

Structural and Functional Annotations

From prediction to validation

2 case stories

200x Illumina + 35x PacBio + 600x Bionano

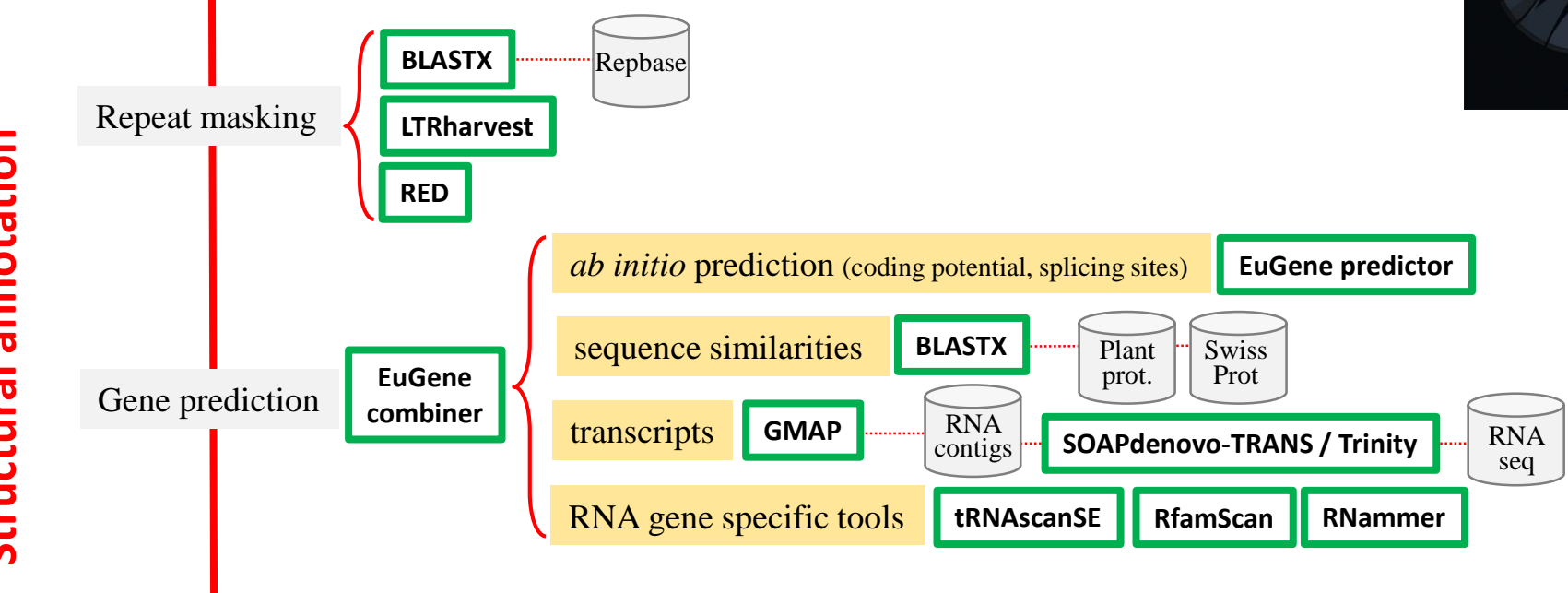
Malus domestica genome

Daccord et al., 2017



2150 contigs / 17 chr. (624,8 Mb)

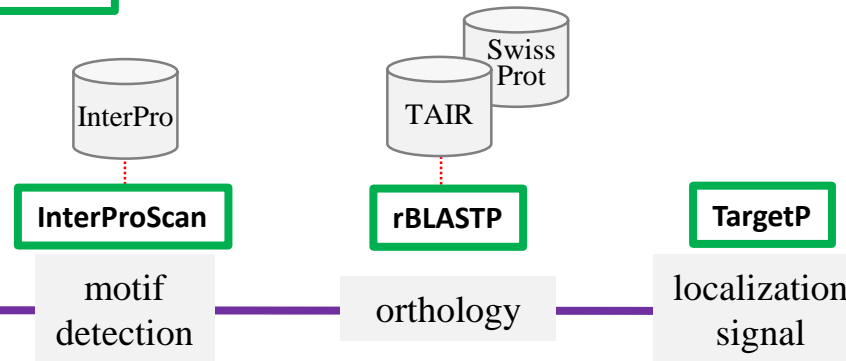
Structural annotation



TE detection/classification **REPET**

TE and gene models

Proteins



Inferred functions
Gene families
GO classification

Functional annotation

coding genes	45116
snoRNA	410
rRNA	4369
tRNA	654
miRNA	141
snRNA	181
other ncRNA	54
undefined ncRNA	1816
total genes	52741

quality ?

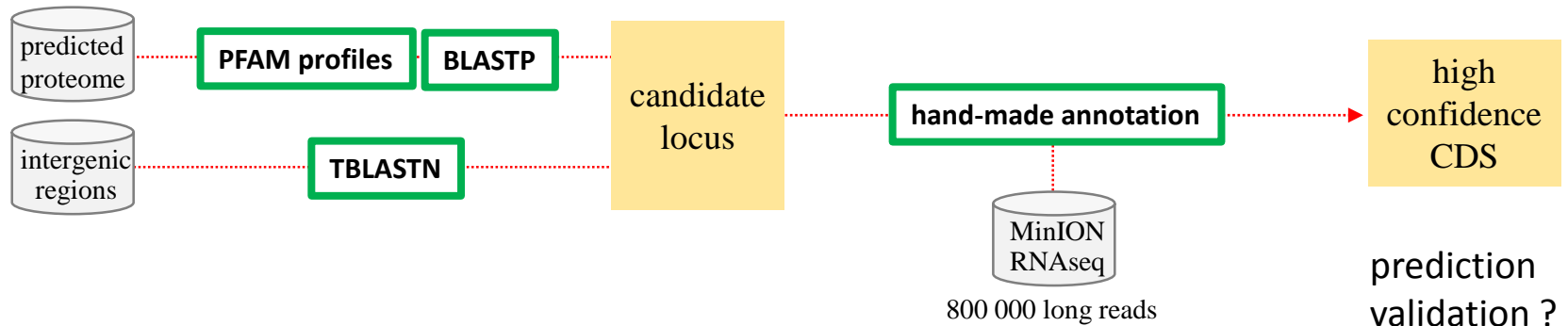
complete	96,8 %
fragmented	1,3 %
missing	1,9 %
transcript support	89,8 %
PFAM signature	74,8 %
TAIR or SwissProt homolog	89,6 %

BUSCO v3

U : 93,2 %

One step further...

