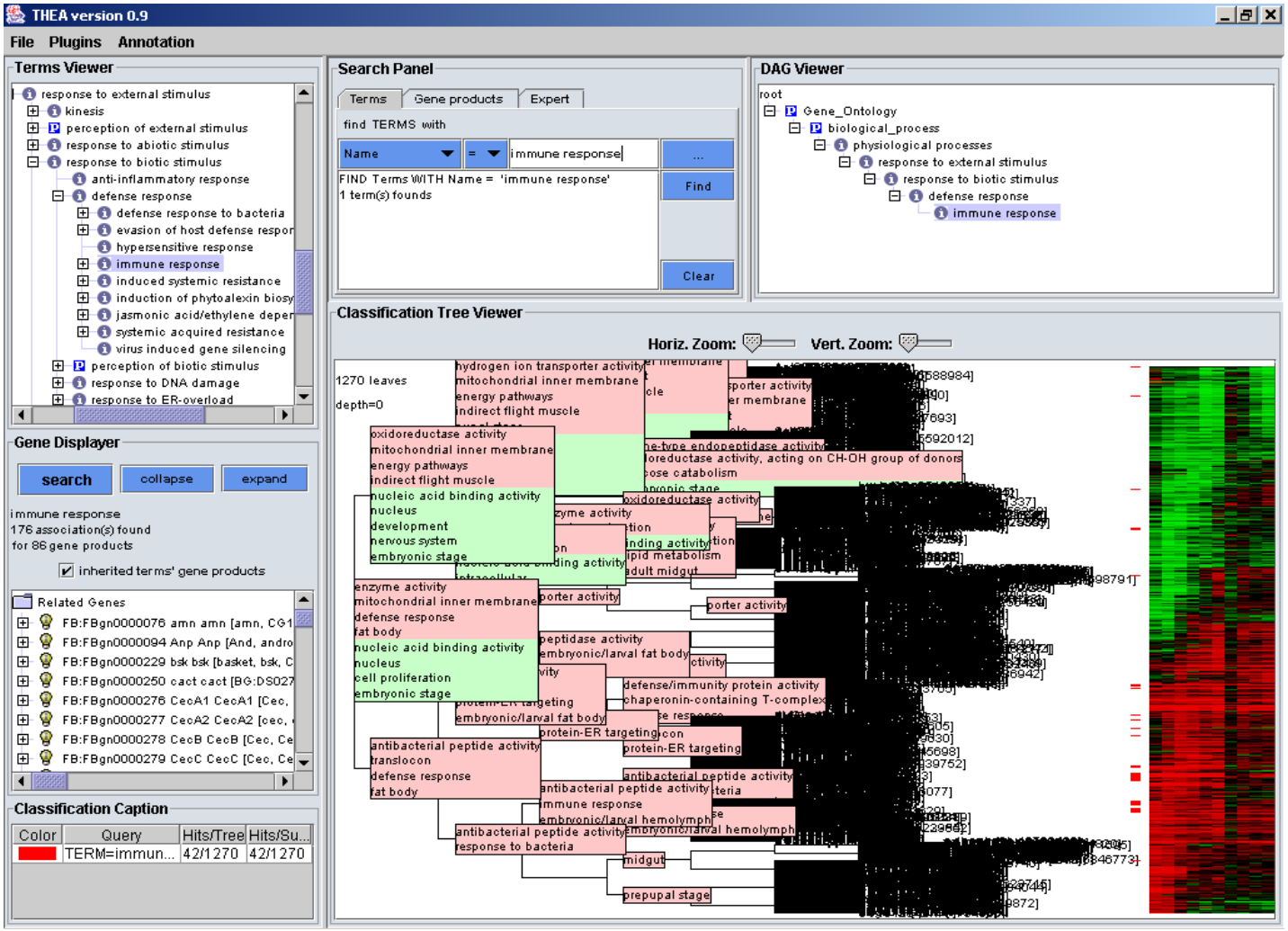
Semantic integration of Biological Data with Allonto

Application to the analysis of gene duplication

Claude Pasquier

Thea (reminder)



Analysis of a dataset consisting of 1270 regulated genes (De Gregorio, E., Spellman, P.T., Rubin, G.M. and Lemaitre, B. (2001) Genome-wide analysis of the Drosophila immune response by using oligonucleotide microarrays. Proc Natl Acad Sci U S A, 98, 12590-12595).

Limitations

- Information integration is difficult
  - Many databases (several hundred) with heterogeneous schemas
  - Traceability problems
    - Identifiers may change (ex.: Unigene IDs)
  - Semantic problems
    - One name ➔ different concepts
    - One concept ➔ different names
- Goal directed search

One possible solution

- Representation of information with a semantic network using ontologies
- Use of OWL (Ontology Web Language) as the exchange format
  - Several databases export information in this format (UniProt, Reactome, …)
- Information integration consists in the merging of several graphs

Partial view of a graph

Partial view of information extracted from SwissProt, Gene Ontology, IntAct and Reactome

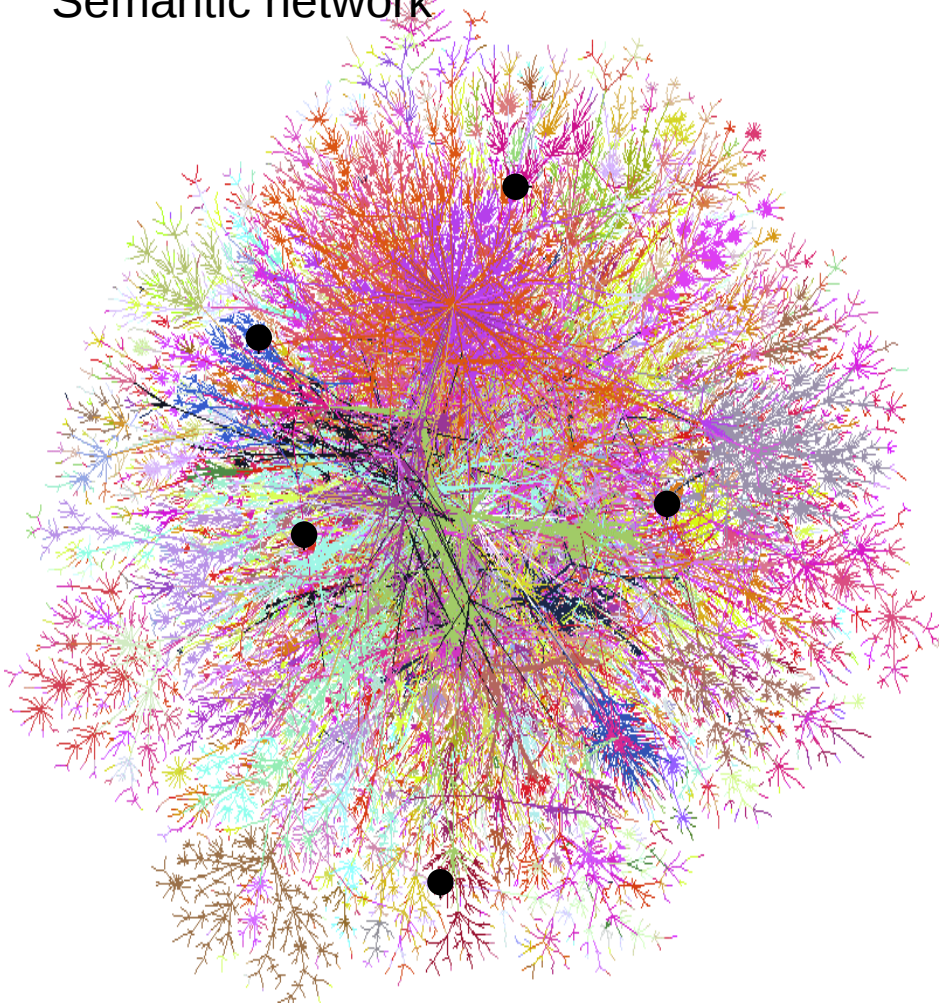
Example of searches

- Goal directed searches
  - Graph queries
    - Ex: what are the proteins annotated « cell proliferation » which interact with CDC6
- Non directed searches
  - Search of a shortest path
    - Ex: what are the paths connecting CDC6 to K17
  - Search of a subgraph
    - Ex: what is the minimal graph connecting CDC6, ERBB2, BRCA1 and K17 (allows to retrieve the shared properties)

