Coexpression and Collaborative Gene Networks

Hairong Wei

Michigan Technological University

References

1) Persson S. **H. Wei** J. Milne G. Page C. Somerville. 2005. Identification of genes required for cellulose synthesis by regression analysis of public microarray data sets. ***Proc Natl Acad Sci USA*** 102: 8633-8638. (Faculty 1000 evaluation)

2) Wei H S. Persson T. Mehta V. Srinivasasainagendra L. Chen G. Page C. Somerville A. Loraine. 2006. Transcriptional coordination of the metabolic network in Arabidopsis thaliana. ***Plant Physiology***. 142(2):762-74.

3) Nie J. R. Stewart F. Ruan J. Thomson H. Zhang X. Cui and **H. Wei**. 2011. TF-Cluster: a pipeline for identifying functionally coordinated transcription factors via network decomposition of the shared coexpression connectivity matrix (SCCM). ***BMC Systems Biology*** 5:53.

4) Ji X. S. Chen J. Li W. Deng Z. Wei and **H. Wei**. 2017. SSGA and MSGA: two seed-growing algorithms for constructing collaborative subnetworks. ***Scientific Reports.*** (in press)

5) Kumari, S., J. Nie,, H.S. Chen, R. Stewart, H. Ma, X. Li, M. Lu, W.M. Taylor, and H. Wei. 2012. Evaluation of gene association methods for coexpression network construction and biological knowledge discovery. ***PLoS ONE*** *7(11): e50411*

What is coexpression network ?

A **gene co-expression network (GCN)** is an undirected graph where each node corresponds to a gene and a pair of nodes is connected with an edge if there is a significant co-expression relationship between them.

Methods for associating co-expressed genes

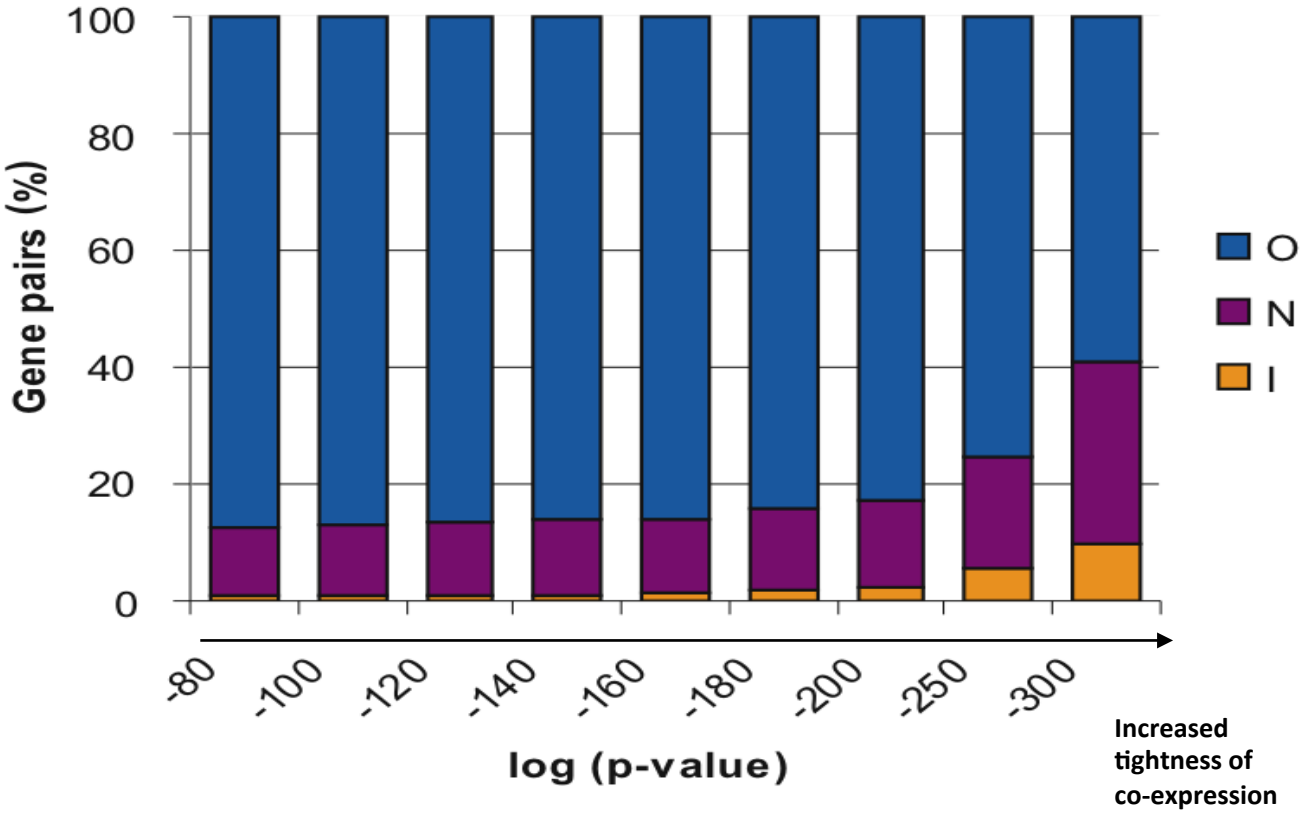
1. Spearman rank correlation
2. Weighted Rank Correlation
3. Kendall rank correlation
4. Hoeffding's D measure
5. Theil-Sen
6. Rank Theil-Sen
7. Distance Covariance
8. Pearson

Kumari et al 2012, PloS ONE. The R codes for above gene associate methods can be found at:

http://journals.plos.org/plosone/article/file?type=supplementary&id=info:doi/10.1371/journal.pone.0050411.s005
Supplemental Doc S1

Are Genes involved in the same biological pathway coexpressed?

Hypothesis: The genes having functional association e.g. in the same pathway are more tightly co-expressed.

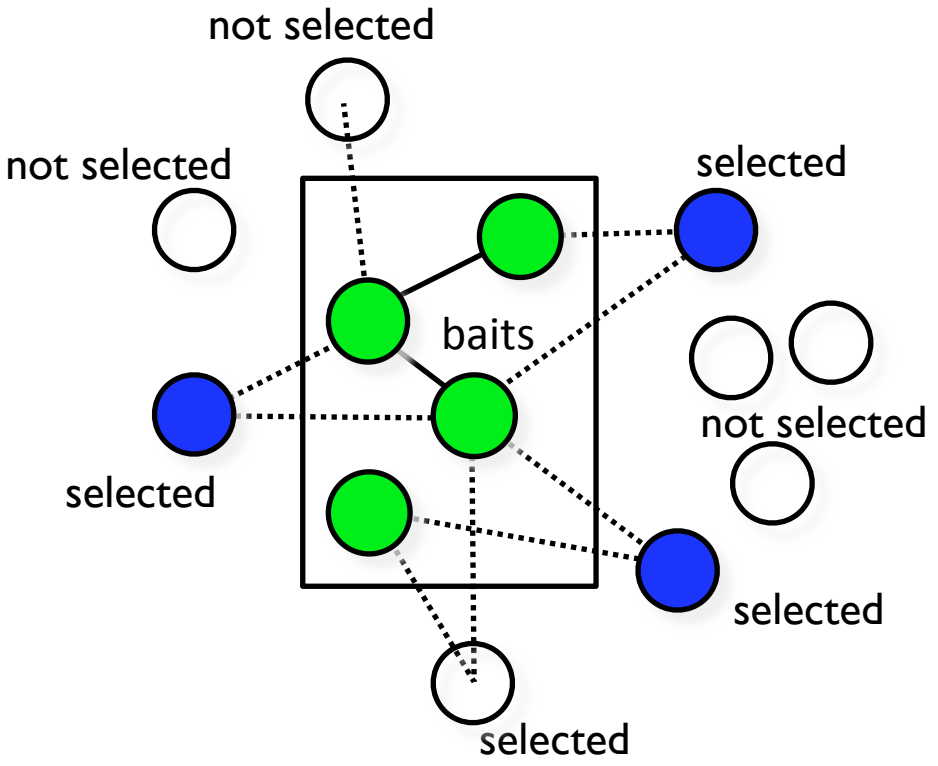


216 biological pathways that contain 1330 genes in Arabidopsis

O-outside
N-different pathway
I-same pathway

(Wei et al. 2006. *Plant Physiol.* 142(2): 762–774)

Intersection of Co-Expression (ICE)



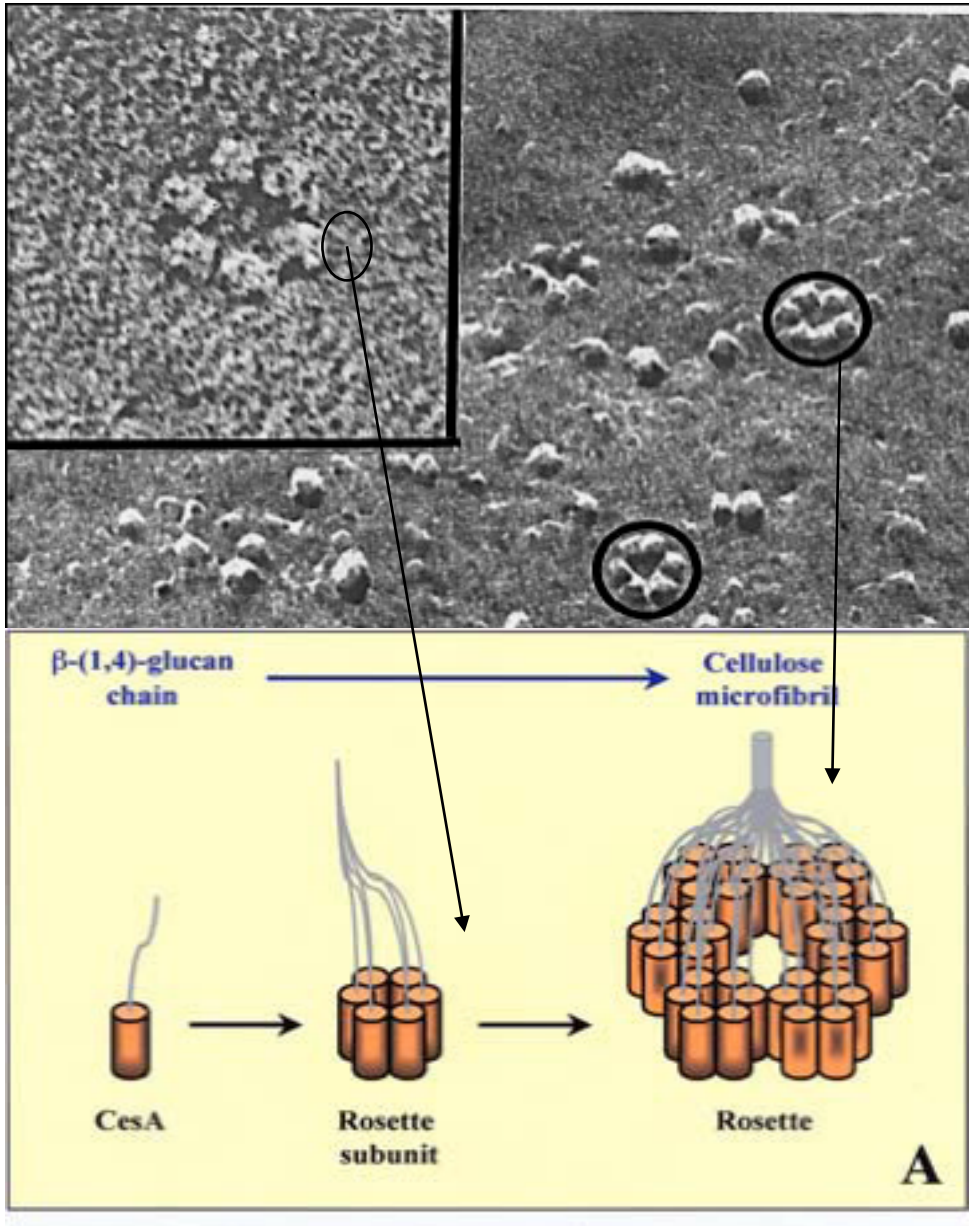
ranked list (|*G*|average p value)

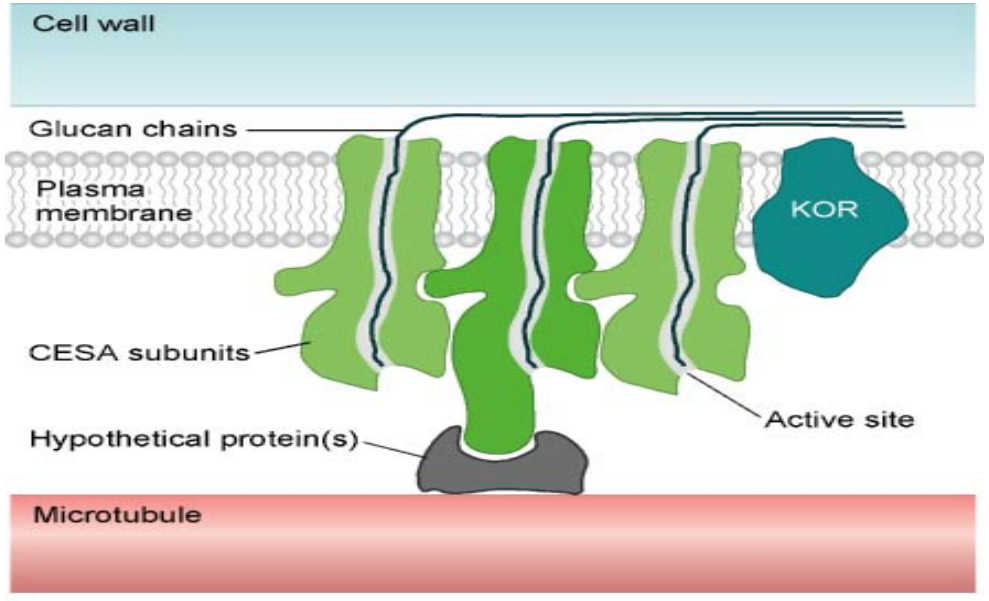
GO analysis homology searches

1. Define co-expression network.
2. Select a group of "bait" genes. (e.g. a pathway a complex). Green nodes
3. Perform genome-wide co-expression analysis (correlation or regression).
3. Select genes outside the group of baits that are linked with >1 group members *G*.
4. Sort them by |G| then average p value (or r^2)

(Wei et al. 2006. *Plant Physiol.* 142(2): 762–774)