Gene families of interest (defense, biocontrol...) tagged by at least one published plant member



50 gene families
1194 genes

Gene Families	gene nb	ok	site	ATG	exons	split	merg	underp	pseudo
Cysteine-rich secretory protein (PR-1)	20	18	0	1	1	0	0	0	5
Glycosyl hydrolase family 17 (PR-2)	76	58	0	7	5	4	2	0	15
Barwin domain protein (PR-4)	3	3	0	0	0	0	0	0	3
Thaumatococin (PR-5)	35	25	1	2	1	1	5	0	6
Chitinase A/Glycosyl hydrolase family 18 (PR-8)	23	18	0	1	1	0	3	0	7
Protease inhibitor/seed storage/LTP family (PR-14)	27	26	0	0	0	0	0	1	4
Cupin/germin/oxalate oxidase (PR-15,PR-16)	33	29	1	1	2	0	0	0	16
Polyphenol oxidase (PPO)	12	11	0	0	1	0	0	0	6
Hydroxymethylglutaryl-CoA reductase (HMGR)	4	4	0	0	0	0	0	0	11
Farnesyl diphosphate synthase (FDS)	4	2	0	0	2	0	0	0	0
Terpene Synthase (TPS)	32	21	2	1	7	1	0	0	20
Pyridoxal phosphate (PLP)-dependent transferase/Alliinase/CSL	8	8	0	0	0	0	0	0	1
Ascorbate Peroxidase (APOX)	9	5	0	4	0	0	0	0	3
Glutathione S-transferase (GST)	50	42	0	1	2	2	3	0	12
Glucan synthase/Callose synthase	14	7	3	1	1	2	0	0	4
Pectinesterase	70	65	1	1	2	1	0	0	22
Cinnamoyl-CoA reductase/Epimerase (CAD)	55	43	4	2	3	2	1	0	21
Tubulin alpha	10	9	0	0	0	1	0	0	1
Actin	11	2	0	9	0	0	0	0	1

CDS ok	site	ATG	exons	split	merge	underpred	pseudo
753	18	55	54	23	20	32	239
78,8 %	1,9 %	5,8 %	5,7 %	2,4 %	2,1 %	3,4 %	20 %

955 functional genes

Pollen Oleoylef desaturase	31	28	0	1	2	0	0	0	8
Pectate lyase C	25	24	0	0	1	0	0	0	5
NPR/NIM family	9	7	1	0	0	1	0	0	2
Salicylic acid carboxyl methyltransferase (SAM)	32	24	1	5	2	0	0	0	13
Wall-associated kinase (WAK)	14	11	0	1	1	1	0	0	8
Agglutinine	24	3	0	0	0	0	0	21	3
Translational Elongation Factor 1-alpha	7	4	0	2	1	0	0	0	0
Genes with transcription affected by <i>Venturia inaequalis</i>	10	8	0	1	1	0	0	0	nd
Genes with transcription affected by <i>Erwinia amylovora</i>	6	6	0	0	0	0	0	0	nd
Aluminium activated malate transporter (ALMT)	25	22	1	1	0	1	0	0	1
Acetoacetyl-CoA thiolase (AACT)	3	2	0	1	0	0	0	0	0
HMG-CoA synthase (HMGS)	2	2	0	0	0	0	0	0	0
Mevalonate kinase (MVK)	2	2	0	0	0	0	0	0	0
Phosphomevalonate kinase (PMK)	2	2	0	0	0	0	0	0	0
Mevalonate diphosphate decarboxylase (MVD)	2	2	0	0	0	0	0	0	0
Isopentenyl diphosphate isomerase (IDI)	2	2	0	0	0	0	0	0	0
Geranyl diphosphate synthase (GDS)	6	5	0	0	1	0	0	0	2



Analysis of CATMA transcriptome data identifies hundreds of novel functional genes and improves gene models in the Arabidopsis genome

[Sébastien Aubourg](#),^{✉1} [Marie-Laure Martin-Magniette](#),^{1,2} [Véronique Brunaud](#),¹ [Ludivine Taconnat](#),¹ [Frédérique Bitton](#),¹ [Sandrine Balzerque](#),¹ [Pauline E Jullien](#),³ [Mathieu Ingouff](#),³ [Vincent Thureau](#),⁴ [Thomas Schiex](#),⁵ [Alain Lechamy](#),^{1,4} and [Jean-Pierre Renou](#)^{✉1}

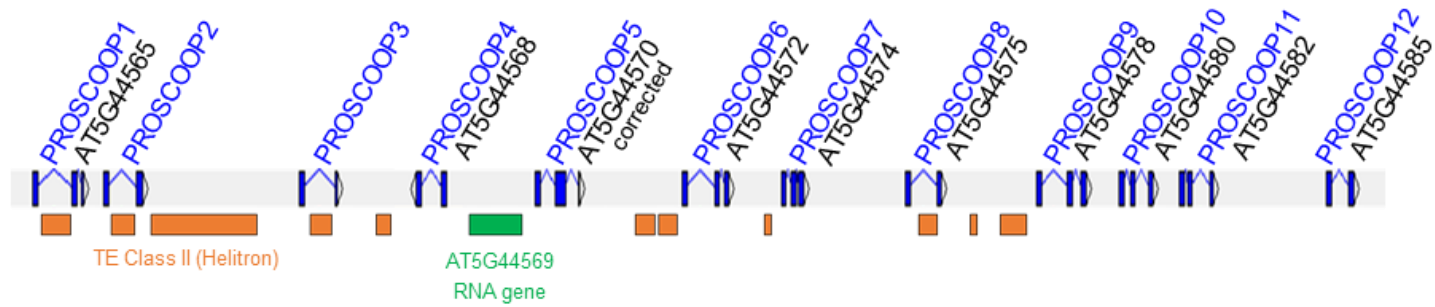
[BMC Genomics](#). 2007; 8: 401.

Focus on a small gene of **unknown function**, strongly induced by the necrogenic bacteria *Erwinia amylovora*

