SCIENCE MEETS LIFE

Unravelling transcription factor functions through integrative inference of transcriptional networks in Arabidopsis

Klaas Vandepoele

Department of Plant Biotechnology and Bioinformatics, UGent

VIB Center for Plant Systems Biology

VIB-UGENT CENTER FOR PLANT SYSTEMS BIOLOGY

UNIVERSITEIT GENT

plaza_genomics

Utrecht

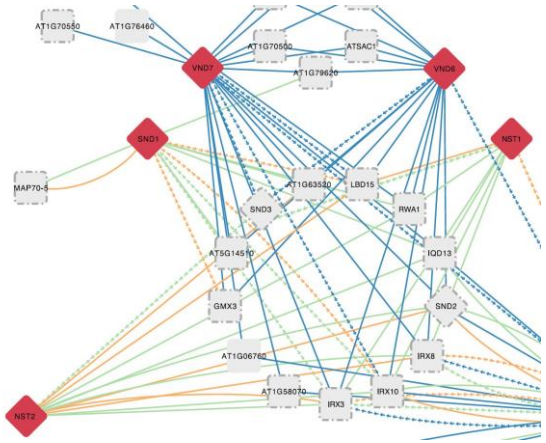
Artificial intelligence in plant science and breeding

24 February 2021

Comparative Network Biology - Vandepoele lab

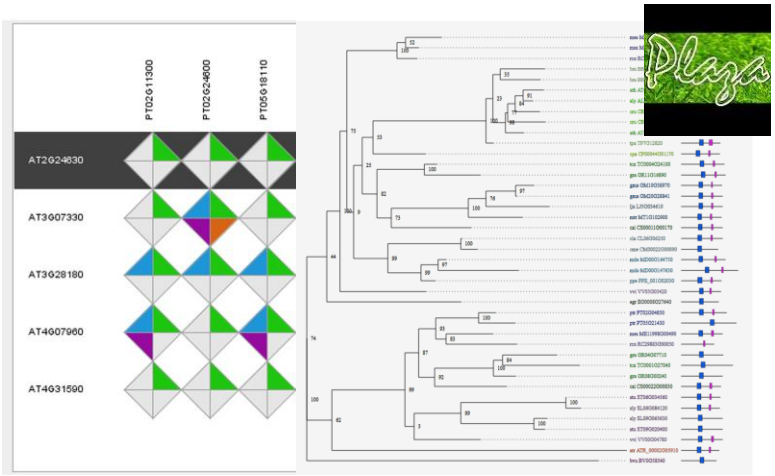
- Extract biological knowledge from large-scale experimental data sets using data integration, comparative sequence & expression analysis, and network biology, to improve our understanding of **gene functions and regulation** in plants and diatoms.

Plant Gene Regulatory Networks



- **TF2Network: predicting transcription factor regulators and gene regulatory networks in Arabidopsis using publicly available binding site information.** Kulkarni et al., 2018, Nucleic Acids Res
- **Enhanced maps of transcription factor binding sites improve regulatory networks learned from accessible chromatin data.** Kulkarni et al., Plant Physiol. 2019

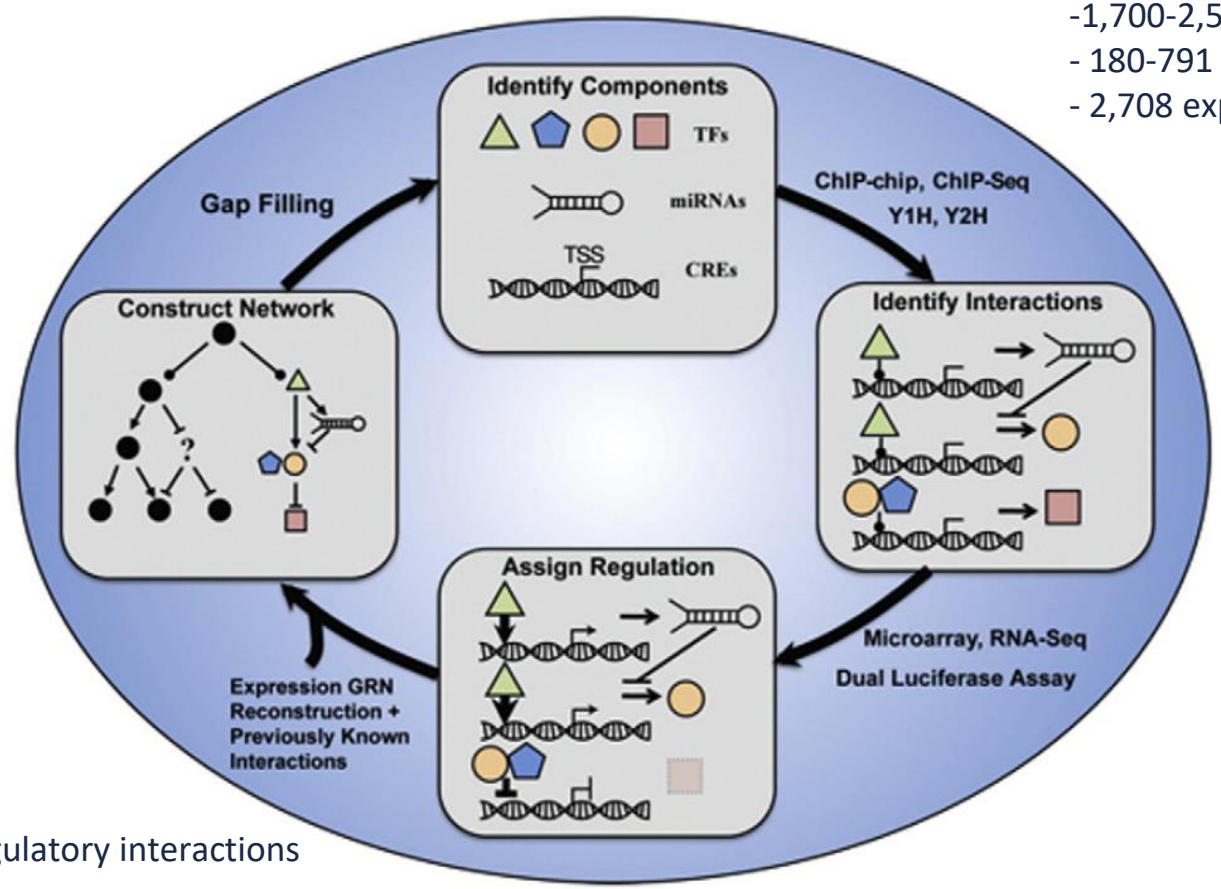
Comparative functional genomics



- **PLAZA 4.0: an integrative resource for functional, evolutionary and comparative plant genomics**. Van Bel et al., 2018, Nucleic Acids Res
- **Curse: building expression atlases and co-expression networks from public RNA-Seq data.** Vaneechoutte D, Vandepoele K. Bioinformatics. 2019

Mapping of Gene Regulatory Networks (GRNs)

Arabidopsis
-1,700-2,500 Transcription Factors
- 180-791 miRNA
- 2,708 expressed lncRNA



49MB non-coding DNA

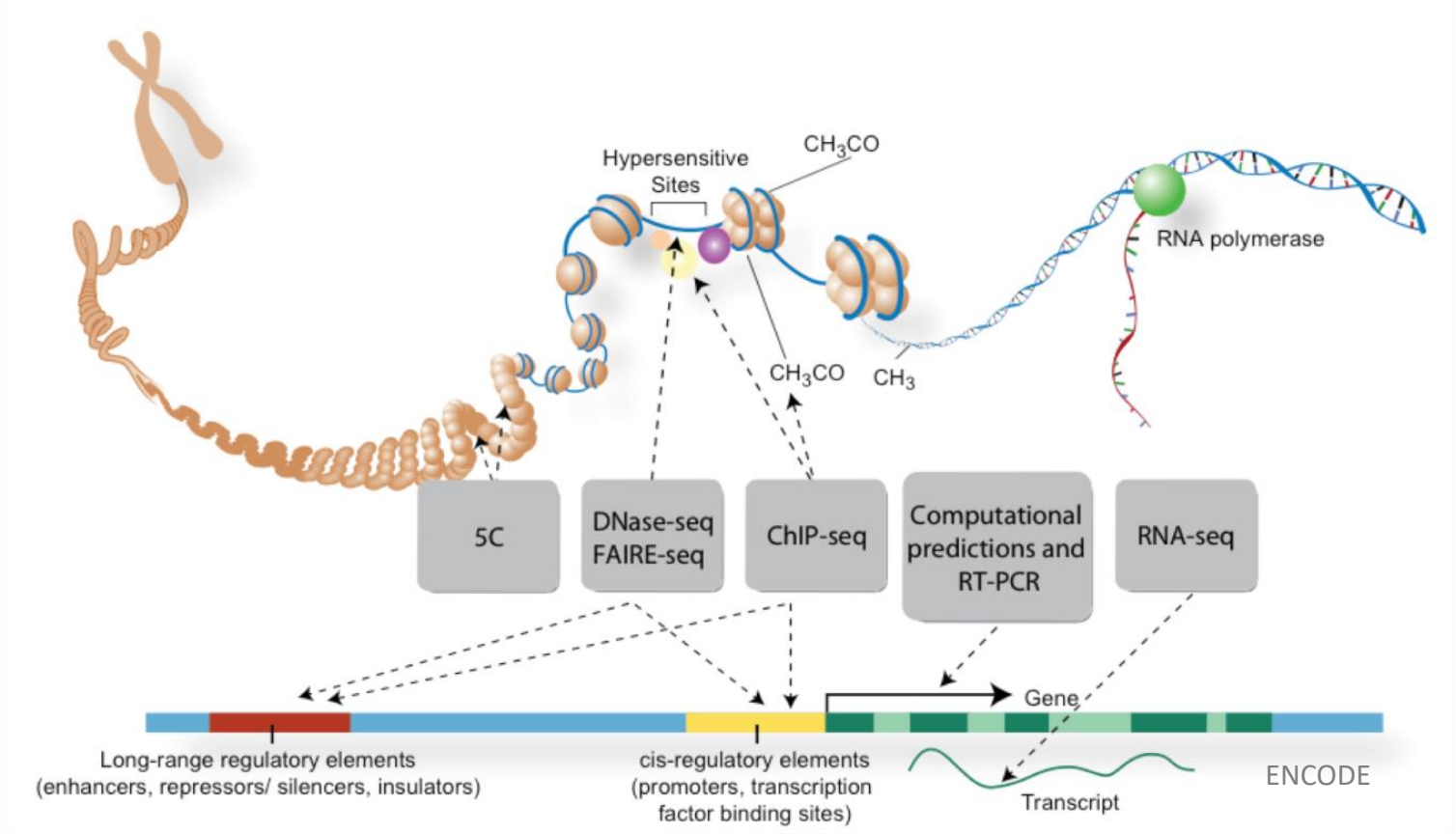
AtRegNet:
17,224 regulatory interactions