ICE Implementation 1: finding functionally associated genes

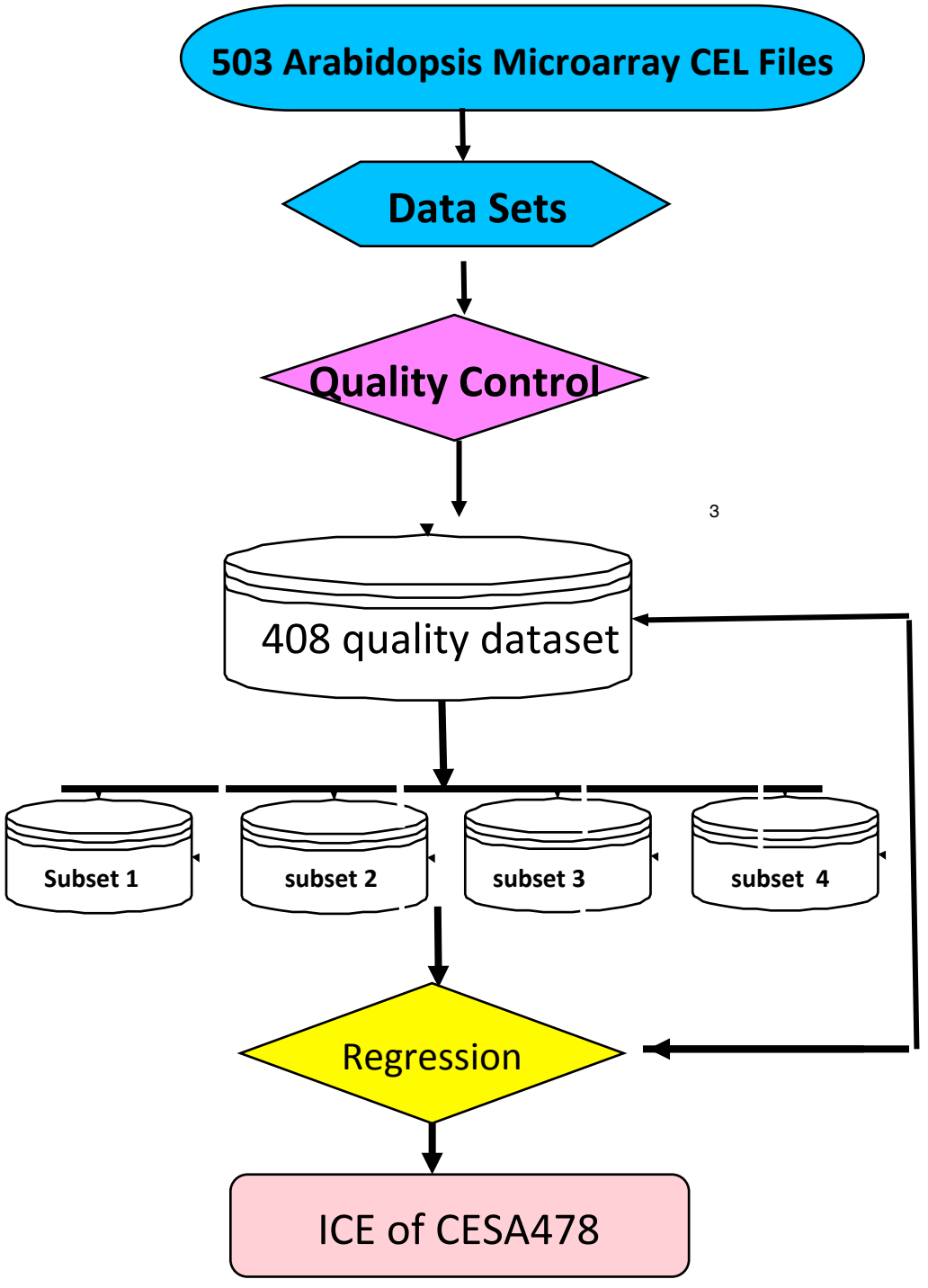


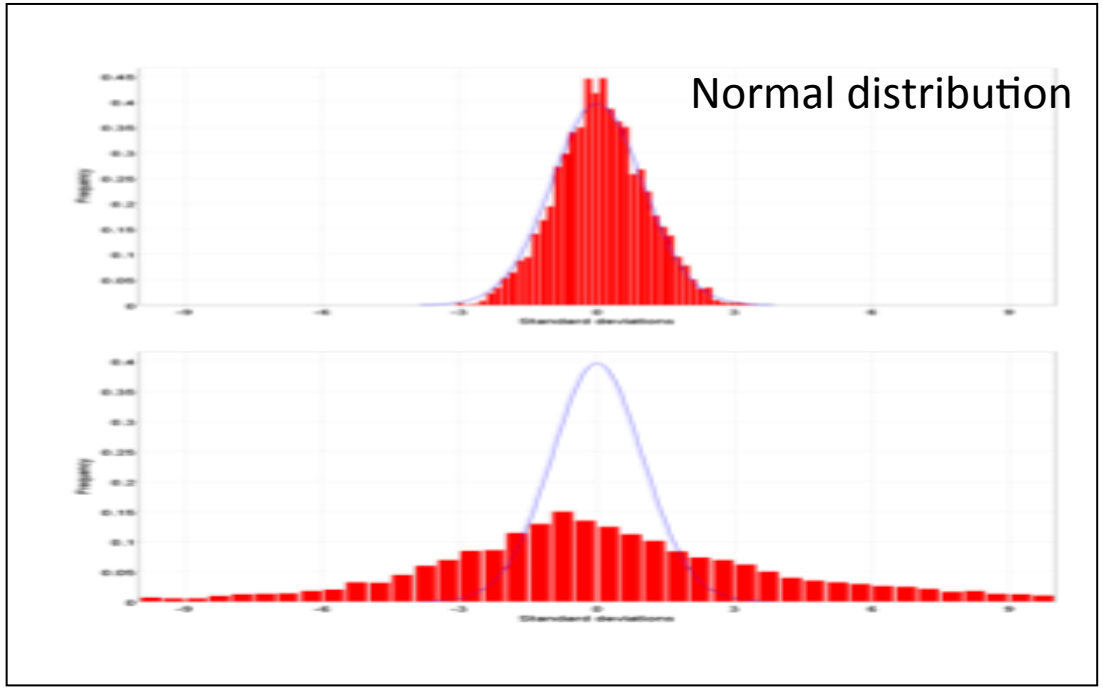


A very challenging problem:

What genes in addition to CEAS genes (4, 7 and 8) are required for secondary cell wall synthesis?

Primary: CESA136 Secondary: CESA4 7 8





For n genes and m data sets deleted residual

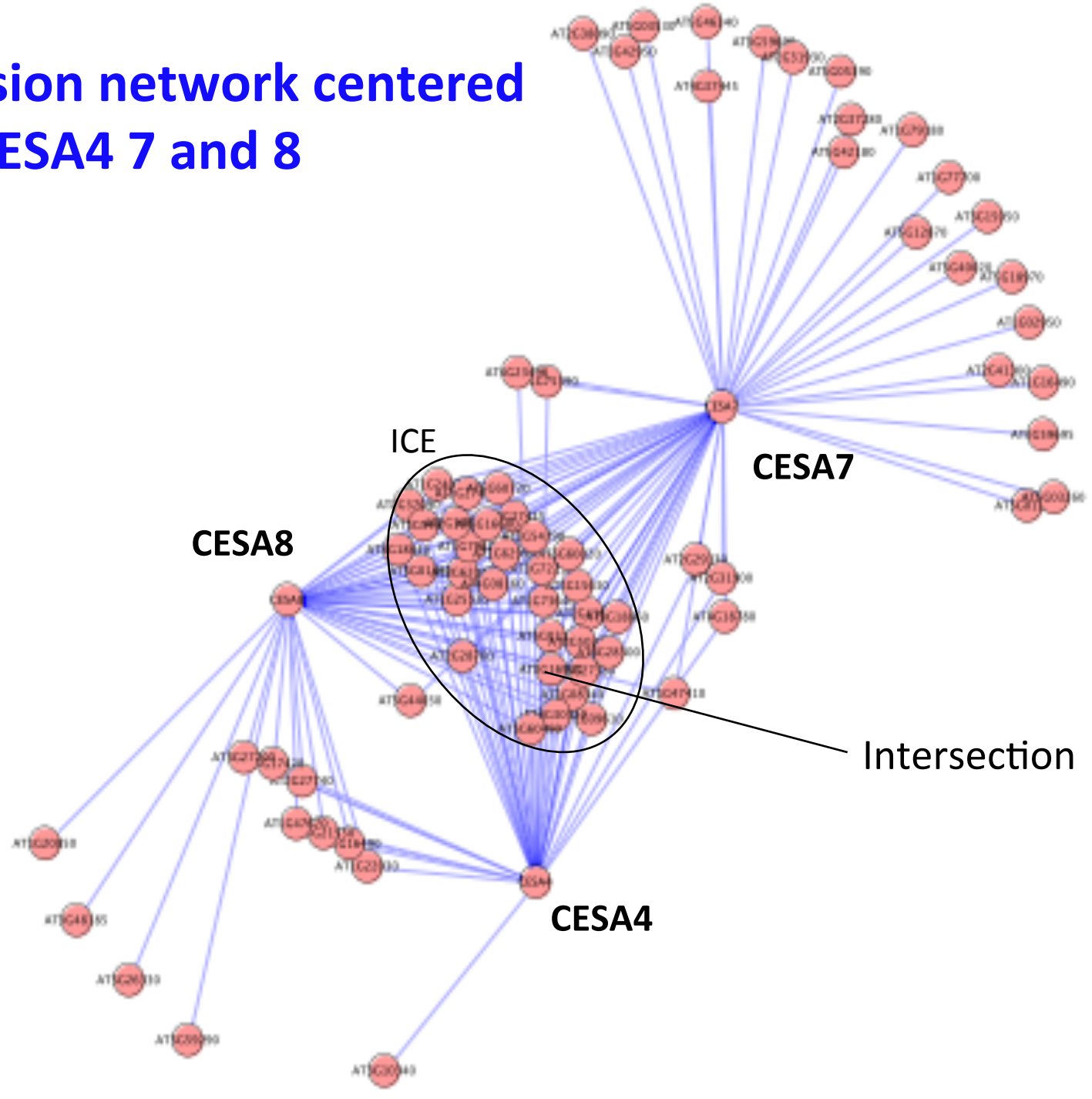
$$d_{ij} = X_{ij} - \bar{X}_{i\neg j},$$

$$d_i^* = d_i / s(d_i)$$

obeys t distribution with m-2 df

(Persson and Wei et al 2005 *PNAS* Vol102:8633)

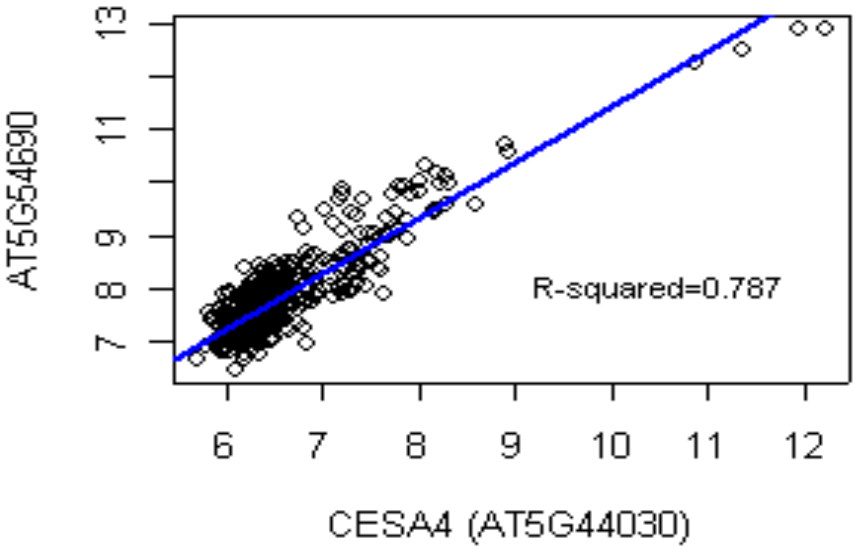
A coexpression network centered around AtCESA4 7 and 8

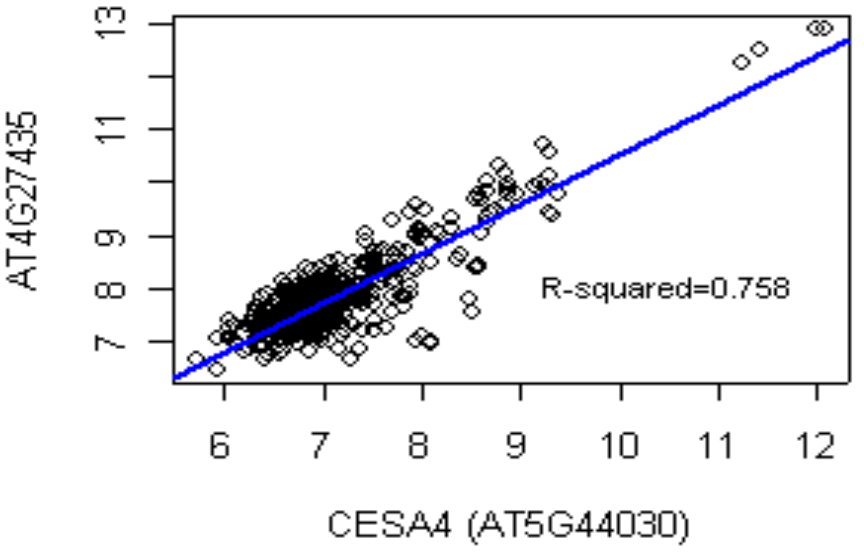


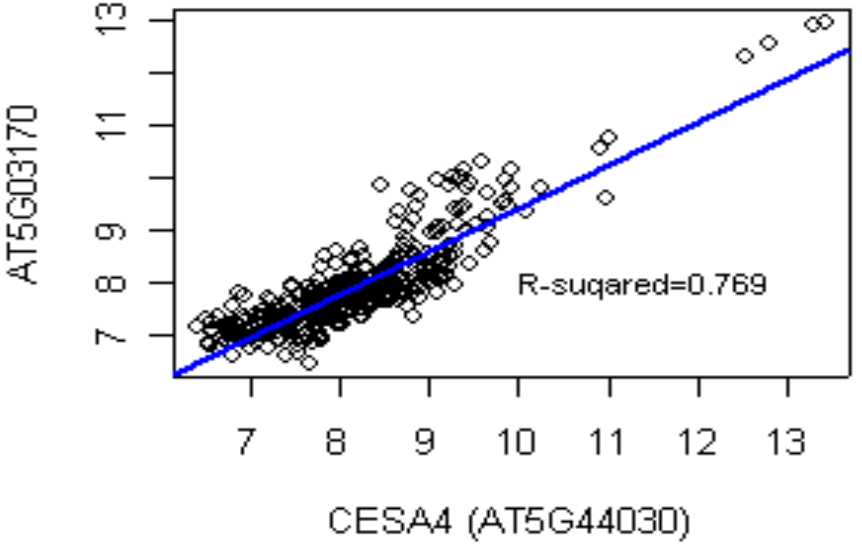
Genes selected from ICE

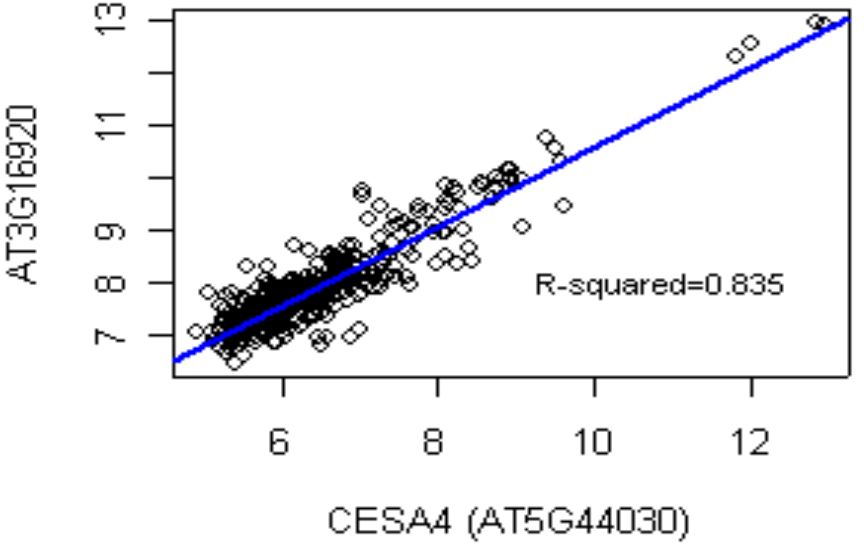
AGI	Gene title	Score*	P-value
AT5G44030	cellulose synthase catalytic subunit (IRX5/CESA4)	4	4.1E-35
AT4G18780	cellulose synthase catalytic subunit (IRX1/CESA8)	5	7.1E-33
AT5G54690*	glycosyl transferase family 8 protein	5	4.1E-35
AT5G17420	cellulose synthase catalytic subunit (IRX3/CESA7)	5	3.7E-30
AT3G16920*	CTL1-like	6	9.3E-33
AT2G38080	laccase putative / diphenol oxidase putative	6	1.4E-34
AT5G15630	COBRA cell expansion protein COBL4	7	1.6E-32
AT5G03170*	fasciclin-like arabinogalactan-protein (FLA11)	8	6.7E-32
AT2G37090	glycosyl transferase family 43 protein	10	2.3E-29
AT3G18660	glycogenin glucosyltransferase (glycogenin)-related	29	4.9E-23
AT4G27435*	expressed protein	30	2.6E-24
AT5G60720	expressed protein	37	1.2E-20
AT3G62020	germin-like protein (GLP10)	42	9.2E-21
AT5G60020	laccase putative / diphenol oxidase putative	58	1.1E-17
AT4G28500	no apical meristem (NAM) family protein	62	9.7E-24
AT5G60490	fasciclin-like arabinogalactan-protein (FLA12)	64	7.9E-21