Partial view of information extracted from SwissProt, Gene Ontology, IntAct and Reactome

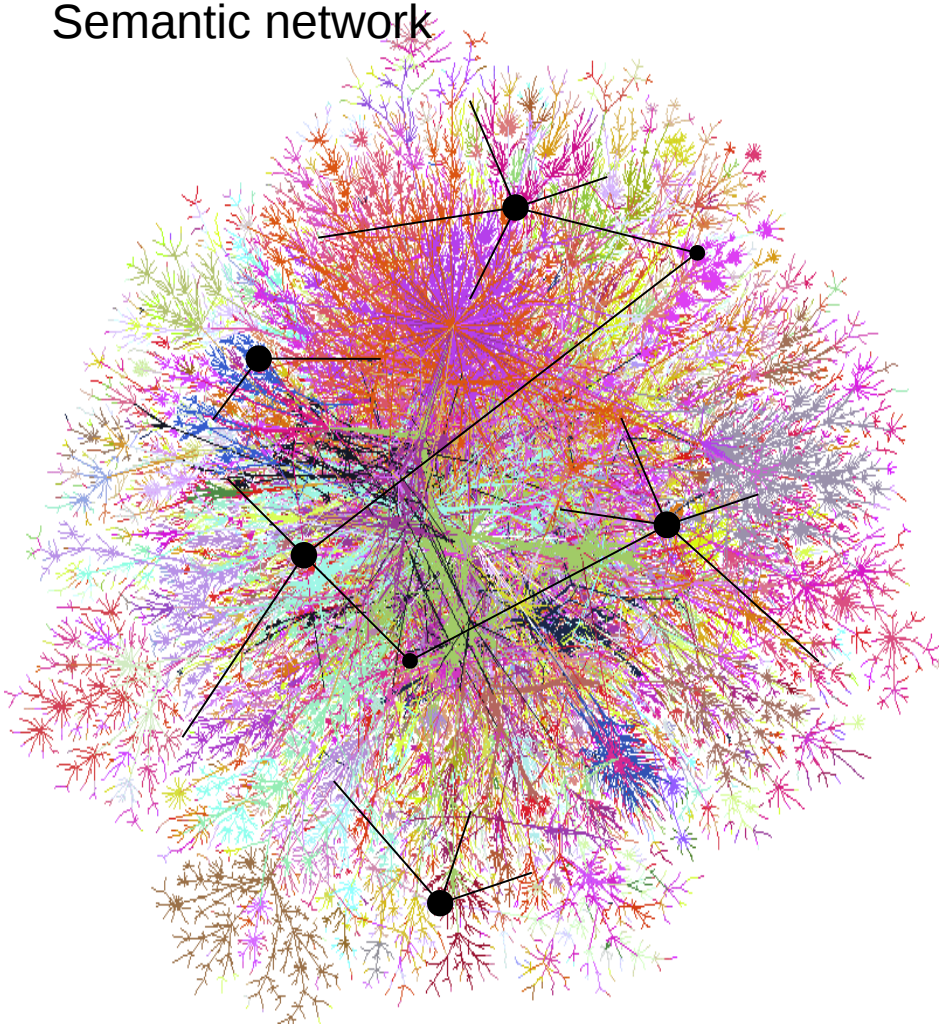
Search for pertinent information characterizing a set of nodes

Semantic network



Search for pertinent information characterizing a set of nodes

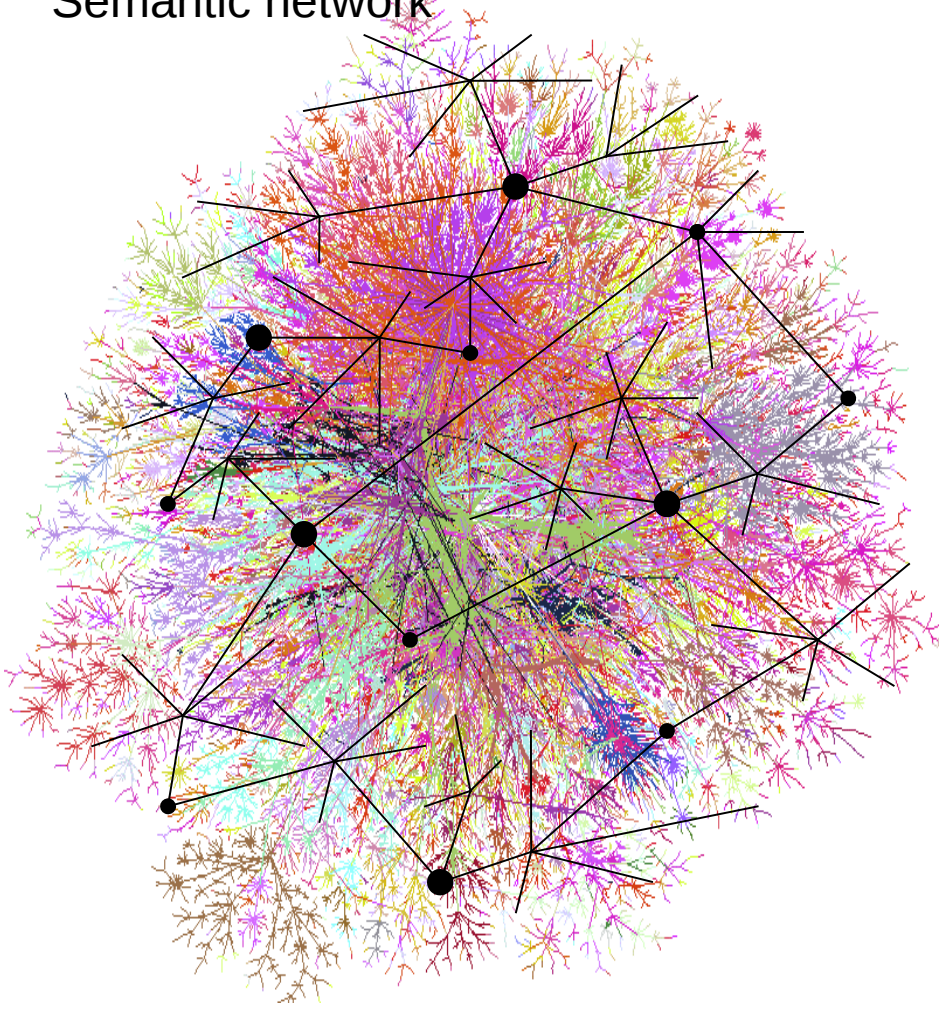
Semantic network



- Extract associated information

Search for pertinent information characterizing a set of nodes

Semantic network



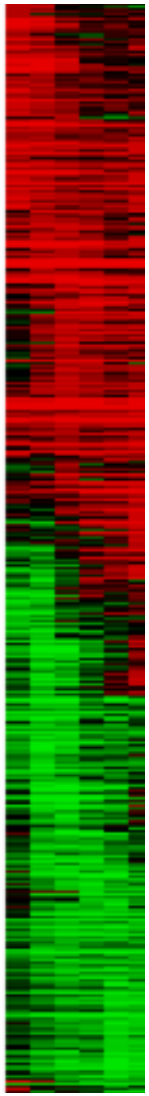
- Extract associated information
- Select pertinent information with an evaluation function
  - Common resources
  - Statistically significant
  - Other

Application of Allonto to Detect DNA copy-number changes

- As copy number and expression level are correlated, it should be possible to detect amplifications (and may be deletions) by using expression levels alone.
- The idea:
  - if an increase of expression level of a set of transcripts is due to an amplification, then, these transcripts should be co-localized.
- If such a method is effective, we will be able to extract very valuable information about cancer related genes by processing the large amount of expression level measurement stored in public databases.