VIB-UGENT CENTER FOR PLANT SYSTEMS BIOLOGY

UNIVERSITEIT GENT

Mejia-Guerra et al., 2012

Experimental characterization of transcriptional activity and regulatory control

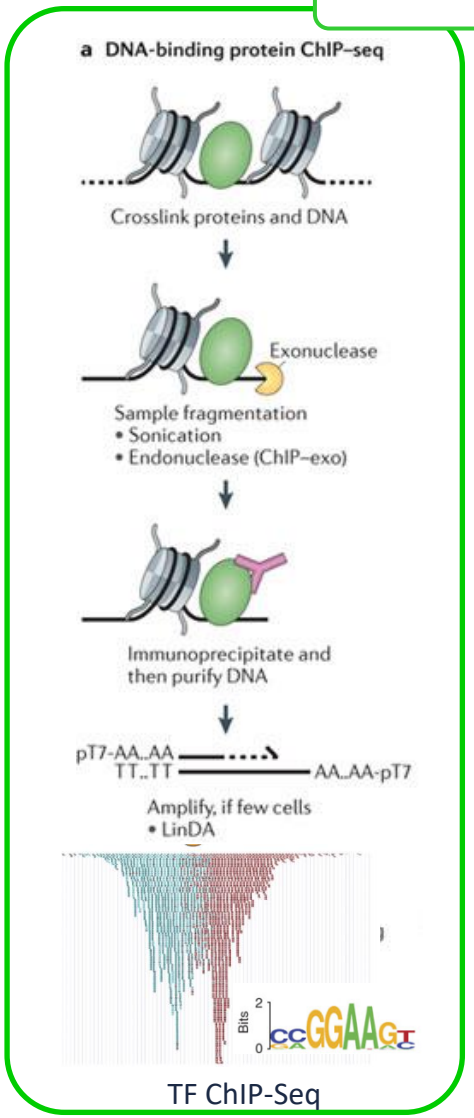


How to integrate the biological knowledge captured by different –omics layers to build better networks reporting functional regulatory interactions?

1. TF ChIP-Seq

- *in vivo* method to measure protein-DNA interactions using chromatin immuno-precipitation
- Different cellular conditions can be profiled