These genes highlighted in blue are four selected for experiment validation

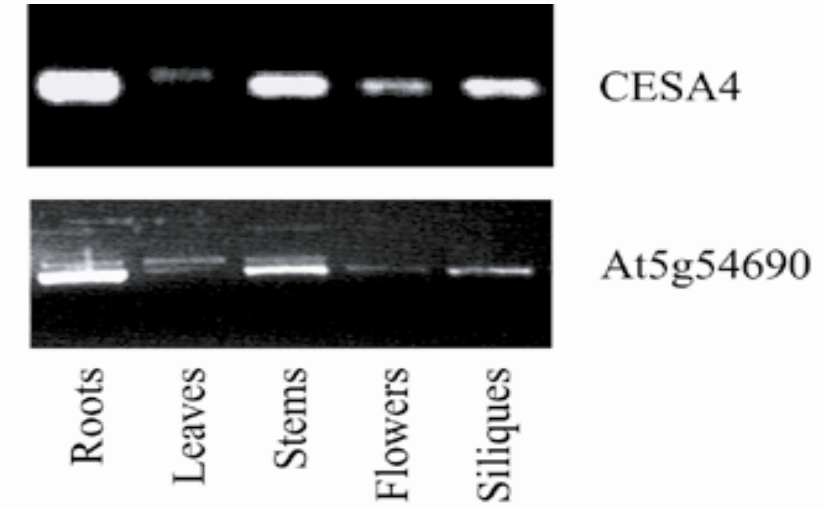

AT5G54690
CESA4 (AT5G44030)
R-squared=0.787

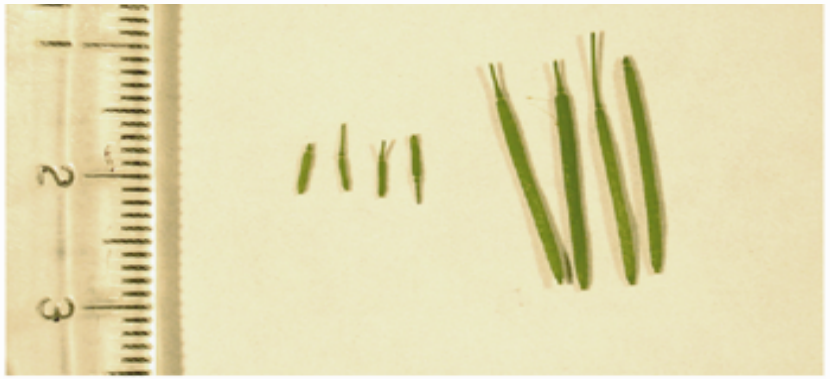

AT4G27435
CESA4 (AT5G44030)
R-squared=0.758


AT5G03170
CESA4 (AT5G44030)
R-suqared=0.769

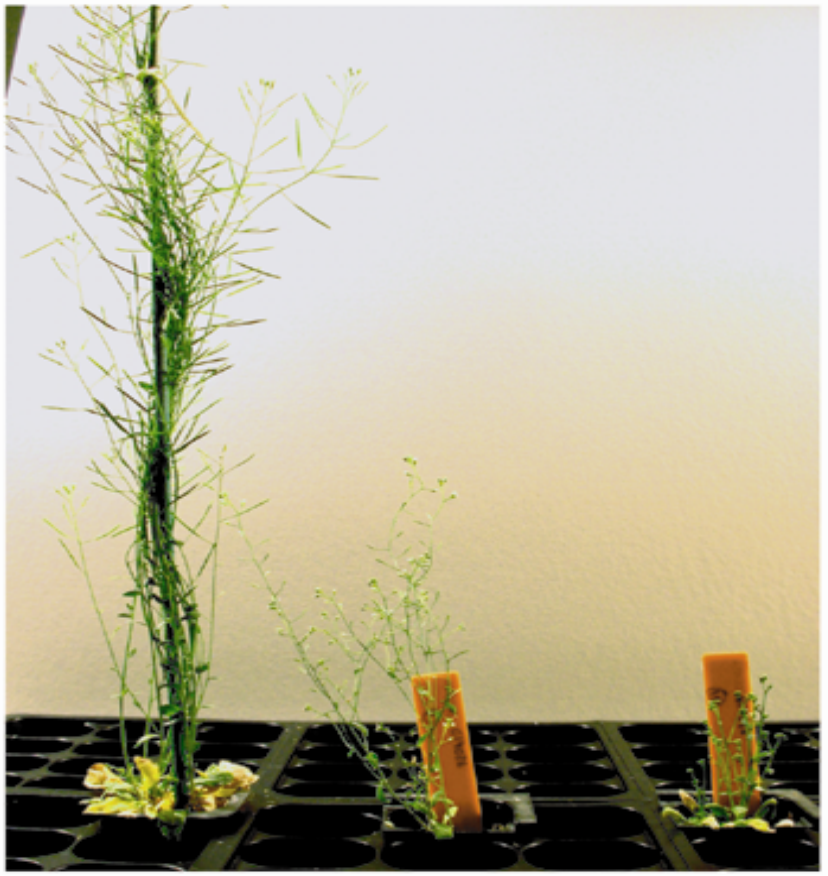

AT3G16920
CESA4 (AT5G44030)
R-squared=0.835

Phenotypic changes in mutant lines

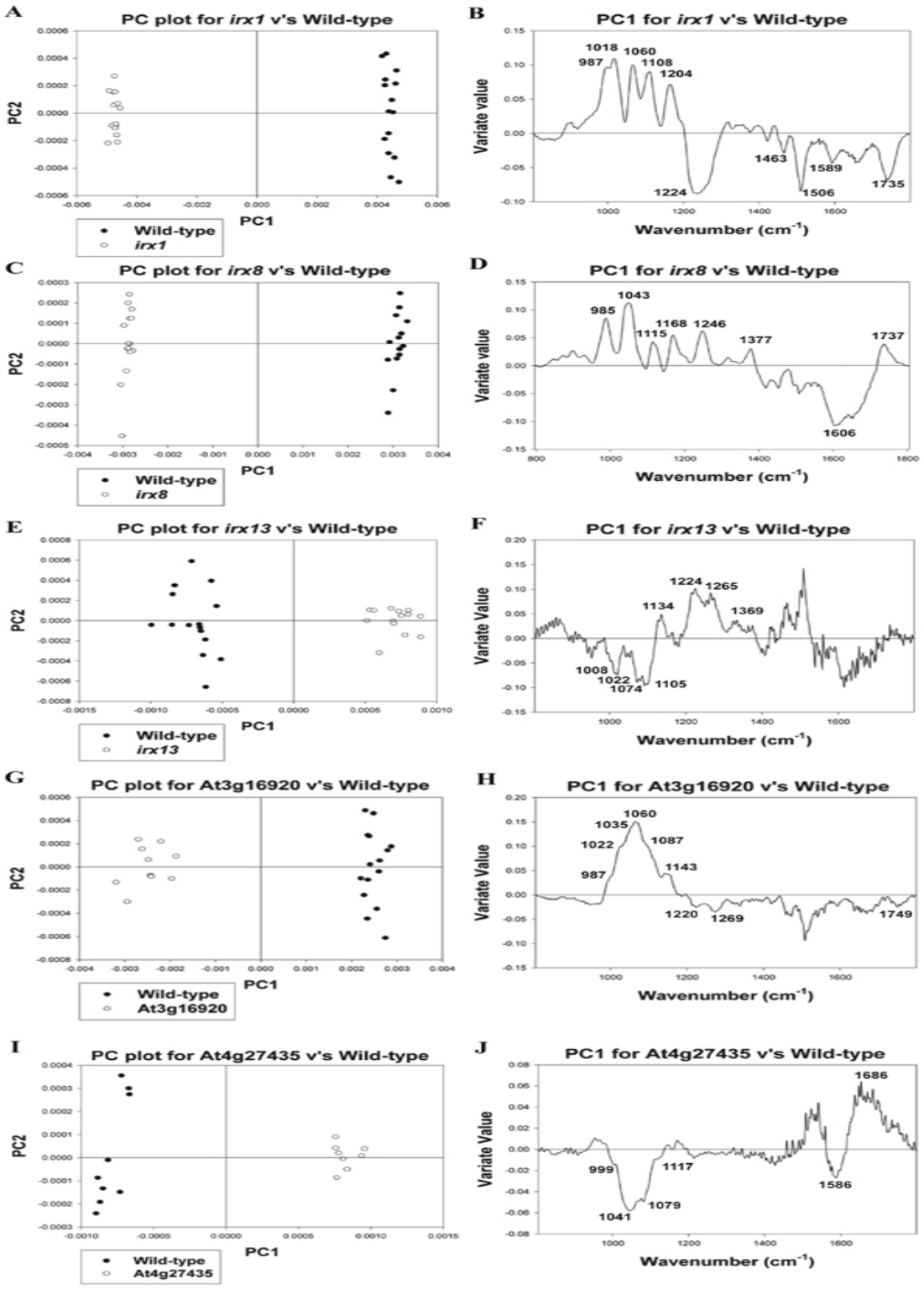




At5g54690 Wild-type



Wild-type At5g54690_1 At5g54690_2

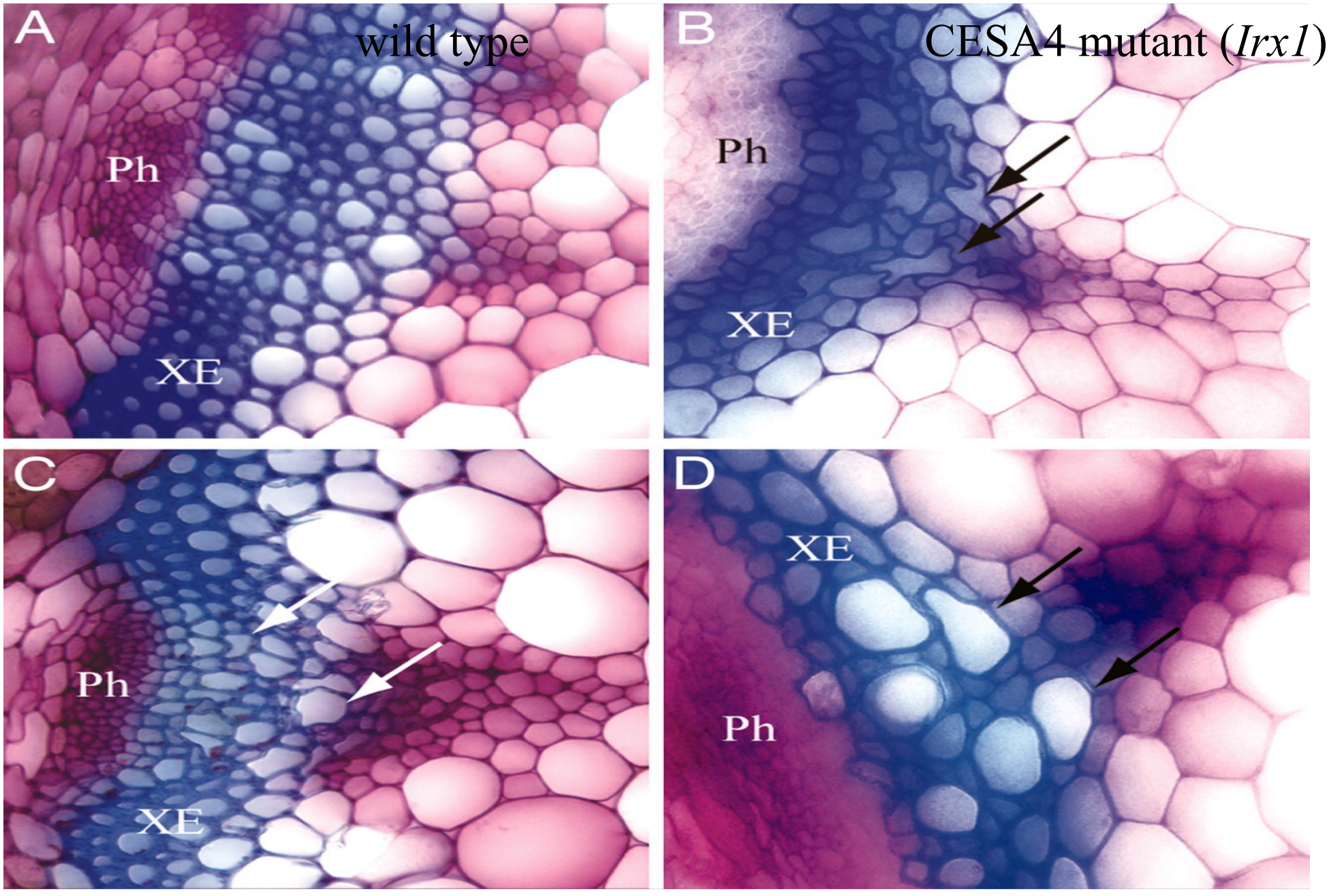


FTIR Analysis

Fourier Transform Infrared Spectroscopy s a technique which is used to obtain an infrared spectrum of absorption or emission of a solid liquid or gas.

Cell wall analyses by FTIR. (*A*-*F*) Principal components plots for mutant [*irx1 (CESA4 mutant) irx8* (SALK_014026 At5g54690) *irx13* (SALK_046976 At5g03170) At3g16920 (SALK_055713) and At4g27435 (SALK_137109)] vs. wild-type spectra.

Spectra for all mutants show a clear separation from wild-type spectra based on principal component (PC) 1. In all cases the mutant spectra show differences from wild-type spectra in the carbohydrate fingerprint regions that correspond to the deformations of cellulosic and noncellulosic polymers.