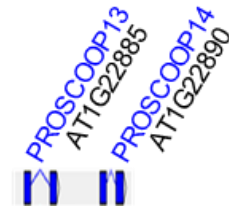


Fireblight on *Rosaceae*

Chromosome 5



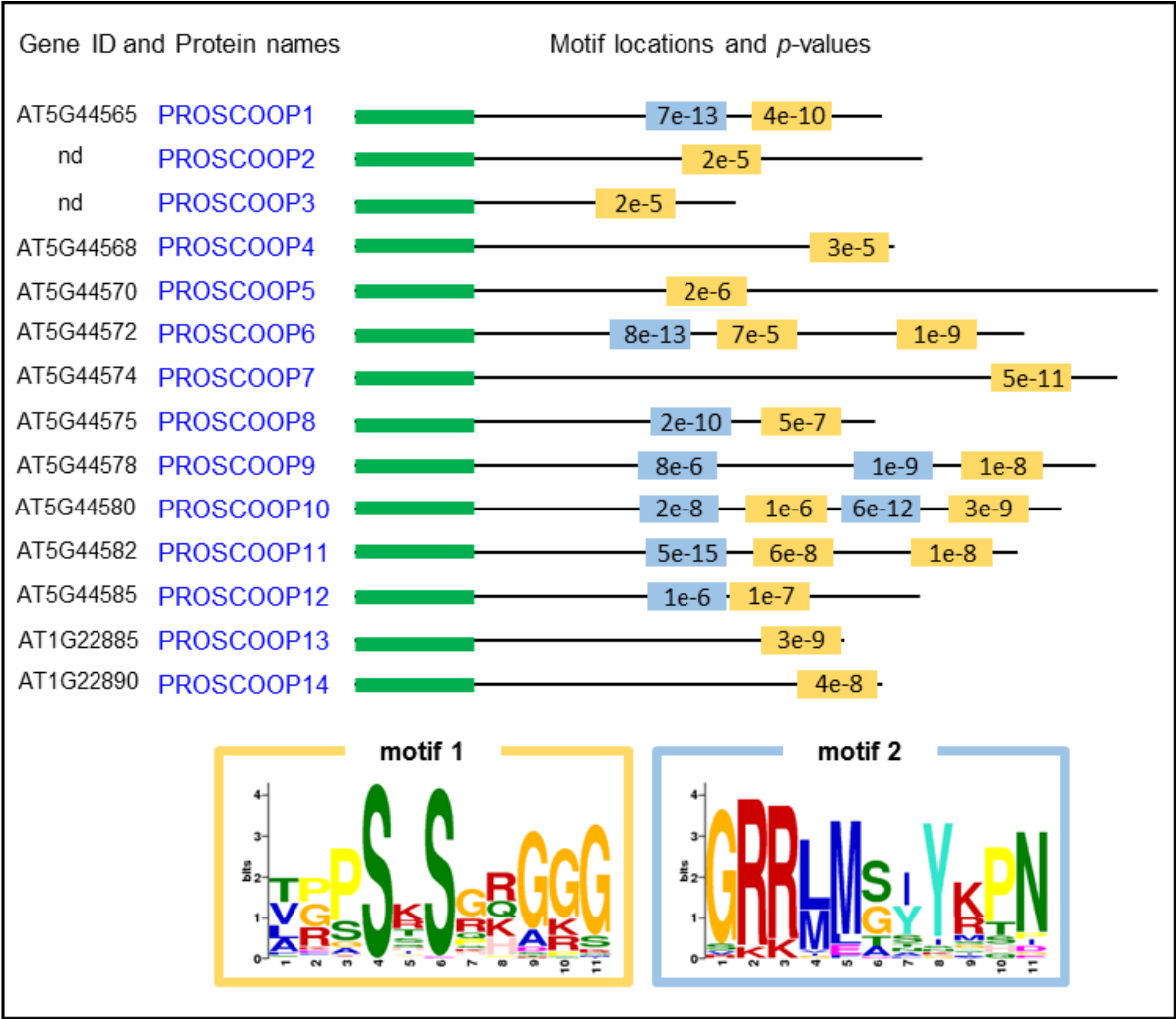
Chromosome 1



AT5G44572	-MGIKCYKRLRQVVVLVLLLVFVPCSLQSAE---SSVIG--RRRMGI-----YPMNGGLIAGSSPSGQAENINNNYHGRRIMISEA-----RFSKSKKGGGREPESPG--
AT5G44574	-MGRKCSKFRQMLVLLVLLLVVFTCL--LSTVAELSVIAKPTSF-KVKARGI-----EDEGQERTHSLNSKKSSRSVEKTHHSEGRRLSNVFNAGIRAGPSKSGQGGGRIPVAAS-----
AT5G44565	-MGMSGSSGLVHVMMLLLLSILFHHTEPTLPSDSEQLSITG--RRMGMN-----YKENTAVET-----PPSRRRGGGGQNTGAD-----
AT5G44578	-MENIFFKRLTQVFIALLC-IFIYRTESAMSSHREQLSLG--RRMGMNIE---Y-PYGGIYVKPPFLKSKDSNQKGRGETYYKENSEYIGT--GPSSHGSGSSIEHVSSE--
AT5G44580	MERKIFSSKFIHLLIVFLLCTFLSRTEALFY-HHELFL-G--RKRMYN-----YKNSAIGTSSSTSD---HAPGNGRKLMSIYRNGDIFT--GPSSHGSGGRTPAF--
AT5G44582	MENNTFSSKGINLLILLLLCTFLCQTESALFS-HQELVITG--RRMMSY-----YKENTDIGTSSSD---RGGGNGRRLMSQMDI--GASSGQGGGRNRH--
AT5G44585	-MRNTISSKMGQVLIVLLLLCTVLCRTESALFSGQHSVLLTG--RRLM-G---SGASGFVRSIQ-----SSQAGGRFNDADEIADYGRK--
AT5G44575	-MEEISASQLG-LVVFLLFLVETHIESLPSQQESLPWIGSSRRLM--STYKTNSEIDFE-----SSISGQAGGGQQNF--
new1	-MAN---NLGLVILLVIVLVSC-SKSSDCALASEQKS-RESEWRRLIIVRSRSRSESEFABKKPPPPPPSEPLSESSSEPSN-----
new2	-MTRKGFNLRLLLLLLVLLPSCS--NCALTSQELRESSE-WRRKM-ITVNSKESY-----
AT5G44568	-MATATS-NLGHLLLSLFILLEIL-SQVGAQAKR-LQQRNK-LRLDCVF--LPPPPFLRGIYKPPFIAS-----FHSASPKDRGF-----
AT1G22890	-MAANTS-NLVALLLSLFLLLSIS-SQVGLGEAKR-NLRNN-LRLDCVS--HPPPPPHRSMA-PPIF-----VPPSTSHKGGQGF-----
AT1G22885	-MATRAS-NLVVFLLSLFLLLIS-FQVGVADATRNKRQGE-QSVDYD--Y-PRPPEAPIYL-----PPSRKRRKRGF-----
AT5G44570	-MATATS-NEVSLRVSIFILLFIS-SQVAIDAKHLQQRRK-LGIIVRRSRSGRQYNEPFLTRVPPPPPPFLPQMSAATPPMPQLSPLQEFMHVSSLQEFQMLYPPPSLFPVASSPTST--