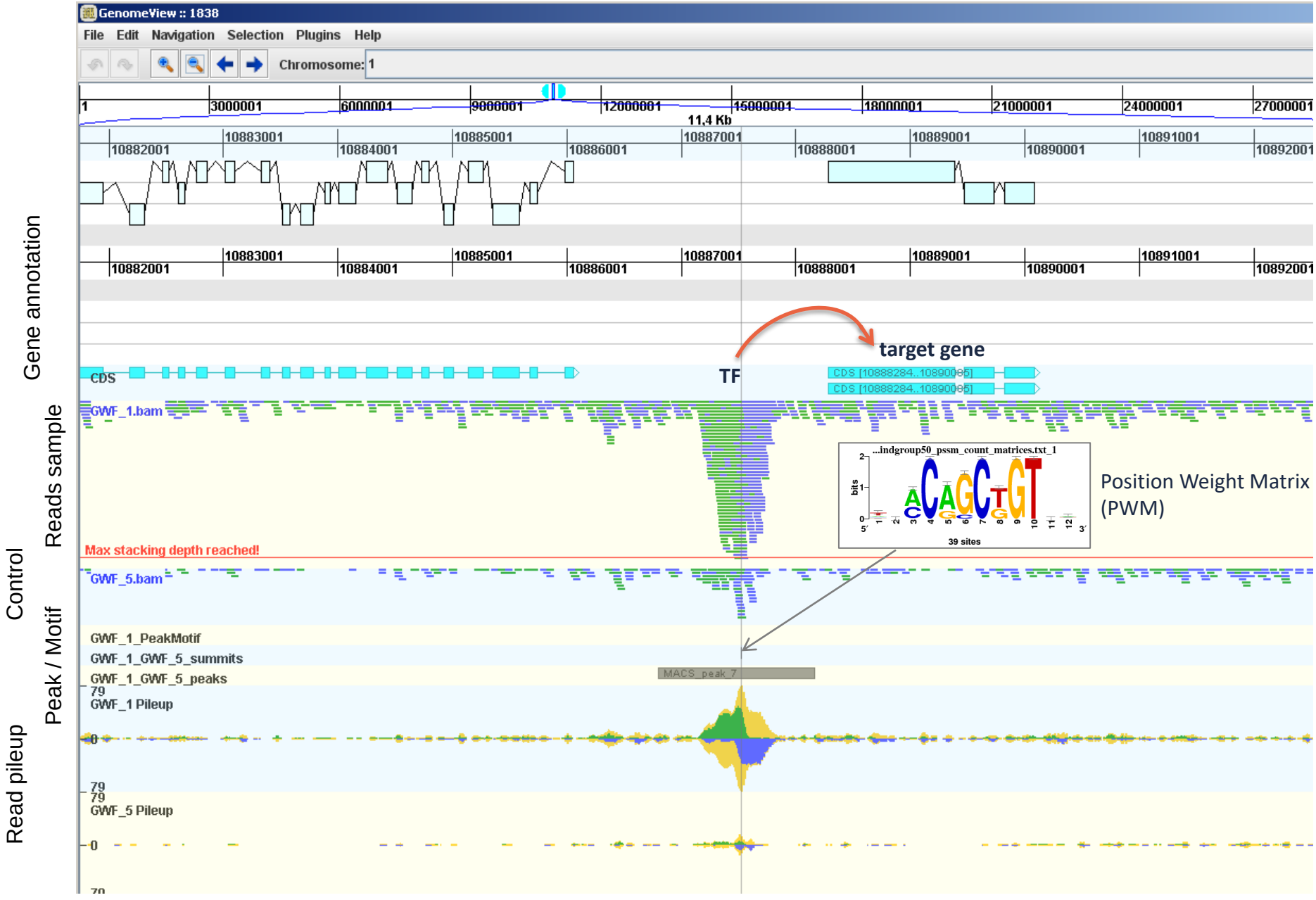
ChIP



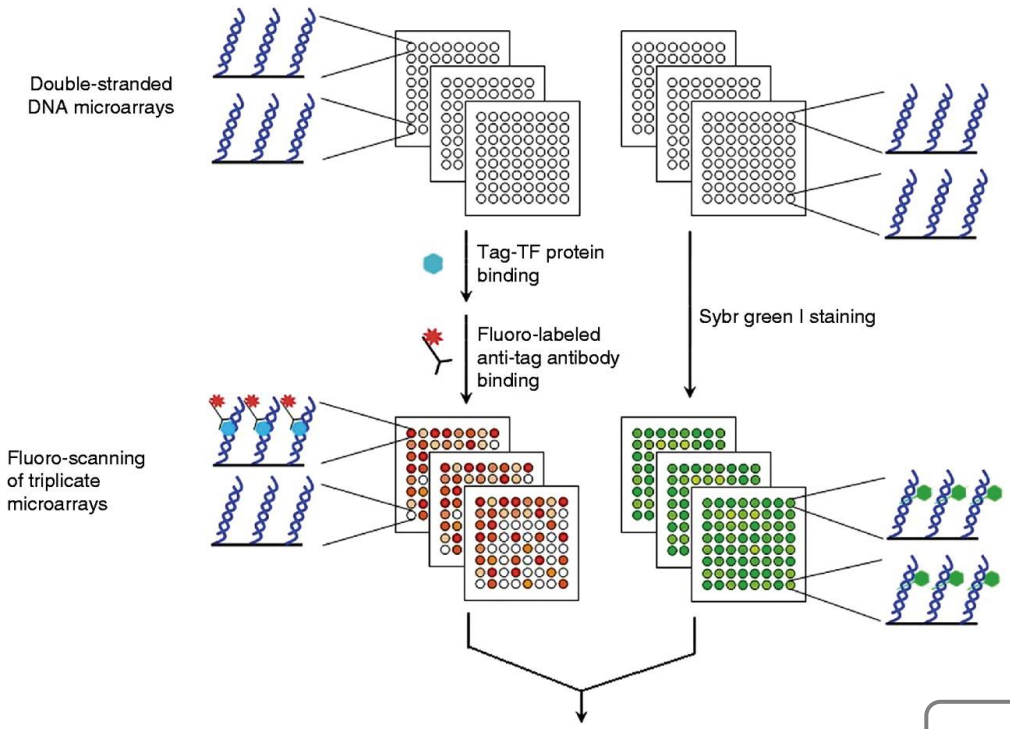
TF ChIP-Seq

Furey et al., 2012

Output ChIP-Seq peak calling procedure displayed in genome browser



2. *in vitro* TF binding specificities



Protein binding microarray

Model TF binding site as Position Weight Matrix (**PWM**) based on k-mer signals

PWM

target gene

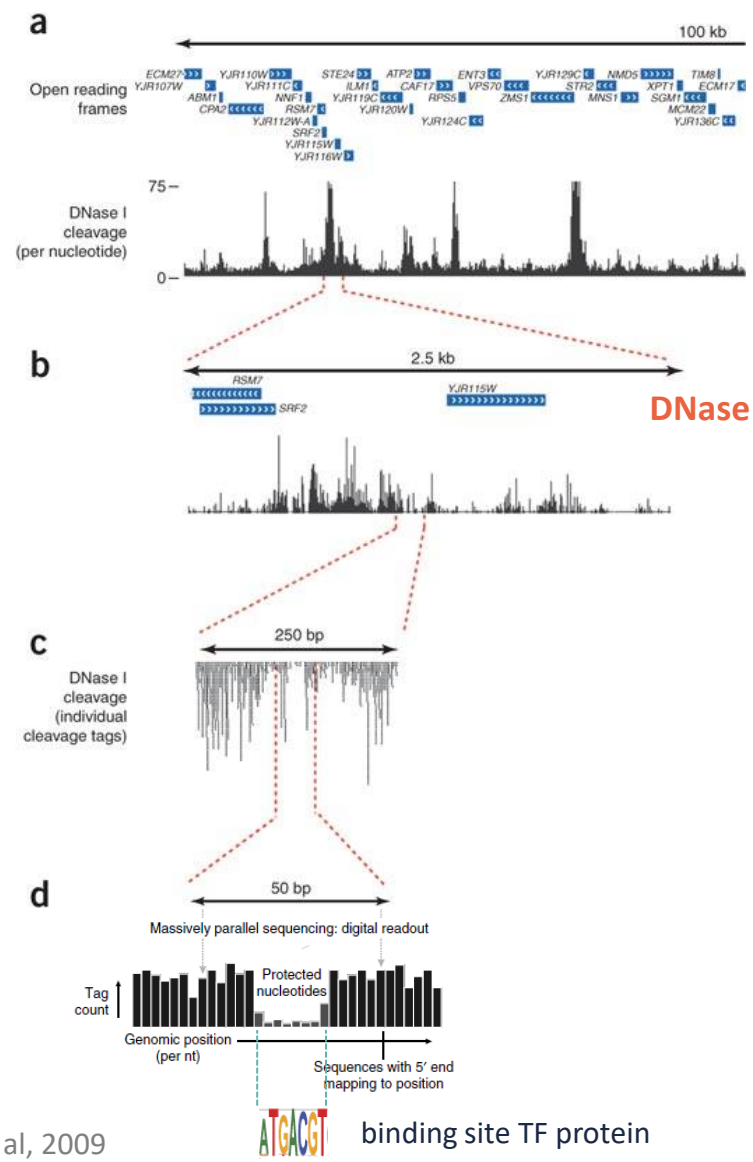
Arabidopsis: PWMs for 990 TFs

VIB-UGENT CENTER FOR PLANT SYSTEMS BIOLOGY

UNIVERSITEIT GENT

Wang et al., 2011

3. DNase-seq - Profiling of accesible chromatin

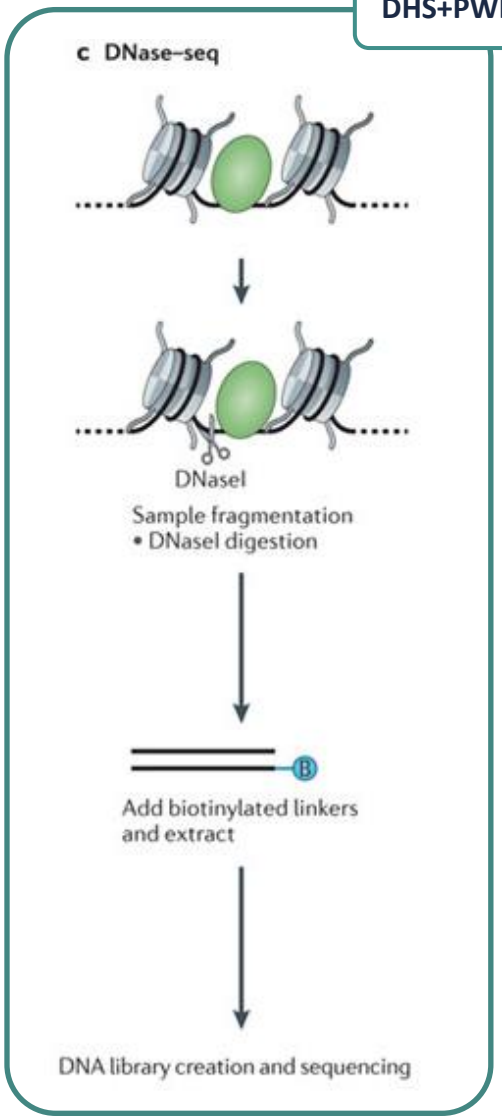


DNase I hypersensitive site (DHS)

DH footprint

binding site TF protein

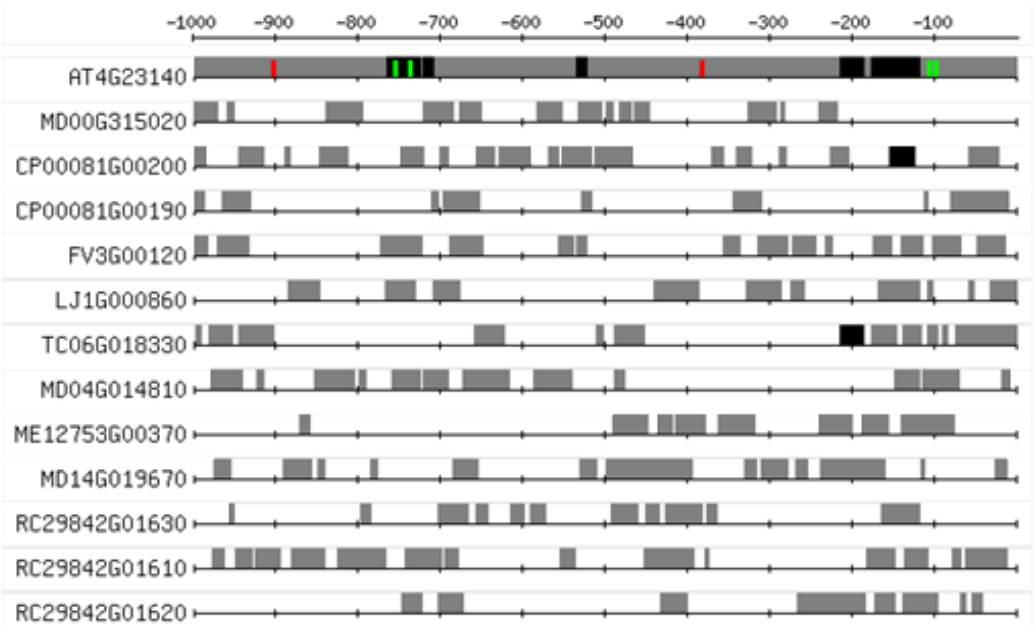
DHS+PWM



Hesselberth et al, 2009

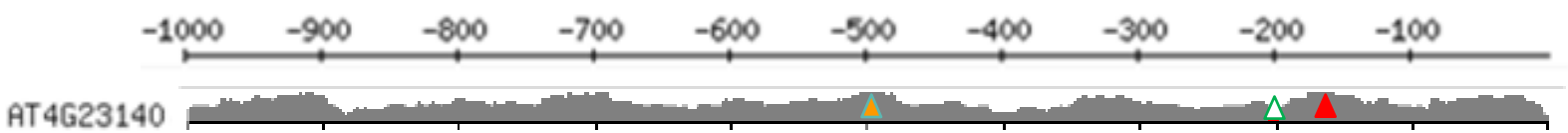
4. Detection of conserved TF binding sites using phylogenetic footprinting

Conserved PWM



- Map all known PWMs on the promoters of the *Arabidopsis* query gene and its orthologs

- Count per PWM position the #species that support a TF binding site
- Significance estimation (FDR<10%)

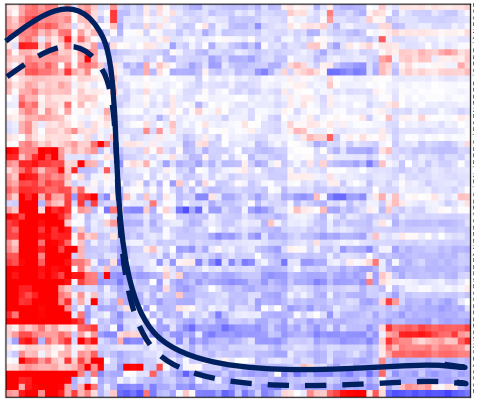


Van de Velde et al., Plant Phys 2016 - A Collection of Conserved Non-Coding Sequences to Study Gene Regulation in Flowering Plants.

