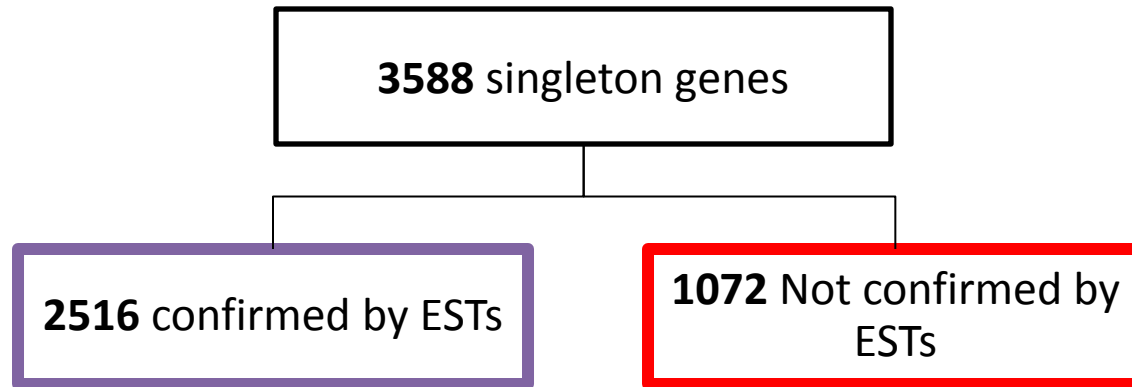


- All species
- Only in *A. thaliana*
- At least in one species

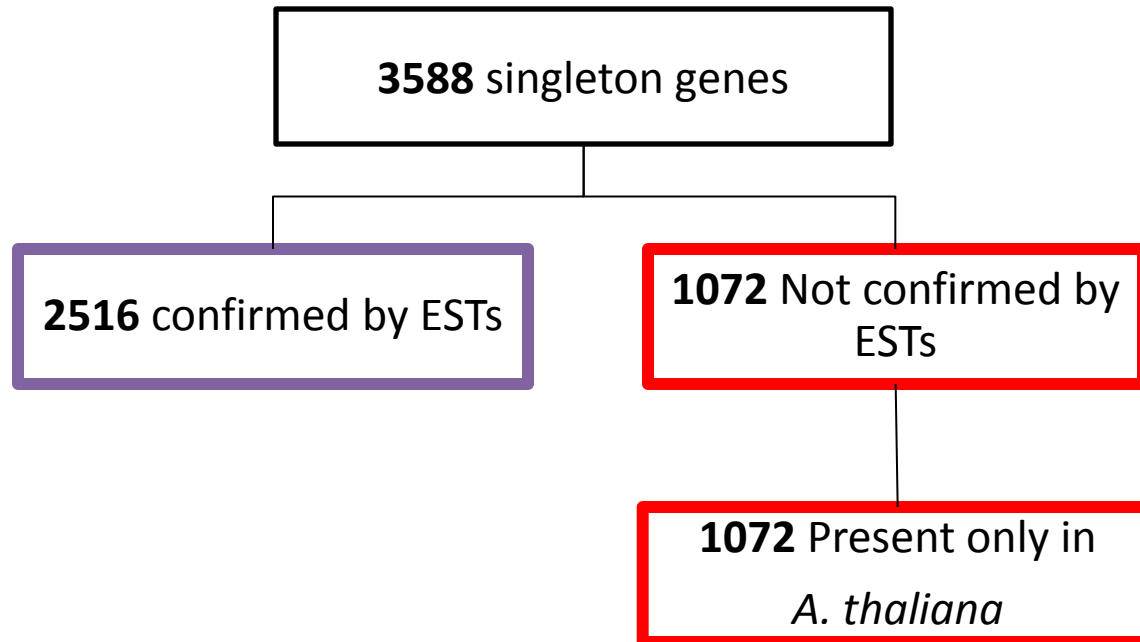