At5g54690 knockout (*Irx8*)

At5g03170 knockout (*Irx13*)

Related studies

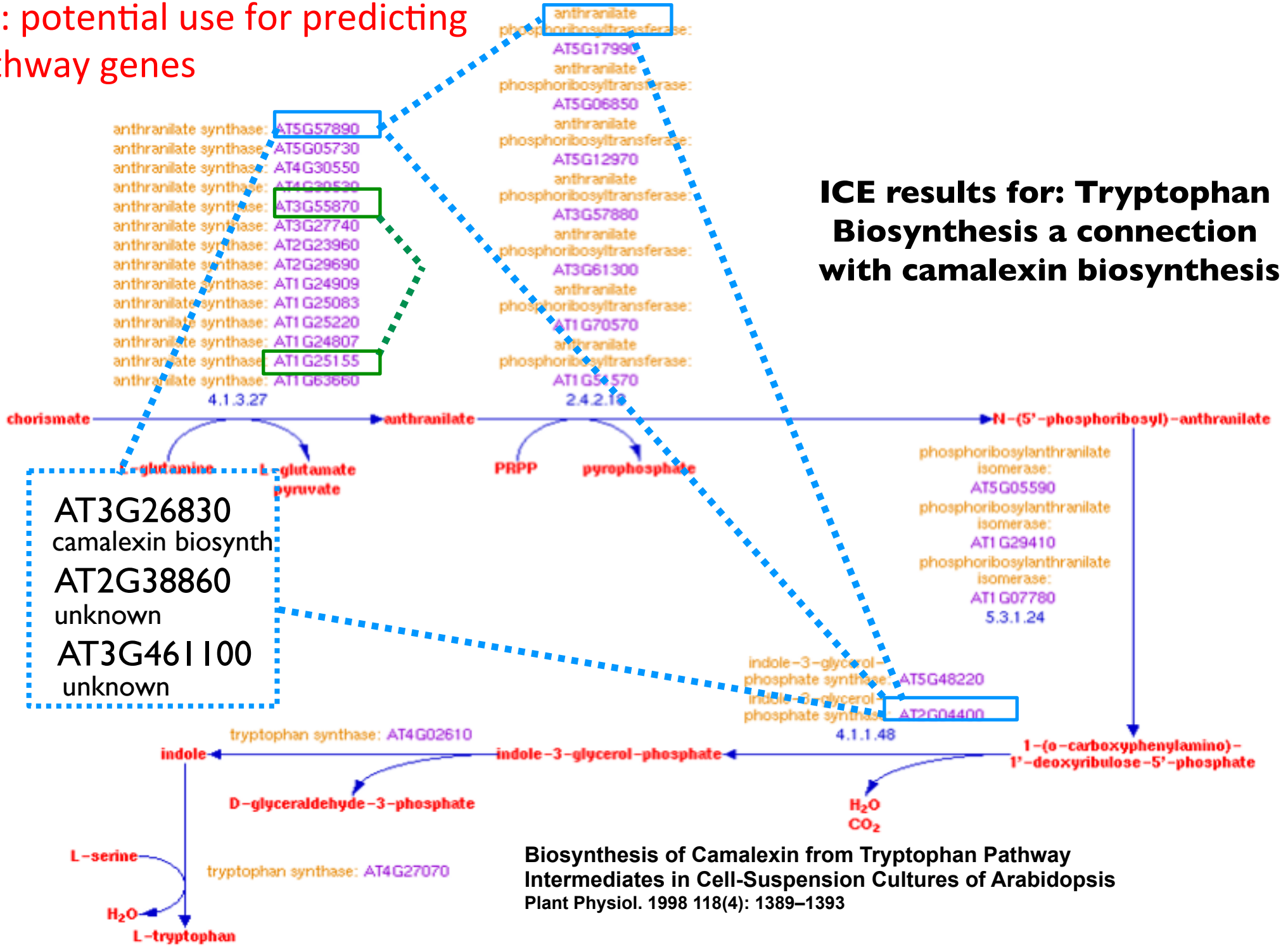
1. **Among the 25 genes 16 are the same as those genes identified by Dr. Simon Turner via forward genetics**

   Brown DM Zeef LA Ellis J Goodacre R Turner SR (2005) Identification of novel genes in Arabidopsis involved in secondary cell wall formation using expression profiling and reverse genetics. *The Plant cell* ***17(8): 2281-95.***

2. ***Irx8*** **(**At5g54690**) protein has been shown to target to Golgi where glucuronoxylans (GXs) are synthesized. *Irx8* is implicated in affecting GX synthesis. GX cellulose and lignin are three major components of secondary cell walls in woody plants.**

   Pena et al Arabidopsis irregular xylem8 and irregular xylem9: Implications for the Complexity of Glucuronoxylan Biosynthesis The Plant Cell 19:549-563 (2007)

ICE: potential use for predicting pathway genes



ICE results for: Tryptophan Biosynthesis a connection with camalexin biosynthesis

Biosynthesis of Camalexin from Tryptophan Pathway Intermediates in Cell-Suspension Cultures of Arabidopsis
Plant Physiol. 1998 118(4): 1389–1393

Summary for coexpression network:

1. Co-expression network can be used to predict genes involved in various pathways and biological processes. It is especially suitable for predicting functionally associated genes metabolic pathway genes structural genes that are usually located at the low hierarchical terminals of gene regulatory networks. These genes are often co-regulated which is why they are often highly correlated.

2. Due to the large number of terminal genes the identification of true genes is often interfered by noise. The prediction of pathway genes in many cases is not accurate. However it indeed works in some cases.

What is gene collaborative network ?

A **gene collaboration network (GCN)** is an undirected graph where each node corresponds to a regulatory gene (e.g. TF) and a pair of nodes is connected with an edge if there is a significant **collaboration** relationship between them.

Microarray or RNA-seq data

Gene	T1	T2	T3	T4	T5	T6	T7	T8	T9	T10
G1	123	790	2300	1597	655	530	321	229	234	354
G2	400	750	790	996	213	350	1504	340	250	1780
G3	250	250	380	150	470	559	890	1250	99	450
…	…	…	…	…	…	…	…	…	…	…
G n-1	550	225	388	650	470	559	899	250	590	460
G n	150	260	950	750	470	585	865	125	99	450

Hypothesis: Regulatory genes (e.g. TFs) controlling a biological process or complex trait are collaborative.

Identification of Regulatory Genes That Collectively Govern a Biological Process or Complex Trait?

NETWORK CONSTRUCTION

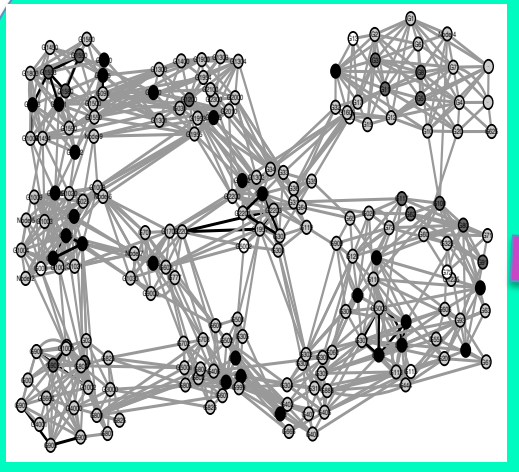
Gene	T1	T2	T3	T4	T5	T6	T7	T8	T9	T10
G1	123	790	2300	1597	655	530	321	229	234	354
G2	400	750	790	996	213	350	1504	340	250	1780
G3	250	250	380	150	470	559	890	1250	99	450
...	...	...	...	...	...	...	...	...	...	...
G n-1	550	225	388	650	470	559	899	250	590	460
G n	150	260	950	750	470	585	865	125	99	450

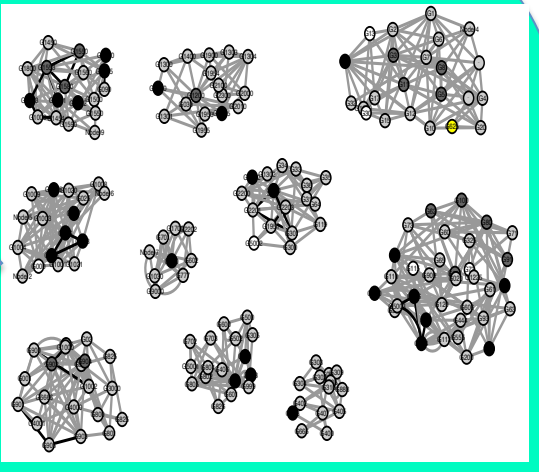
Expression data

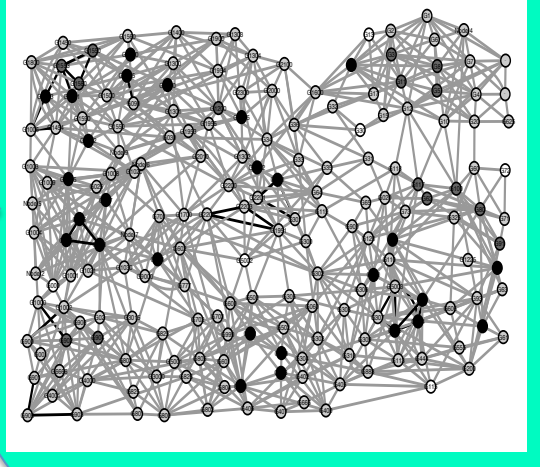
	TF1	TF2	TF3	TF4	TF5	TF6	TF7	TF8	TF9	TF10	TF11	TF12
TF1	00	50	00	00	00	30	00	00	40	00	00	00
TF2	50	00	00	00	32	00	00	00	00	00	00	55
TF3	00	00	00	00	00	00	39	00	00	00	00	00
TF4	00	00	00	00	00	00	00	00	60	00	00	00
TF5	00	32	00	00	00	00	00	00	00	36	00	00
TF6	30	00	00	00	00	00	00	00	38	00	00	00
TF7	00	00	39	00	00	00	00	00	00	00	00	40
TF8	00	00	00	00	00	00	00	00	00	00	50	00
TF9	40	00	00	00	00	38	00	00	00	00	00	00
TF10	00	00	00	00	36	00	00	00	00	00	00	00
TF11	00	00	00	00	00	00	50	00	00	00	00	00
TF12	00	55	00	00	00	00	40	00	00	00	00	00

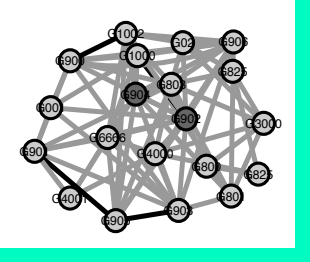
SCCM--Shared Coexpression Connectivity Matrix

NETWORK DECOMPOSITION









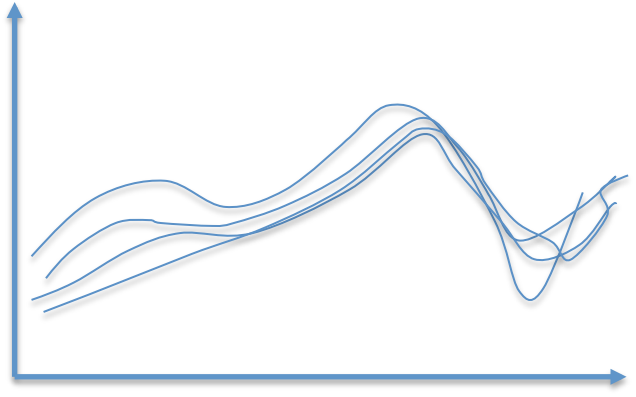
TFs control a trait or a biological process

How can we identify regulatory genes that are collaborative?

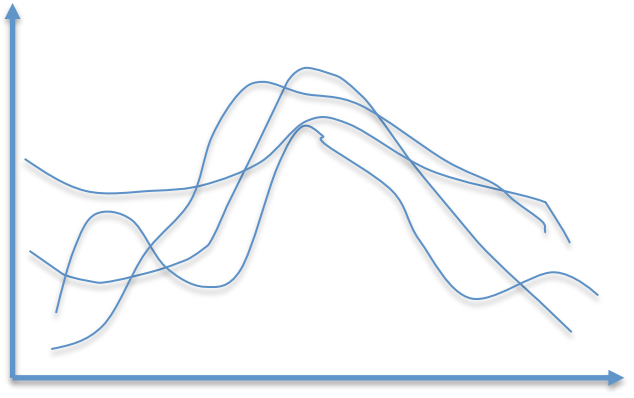
Hypothesis

Regulatory genes that collaboratively control the same complex trait are **coordinated in gene expression profiles**.

Coexpressed genes refer to those that have higher concordance in expression profiles whereas coordinated genes refer to those that have relatively looser concordance in expression profiles.