MEME →

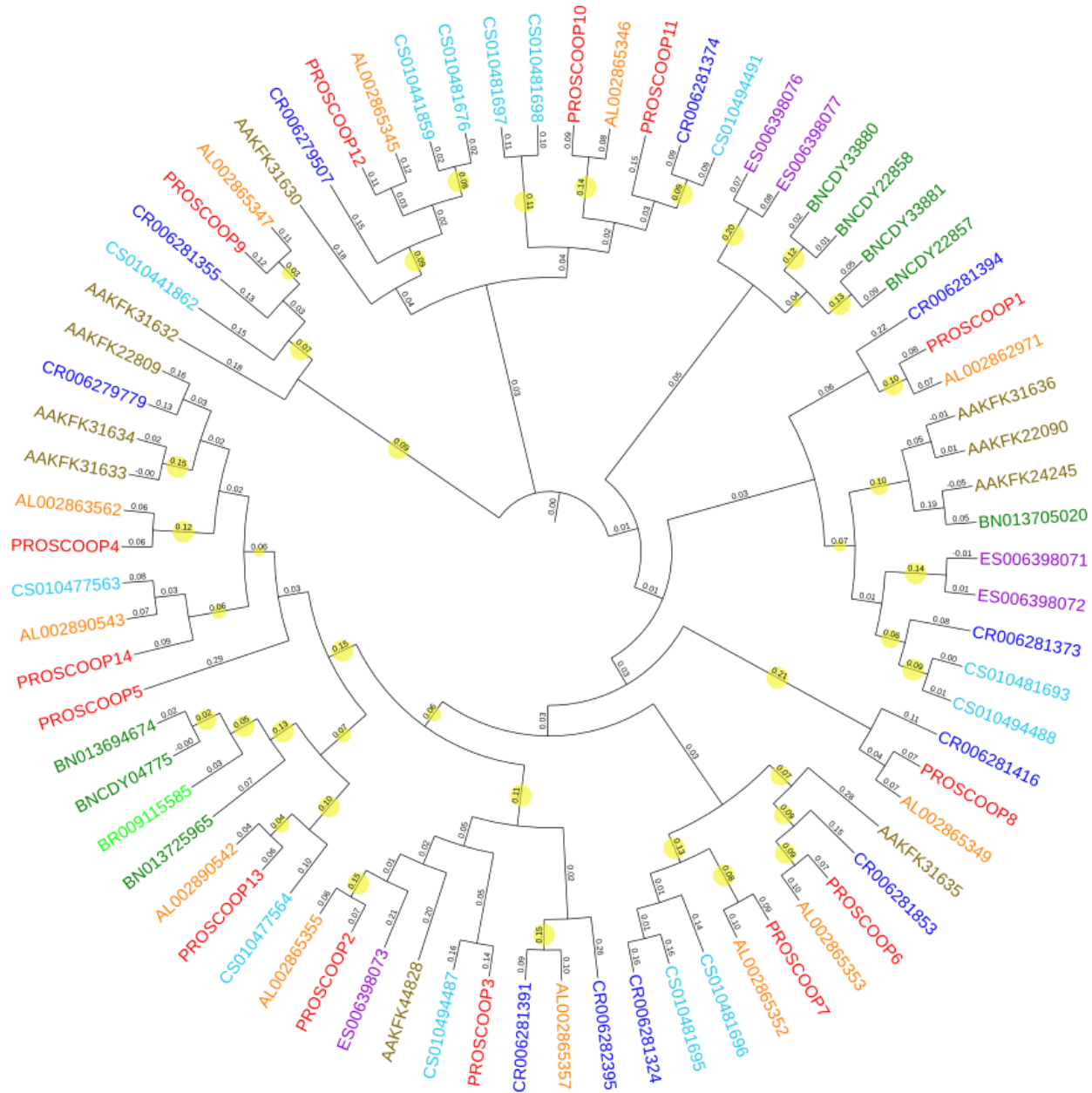
Motif searching



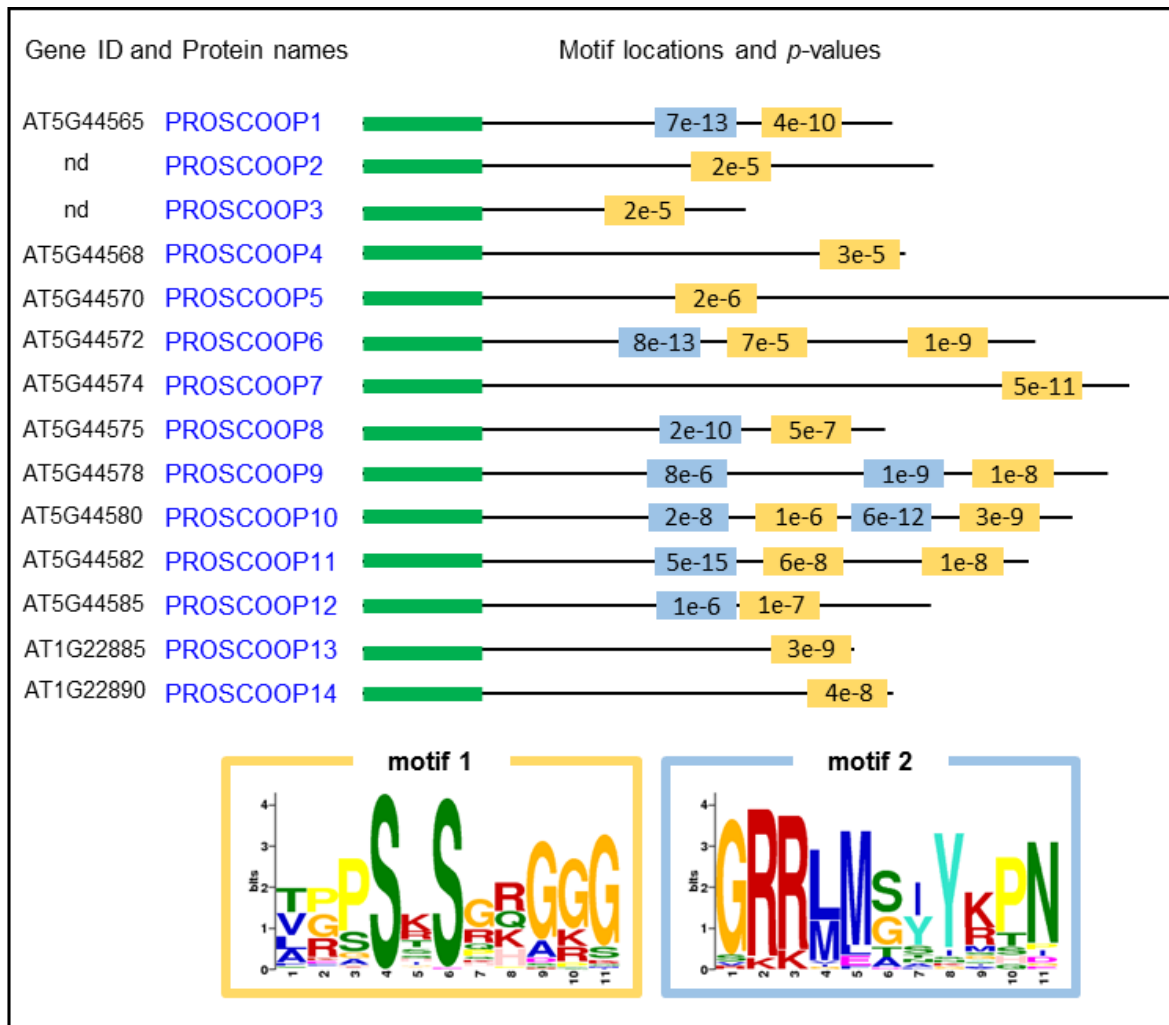
→ **MAST**

Uniprot screening

Arabidopsis thaliana
Arabidopsis lyrata
Arabis alpina
Capsella rubella
Camelina sativa
Eutrema salsugineum
Brassica napus
Brassica rapa