SINGLETON GENES ANALYSIS RESULTS

3588 singleton genes

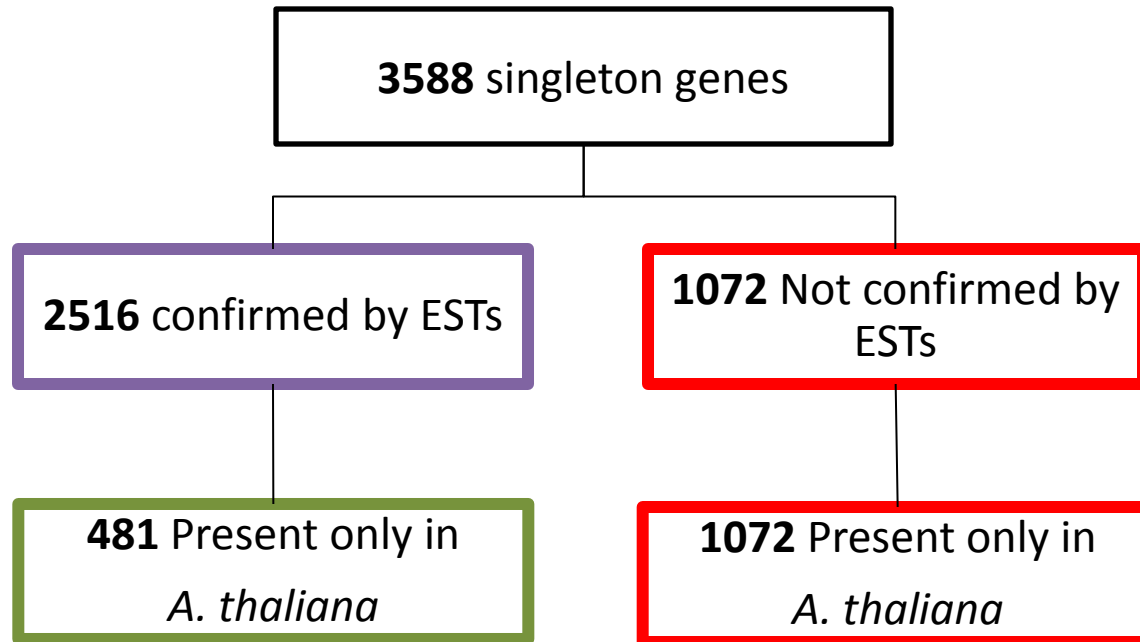
SINGLETON GENES ANALYSIS RESULTS