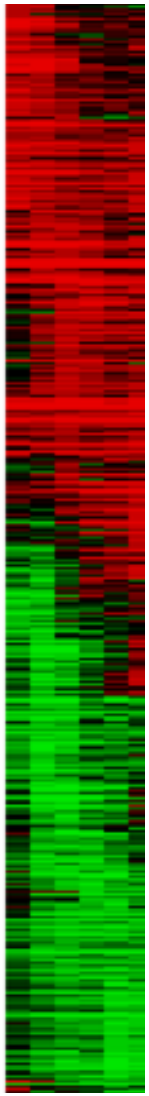
The algorithm

Transcripts sorted by fold-change (cancer vs. normal)



The algorithm

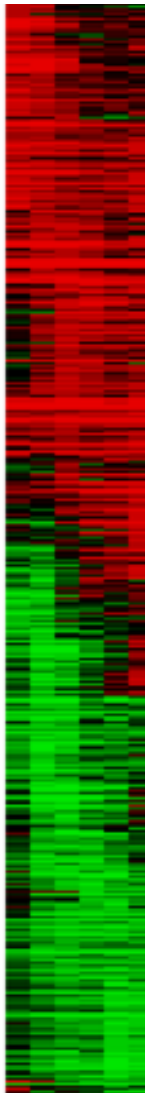
Transcripts sorted by fold-change (cancer vs. normal)



- Choose a set of transcripts that are over-expressed and co-expressed

The algorithm

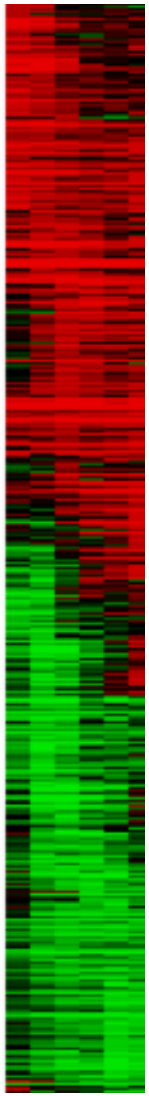
Transcripts sorted by fold-change (cancer vs. normal)



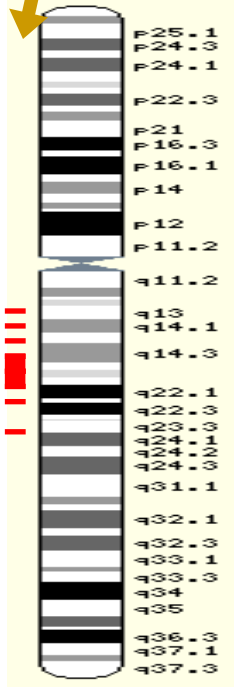
- Choose a set of transcripts that are over-expressed and co-expressed
- Select the corresponding resources on the semantic network

The algorithm

Transcripts sorted by fold-change (cancer vs. normal)

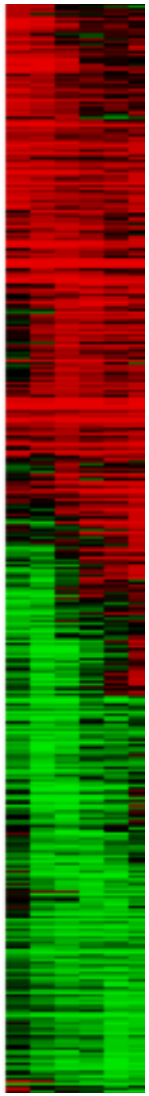


- Choose a set of transcripts that are over-expressed and co-expressed
- Select the corresponding resources on the semantic network
- Search in the network for the localization of the corresponding genes on chromosomes



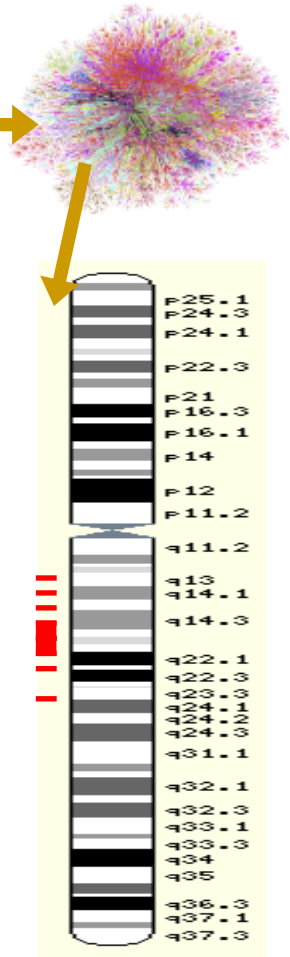
The algorithm

Transcripts sorted by fold-change (cancer vs. normal)



- Choose a set of transcripts that are over-expressed and co-expressed
- Select the corresponding resources on the semantic network
- Search in the network for the localization of the corresponding genes on chromosomes
- Evaluate the proximity of these genes by computing:
  - **p-value** = probability for these genes to be grouped in a portion of a chromosome by chance
  - **IF p-value** < given **cutoff**

    THEN select this group of genes (over-expression might be due to an amplification)



Dataset and parameters

- Use of a semantic network integrating data from Ensembl, SwissProt, IntAct, Pubmed and Reactome
- Analysis of the dataset from:
  - Pollack JR, Sorlie T, Perou CM, Rees CA, Jeffrey SS, Lonning PE, Tibshirani R, Botstein D, Borresen-Dale AL, Brown PO. Microarray analysis reveals a major direct role of DNA copy number alteration in the transcriptional program of human breast tumors. Proc Natl Acad Sci U.S.A. 2002 Oct 1;99(20):12963-8
  - Experiments done for the 4 cell lines and 37 tumors
  - 6096 Unigenes IDs ➔ 2898 ENSEMBL IDs
- parameters
  - P-value < 10^{-5}
  - At least 3 co-localized genes
  - Fold-change (cancer vs. normal) > 1.5

Amplifications and deletions in BT474 cell line

High-level alterations detected by array CGH					
Cell line	Locus	Size (kb)		Locus	Size (kb)
BT474	**Amplifications**			**Amplifications**	
	1q21.2-q25.1			15q11.2 -q12	5030
Peaks:	*1q21.2-q21.3*	*366*	*Peaks:*	*15q11.2*	*949*
	1q22-q23.1	*510*		*15q11.2*	*624*
	1q24.2	*540*		17q12-21.2	4090
	1q31.3	***1661***		17q21.32-23.2	13800
	1q32.1	**936**	*Peaks:*	*17q21.32-q21.33*	*2500*
	1q42.12-q42.13	*500*		*17q22-23.2*	*2180*
	1q43	449		*17q23.2*	*1170*
	1q44-q43	865		17q24.1-24.3	5300
	1q44	1720	*Peak:*	*17q24.1*	*482*
	4p16.1-15.33	1210		19p13.2-13.12	6310
	4q21.1	**2700**		20p12.1	*800*
	9p13.3	**2060**		20q11.22	1300
	9q33.1-34.13	**12600**		20q13.11-13.32	14800
	11q13.1-13.5	19800		20q13.33	305
Peaks:	*11q13.1*	*704*		20q13.33-tel	1380
	11q13.4	*1030*		**Deletions**	
	11q22.1-22.2	**3560**		5q14.1-14.3	5390
	14q11.2-q21.1	**21230**		**6q24.1**	**3200**
	14q31.3-32.12	**3080**		**20q11.22**	**670**
				20p11.23-13.11	**7140**

Shadeo A, Lam WL. Comprehensive copy number profiles of breast cancer cell model genomes. Breast Cancer Res. 2006 Jan 3;8(1):R9

Amplifications and deletions in BT474 cell line

High-level alterations detected by array CGH

Cell line	Locus	Size (kb)		Locus	Size (kb)
BT474	**Amplifications**				
					5030
		366	*Peaks:*	*15q11.2*	*949*
	1q22-q23.1	*510*			*624*
	1q24.2	*540*		17q12-21.2	4090
	1q31.3	***1661***			13800
	1q32.1	**936**	*Peaks:*	*17q21.32-q21.33*	*2500*
	1q42.12-q42.13	*500*			*2180*
	1q43			*17q23.2*	*1170*
	1q44-q43	865			5300
	1q44	1720	*Peak:*	*17q24.1*	*482*
	4p16.1-15.33			19p13.2-13.12	6310
	4q21.1	**2700**			*800*
	9p13.3	**2060**		20q11.22	1300
	9q33.1-34.13	**12600**		20q13.11-13.32	14800
		19800			305
		704			
	11q13.4	*1030*	**Deletions**		
	11q22.1-22.2	**3560**	5q1		
	14q11.2-q21.1	**21230**	**6q2**		
	14q31.3-32.12	**3080**	**20q11.22**		**670**
			20p11.23-13.11		**7140**

3q25.2-3q26.2 (16030 kb)

17q12-17q21.2 (6.650 kb)

17q23.2 (1311 kb)

19p13.2 (1285 kb)

11q13.4 (1468 kb)

20q11.22-20q13.13 (13980 kb)

20q13.2-20q13.31 (9002 kb)

Shadeo A, Lam WL. Comprehensive copy number profiles of breast cancer cell model genomes. Breast Cancer Res. 2006 Jan 3;8(1):R9