5-6. Network inference based on expression data

COE TF-target

TF target



Expression-based network inference

Bioinformatics, 2019, 1–2
doi: 10.1093/bioinformatics/bty1052
Advance Access Publication Date: 24 December 2018
Applications Note

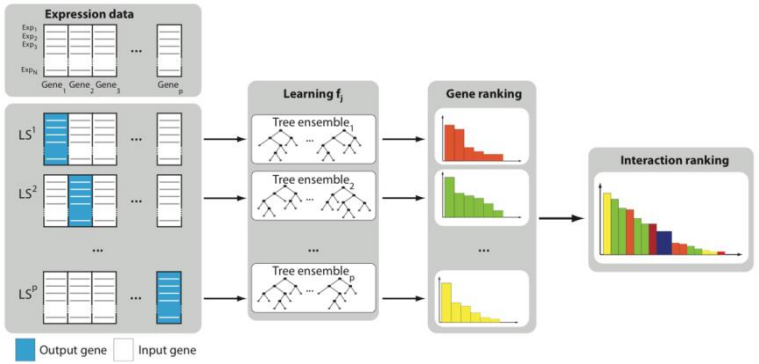
Data and text mining

Curse: building expression atlases and co-expression networks from public RNA-Seq data

Dries Vaneechoutte [1,2] and Klaas Vandepoele [1,2,3,*]

GENIE3

GENIE3



VIB-UGENT CENTER FOR PLANT SYSTEMS BIOLOGY

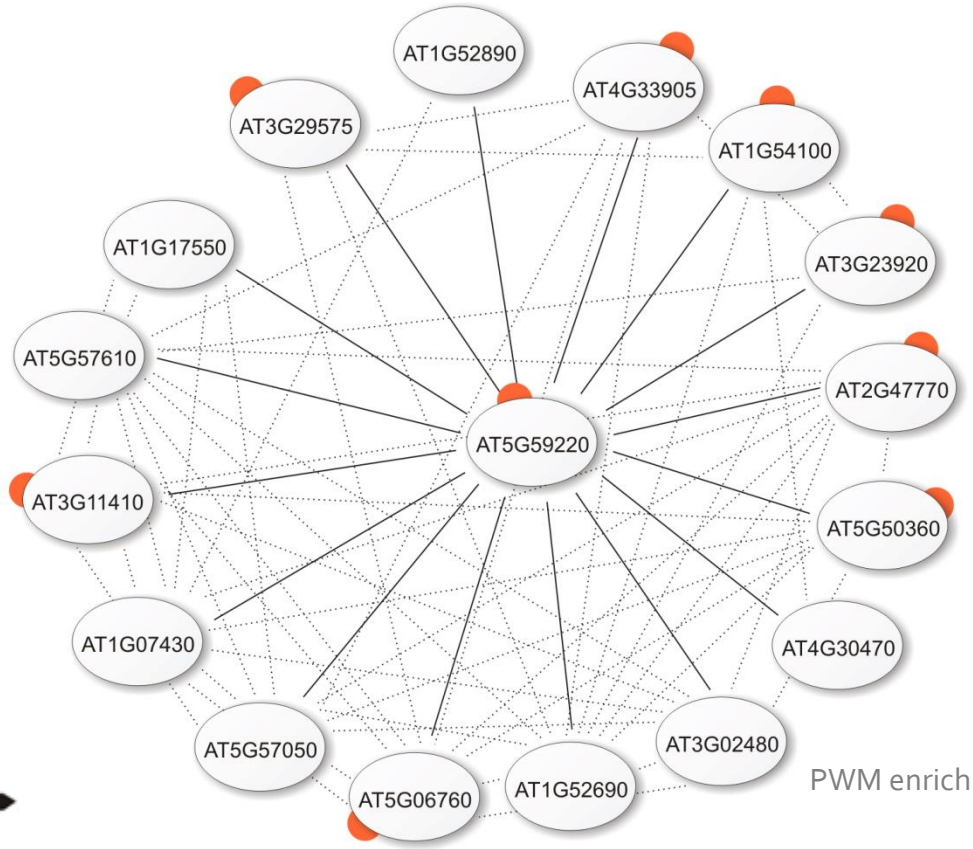
UNIVERSITEIT GENT

GENIE3 - Huynh-Thu et al., 2013

7. Co-expression + PWM enrichment

- Integrate co-regulatory gene expression data with TF binding sites (PWMs)

COE+PWM



PWM enrichment in kNN co-expression cluster
(hypergeometric distribution)

VIB-UGENT CENTER FOR PLANT SYSTEMS BIOLOGY

UNIVERSITEIT GENT

Benchmarking of different methods to map gene regulatory networks

Approach	Number of Interactions	Recall of True Positives	Enrichment test set
Y1H	1,801	11 %	57.0
TF ChIP−Seq	46,619	29 %	19.5
PWM map	17,053,785	60 %	1.0
PWM map in Conserved region	340,893	11 %	5.0
PWM map in Open Chromatin	5,231,258	34 %	1.3
Co−Expression + PWM enrichment	324,864	3 %	1.9
Co−Expression TF- target	1,822,002	10 %	2.9
GENIE3	3,000,000	17 %	1.7

Gold standard: 5.7k interactions covering

Test set: 20% of gold standard (80% used

VIB-UGENT CENTER FOR PLANT SYSTEMS BIOLOGY

UNIVERSITEIT GENT

Benchmarking of different methods to map gene regulatory networks

Approach	Number of Interactions	Recall of True Positives	Enrichment test set
Y1H	1,801	11 %	57.0
TF ChIP−Seq	46,619	29 %	19.5
PWM map	17,053,785	60 %	1.0
PWM map in Conserved region	340,893	11 %	5.0
PWM map in Open Chromatin	5,231,258	34 %	1.3
Co−Expression + PWM enrichment	324,864	3 %	1.9
Co−Expression TF- target	1,822,002	10 %	2.9
GENIE3	3,000,000	17 %	1.7

Gold standard: 5.7k interactions covering 522 TFs (AtRegNet+literature)

Test set: 20% of gold standard (80% used for training later)

Jan Van de Velde

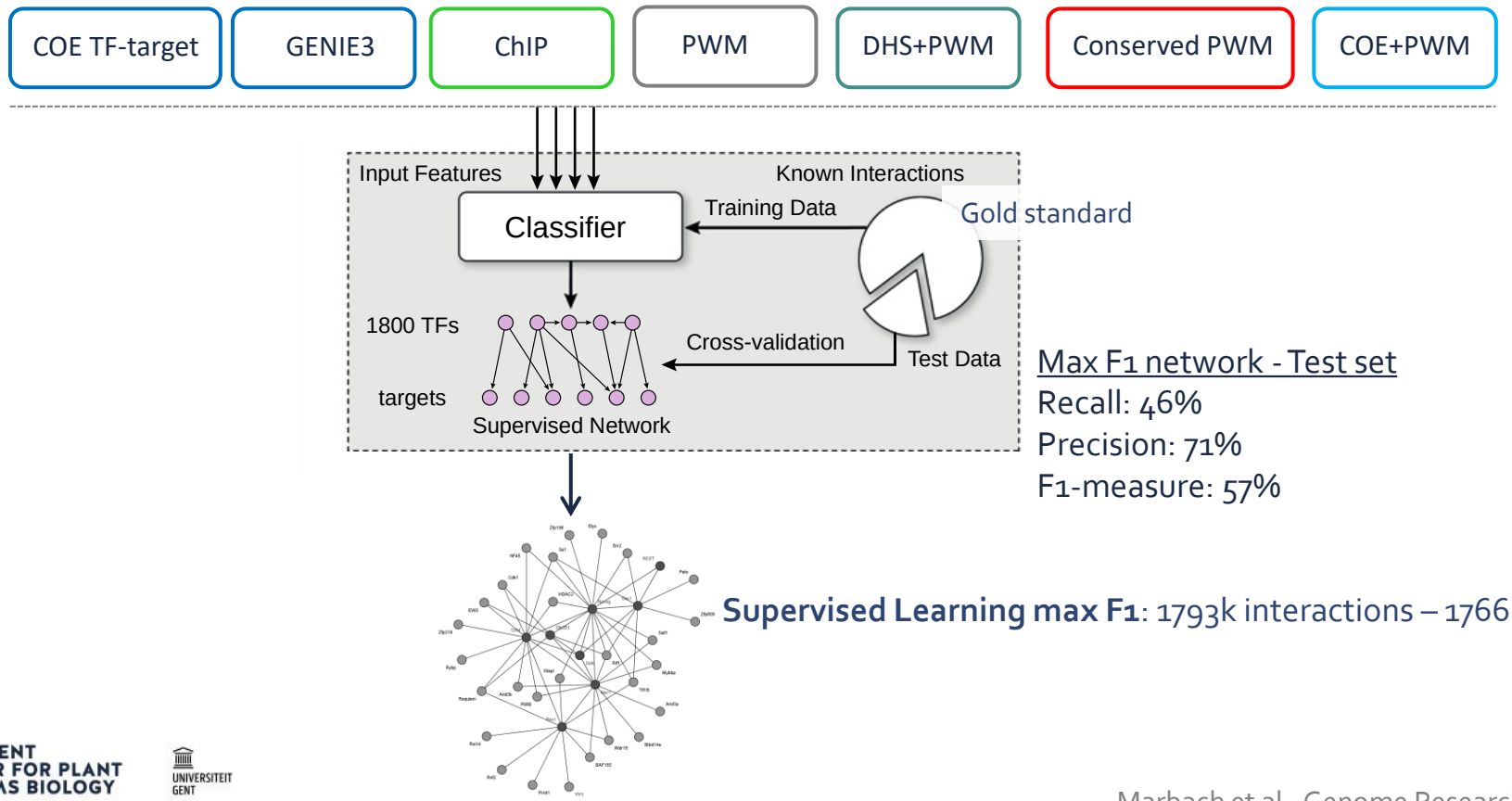
VIB-UGENT CENTER FOR PLANT SYSTEMS BIOLOGY

UNIVERSITEIT GENT

Supervised learning: a network-based approach for large-scale functional data integration