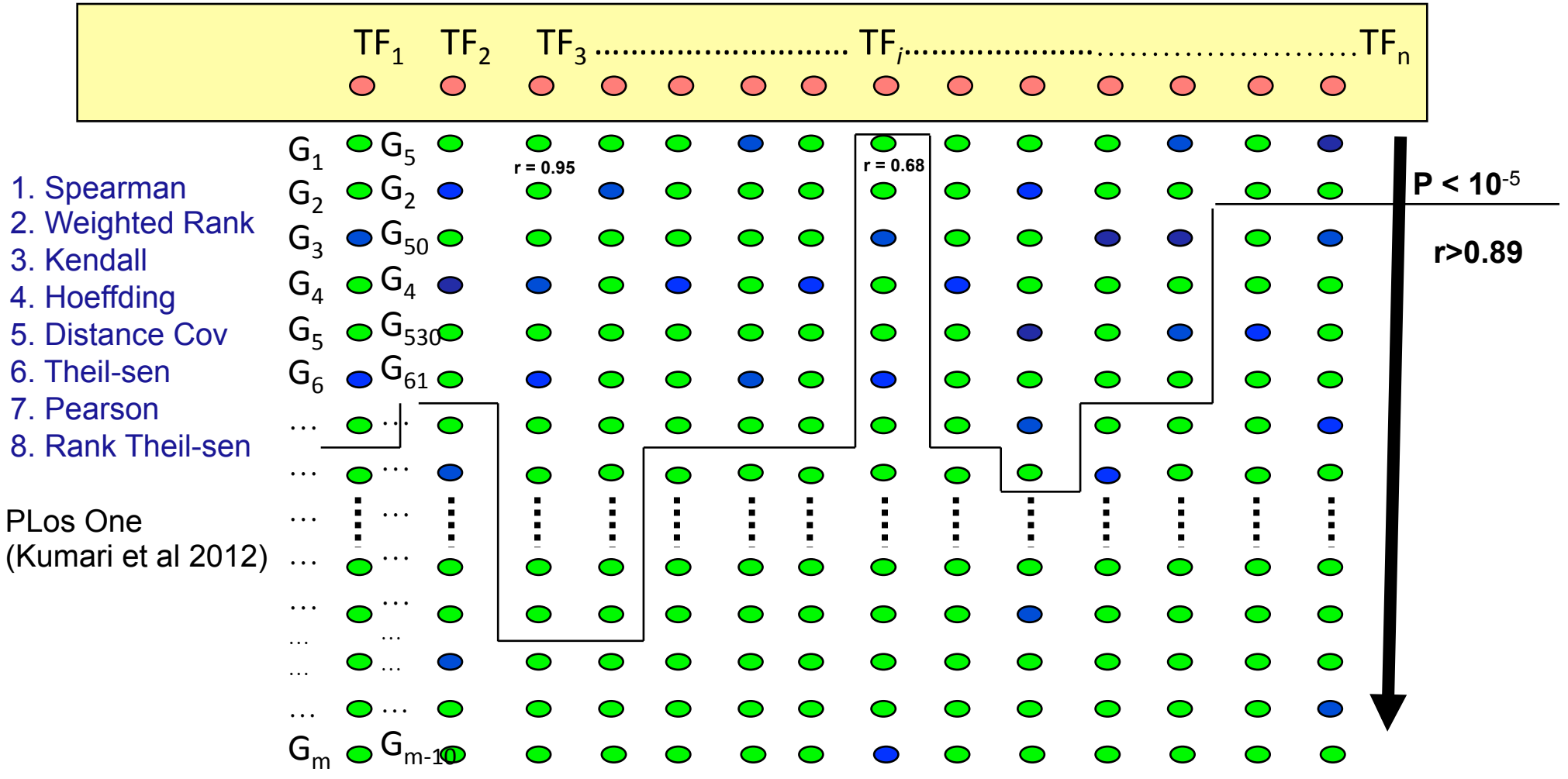


Coexpressed genes



Collaborative genes

Genome-wide coexpression analysis



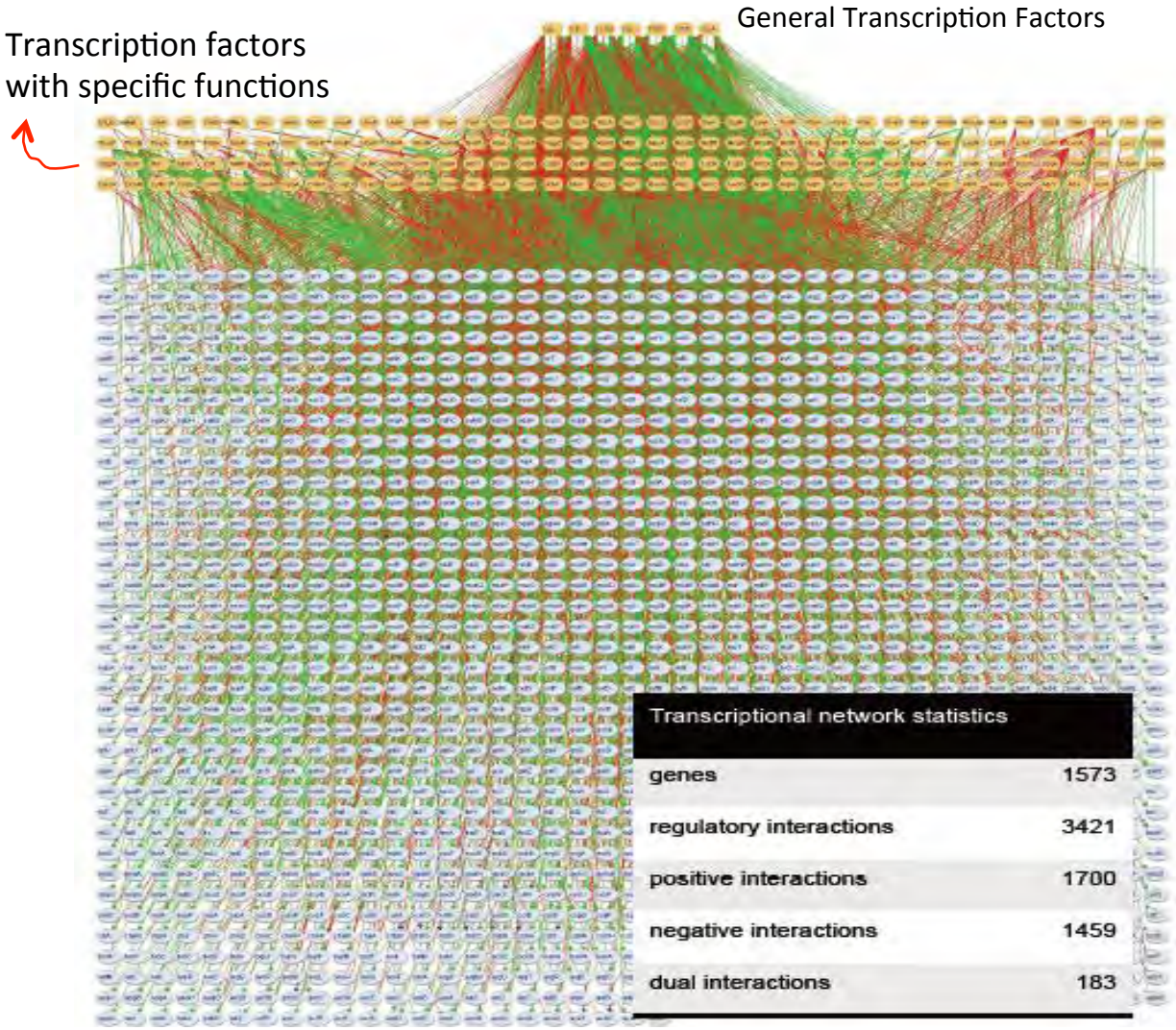
Where m > 20000 while n = 1600 ~ 5000

Functionally coordinated genes Others

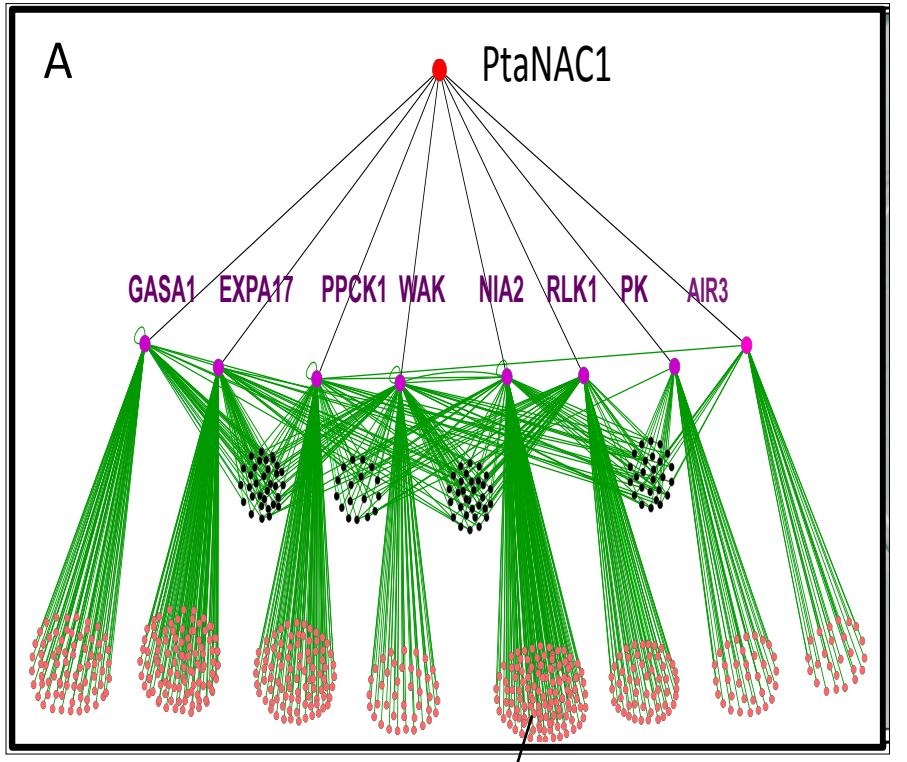
Examples for illustration of TFs' hierarchies and their possible impact on coexprssion

Escherichia coli
transcriptional regulatory network

Network Biology 2011, 1(1): 21-33

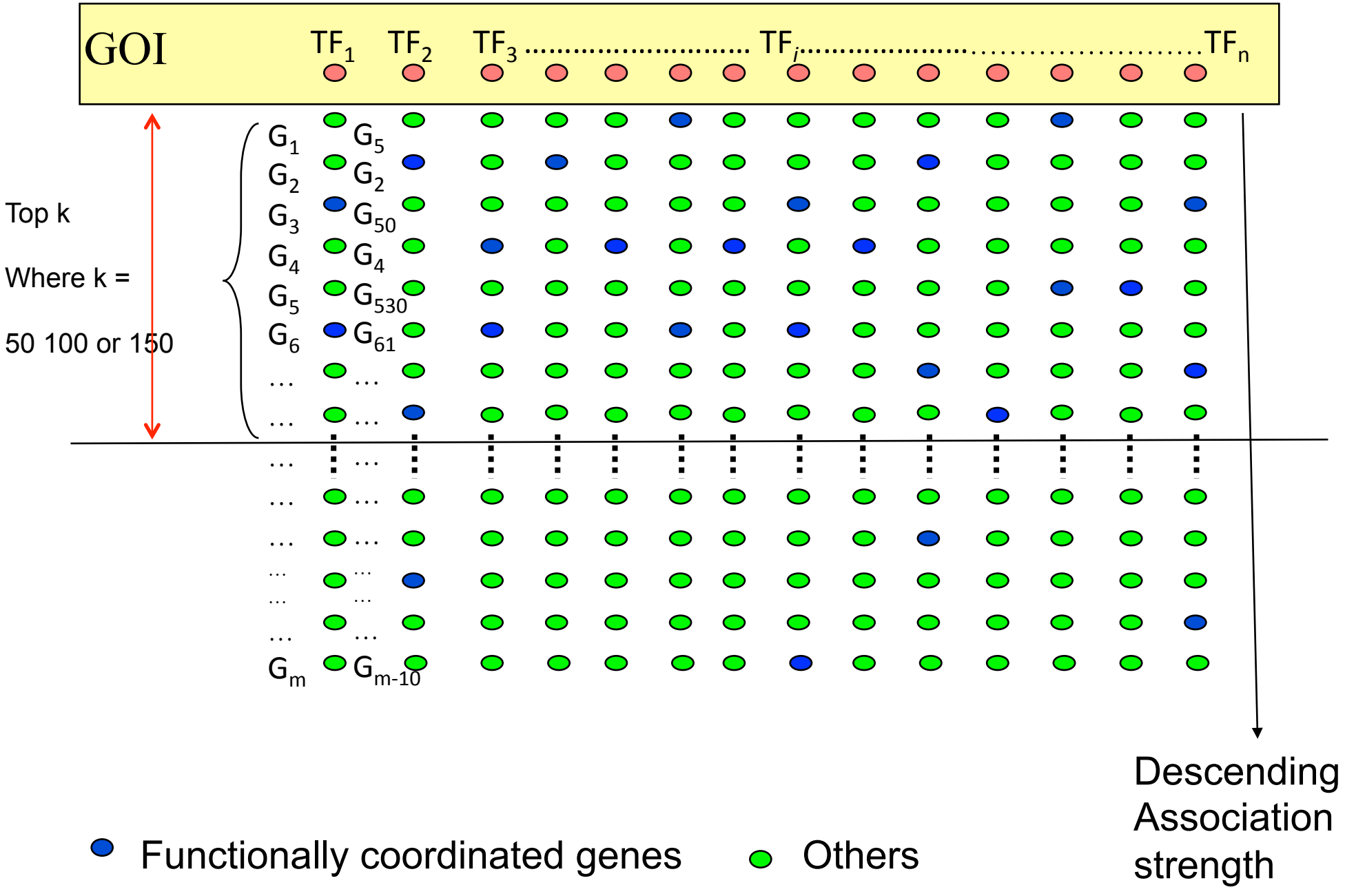


A Hierarchical Gene Regulatory Network in poplar



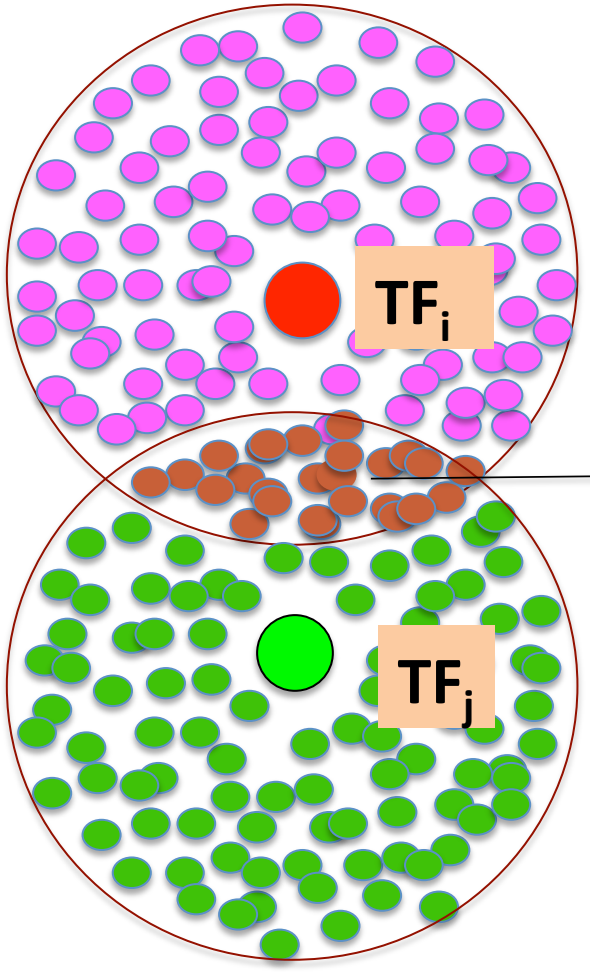
Terminal genes

Genome-wide coexpression analysis



How to measure the loosely coordination between two TFs?

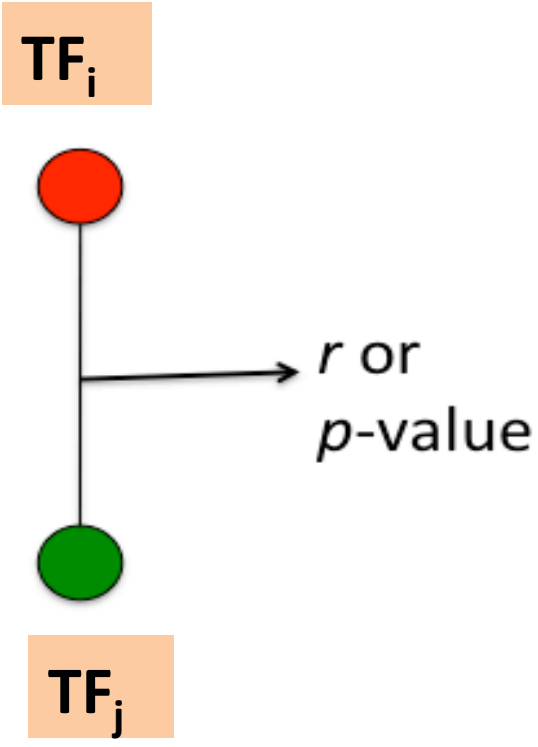
100 genes that are
most tightly coexpressed to TF_1



100 genes that are
most tightly coexpressed to TF_j

n_c

Number of
shared genes



Kumari S Nie J Chen H-S Ma H et al. (2012) Evaluation of Gene Association Methods for Coexpression Network Construction and Biological Knowledge Discovery. PLoS ONE 7(11): e50411. doi:10.1371/journal.pone.0050411
http://www.plosone.org/article/info:doi/10.1371/journal.pone.0050411

Shared Coexpression Connectivity Matrix (SCCM)

TFs

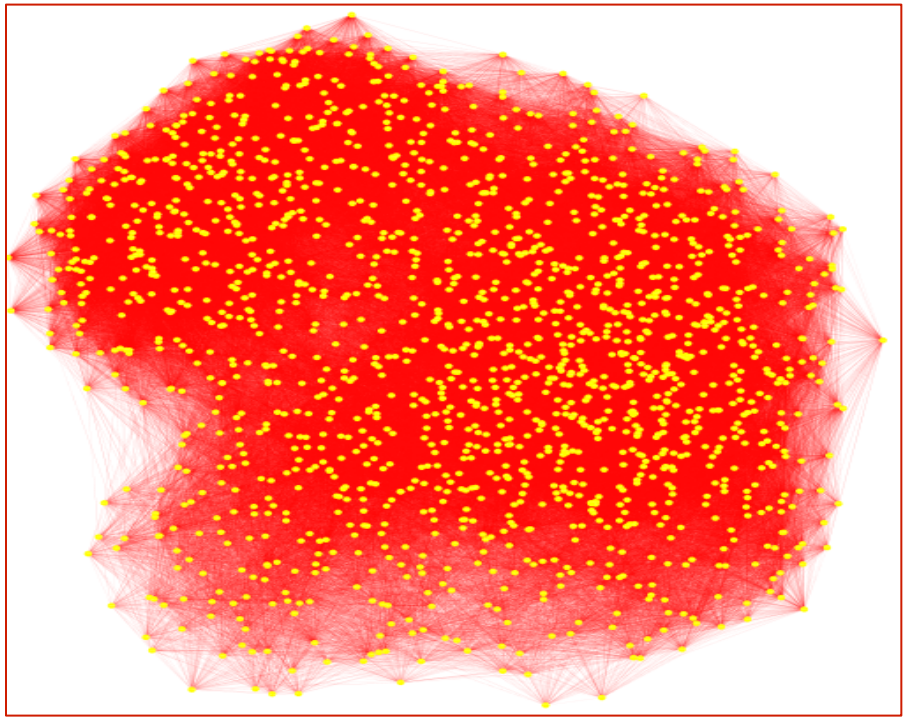
TFs

	AATF	ABCA11	ABCA11	ABCA11	ABT1	ACVR2A	ADNP	ADPGK	AEBP1	AEBP2	AFF1
AATF	100	0	0	0	3	0	0	0	0	0	0
ABCA11	0	100	0	13	0	0	0	0	0	0	0
ABCA11	0	0	100	11	0	0	0	0	0	0	0
ABCA11	0	13	11	100	0	0	0	0	0	0	0
ABT1	3	0	0	0	100	0	0	0	0	0	11
ACVR2A	0	0	0	0	0	100	8	8	0	0	0
ADNP	0	0	0	0	0	8	100	0	0	0	0
ADPGK	0	0	0	0	0	8	0	100	0	0	0
AEBP1	0	0	0	0	0	0	0	0	100	0	0
AEBP2	0	0	0	0	0	0	0	0	0	100	0
AFF1	0	0	0	0	11	0	0	0	0	0	100
AFF2	0	0	0	0	3	0	0	0	0	0	11
AFF3	0	0	0	0	0	0	0	0	0	0	0
AHR	0	0	0	0	0	0	0	0	0	0	0
AIP	3	0	0	0	13	0	0	0	0	0	15
AIP	0	0	0	0	0	0	0	0	0	4	0
AIRE	0	0	0	0	0	0	0	0	0	0	0
AK096318	0	0	0	0	0	0	0	0	0	0	0
AK126463	0	0	0	0	0	0	0	0	5	0	0
AL834146	0	0	0	0	0	4	11	3	0	0	0
ALF	0	0	0	0	0	0	0	0	0	0	0
ALX3	0	0	0	0	0	0	0	0	6	0	0
ALX4	0	0	0	0	0	0	0	0	0	0	0
AMOT	0	0	0	0	0	0	0	0	0	0	0
ANKRD30A	0	0	0	0	0	0	0	0	0	0	0
ANP32A	0	0	0	0	0	0	0	0	0	0	0
APBA2BP	0	0	0	0	0	0	0	0	0	0	0
APEX1	0	0	0	0	0	0	0	0	0	0	0
AR	0	0	0	0	0	0	0	0	0	0	0
ARC	0	0	0	0	0	0	0	0	0	0	0
ARID1A	0	0	0	0	0	0	17	0	0	0	0
ARID1B	0	0	0	0	8	0	0	0	0	0	19