Amplifications and deletions in BT474 cell line

High-level alterations detected by array CGH

Cell line	Locus	Size (kb)	Locus	Size (kb)
BT474	**Amplifications**			
				5030
		366	*Peaks:* *15q11.2*	*949*
	1q22-q23.1	*510*		*624*
	1q24.2	*540*	17q12-21.2	4090
	1q31.3	***1661***		13800
	1q32.1	**936**	*Peaks:* *17q21.32-q21.33*	*2500*
	1q42.12-q42.13	*500*		*2180*
	1q43		*17q23.2*	*1170*
	1q44-q43	865		5300
	1q44	1720	*Peak* *17q24.1*	*482*
	4p16.1-15.33		19p13.2-13.12	6310
	4q21.1	**2700**		*800*
	9p13.3	**2060**	20q11.22	1300
	9q33.1-34.13	**12600**	20q13.11-13.32	14800
		19800		305
		704		
	11q13.4	*1030*	**Deletions**	
	11q22.1-22.2	**3560**	5q1	
	14q11.2-q21.1	**21230**	**6q2**	
	14q31.3-32.12	**3080**	**20q11.22**	**670**
			20p11.23-13.11	**7140**

3q25.2-3q26.2 (16030 kb) 5

?

17q12-17q21.2 (6.650 kb) 15

17q23.2 (1311 kb) 5

19p13.2 (1285 kb) 3

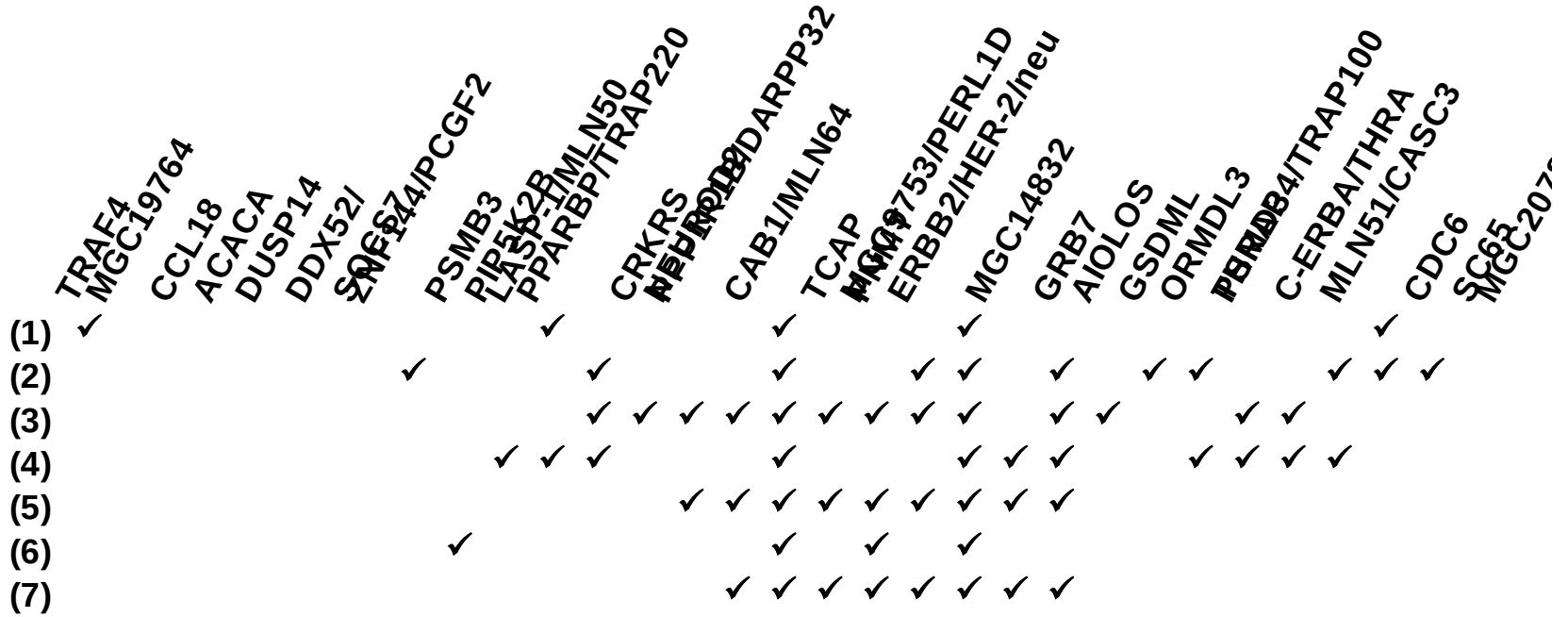
11q13.4 (1468 kb) 3

20q11.22-20q13.13 (13980 kb) 4

20q13.2-20q13.31 (9002 kb) 4

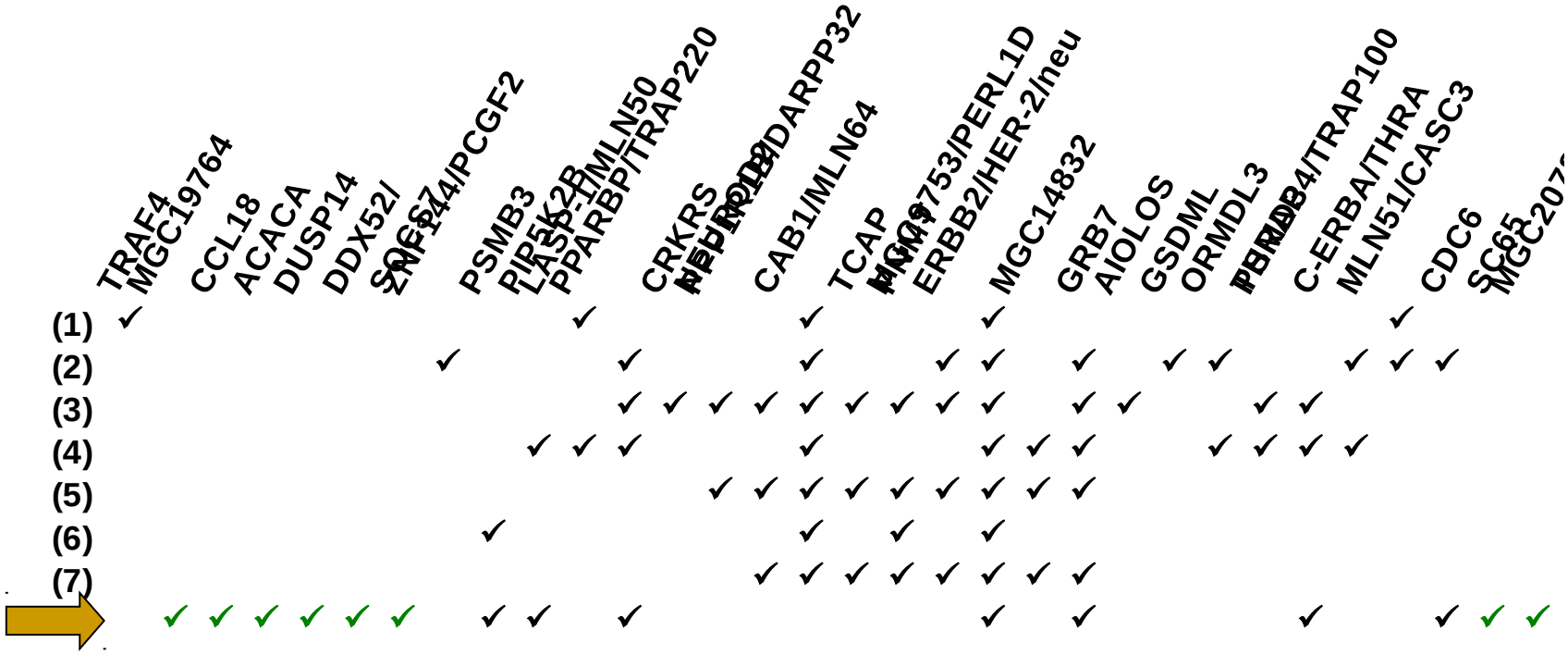
Shadeo A, Lam WL. Comprehensive copy number profiles of breast cancer cell model genomes. Breast Cancer Res. 2006 Jan 3;8(1):R9