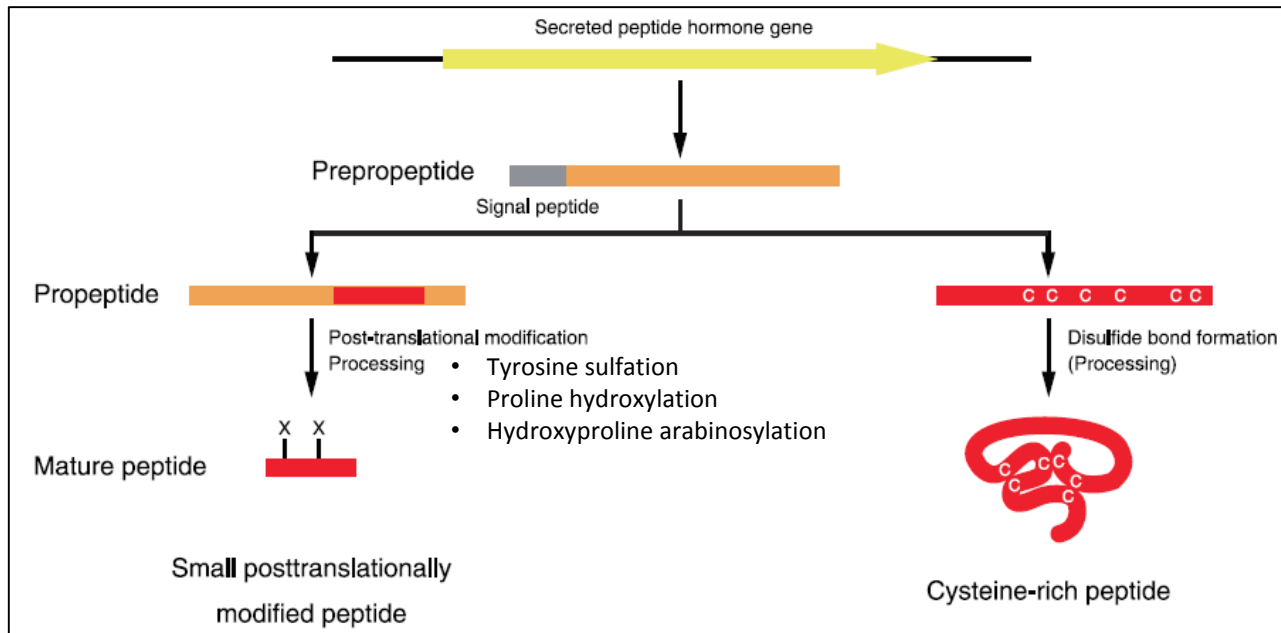
Brassicaceae specific



72 sites
(e-value of 9.8e-213)

39 sites
(e-value of 3.4e-179)

SCOOP motif Serine riCh endOGenOus Peptide



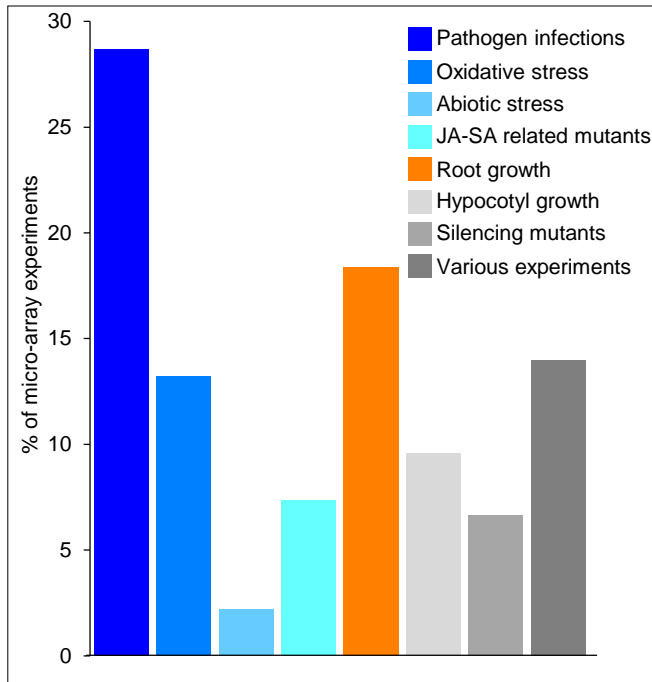
from Matsubayashi, 2011

Structural features of small post-translationally modified secreted peptides in plants :

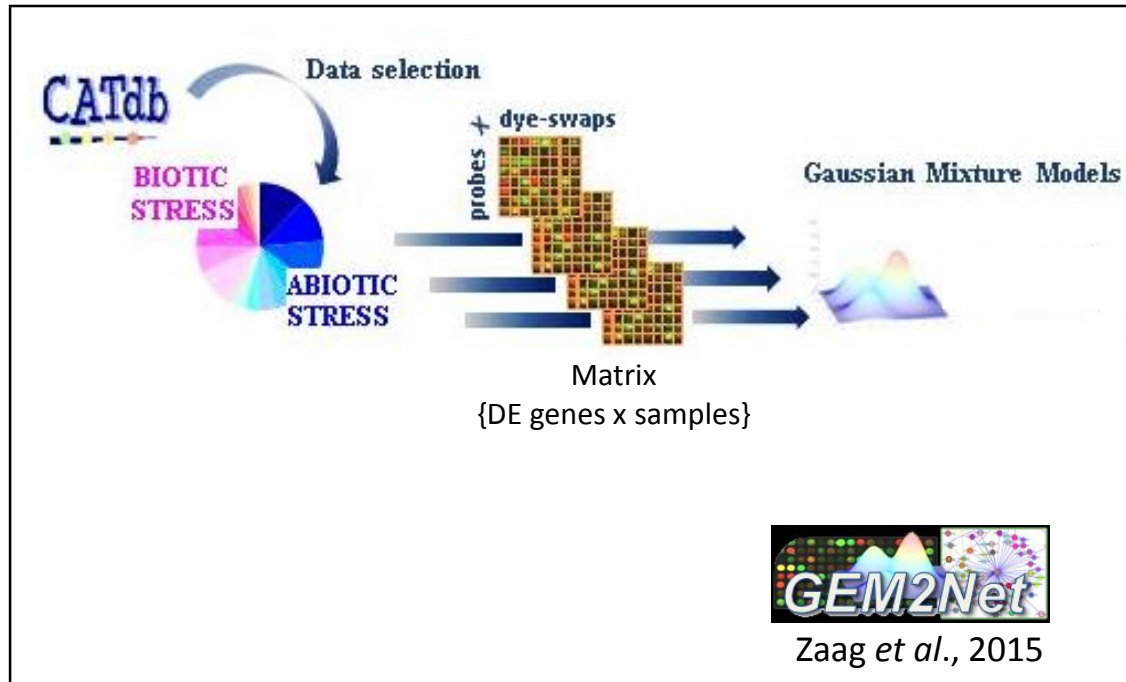
- ✓ Multiple paralogous genes (from 3 to 30)
- ✓ Encode 70 to 120 aa pre-propeptides
- ✓ Contain N-terminal secretion signal (RE)
- ✓ No or few Cysteine residues
- ✗ Sequence diversity excepted in the C-terminus corresponding to 5 to 15 aa mature peptide