Gradient Boosting Machine

- 1000 trees (shrinkage of 0.01, interaction depth 3, 10-fold CV training)
- 80% training data with True:False sampling ratio of 3:1
- 7 input networks



Max F1 network - Test set
Recall: 46%
Precision: 71%
F1-measure: 57%

Supervised Learning max F1: 1793k interactions – 1766 TFs

VIB-UGENT CENTER FOR PLANT SYSTEMS BIOLOGY

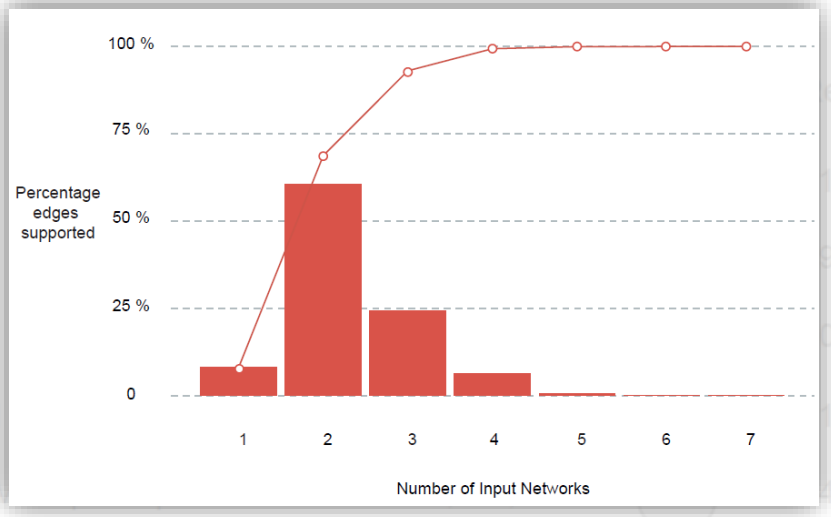
UNIVERSITEIT GENT

Marbach et al., Genome Research 2012

Performance supervised learning network (iGRN)

Approach	Number of Interactions	Recall of True Positives	Enrichment test set
Y1H	1,801	11 %	57.0
TF ChIP−Seq	46,619	29 %	19.5
PWM map	17,053,785	60 %	1.0
PWM map in Conserved region	340,893	11 %	5.0
PWM map in Open Chromatin	5,231,258	34 %	1.3
Co−Expression + PWM enrichment	324,864	3 %	1.9
Co−Expression TF- target	1,822,002	10 %	2.9
GENIE3	3,000,000	17 %	1.7
Supervised Learning max F1	1,793,435	46 %	8.2

Different support of input networks for iGRN

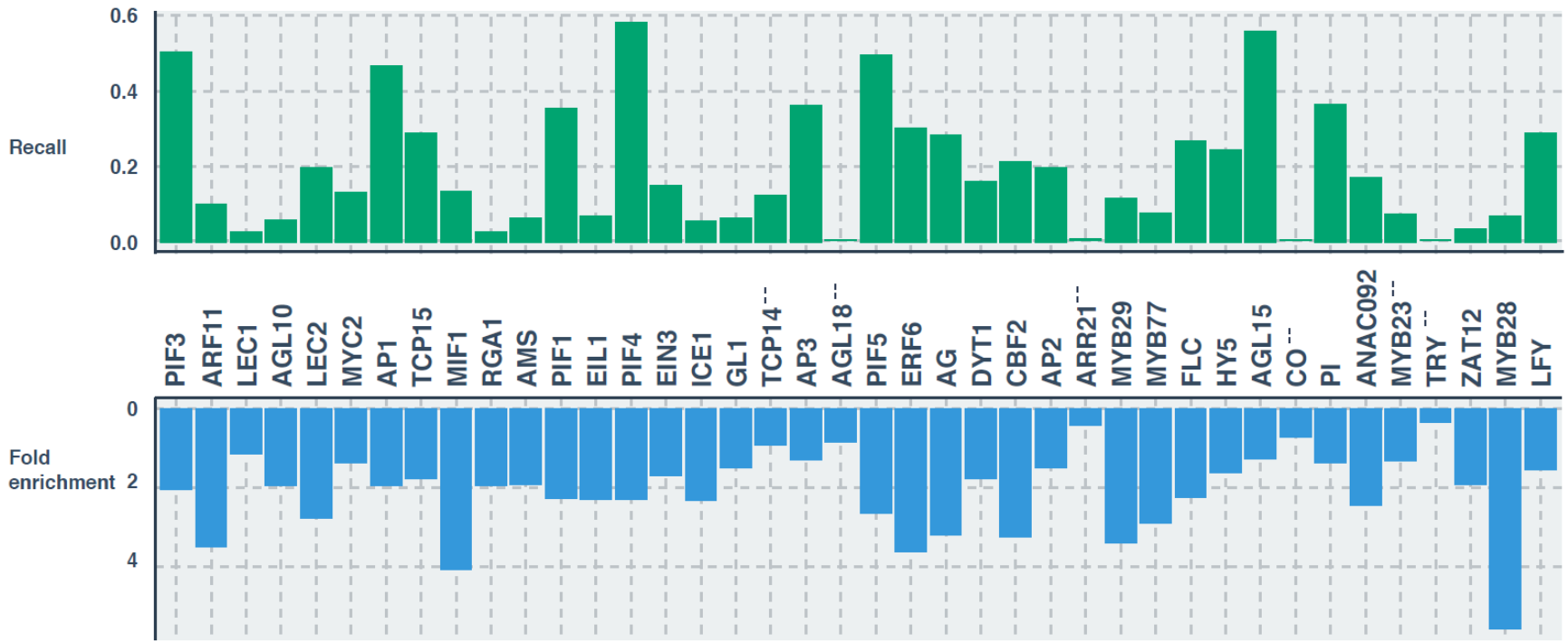


Supervised Learning max F1

Approach	Number of Interactions	Network contribution	Percentage of iGRN
TF ChIP–Seq	46,619	80 %	2 %
PWM map	17,053,785	7 %	70 %
PWM map in Conserved region	340,893	91 %	17 %
PWM map in Open Chromatin	5,231,258	17 %	49 %
Co–Expression + PWM enrichment	324,864	18 %	3 %
Co–Expression TF- target	1,822,002	38 %	38 %
GENIE3	3,000,000	30 %	50 %

VIB-UGENT CENTER FOR PLANT SYSTEMS BIOLOGY

iGRN captures functional TF – target gene interactions



-- overlap target genes <u>not</u> significant (p-value hypergeometric distribution > 0.05)

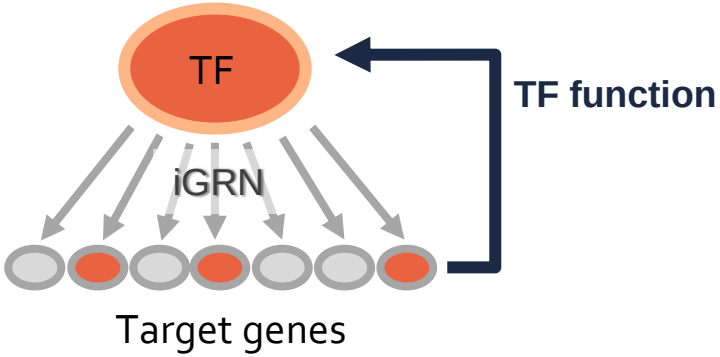
34/40 TFs have significant overlap between predicted target genes and DE genes after TF perturbation