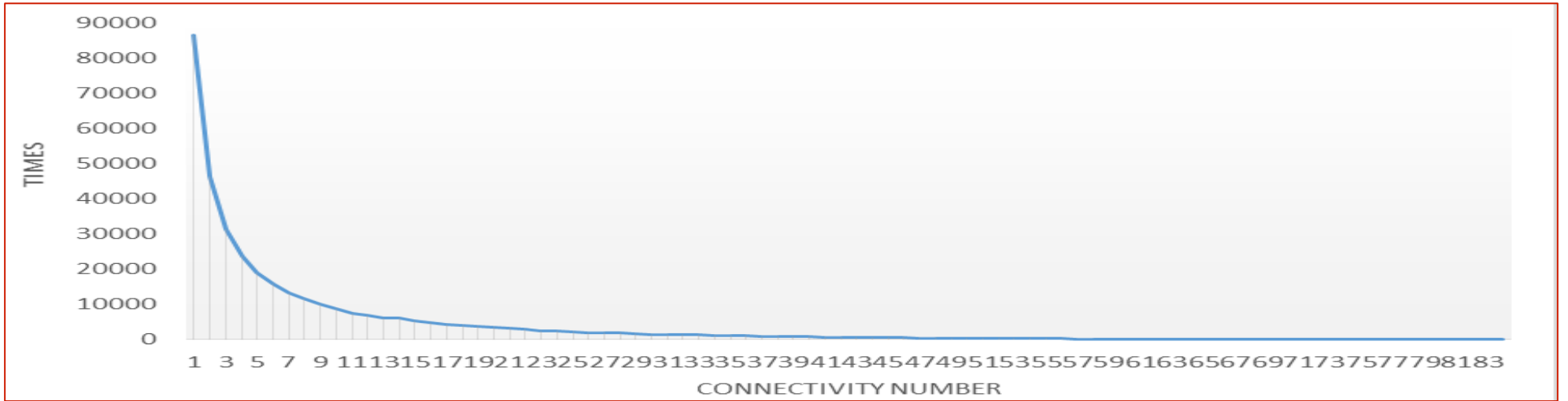
Collaborative Network of TF

	AATF	ABCA11	ABCA11	ABCA11	ABT1	ACVR2A	ADNP	ADPGK	AEBP1	AEBP2	AFF1
AATF	100	0	0	0	3	0	0	0	0	0	0
ABCA11	0	100	0	13	0	0	0	0	0	0	0
ABCA11	0	0	100	11	0	0	0	0	0	0	0
ABCA11	0	13	11	100	0	0	0	0	0	0	0
ABT1	3	0	0	0	100	0	0	0	0	0	11
ACVR2A	0	0	0	0	0	100	8	8	0	0	0
ADNP	0	0	0	0	0	8	100	0	0	0	0
ADPGK	0	0	0	0	0	8	0	100	0	0	0
AEBP1	0	0	0	0	0	0	0	0	100	0	0
AEBP2	0	0	0	0	0	0	0	0	0	100	0
AFF1	0	0	0	0	11	0	0	0	0	0	100
AFF2	0	0	0	0	3	0	0	0	0	0	11
AFF3	0	0	0	0	0	0	0	0	0	0	0
AHR	0	0	0	0	0	0	0	0	0	0	0
AIP	3	0	0	0	13	0	0	0	0	0	15
AIP	0	0	0	0	0	0	0	0	0	4	0
AIRE	0	0	0	0	0	0	0	0	0	0	0
AK096318	0	0	0	0	0	0	0	0	0	0	0
AK126463	0	0	0	0	0	0	0	0	5	0	0
AL834146	0	0	0	0	0	4	11	3	0	0	0
ALF	0	0	0	0	0	0	0	0	0	0	0
ALX3	0	0	0	0	0	0	0	0	6	0	0
ALX4	0	0	0	0	0	0	0	0	0	0	0
AMOT	0	0	0	0	0	0	0	0	0	0	0
ANKRD30A	0	0	0	0	0	0	0	0	0	0	0
ANP32A	0	0	0	0	0	0	0	0	0	0	0
APBA2BP	0	0	0	0	0	0	0	0	0	0	0
APEX1	0	0	0	0	0	0	0	0	0	0	0
AR	0	0	0	0	0	0	0	0	0	0	0
ARC	0	0	0	0	0	0	0	0	0	0	0
ARID1A	0	0	0	0	0	0	17	0	0	0	0
ARID1B	0	0	0	0	8	0	0	0	0	0	19

shared coexpression connectivity matrix (SCCM)



Collaborative network of all TFs



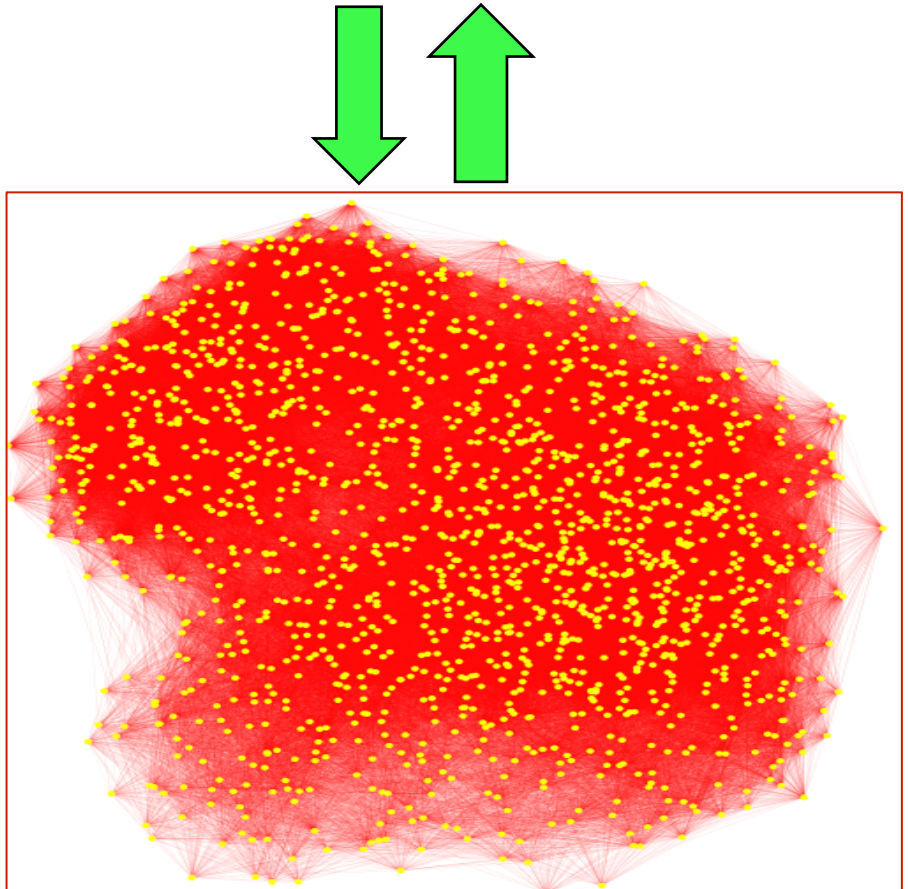
Original Gene Expression Data

#Gene	N6_R1	N6_R2	N12_R1	N12_R2	N24_R1	N24_R2	N48_R1	N48_R2	N96_R1	N96_R2	N504_R1	N504_R2	CTRL0_R1
PtpAffx.1000.1.A1_at	7.530576895	7.385883262	7.29217316	7.469847873	7.641162094	7.288071234	7.390286341	7.449166454	7.466283615	7.241016077	7.342268529	7.654158945	7.033911309
PtpAffx.1000.24.1.A1_at	2.551519352	2.938757052	2.533862853	2.899883611	2.720421475	2.609122479	2.715582978	2.76047964	2.400742676	3.171871681	2.629486858	2.875408787	3.022726534
PtpAffx.1000.12.1.A1_at	9.71617722	9.308978126	9.950072334	9.618868144	9.118708133	8.927187626	8.781785936	9.047149972	9.192139595	9.20270534	9.390355895	9.728563805	9.646360527
PtpAffx.1000.40.1.A1_at	8.679376124	8.609584391	8.48099769	8.41896876	8.179562356	7.871582665	7.974005483	8.109956171	7.992523786	7.969483292	8.119931061	7.973795717	7.673801279
PtpAffx.1000.48.1.A1_at	11.06439533	11.04086855	11.22723414	10.85064109	10.26732642	10.3607159	9.775006293	10.02859713	9.542576329	9.973979947	9.638862316	9.660750846	9.048866151
PtpAffx.1000.51.1.A1_at	7.063140781	6.884067457	7.569984454	7.212236568	7.379439089	7.3168531	7.242005678	7.548254117	7.056245364	6.527081951	6.834298603	6.713576743	6.775959154
PtpAffx.1000.56.1.A1_at	3.445765765	3.114172568	3.020712261	2.790946197	3.245034643	2.876452768	3.181715363	3.120077584	3.110975379	2.861430184	3.196107627	3.212769326	3.196107627
PtpAffx.1000.58.1.A1_at	2.501403096	2.581036102	2.395210355	2.816308622	2.404598618	2.371276232	2.469089445	2.689634178	2.785033071	2.447731635	2.574245602	2.185271826	2.503766491
PtpAffx.1000.59.1.A1_at	2.675794623	2.959692346	2.395005475	2.881104163	2.79930837	2.370288103	2.678930974	2.673482303	2.684552145	2.702957057	2.842096724	2.776562628	2.619820015
PtpAffx.1000.75.1.A1_at	3.139442119	3.028297298	3.130116851	3.391364565	2.719676631	3.403120629	3.644620603	3.161831441	2.892081724	3.145863654	2.739489279	2.998165125	3.052749588
PtpAffx.1000.80.1.A1_at	8.935888629	9.460742664	7.973840189	8.703325099	7.144191177	7.9746469	7.150199268	7.686963065	7.033531154	6.939392532	7.641943859	8.008344292	7.520134468
PtpAffx.1000.88.1.A1_at	9.827052126	9.512160252	9.609156912	9.766238434	9.724248847	9.334555522	9.533836618	9.58249648	9.527103481	9.494688253	9.138534446	9.530988131	9.15565086
PtpAffx.1000.8.1.A1_at	9.102721175	8.934163856	8.820337561	8.964918117	8.449918169	8.674228267	8.34166232	8.440643853	8.480705983	8.293508443	8.466935387	8.723358619	8.760335956
PtpAffx.1000.93.1.A1_at	5.614092034	5.76134207	5.895129911	6.64836244	6.177825367	5.649230504	5.983133631	5.831787812	5.437203311	5.628615495	5.869795331	5.755178873	5.1835173
PtpAffx.1000.93.1.A1_s_at	6.17604046	6.484205244	6.207913129	6.920469098	6.578339387	6.342432242	6.332950879	6.405636175	6.310785209	6.046040856	6.425189532	5.95524848	5.911994568
PtpAffx.1000.104.1.A1_at	4.030830458	4.184088293	3.764445873	4.377386088	4.492888495	3.923369349	4.268873836	4.151656223	4.602773795	4.242018351	3.997113701	4.164297217	4.138031358
PtpAffx.1000.107.1.A1_at	6.875218306	6.76247355	6.711435038	6.96288432	6.771406546	6.451405679	6.665955609	6.574998485	6.611927423	6.636686741	6.591997023	6.78451425	6.5615729
PtpAffx.1000.112.1.A1_at	9.113798289	9.153960892	9.371940497	9.439338664	8.460504437	8.728748299	8.428371417	8.348964054	8.317411057	8.914960003	9.20111242	9.383228788	8.587227886
PtpAffx.1000.116.1.A1_at	4.088766462	4.276171857	3.818106466	4.245206767	3.925822355	4.487977089	4.023054708	3.627705516	4.327228872	3.891295908	3.519474212	3.432750875	4.272216443
PtpAffx.1000.128.1.A1_at	6.570911513	6.485499494	6.399803437	6.761483013	6.983551934	6.447704353	6.041106058	6.648523919	6.623907425	6.729391248	6.638279437	6.758376307	6.318836876
PtpAffx.1000.134.1.A1_s_at	11.59410661	11.55570837	11.56204021	11.56690963	11.59974815	11.37090359	11.38378841	11.38961797	11.34116723	11.44854298	11.36564949	11.52630257	11.05337273
PtpAffx.1000.18.1.S1_a_at	6.231082673	6.583387171	6.144500362	6.738590362	5.892426893	6.61116439	6.754491041	5.980216405	6.484767774	6.404222377	6.888236257	6.652642354	6.882632846
PtpAffx.1000.195.1.S1_a_at	2.940794696	2.627804089	2.875355526	3.032833702	2.767189463	2.910219516	3.171061943	2.69810815	2.458075975	2.757584252	2.955458763	2.897414561	3.033996943
PtpAffx.1000.195.2.S1_s_at	4.017458429	3.829396745	3.829396745	4.371962306	3.823360311	3.992254476	4.001399412	3.686187174	3.917954727	4.261513083	4.515946982	4.550802594	4.533034651
PtpAffx.100.1.S1_at	9.970490511	9.788203279	9.948562617	10.06086854	10.36069909	10.03070437	10.14015215	10.12312172	10.04553731	10.13028529	9.995158186	10.09805478	9.842374041
PtpAffx.10021.1.A1_at	7.619323235	7.723760222	7.621915007	7.624226215	7.877959754	7.638419453	7.730662586	7.93154864	7.810474908	7.778023079	7.53834469	7.709431077	7.843305314
PtpAffx.1002.2.A1_a_at	8.809016286	9.13913662	9.321093711	8.859204391	8.926774598	9.354924913	9.080036876	9.104038507	9.063473788	9.011348394	9.386738787	9.372702265	9.386239367
PtpAffx.100255.1.S1_s_at	6.050833999	6.066638338	6.38420101	6.01470608	6.58090111	7.162710354	6.425037845	6.945826465	6.996668924	6.632604486	7.46756278	7.149857821	7.270355703
PtpAffx.10026.1.S1_at	9.36907264	9.332941709	9.468359427	9.328311728	9.574827428	9.435494017	9.707094899	9.573732821	9.80539335	9.71175671	9.220819001	9.18230644	9.456198049

Collaborative Network

TF \	TF1	2	3	4	5	6	7	8	9	10	…
1	0	0	0	0	0	0	0	0	0	0	
2	0	0	0	0	0	0	0	82	0	0	
3	0	0	0	0	0	88	0	0	0	0	
4	0	0	0	0	22	0	18	0	24	9	
5	0	0	0	22	0	0	0	0	7	8	
6	0	0	88	0	0	0	0	0	0	0	
7	0	0	0	18	0	0	0	0	20	10	
8	0	82	0	0	0	0	0	0	0	0	
9	0	0	0	24	7	0	20	0	0	45	
10	0	0	0	9	8	0	10	0	45	0	
…											

How to decompose this collaborative network into subnetworks each of which controls a biological process or complex trait?