Transcriptome mining

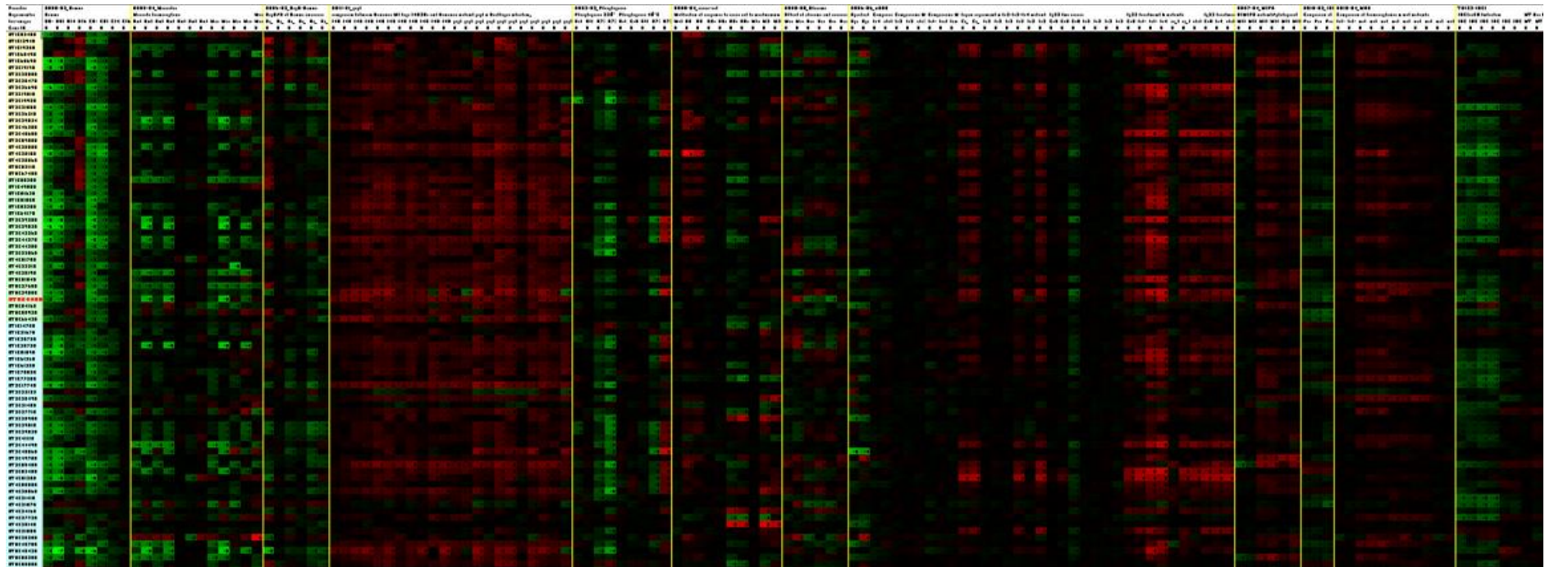
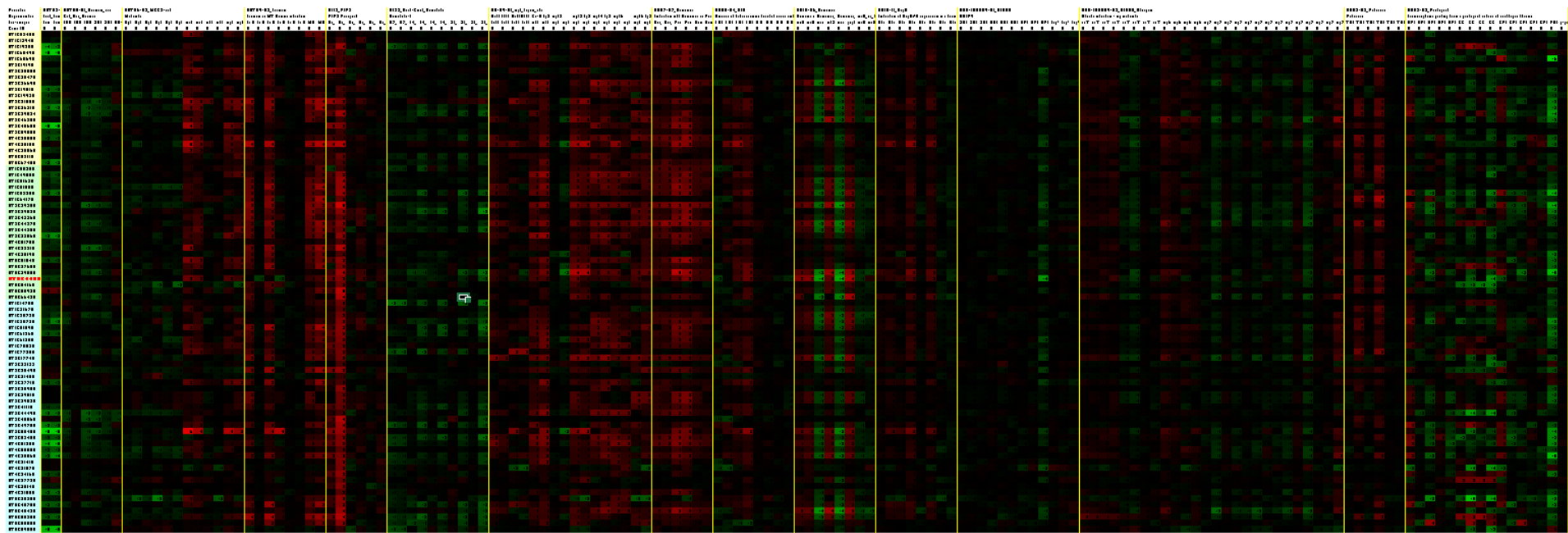
Differential expression - Metadata