VIB-UGENT CENTER FOR PLANT SYSTEMS BIOLOGY

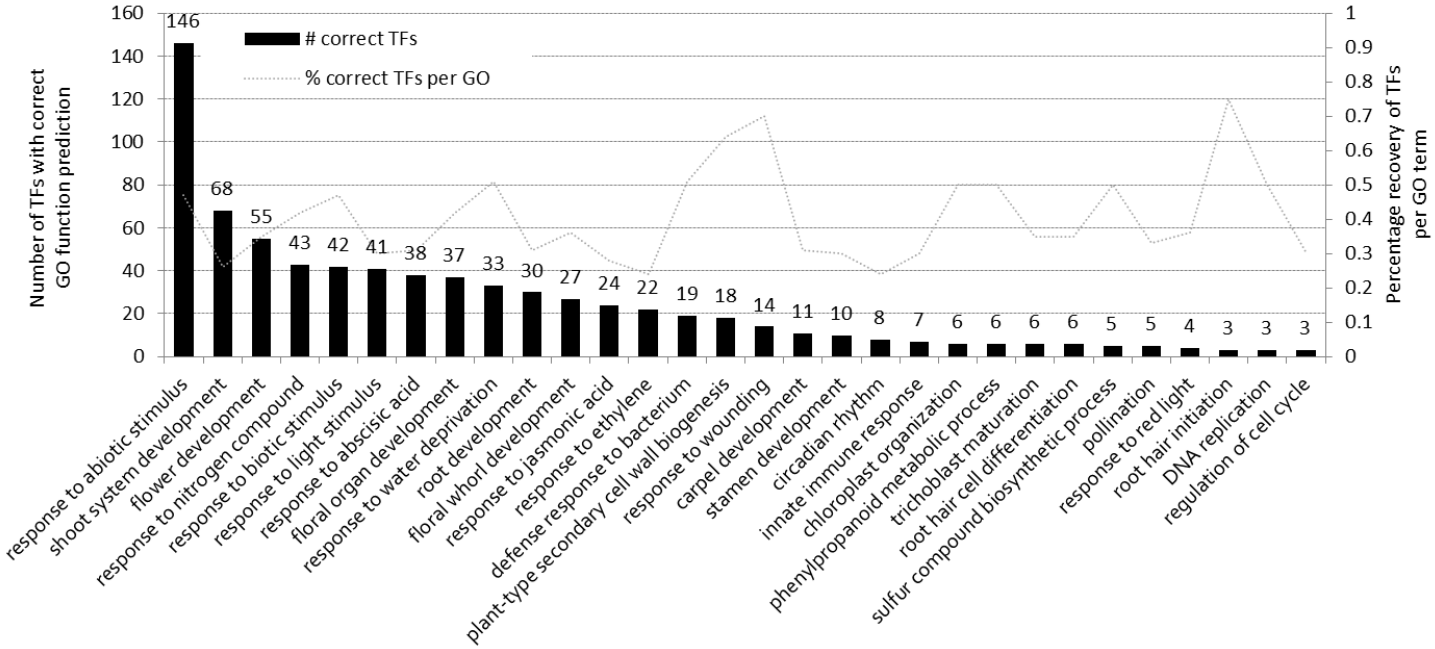
UNIVERSITEIT GENT

iGRN-based functional annotation of TFs

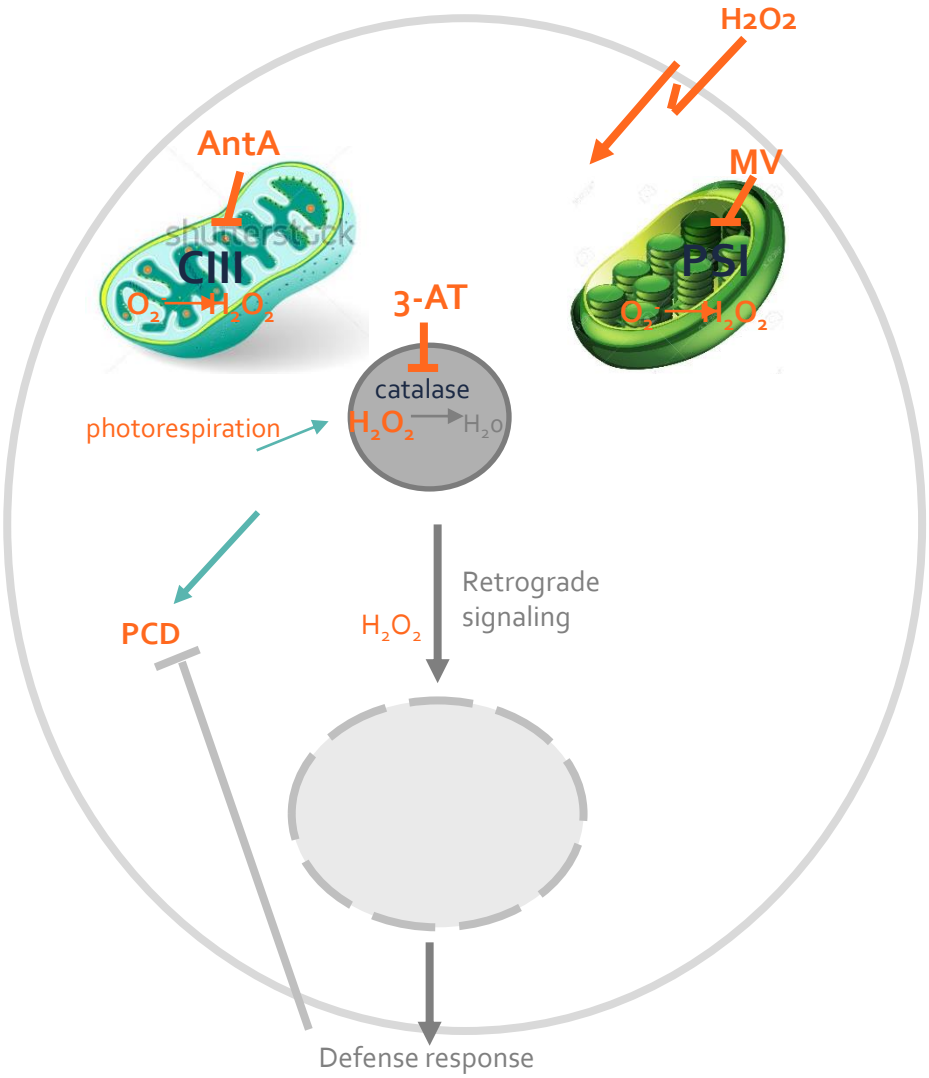
Recovery of experimental Gene Ontology Biological Process annotations for TFs with known function



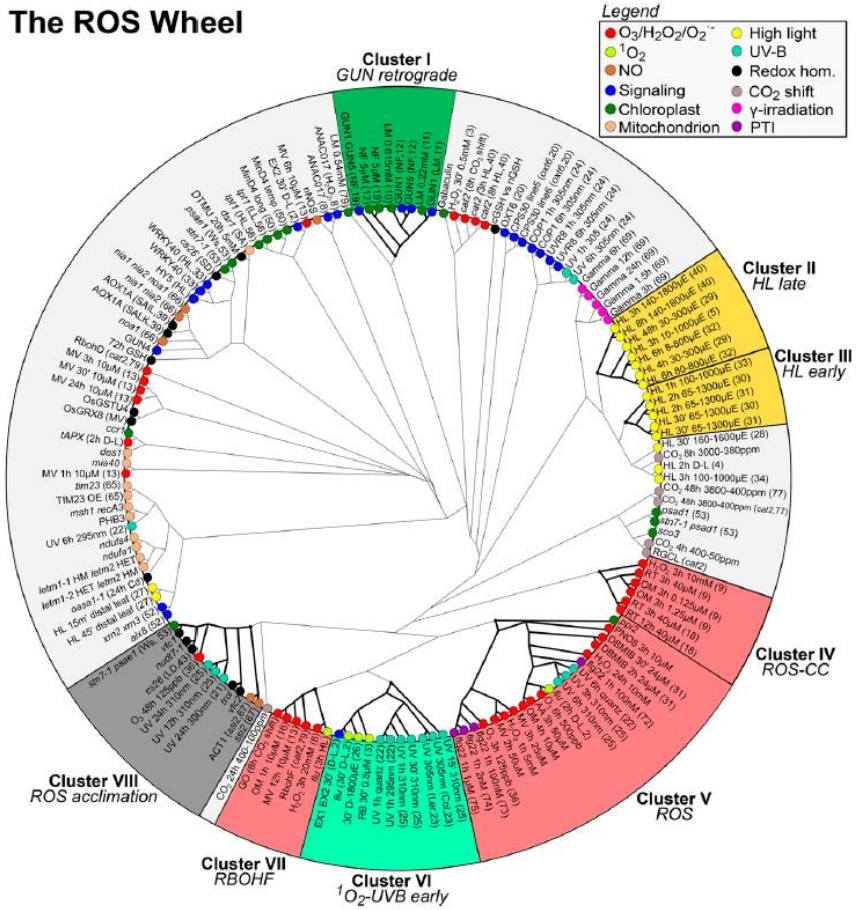
- Recovery of known experimentally-supported functions for >600 TFs
- Novel functional predictions for 268 unknown TFs
- Highly complementary with AraNet v2



Oxidative stress signaling



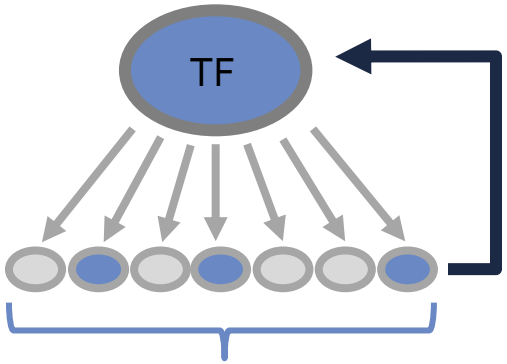
In house dataset of ROS marker genes



Van Breusegem lab - VIB

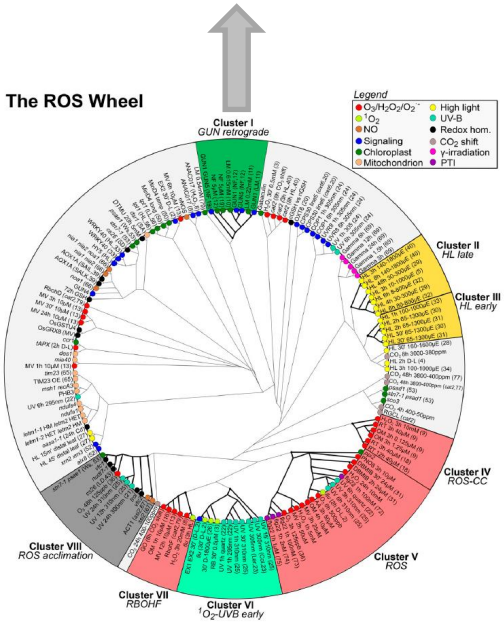
Willems et al., 2016, Plant Physiology

Prediction and evaluation of novel oxidative stress TFs ("ROS-TFs")



Novel TF functions
(e.g. oxidative stress responses)

Target gene enrichment:
- ROS wheel
- GO-BP 'response to oxidative stress'