Detail of the amplicon 17q11.2-21.2



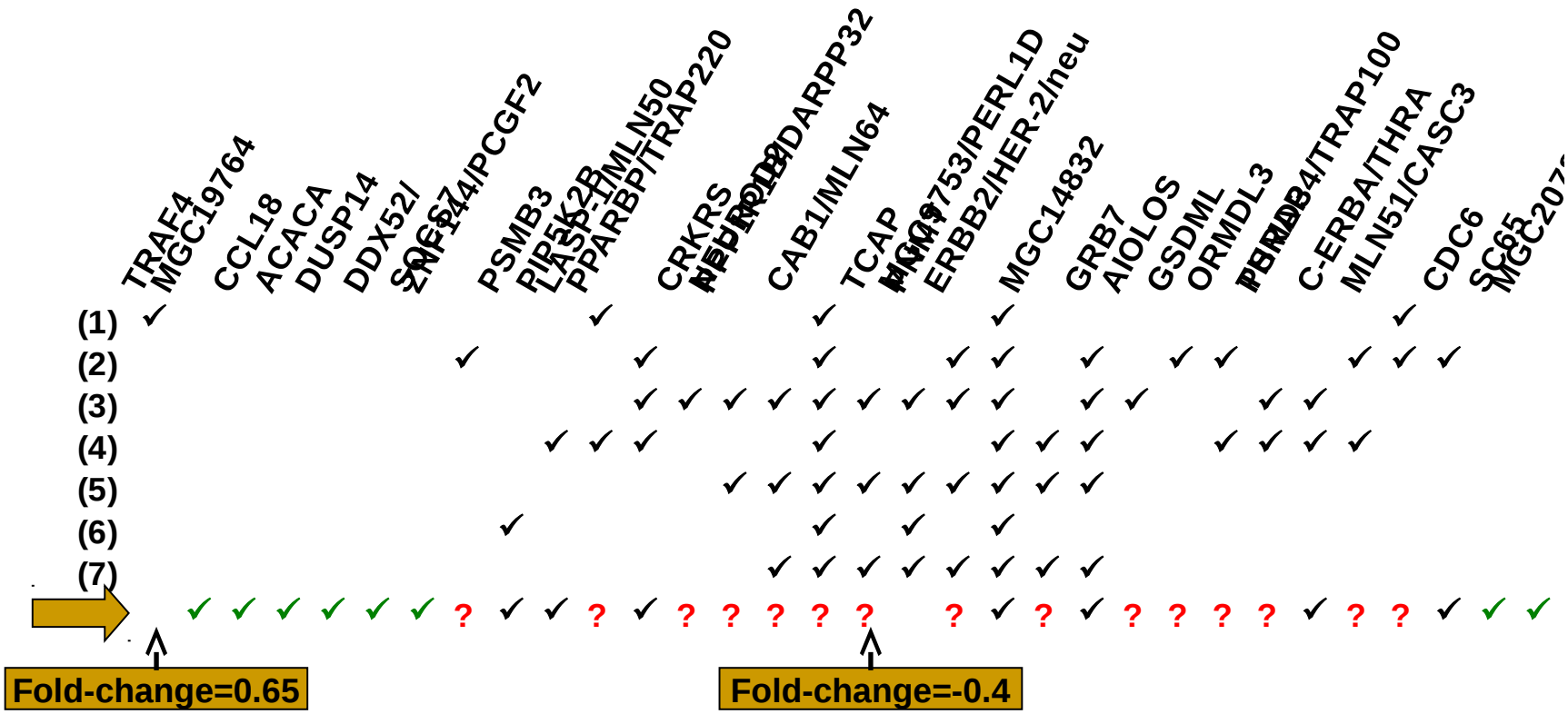
(1) Bieche et al..Cancer Res. 1996 Sep 1;56(17):3886-90
(2) Kauraniemi et al., 2001 Cancer Res. 2001 Nov 15;61(22):8235-40
(3) Luoh, 2002 Cancer Genet Cytogenet. 2002 Jul 1;136(1):43-7
(4) Clark et al., 2002 Genes Chromosomes Cancer. 2002 May;34(1):104-14.
(5) Kauraniemi et al., 2003 Am J Pathol. 2003 Nov;163(5):1979-84
(6) Dressman et al., 2003 Cancer Res. 2003 May 1;63(9):2194-9
(7) Katoh et al., 2004 Int J Oncol. 2004 Apr;24(4):757-63

Detail of the amplicon 17q11.2-21.2



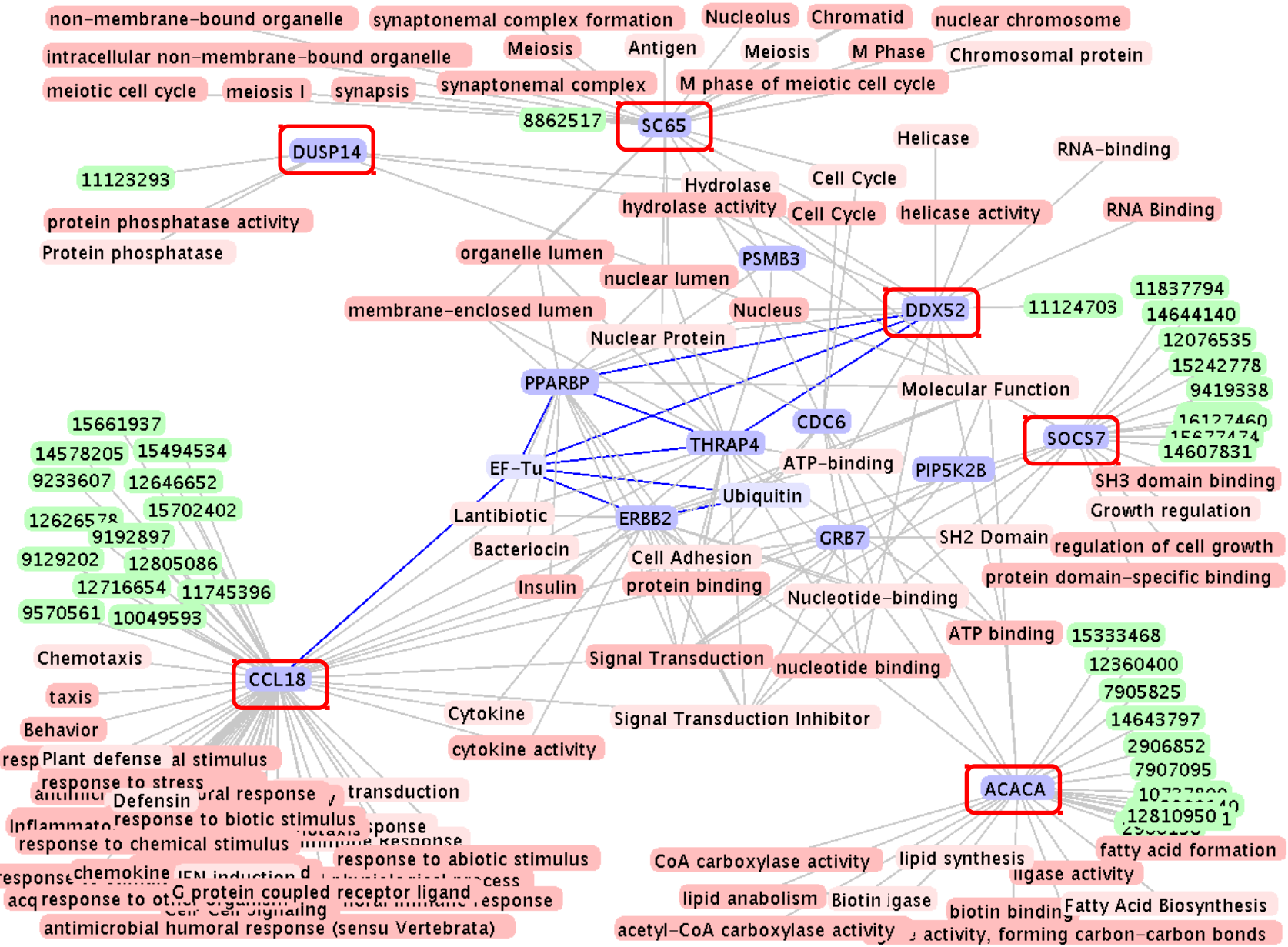
(1) Bieche et al..Cancer Res. 1996 Sep 1;56(17):3886-90
(2) Kauraniemi et al., 2001 Cancer Res. 2001 Nov 15;61(22):8235-40
(3) Luoh, 2002 Cancer Genet Cytogenet. 2002 Jul 1;136(1):43-7
(4) Clark et al., 2002 Genes Chromosomes Cancer. 2002 May;34(1):104-14.
(5) Kauraniemi et al., 2003 Am J Pathol. 2003 Nov;163(5):1979-84
(6) Dressman et al., 2003 Cancer Res. 2003 May 1;63(9):2194-9
(7) Katoh et al., 2004 Int J Oncol. 2004 Apr;24(4):757-63