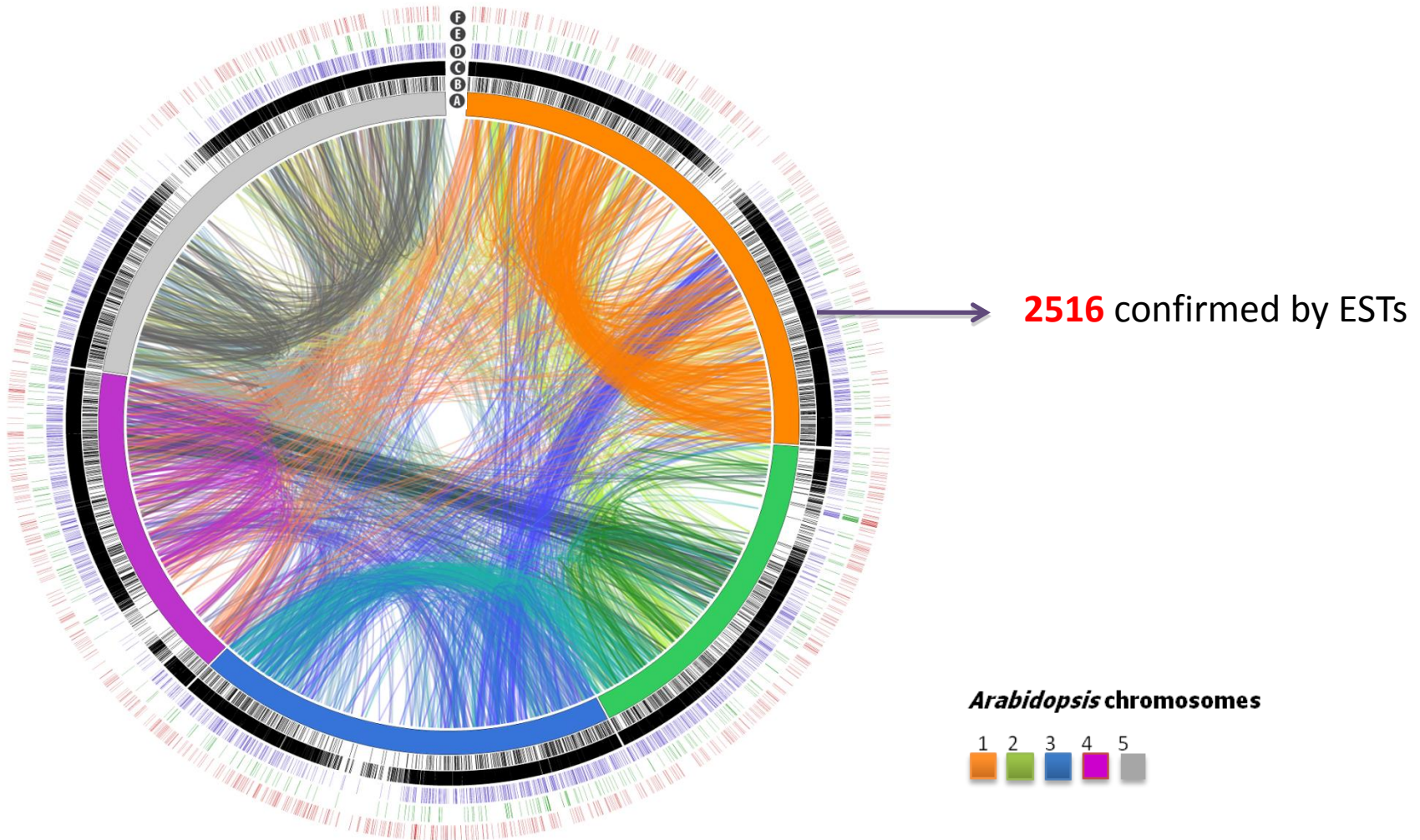
SINGLETON GENES ANALYSIS RESULTS



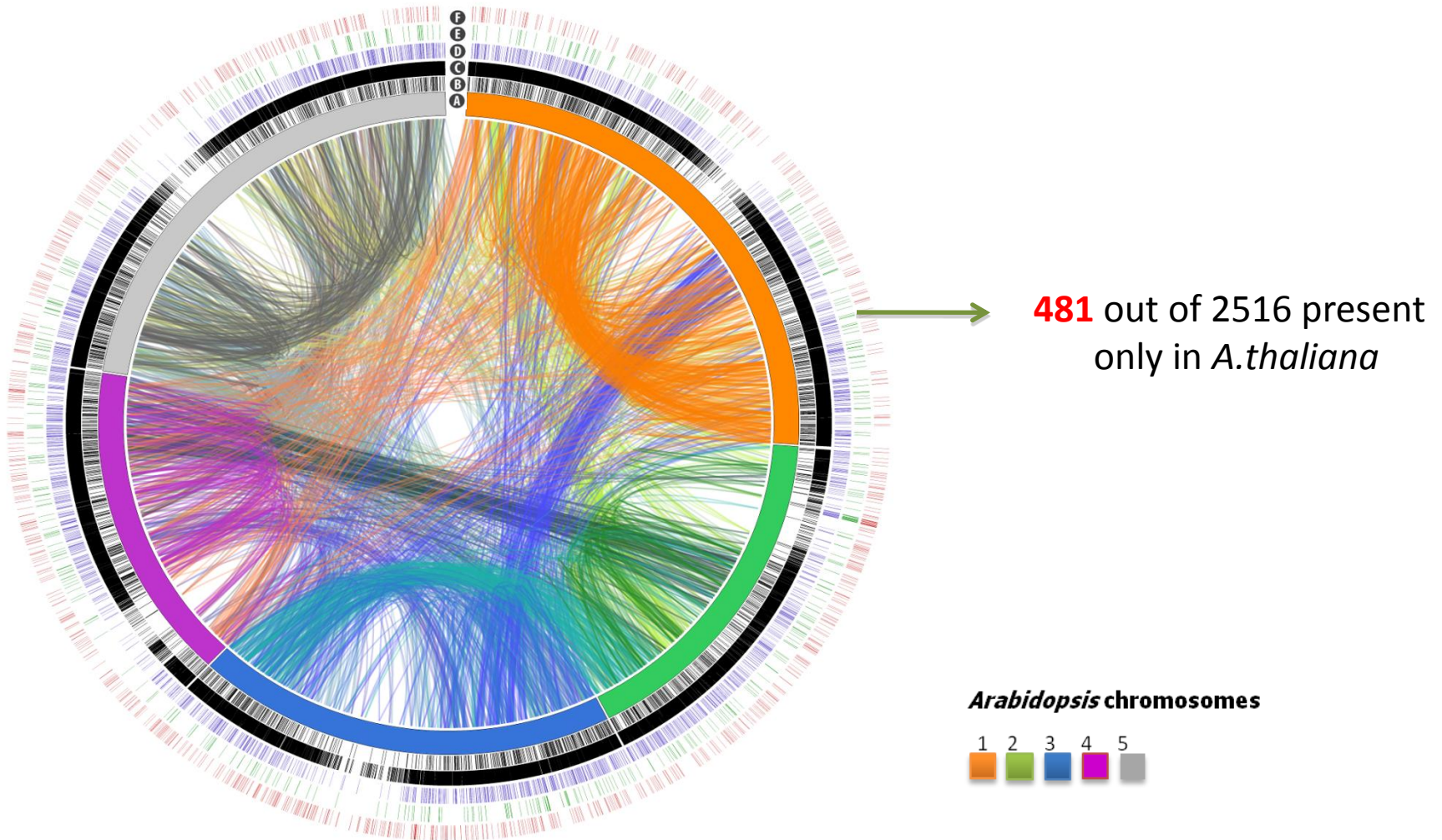
SINGLETON GENES ANALYSIS RESULTS