Unknown or no stress-related function

Oxidative stress function

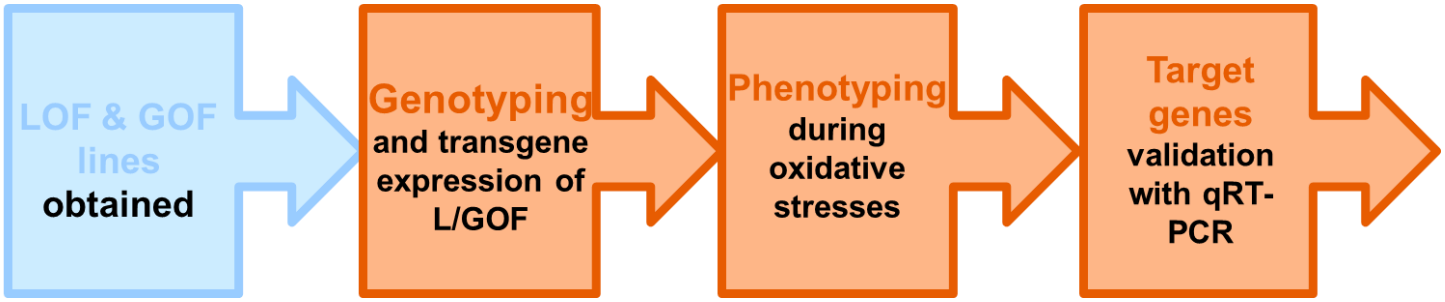
Other (a/biotic) stress function

Ranking based on ROS wheel target genes enrichment (n=124 TFs)

Rank	ID	Gene name	q-val	enrichment	Phenotype
1	AT5G63790	ANAC102,NAC102	1,84E-35	22,74	Oxidative (our data)
2	AT3G55980	ATSZF1,SZF1	5,10E-26	32,71	Salt
3	AT2G37430	ZAT11	2,81E-25	14,41	Oxidative (paraquat, Ni)
4	AT2G40140	ATSZF2,CZF1	6,41E-25	17,45	Salt
5	AT5G59820	AtZAT12,RHL41,ZAT12	8,70E-25	18,64	Oxidative and abiotic
6	AT5G24110	ATWRKY30,WRKY30	1,42E-24	6,82	oxidative, salt (at early developmental stage)
7	AT2G38470	ATWRKY33,WRKY33	5,03E-24	6,85	Pathogen, salt, heat stress
8	AT2G46400	ATWRKY46,WRKY46	9,78E-24	5,79	Osmotic/salt
9	AT4G17500	ATERF-1,ERF-1	2,57E-23	9,42	Biotic …
10	AT1G28370	ATERF11,ERF11	3,10E-23	6,75	Osmotic, ET signaling
11	AT4G18880	AT-HSFA4A,HSF21	5,09E-23	6,39	oxidative, salt
12	AT2G23320	AtWRKY15,WRKY15	7,46E-23	6,31	Oxidative, salt
13	AT5G04340	C2H2,CZF2,ZAT6	7,52E-23	14,09	Cadmium, salt, osmotic stress, P deficiency, pathogen, drought, cold
14	AT1G80840	ATWRKY40,WRKY40	8,93E-23	6,51	Biotic, MRS
15	AT3G23250	ATMYB15,ATY19	1,86E-22	5,71	Drought, cold
16	AT1G27730	STZ,ZAT10	4,90E-22	9,99	Oxidative transcripts, abiotic, ….
17	AT5G49520	ATWRKY48,WRKY48	5,36E-22	11,26	Biotic
18	AT5G13080	ATWRKY75,WRKY75	1,40E-21	5,81	Phosphate starvation, biotic
19	AT4G17230	5CL13	3,76E-21	16,60	Phytochrome dependent light signaling
20	AT5G59450	AT5G59450	5,14E-21	19,95	cell division
21	AT1G42990	ATBZIP60,BZIP60	1,36E-20	5,46	ER, unfolded protein
22	AT1G18570	ATMYB51,BW51A	1,84E-20	5,82	glucosinolate biosynthesis
23	AT4G23810	ATWRKY53,WRKY53	6,64E-20	7,56	leaf senescence and regulation of oxidative stress genes
24	AT1G66550	ATWRKY67,WRKY67	3,25E-19	8,88	None; no lines available
25	AT3G23220	ESE1	3,54E-19	7,50	Salt, ET signaling
26	AT5G47230	ATERF5,ATERF-5	3,60E-19	6,00	Biotic
27	AT3G54810	BME3,BME3-ZF,GATA8	2,01E-18	14,92	Germination, salt/drought
28	AT5G47220	ATERF2,ATERF-2,ERF2	4,16E-18	5,57	Biotic
29	AT2G40740	ATWRKY55,WRKY55	6,69E-18	5,70	None
30	AT4G36990	ATHSF4,AT-HSFB1,HSF4	1,94E-17	5,98	similar to heat shock factor, no known phenotype
31	AT2G30250	ATWRKY25,WRKY25	1,59E-16	4,81	Biotic, salt
32	AT4G31550	ATWRKY11,WRKY11	2,03E-16	5,33	Biotic
33	AT5G01380	GT3a	2,15E-16	4,93	None
34	AT5G22570	ATWRKY38,WRKY38	2,19E-16	6,01	Biotic
35	AT3G23240	ATERF1,ERF1	2,36E-15	4,16	ET
36	AT4G17490	ATERF6,ERF6,ERF-6-6	3,13E-15	5,26	Oxidative
37	AT4G32950	AGL19,SEP3	3,37E-15	15,26	flowering
38	AT3G10500	ANAC053,NAC053,NTL4	3,63E-15	5,02	Oxidative/ROS
39	AT3G49530	ANAC062,NAC062,NTL6	1,76E-14	7,24	ER, unfolded protein
40	AT5G62020	AT-HSFB2A,HSF6,HSFB2A	5,87E-14	5,87	
41	AT1G22070	TGA3	6,60E-14	5,33	biotic
42	*AT1G67970*	*AT-HSFA8,HSFA8*	*7,97E-14*	*11,40*	*redox dependent nucleus translocation*

⋮

Functional validation of the predicted ROS-TFs



Stress treatments

MV

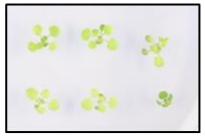
H2O2

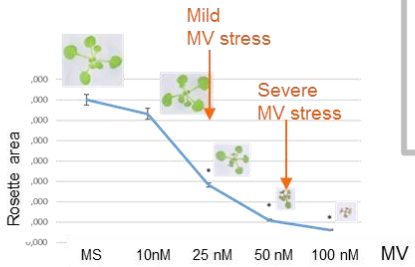
3-AT/RG-CL

AntA

mild

severe

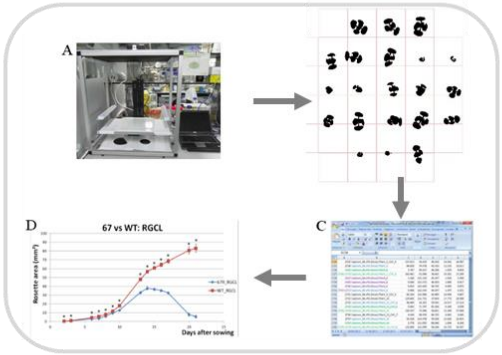




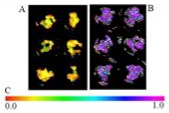
Scoring

Rosette area

PRAC: semi-automated pipeline for area measurement



F_v/F_m



Germination rate, bleaching, survival

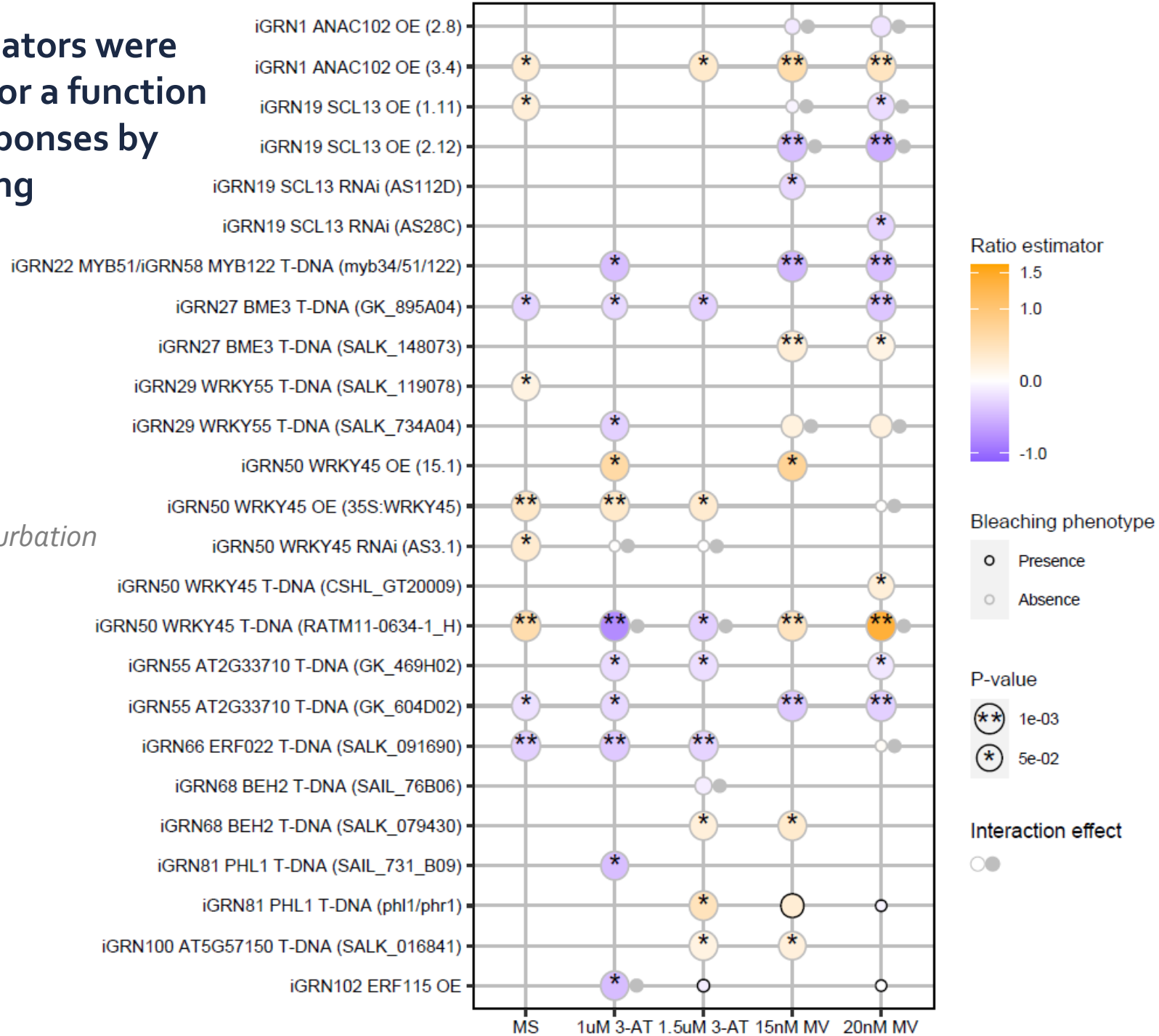
VIB-UGENT CENTER FOR PLANT SYSTEMS BIOLOGY