Detail of the amplicon 17q11.2-21.2

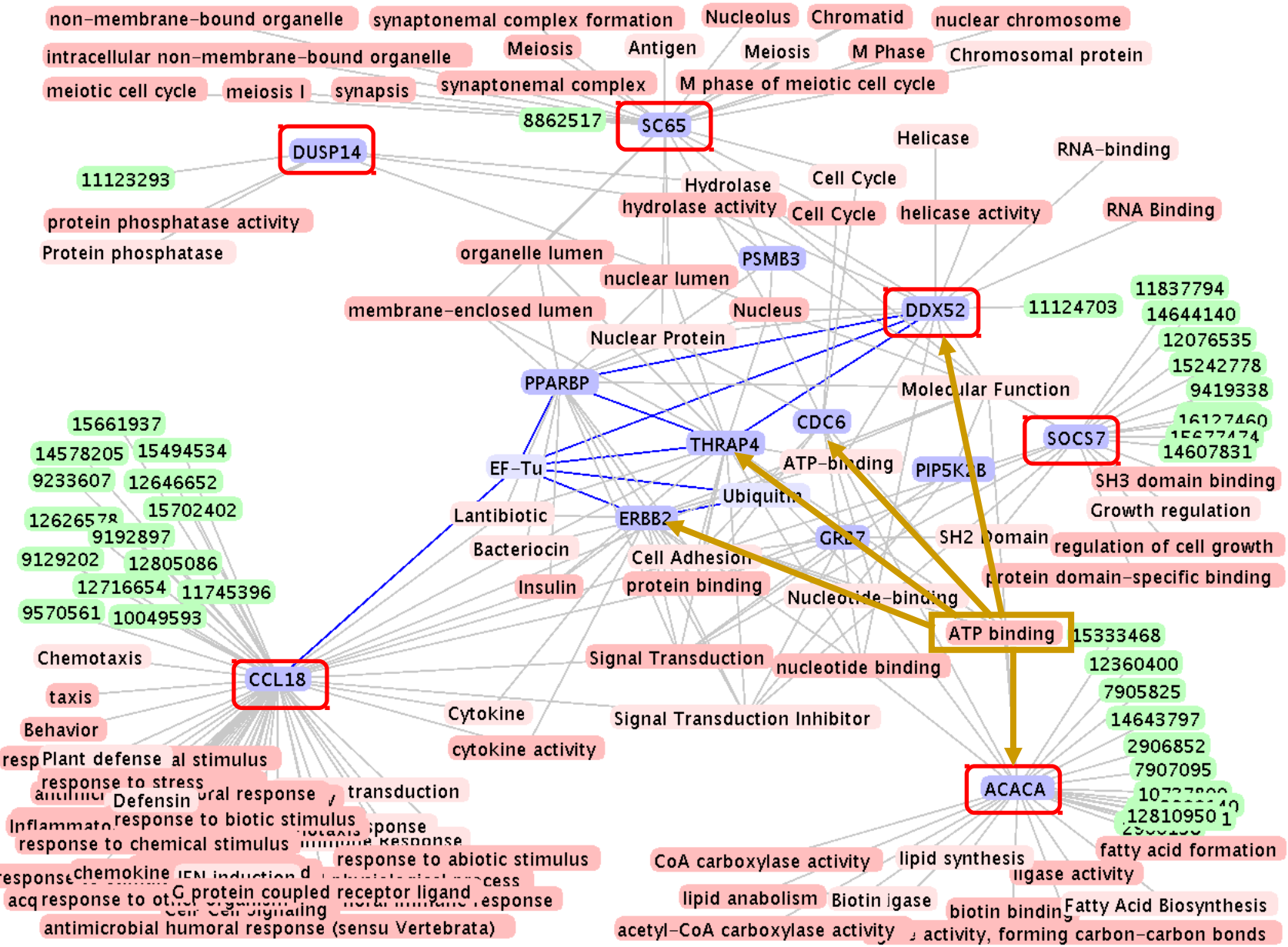


(1) Bieche et al..Cancer Res. 1996 Sep 1;56(17):3886-90
(2) Kauraniemi et al., 2001 Cancer Res. 2001 Nov 15;61(22):8235-40
(3) Luoh, 2002 Cancer Genet Cytogenet. 2002 Jul 1;136(1):43-7
(4) Clark et al., 2002 Genes Chromosomes Cancer. 2002 May;34(1):104-14.
(5) Kauraniemi et al., 2003 Am J Pathol. 2003 Nov;163(5):1979-84
(6) Dressman et al., 2003 Cancer Res. 2003 May 1;63(9):2194-9
(7) Katoh et al., 2004 Int J Oncol. 2004 Apr;24(4):757-63