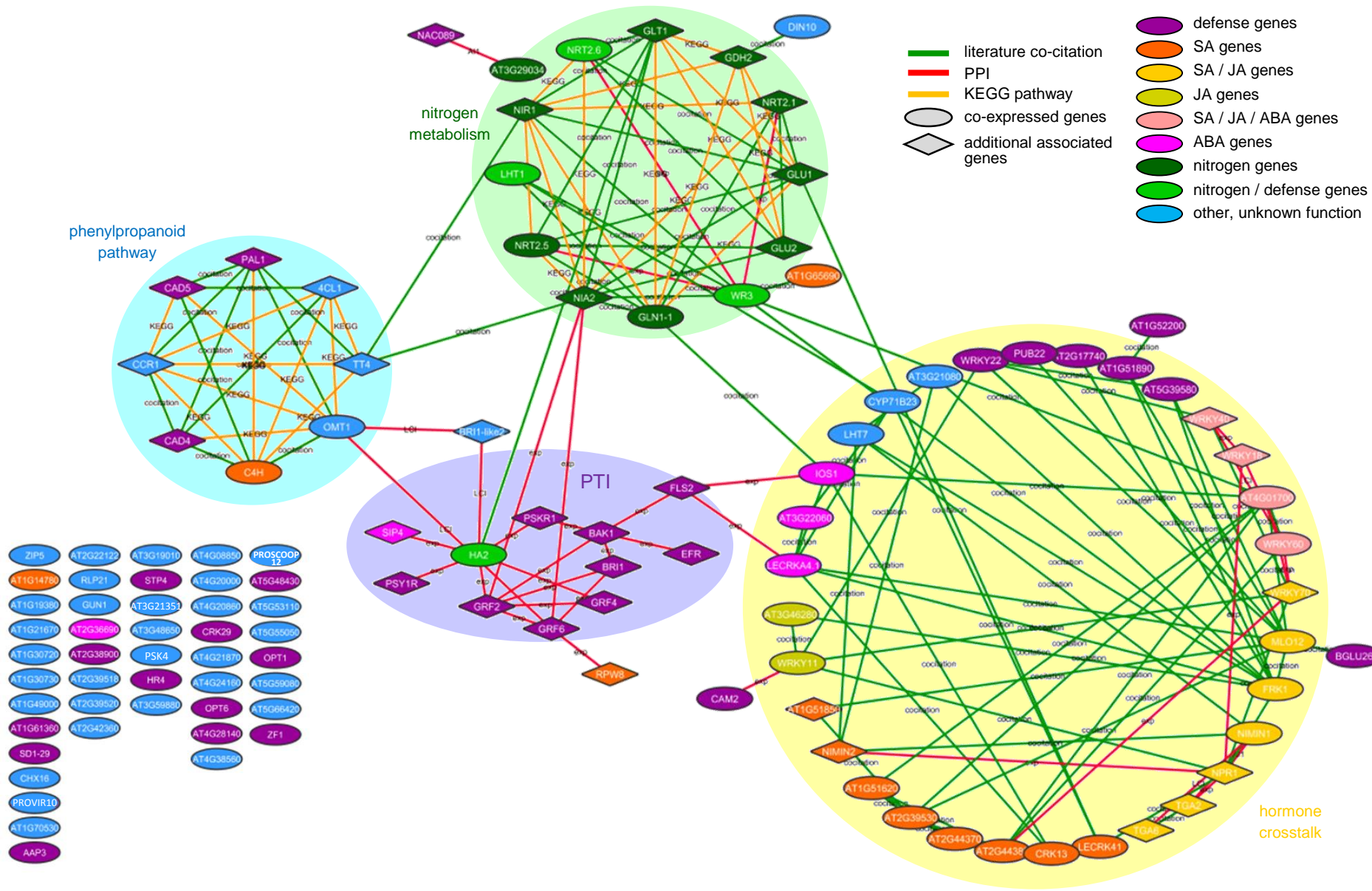


Clustering



83 genes co-regulated with *PROSCOOP12* (response to necrotrophic bacteria and fungi)

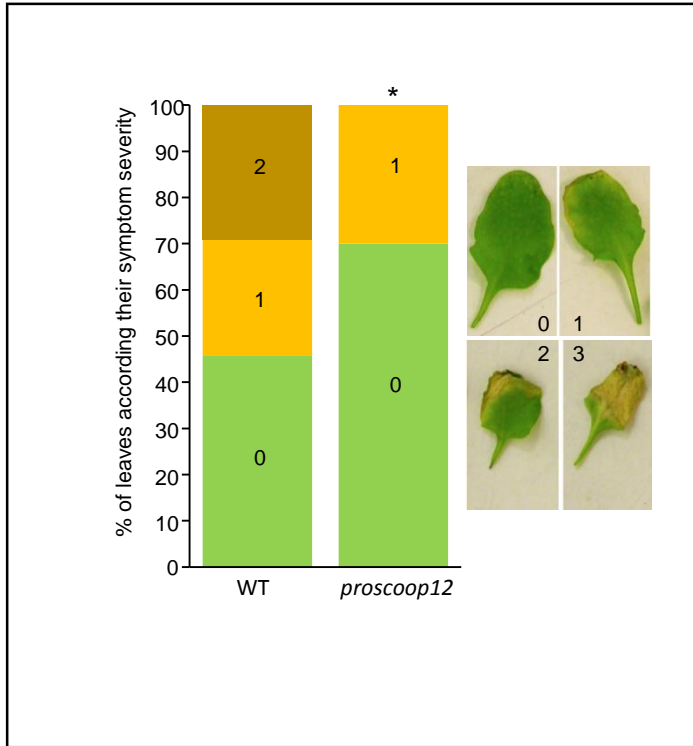




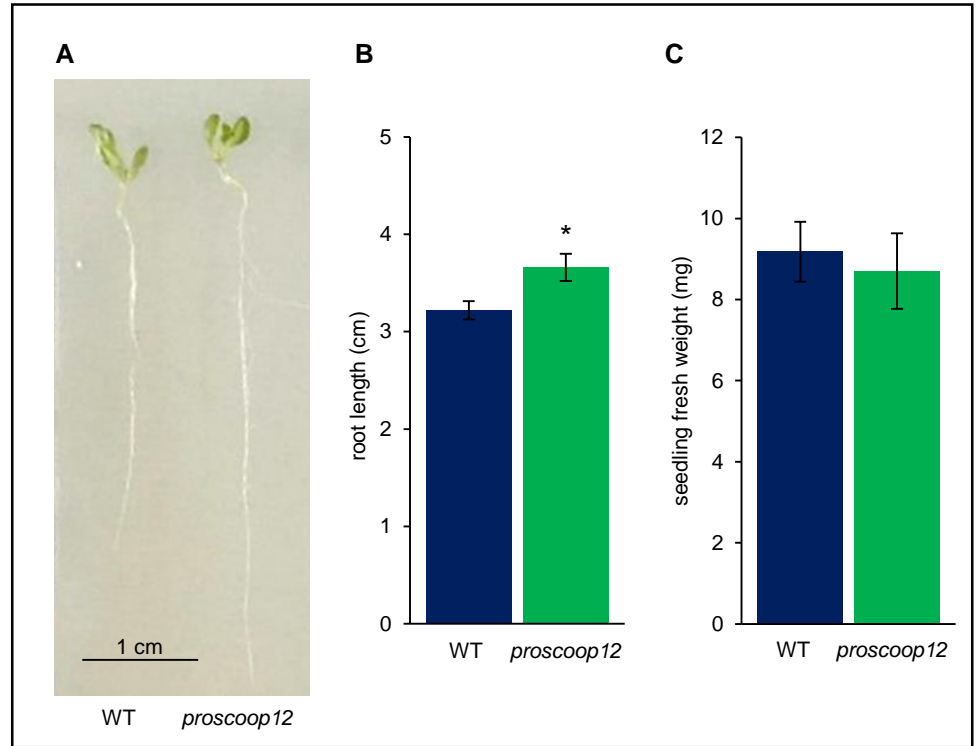
Relational annotation of *PROSCOOP12*
117 genes

Validation 1 : mutant phenotyping (T-DNA and CRISPR-Cas9 lines)