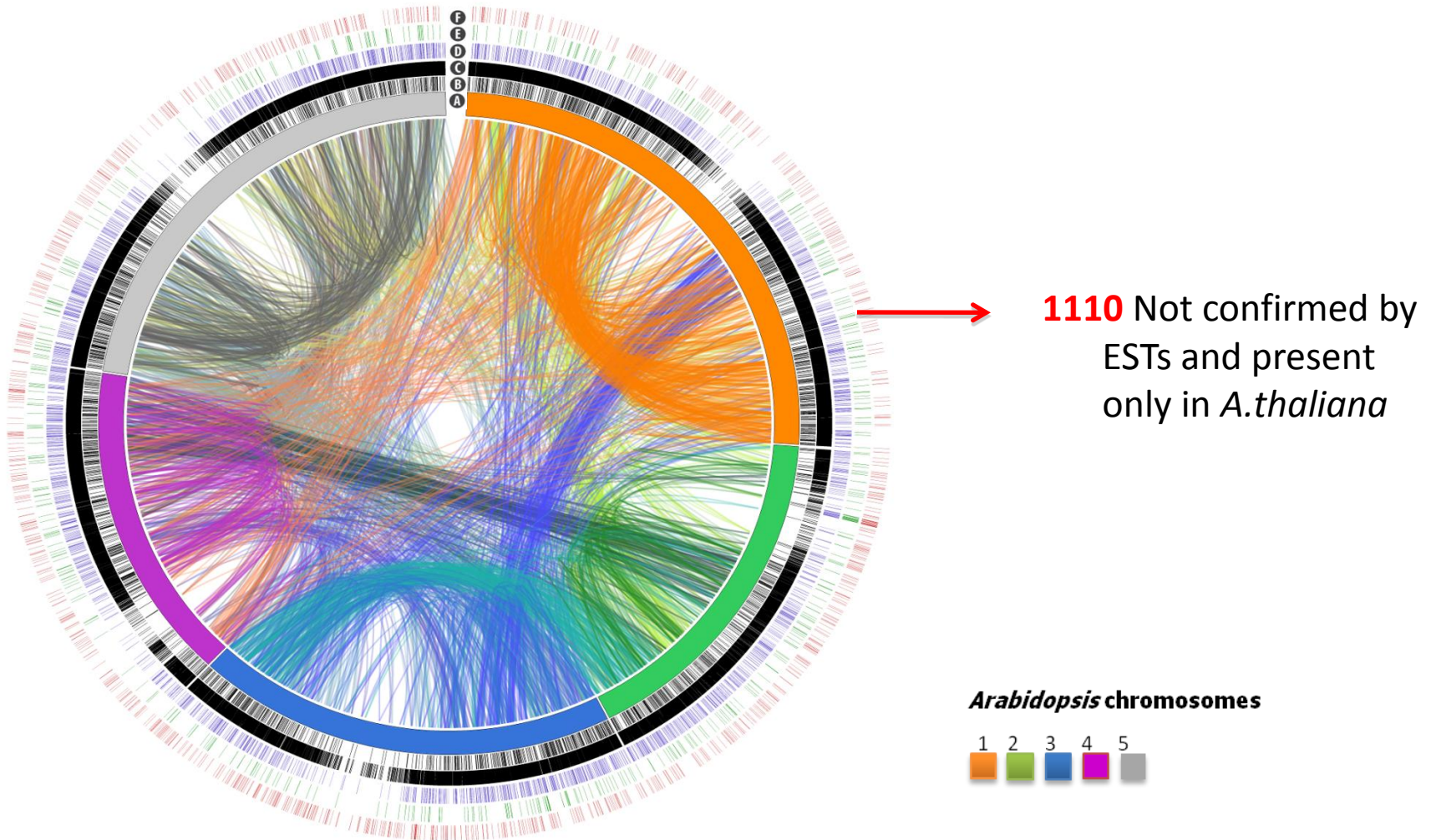
SINGLETON GENES ANALYSIS RESULTS



SINGLETON GENES ANALYSIS RESULTS



SINGLETON GENES ANALYSIS RESULTS



RESULTS: Singleton genes analysis

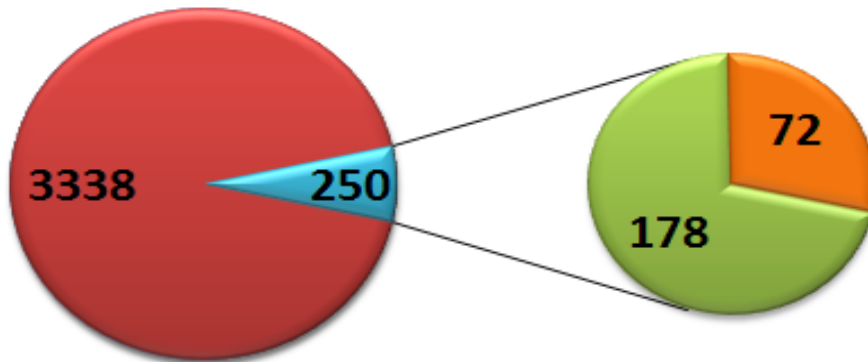
Singleton genes
not having protein-coding
paralogs