UNIVERSITEIT GENT

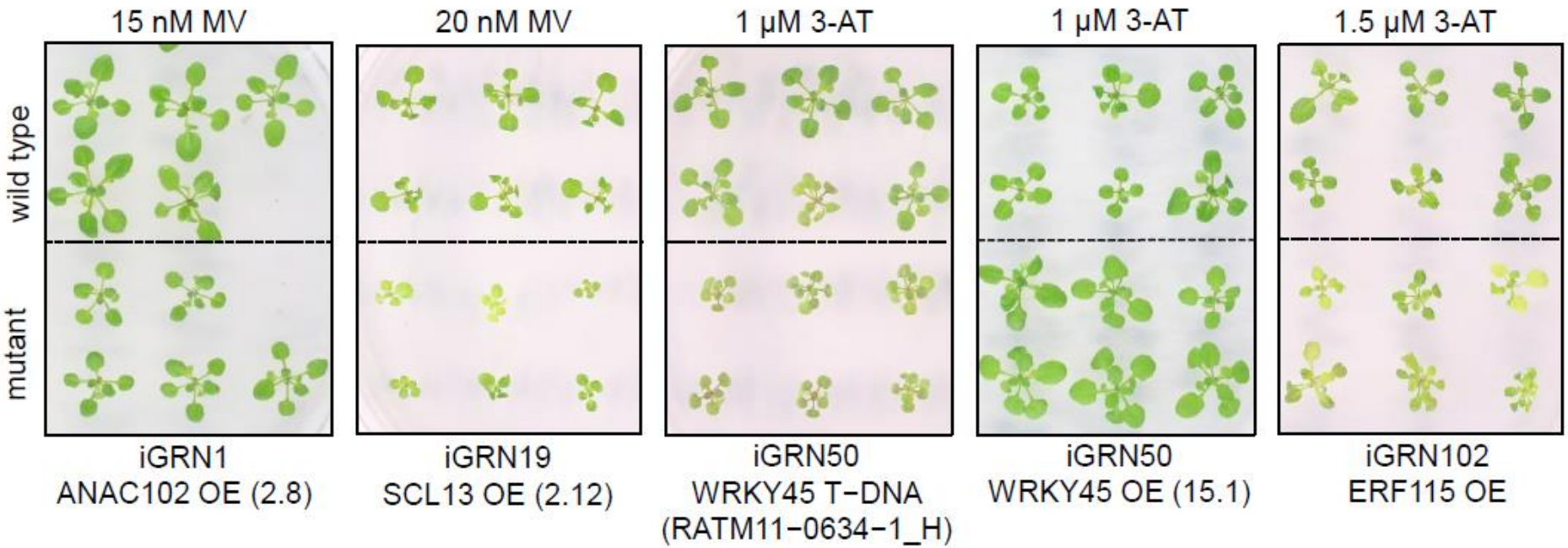
Inge De Clercq

13/32 regulators were validated for a function in ROS responses by phenotyping

Rank – TF - perturbation



Phenotypes for predicted ROS-TFs



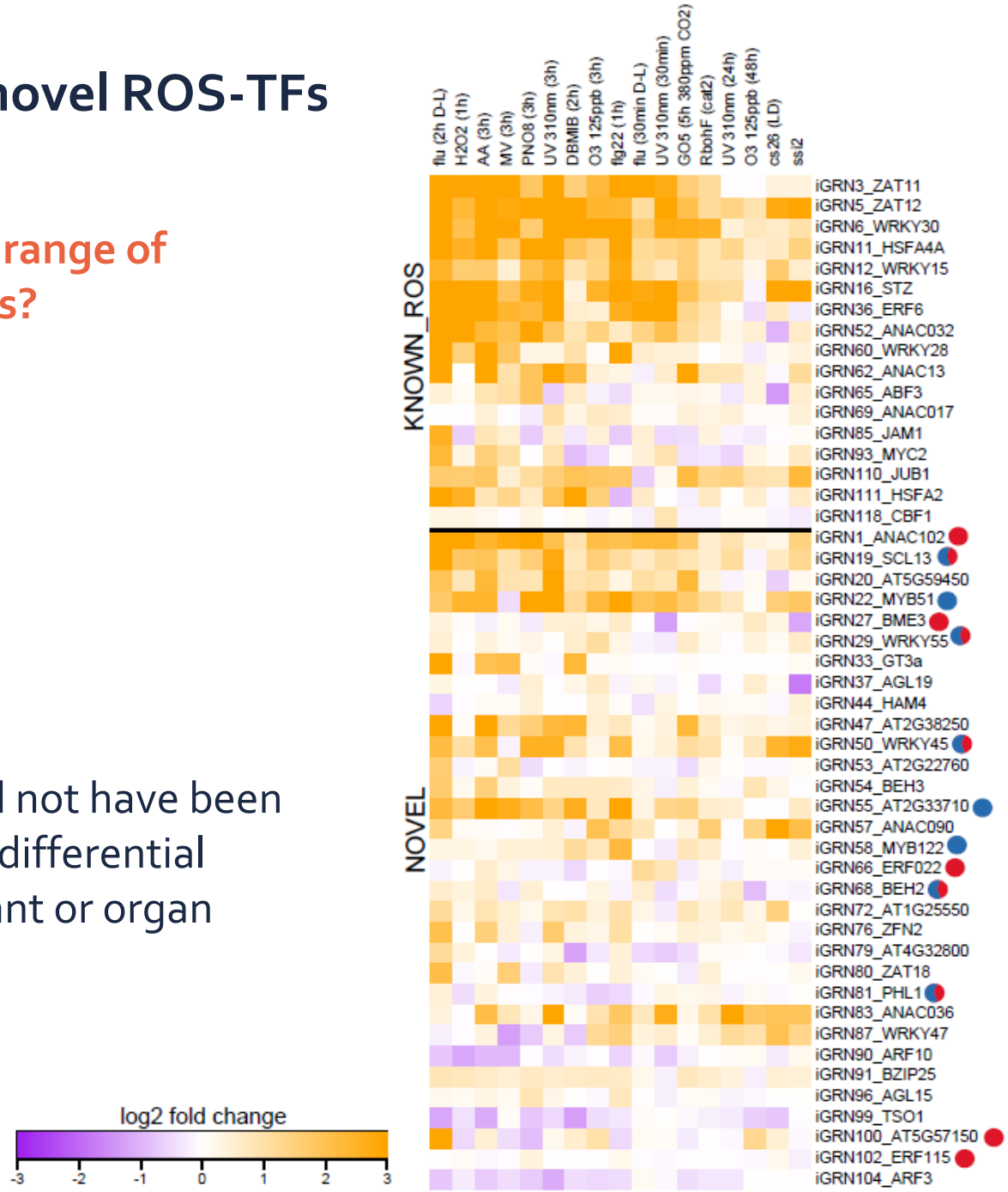
iGRN identified novel ROS TFs from the GRAS, BES1 and GATA families

Expression patterns for novel ROS-TFs

Responsiveness to a wide range of oxidative stress conditions?

- 14/17 known ROS TFs
- 6/13 novel ROS TFs

Many novel ROS TFs would not have been predicted solely relying on differential expression at the whole plant or organ level!



Conclusions

- Different regulatory –omics data types as well as advanced computational integration methods contribute significantly to the improved delineation of high-quality gene regulatory networks

- TF binding site-based as well as expression-based regulatory networks offer a complementary view on functional gene regulatory interactions

- Gene regulatory networks obtained by supervised learning are a starting point for
  - the systematic functional/regulatory annotation of all Arabidopsis genes
  - new biological discoveries

Jan Van de Velde

Inge De Clercq

Li Liu, Dries Vaneechoutte
Robin Pottie, Xiaopeng Liu, Frank Van Breusegem

Further reading

Unraveling transcription factor functions through integrative inference of transcriptional networks in *Arabidopsis thaliana* yields novel regulators involved in reactive oxygen species stress signaling

Inge De Clercq, Jan Van de Velde, Xiaopeng Luo, Li Liu, Veronique Storme, Robin Pottie, Dries Vaneechoutte, Frank Van Breusegem, Klaas Vandepoele

doi: https://doi.org/10.1101/2020.08.11.245902

Curse: Building expression atlases and co-expression networks from public RNA-Seq data. Vaneechoutte and Vandepoele (2019) Bioinformatics

TF2Network: predicting transcription factor regulators and gene regulatory networks in Arabidopsis using publicly available binding site information Kulkarni, Vaneechoutte, Van de Velde and Vandepoele (2018). Nucleic Acids Research