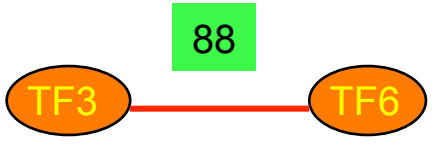
TF	TF1	2	3	4	5	6	7	8	9	10 …
1	0	0	0	0	0	0	0	0	0	0
2	0	0	0	0	0	0	0	82	0	0
3	0	0	0	0	0	88	0	0	0	0
4	0	0	0	0	22	0	18	0	24	9
5	0	0	0	22	0	0	0	0	7	8
6	0	0	88	0	0	0	0	0	0	0
7	0	0	0	18	0	0	0	0	20	10
8	0	82	0	0	0	0	0	0	0	0
9	0	0	0	24	7	0	20	0	0	45
10	0	0	0	9	8	0	10	0	45	0
…										

$\alpha = \mu + \delta$

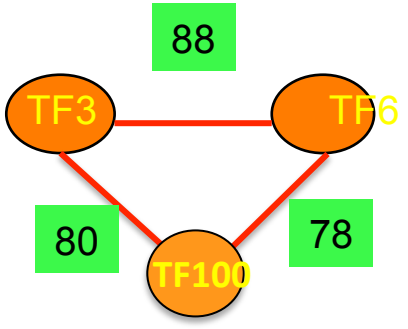
$\beta = \mu + 2\delta$

$\gamma = \mu + 3\delta$

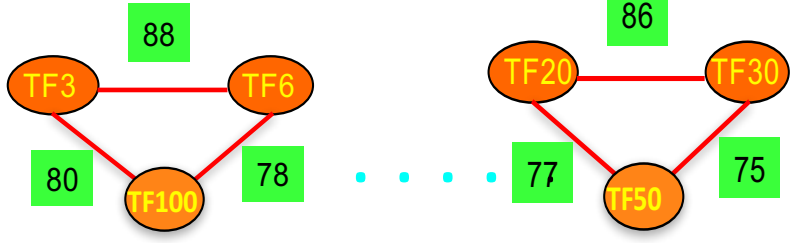
μ: mean
δ: standard deviation



Triple-Link algorithm



Single-Seed Growing Algorithm



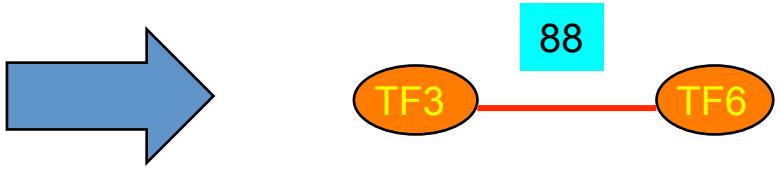
Multi-Seed Growing Algorithm

Nie et al BMC Systems Biology

Ji et al. ***Scientific Reports*** (2017)

Triple-link algorithm

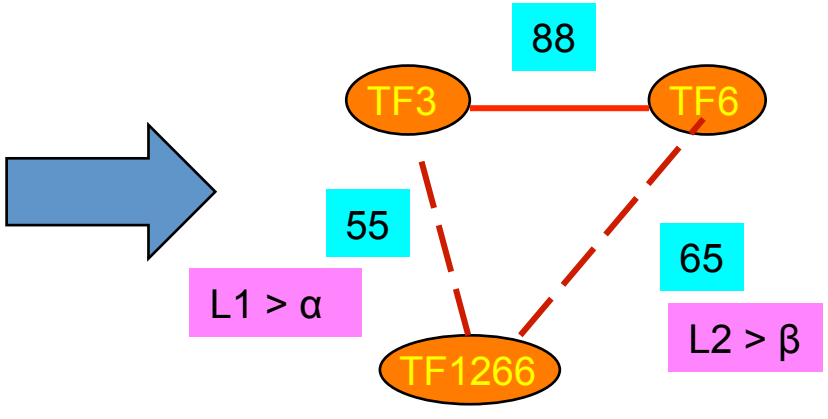
TF	TF1	2	3	4	5	6	7	8	9	10 ...
1	0	0	0	0	0	0	0	0	0	0
2	0	0	0	0	0	0	0	82	0	0
3	0	0	0	0	0	88	0	0	0	0
4	0	0	0	0	22	0	18	0	24	9
5	0	0	0	22	0	0	0	0	7	8
6	0	0	88	0	0	0	0	0	0	0
7	0	0	0	18	0	0	0	0	20	10
8	0	82	0	0	0	0	0	0	0	0
9	0	0	0	24	7	0	20	0	0	45
10	0	0	0	9	8	0	10	0	45	0
...										

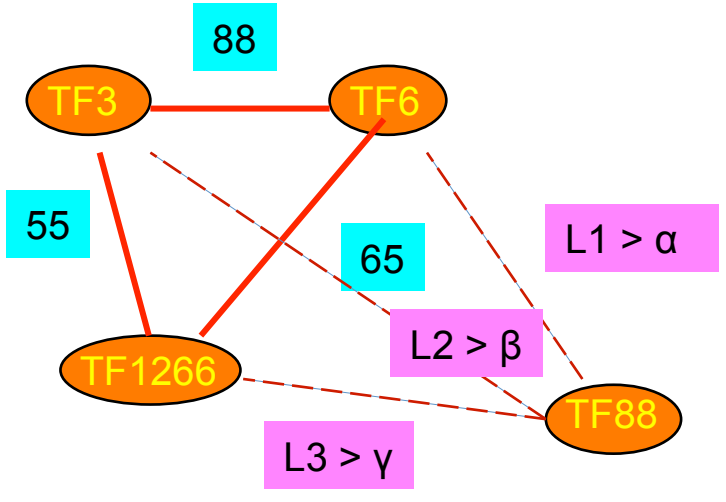


$\alpha = \mu + \delta$

$\beta = \mu + 2\delta$

$\gamma = \mu + 3\delta$





Nie et al. 2011. BMC Systems Biology 5:53

Test 1: Arabidopsis Data

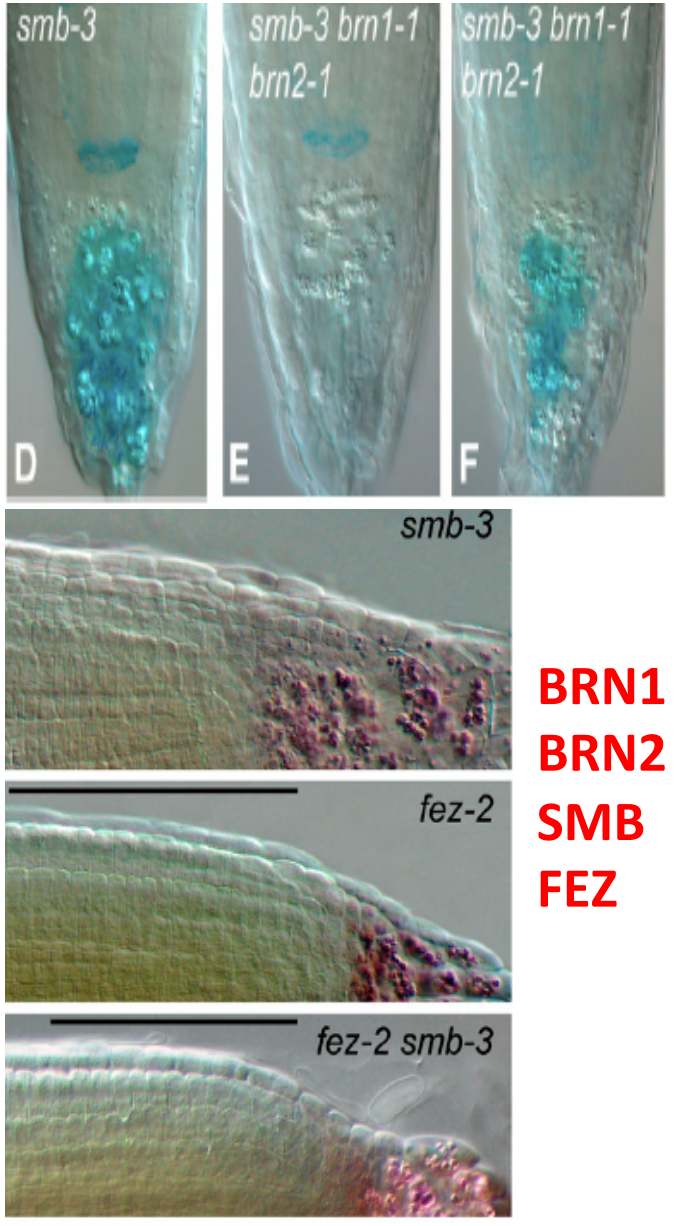
Arabidopsis salt stress experiments

- 6 Microarray experiments of Arabidopsis
- 108 chips (downloaded from GEO)

Question: What TFs control root growth under salt stress condition in Arabidopsis?

Nie et al: TF-Cluster (BMC Systems Biology 2011)

Cluster 2: A cluster of TFs control root cap development (stem cells of roots)			
AT1G33280	BRN1	BRN1, SMB control root cap maturation	[66]
AT4G10350	BRN2	BRN2, SMB control root cap maturation	[66]
AT1G79580	SMB	FEZ and SMB control root stem cells	[67]
AT5G39820	ANAC094	Apical meristem protein, function unknown	[59]
AT1G26870	FEZ	FEZ and SMB control root stem cells in cap	[67]
AT1G74500	TOM7	Embryonic root initiation	[68]
AT3G27010	TCP20	Postembryonic cell division in root	[70]
AT2G30340	LBD13	Expressed in cells at the adaxial base of lateral roots	[104]
AT2G40470	LBD15	Expressed in cells at the adaxial base of lateral roots	[104]
AT1G51190	PLT2	Control root stem cell activity near cap	[69]
AT1G66350	RGL1	Root epidermal differentiation	[105]
AT2G37260	TTG2	Differentiation of trichomes and root hairless cells	[106]
AT5G57420	IAA33	IAA is involved in root development	[107, 108]
AT2G29060		scarecrow transcription factor family protein	



BRN1
BRN2
SMB
FEZ

The Plant Cell, Vol. 22: 640–654, March 2010,

Test 2: Data Set from humans

What TFs are necessary for pluripotency maintenance in human stem cells?

Present knowledge about regulation of human stem cell pluripotency

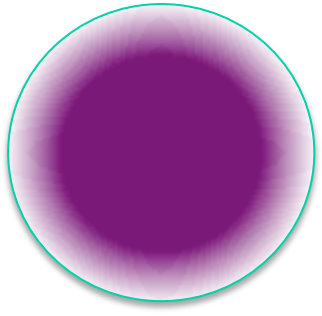
NANOG + POU5F1 + SOX2

SALL2
PHC1
PRDM14
DNMT3A

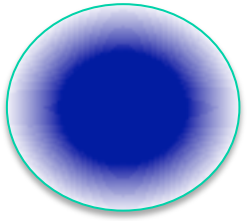
CSRP1 LIN28
NANOGP8 MYCL1
LOC653441 DNMT3B
SOX3 RORB

Identified by many biologists from 1998 to 2008

The Power of TFs: Reprogramming in somatic cells



Skin Cell

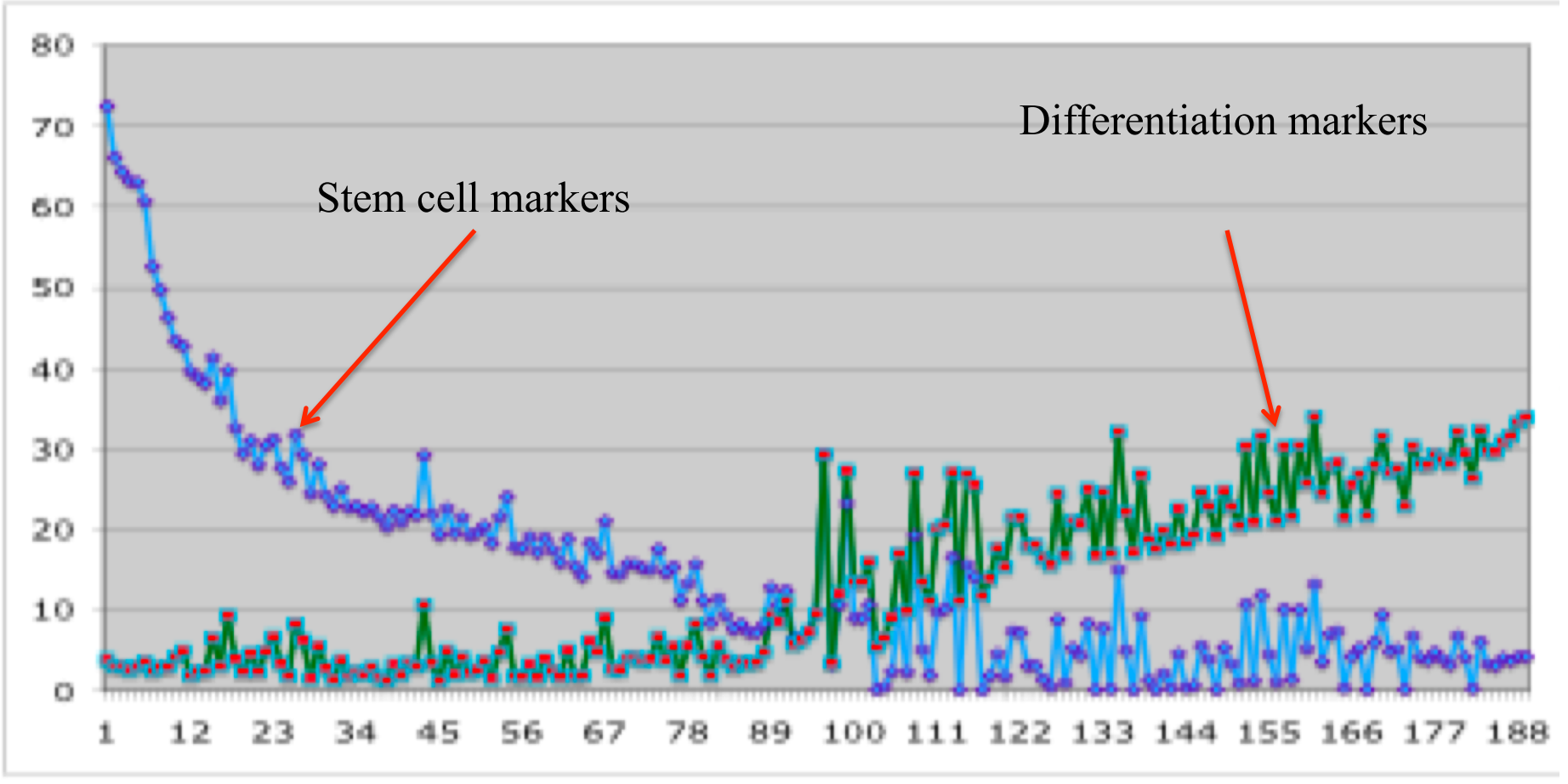


NANOG
POU5F1
SOX2

+ MYC2

Induced Pluripotency Cell (IPC)

Input data: 189 microarray chips from human stem cells that underwent differentiation from 0 ~ 96 hours.