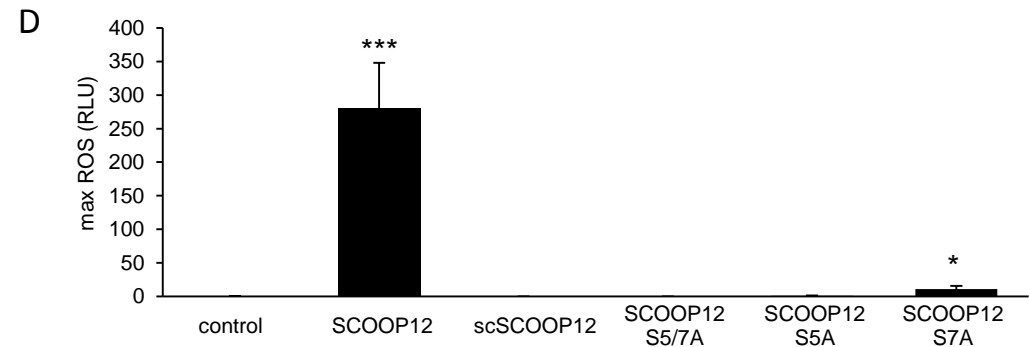
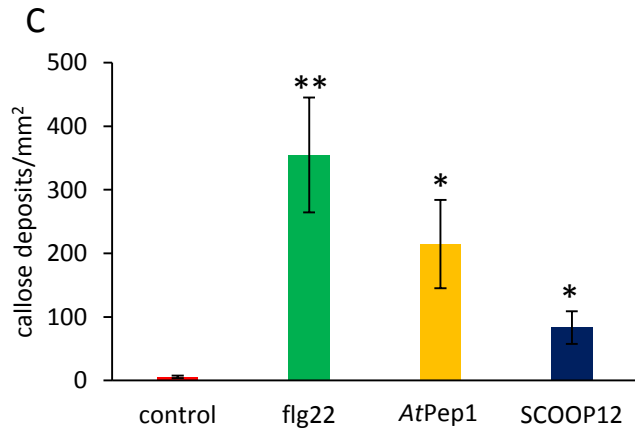
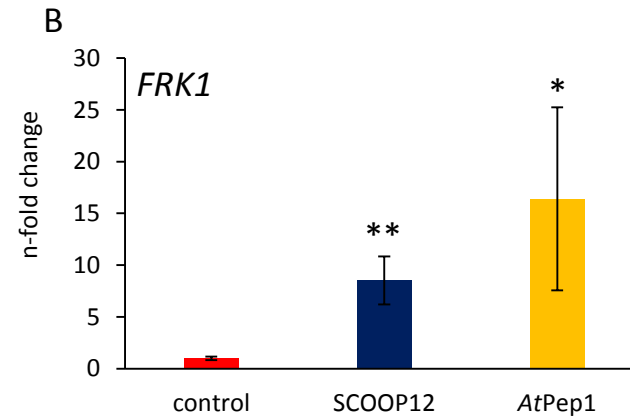
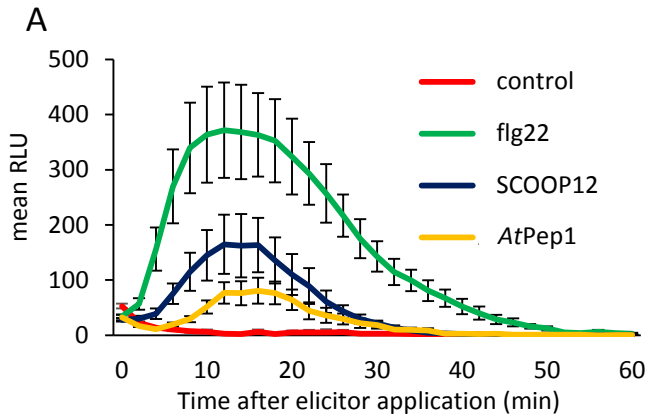
Erwinia amylovora infection



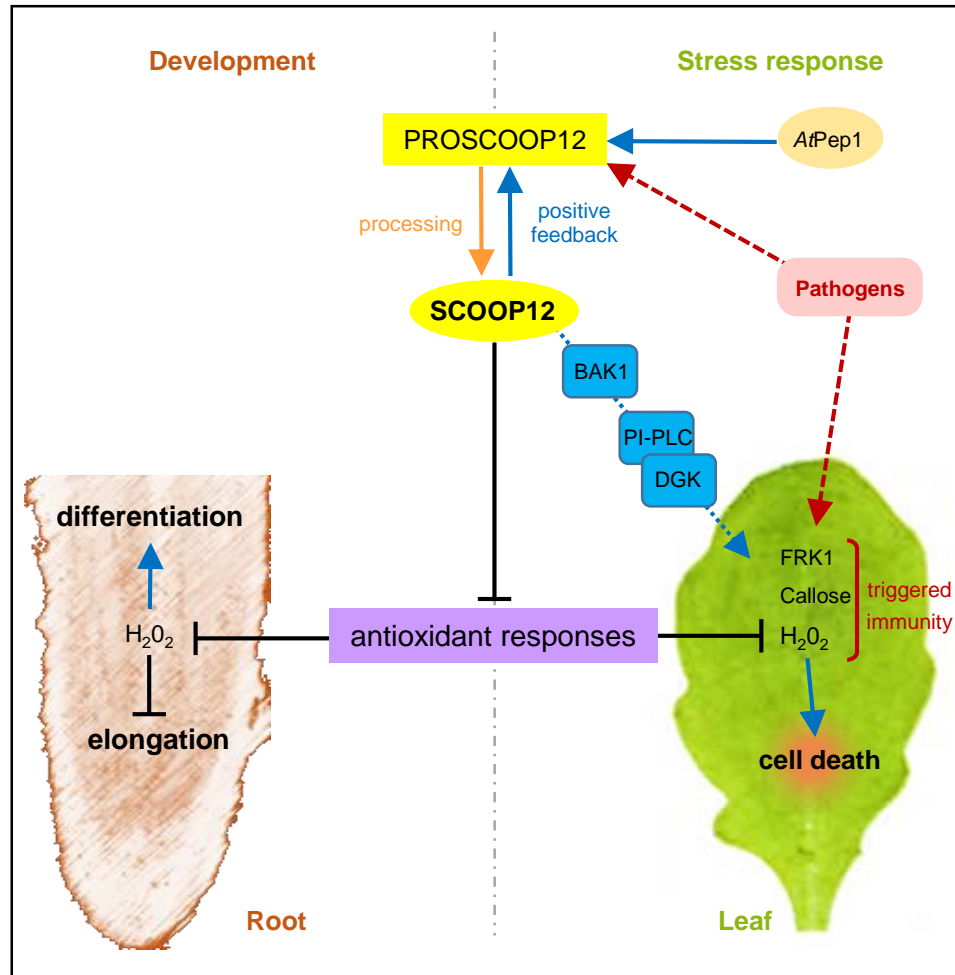
root development



Validation 2 : synthetic peptide



Functional model



The SCOOP12 peptide regulates defense response and root elongation in *Arabidopsis thaliana*.

K. Gully, S. Pelletier, M.-C. Guillou, M. Ferrand, S. Aligon, I. Pokotylo, A. Perrin, E. Vergne, M. Fagard, E. Ruelland, P. Grappin, E. Bucher, J.-P. Renou, S. Aubourg. J. Exp. Bot. (in press)

Acknowledgements



Nicolas Daccord (PhD student)
Florian Barbazange (M2 student)
Sylvain Gaillard
Charles-Eric Durel
Etienne Bucher

Jérôme Guzy (LIPM, Toulouse)



Kay Gully (PhD student)
Sandra Pelletier
Marie-Charlotte Guillou (M2 student)
Philippe Grappin
Jean-Pierre Renou

Eric Ruelland (iEES, Paris)

Mathilde Fagard (IJPB, Versailles)