3588

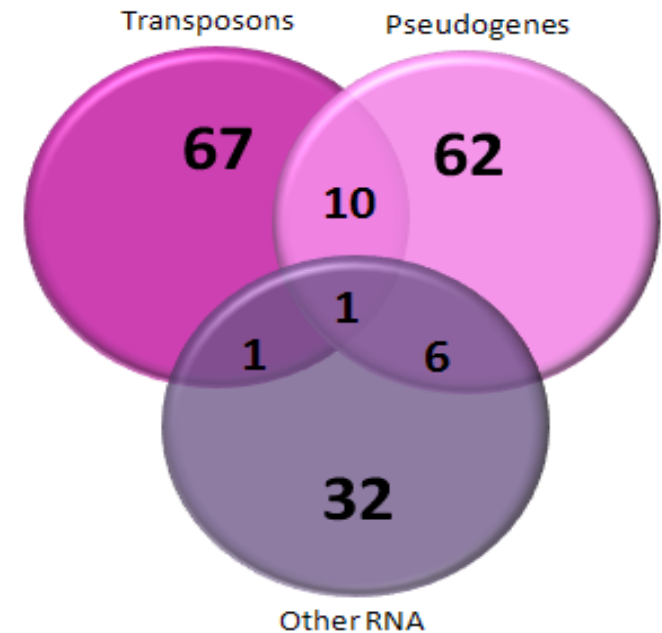
BLASTn vs:
Pseudogenes
Transposons
Other RNA
Intergenic regions

Singleton genes
not having protein-coding/
non protein-coding paralogs
and any significant hit in the
genome

250



- Significant hit
- No significant hit (discarded by E-value cut-off)
- Singleton genes with non protein-coding paralogs
- Matching the query within the intron



RESULTS: Singleton genes analysis

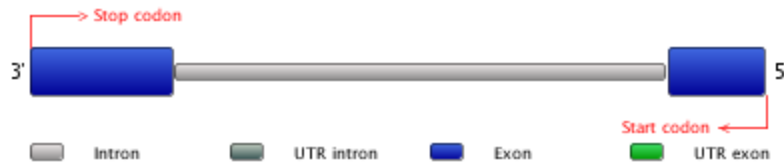
Singleton genes
not having protein-coding
paralogs

3588

Blastx protein
sequences against
transcript

Searching for ORF annotation
errors

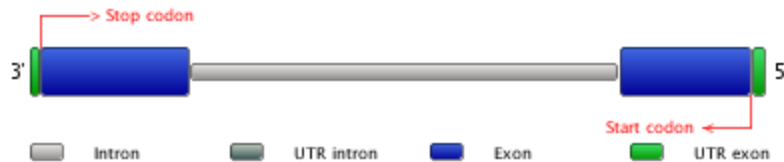
MT1A (METALLOTHIONEIN 1A)



Gene family: **metallothionein**

binds to and detoxifies excess copper and other metals, limiting oxidative damage.
copper ion binding / metal ion binding

MT1B (METALLOTHIONEIN 1B)



This work was helpful to highlight several crucial issues related to gene annotation

Networks of paralogs

- Are useful for the investigation of the gene annotation of reference and novel genomes
- Provide a useful tool for deeper investigations of gene families
- Reveal intriguing issues on the genome organization of *Arabidopsis thaliana*



Prof. Luigi Frusciante

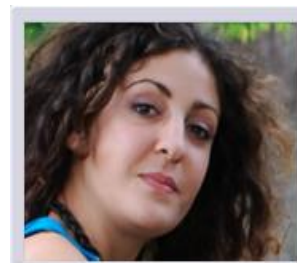
Progetto
Agronanotech



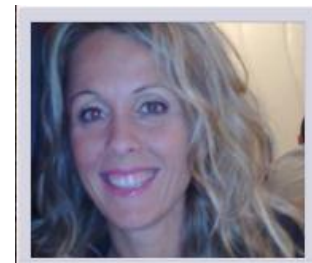
Maria Luisa Chiusano
PRINCIPAL INVESTIGATOR



Nunzio D'Agostino
BIOLOGIST



Alessandra Traini
PHYSICIST



Miriam Di Filippo
BIOLOGIST



Alessandra Vigilante
BIOLOGIST



Mara Sangiovanni
COMPUTER SCIENTIST



Chiara Colantuono
STUDENT



Mario Aversano
TECHNICAL ASSISTANT



THANK YOU FOR THE ATTENTION!