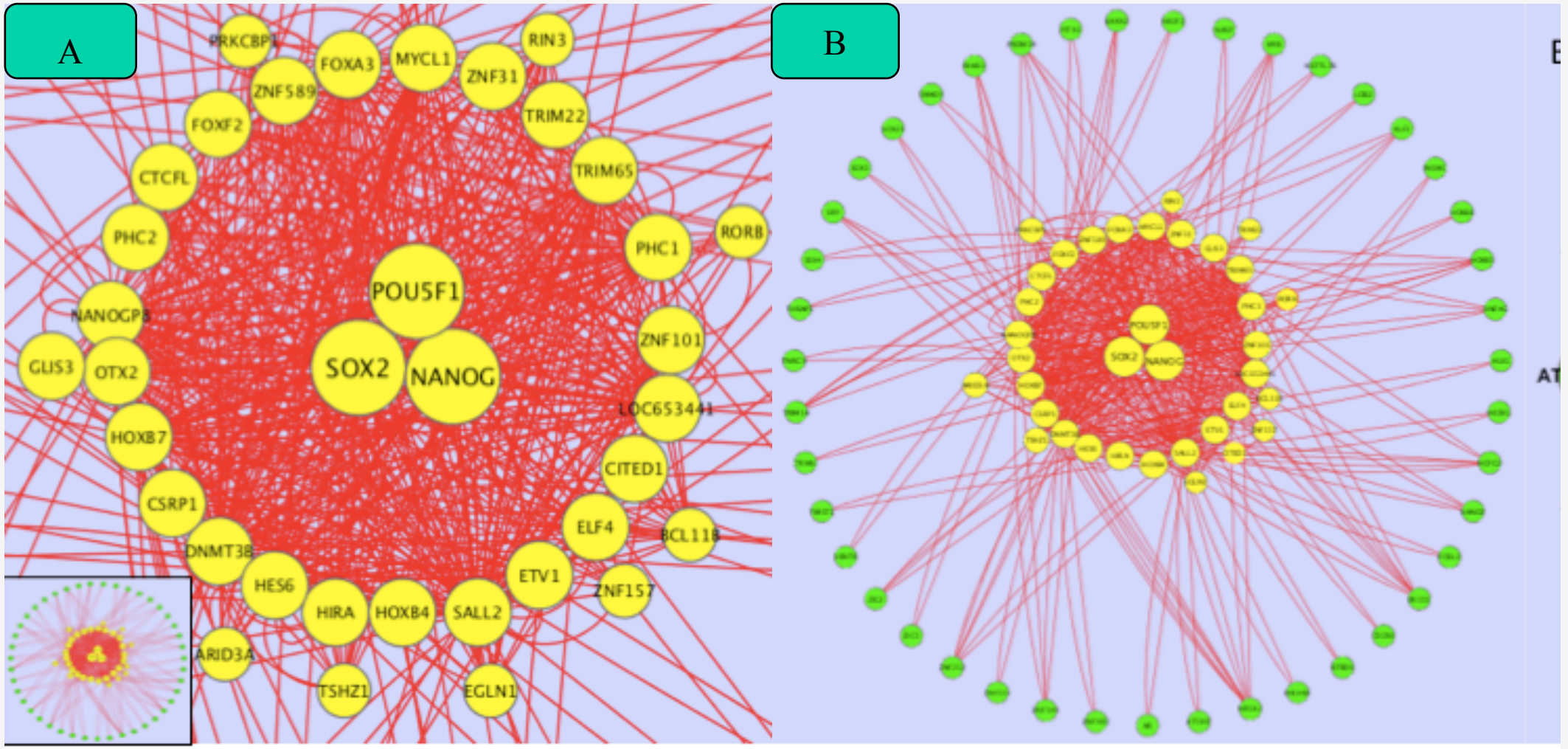
Microarray chips (189) provided by James Thomson @ U of Wisconsin

Subnetwork 1 : Output from TF-cluster

Genes	Symbol	Description	Evidence
NM_024865	NANOG	Nanog homeobox	[14]
BC099704	NANOGP8	Nanog homeobox pseudogene 8	Pseudogene with similarity to Nanog.
NM_003106	SOX2	SRY box 2	[14]
NM_002701	POU5F1	POU class 5 homeobox 1	[14]
NM_006892	DNMT3B	DNS methyltransferase 3 beta	[15, 16]
NM_004078	CSRP1	cysteine-rich protein	
NM_080618	CTCFL	CCCTC-binding factor (zinc finger protein)-like	[17]
NM_016089	ZNF589	Zinc finger 589	[18, 19]
NM_004426	PHC1	Polyhomeotic homolog 1	[20]
NM_005407	SALL2	SAL2 like	[18]
NM_018645	HES6	hairy and enhancer of split 6	
NM_173547	TRIM65	Tripartite motif containing 65	
NM_004427	PHC2	Polyhomeotic homolog 2	[21]
NM_032805	ZFP206	Zinc finger protein 206 (ZSCAN10)	[22, 23]
NM_001421	ELF4	ETS domain TF	
NM_003325	HIRA	HIR Histone Cell Cycle regulator	[24]
NM_033204	ZNF101	Zinc finger protein 101	
BC098403	ETV1	ets variant 1	[25, 26]
NM_006079	CITED2	Cbp/p300-interacting transactivator	[27, 28]
NM_021728	OTX2	orthodenticle homeobox 2	
NM_024015	HOXB4	homeobox B4	
NM_006074	TRIM22	tripartite motif-containing 22	[19]
XM_929986	LOC653441	similar to polyhomeotic 1-like	Gene with sequence similarity to PHC1
NM_004497	FOXA3	Forkhead box 3	

Among 24 predicted TFs 17 have literature evidence(~70%)

The coordination between TFs within Subnetwork 1



A. Intra-connections among the genes within Collaborative Subnetwork 1
B. Comparison of intra-subnetwork connections with outward-connections

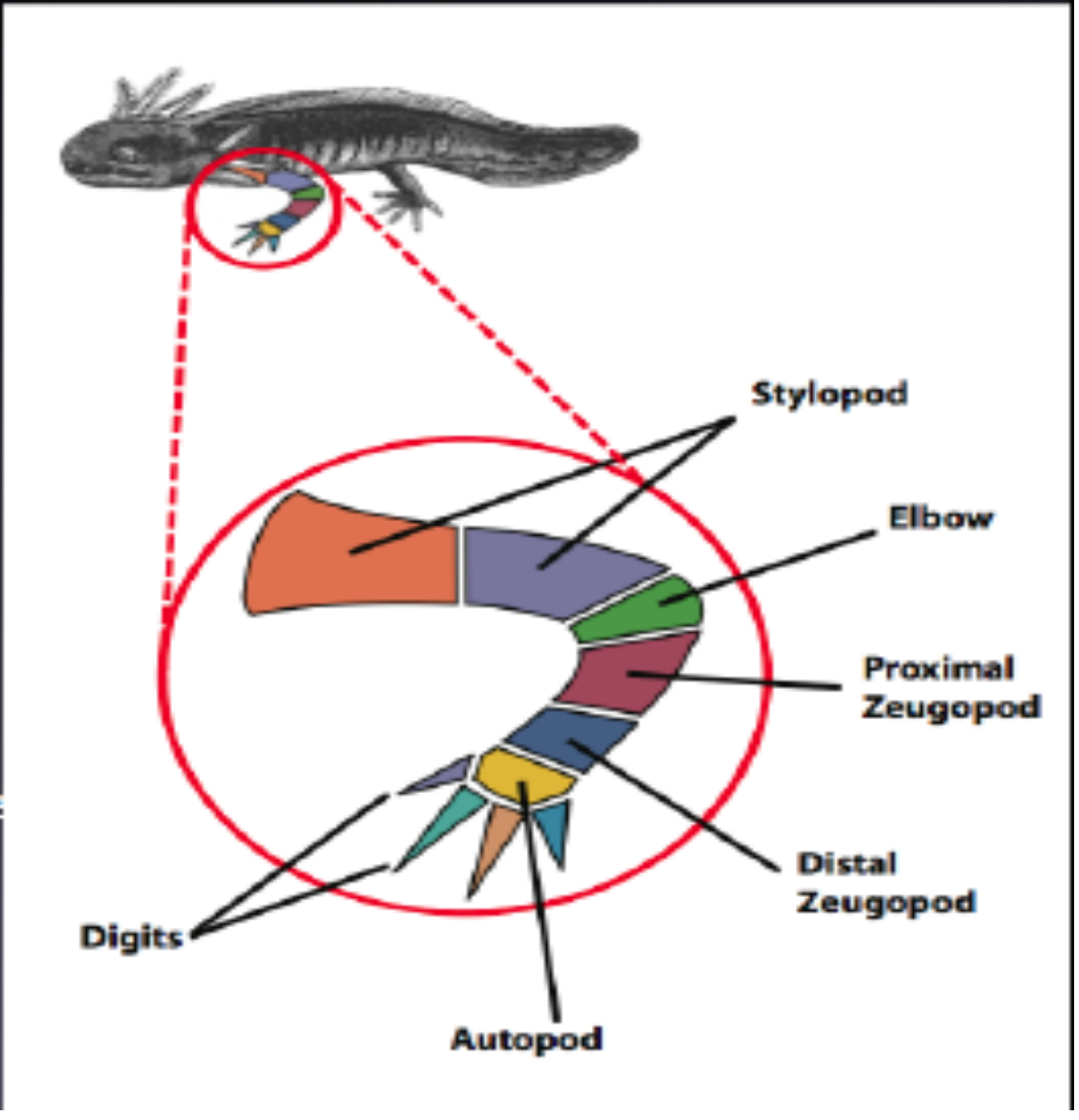
Test 3: Axolotl RNA-Seq

Total 74 samples

Axolotl transcriptome

S, Z, A, Digits,
30 day blastema
RNA-Seq (on GA IIX)
Eval for patterns in adult limb
and Blastema-specific genes

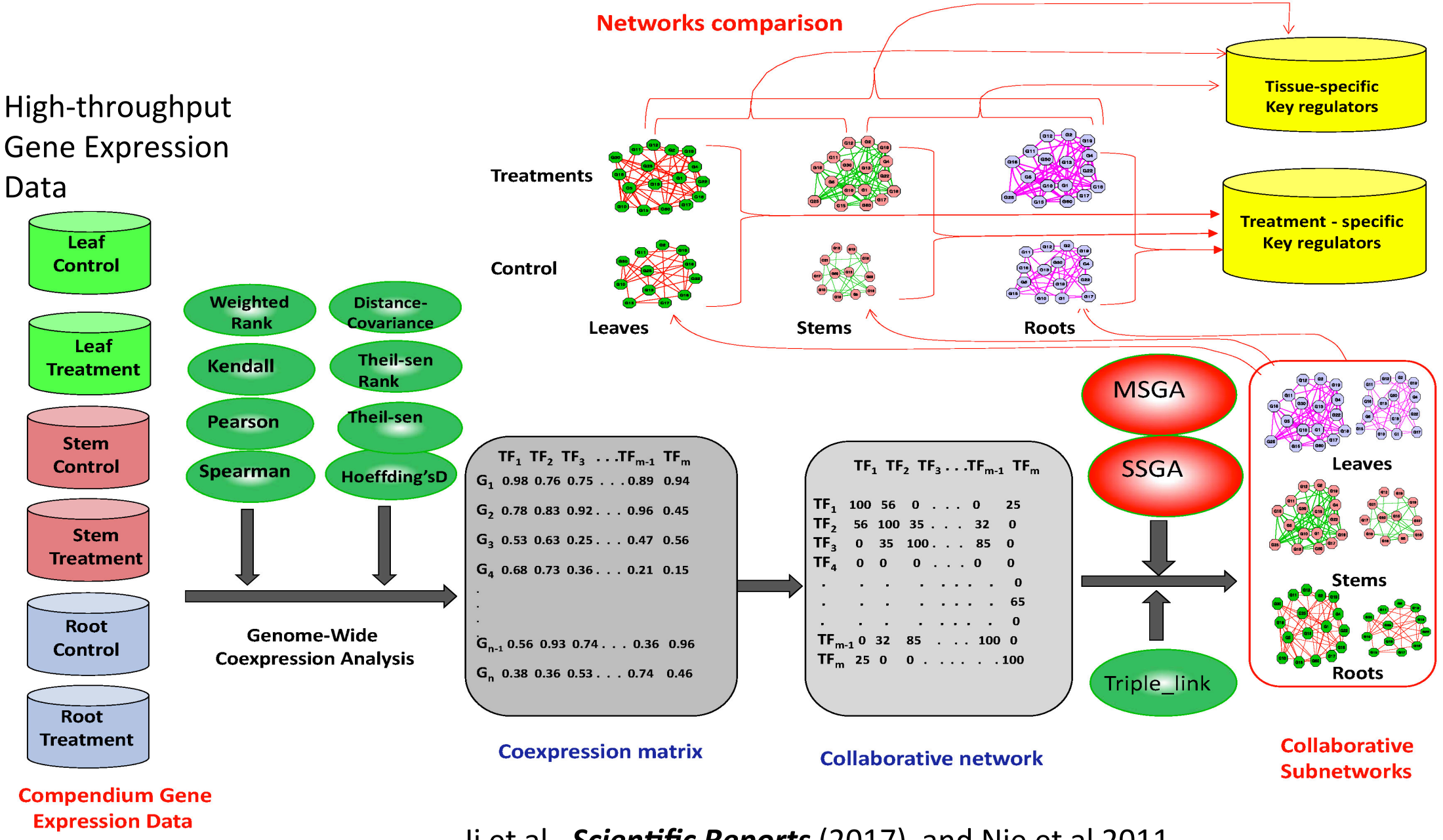
Challenges
No Genome,
Poor gene annotation
Genome 30GB,
Evol distance to Human/Mouse



Cluster 2: Limb development

1. HES4: multicellular organismal development
2. HOXB1: anterior/posterior pattern formation
3. BARHL2: neuron differentiation
4. CDX2: caudal type homeobox 2
5. CIC
6. FOXH1 nodal expression; Lower-limb Deep Vein Thrombosis
7. FOXJ1: motile cilia lupus erythematosus
8. FOXN1: skin nail development
9. GLIS3: limb mesenchymal patterns
10. HES6: limb buds
11. HESX1: multicellular organismal development
12. HNF4A: limb mesoderm
13. HOXA5: skeletal and organ morphogenesis
14. HOXB6: Development
15. OLIG1: oligodendroglial differentiation
16. PAX2:
17. PLAGL1
18. PPARGC1B
    RIN3
20. SCML4
21. T: anterior/posterior pattern formation
    tail morphogenesis notochord development
22. TAF6L
23. TAL2
24. TGIF2
25. TRAF4: development of the axial skeleton
    the closure of the neural tube
26. ZBTB16
27. ZBTB4
28. ZFPM2: hematopoiesis and cardiogenesis
29. ZNF154

Collaborative Network Pipeline



Ji et al. ***Scientific Reports*** (2017) and Nie et al 2011

Summary for Collaborative Gene Network:

1. Construction of collaborative subnetworks of TFs is novel approach for identifying collaborative TFs that govern a biological process or complex trait using gene expression data. Compared to markers-assistant approaches for identifying genes that control a trait, collaborative subnetwork construction is a more direct one. It appears to be highly efficient because it is based on the footprints of regulatory events in gene expression profiles rather than the information encoded in DNA sequences, for instance, genome-wide gene association by SNPs.

2. As gene expression data explode we believe construction of collaborative network will find its way into a wide range of applications for novel biological knowledge discovery.