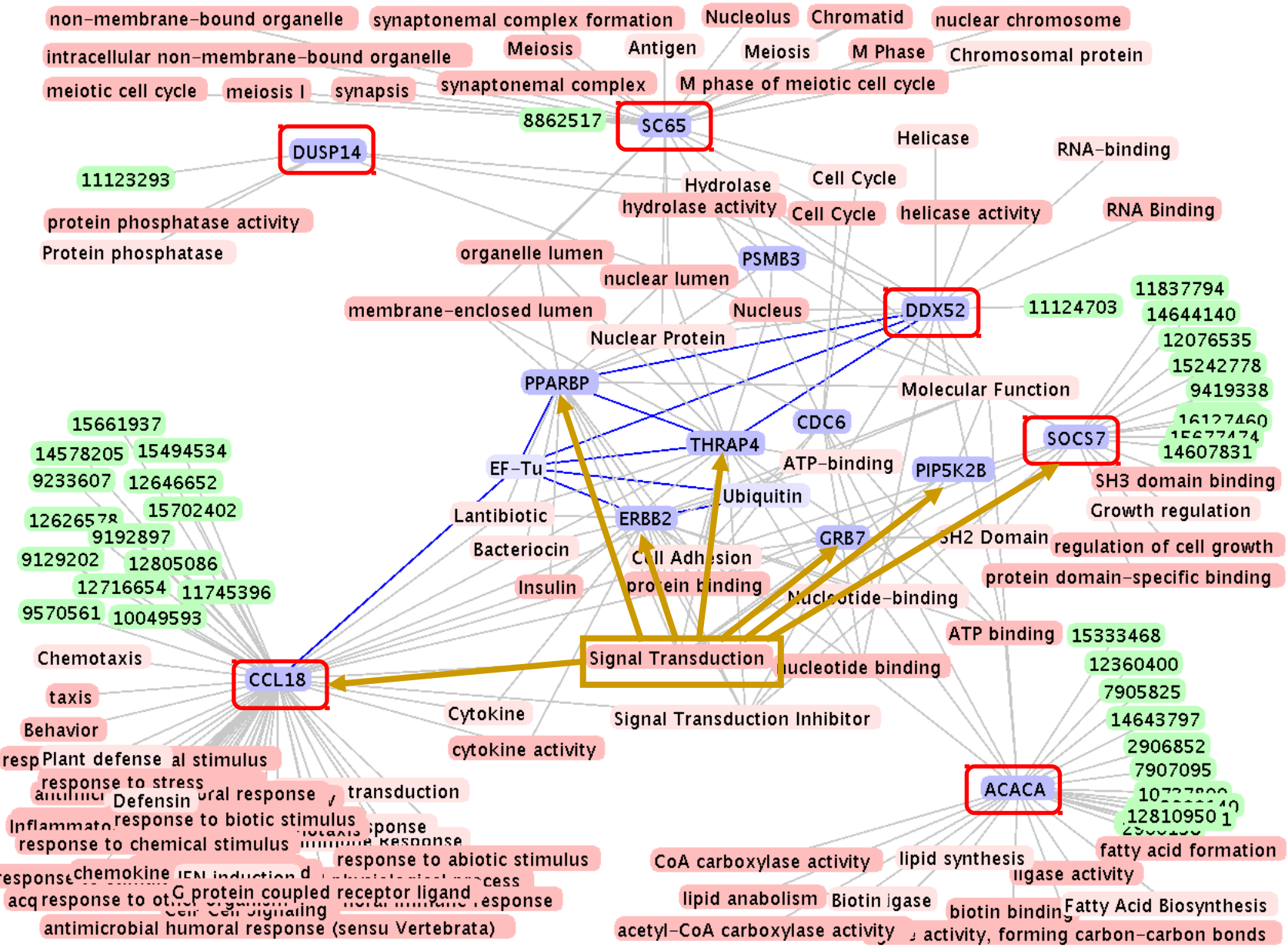
Annotations of new identified proteins



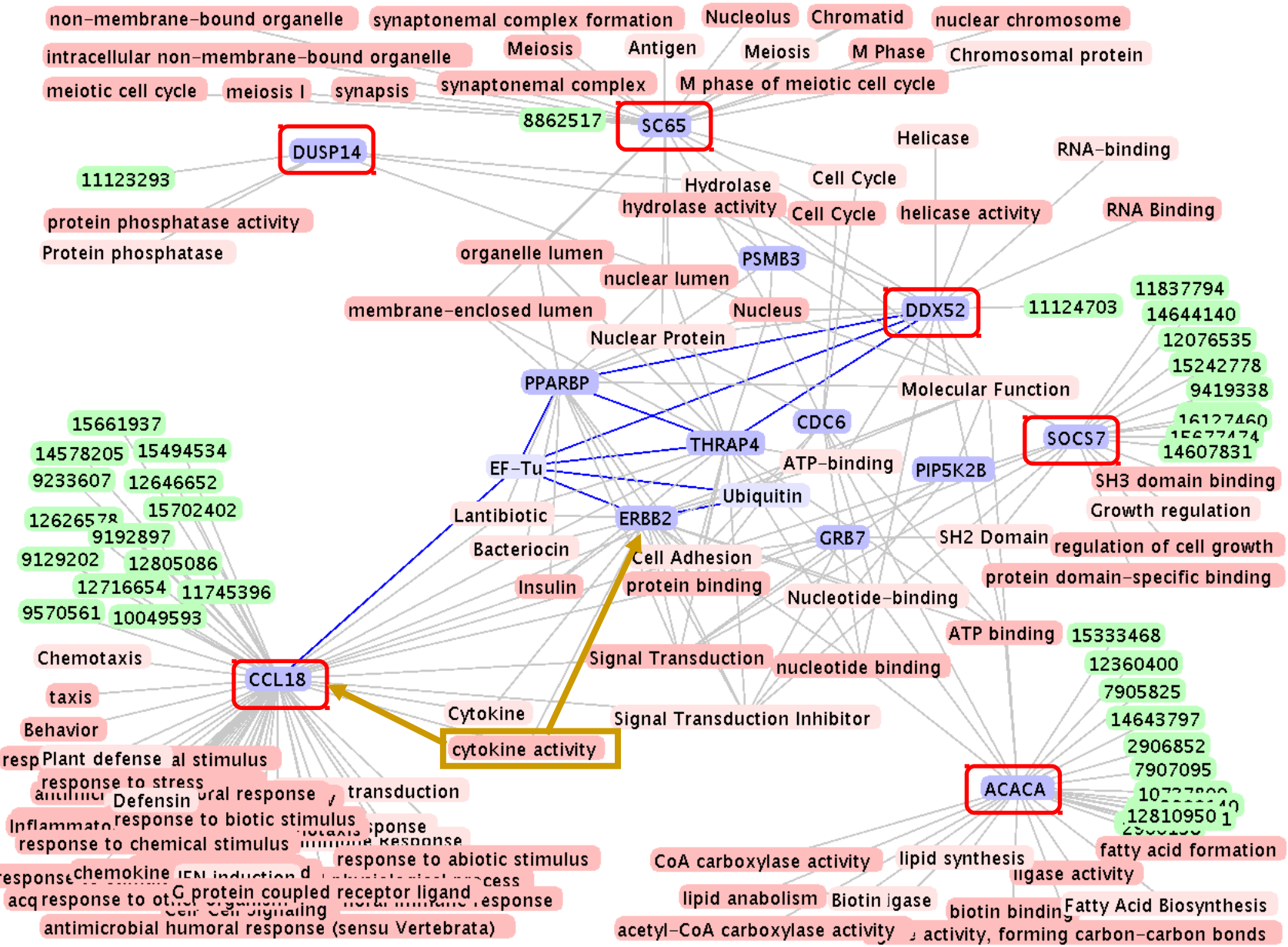
Annotations of new identified proteins



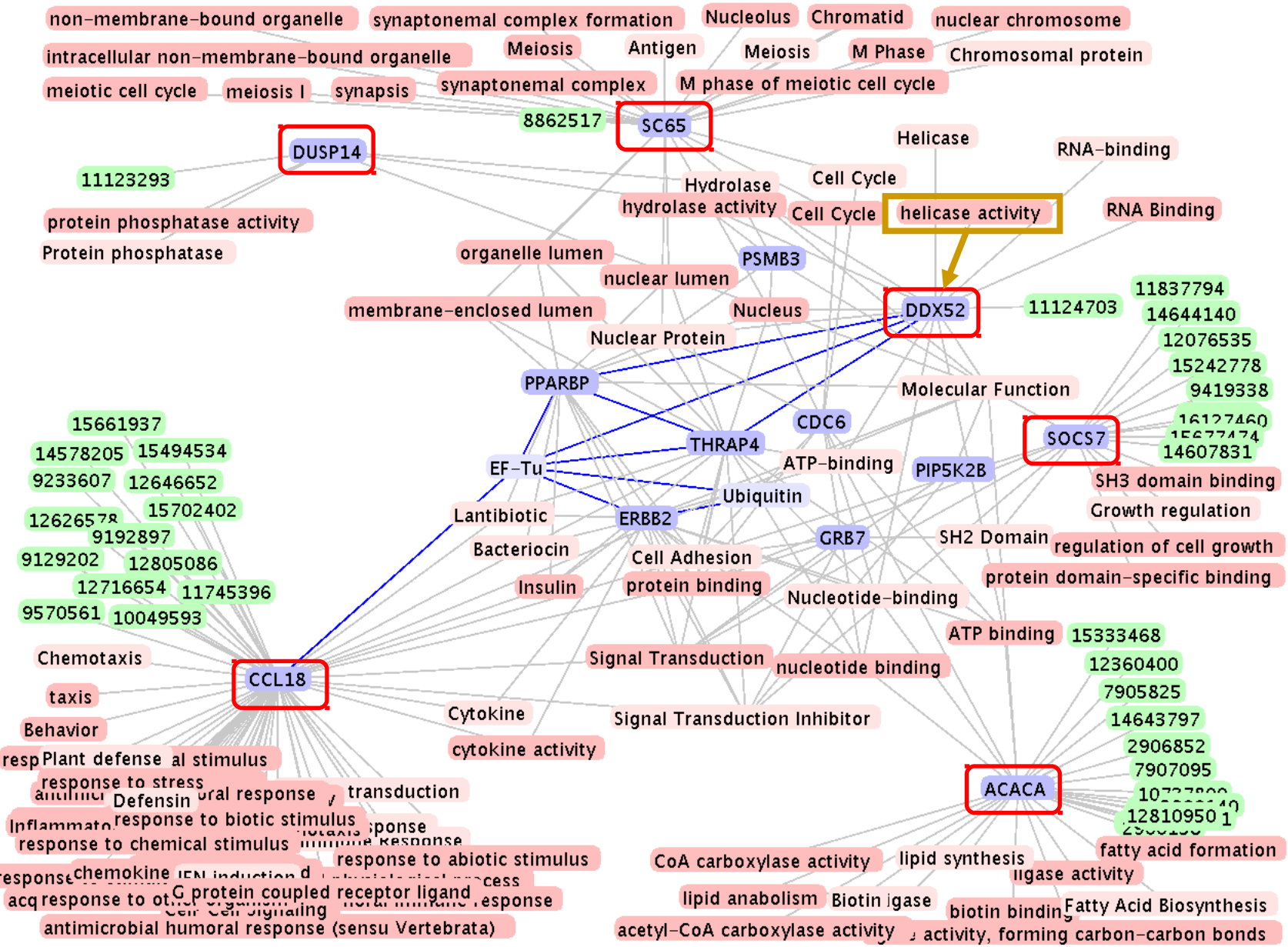
Annotations of new identified proteins



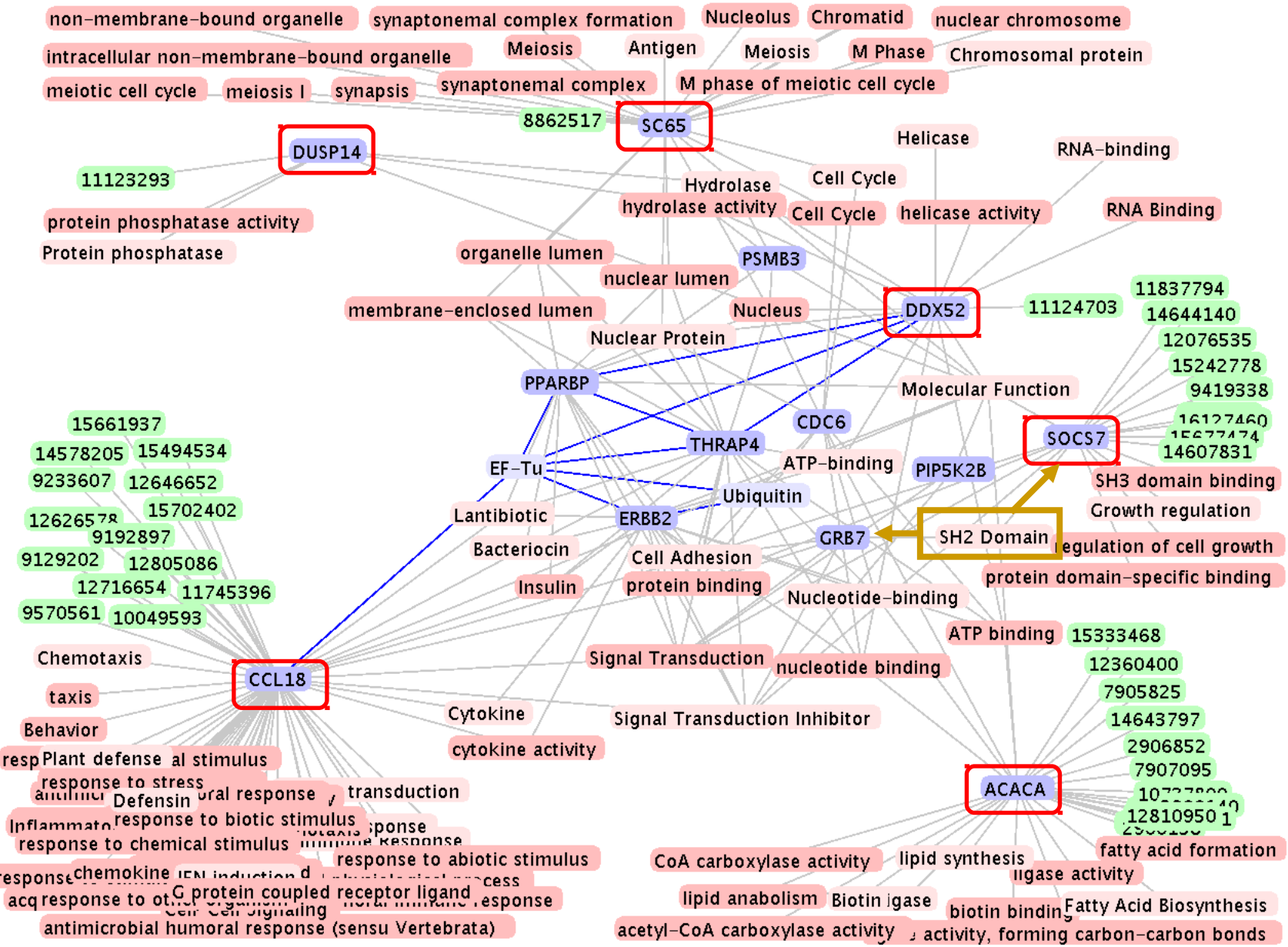
Annotations of new identified proteins



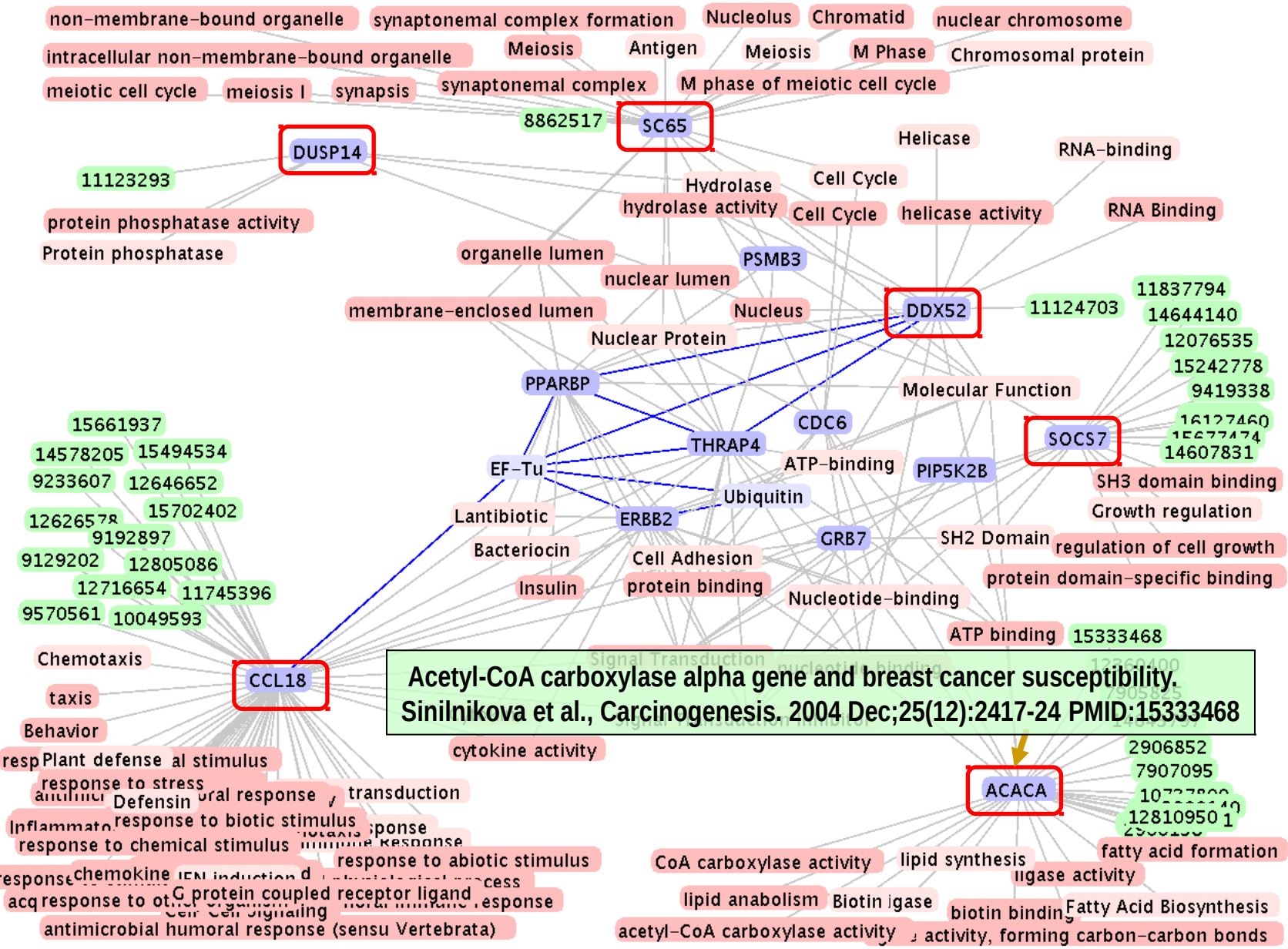
Annotations of new identified proteins



Annotations of new identified proteins



Annotations of new identified proteins



Acetyl-CoA carboxylase alpha gene and breast cancer susceptibility. Sinilnikova et al., Carcinogenesis. 2004 Dec;25(12):2417-24 PMID:15333468

Conclusion

- Semantic networks can be used:
  - to facilitate information integration
  - to extract valuable data
- An application to the analysis of expression values, allowed us:
  - to detect highly amplified zones
  - to highlight several genes potentially implied in breast cancer