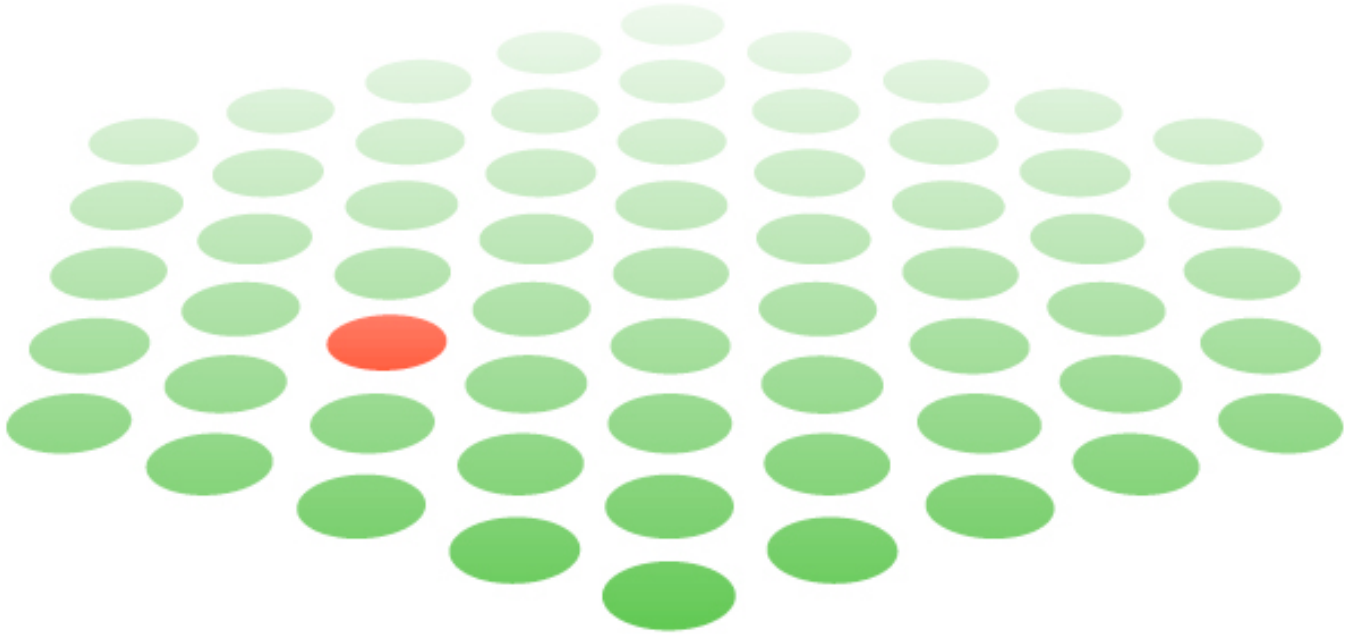
# Open access – making the most of biomedical literature mining



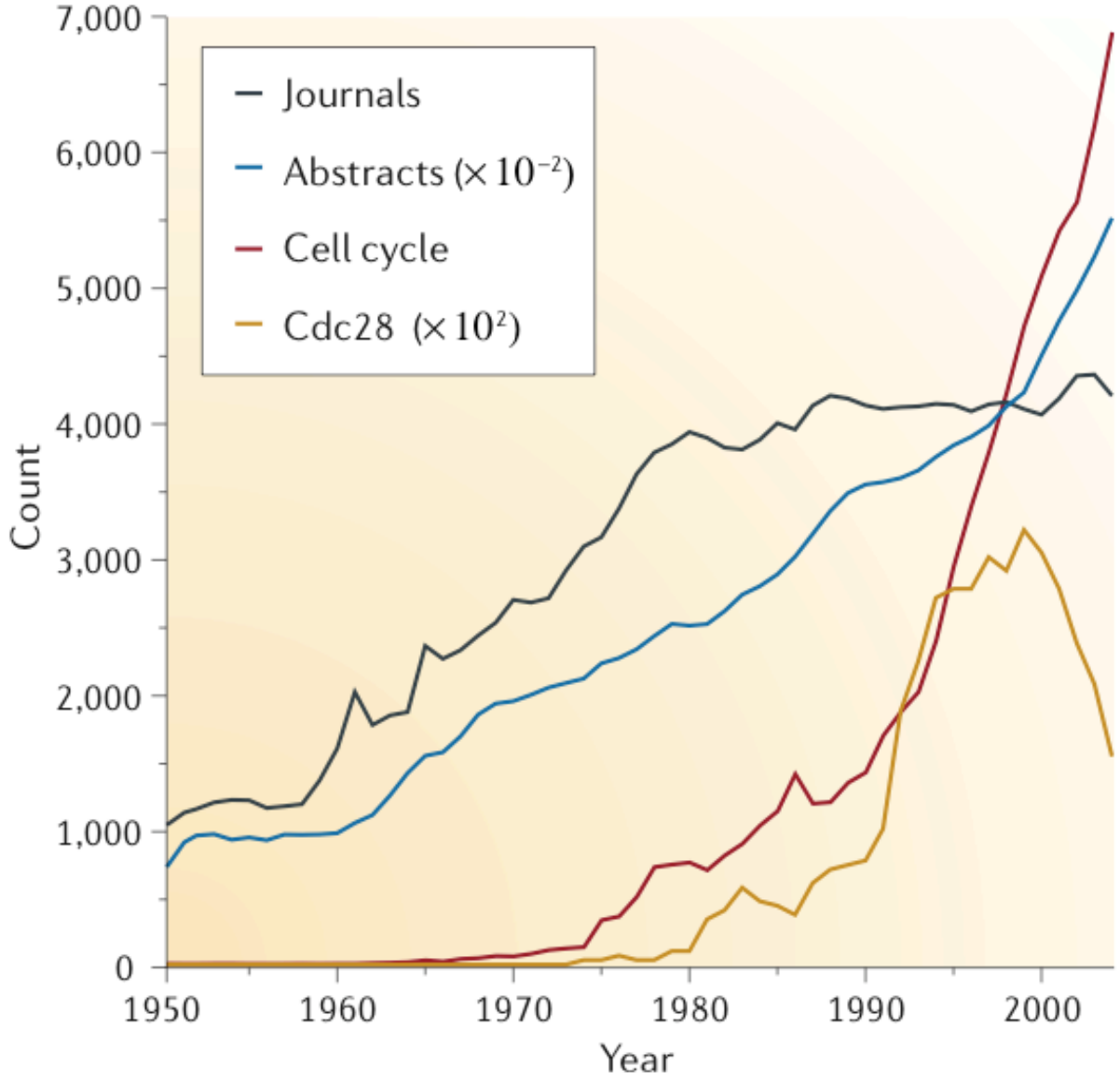
Lars Juhl Jensen
EMBL Heidelberg

# why open access?

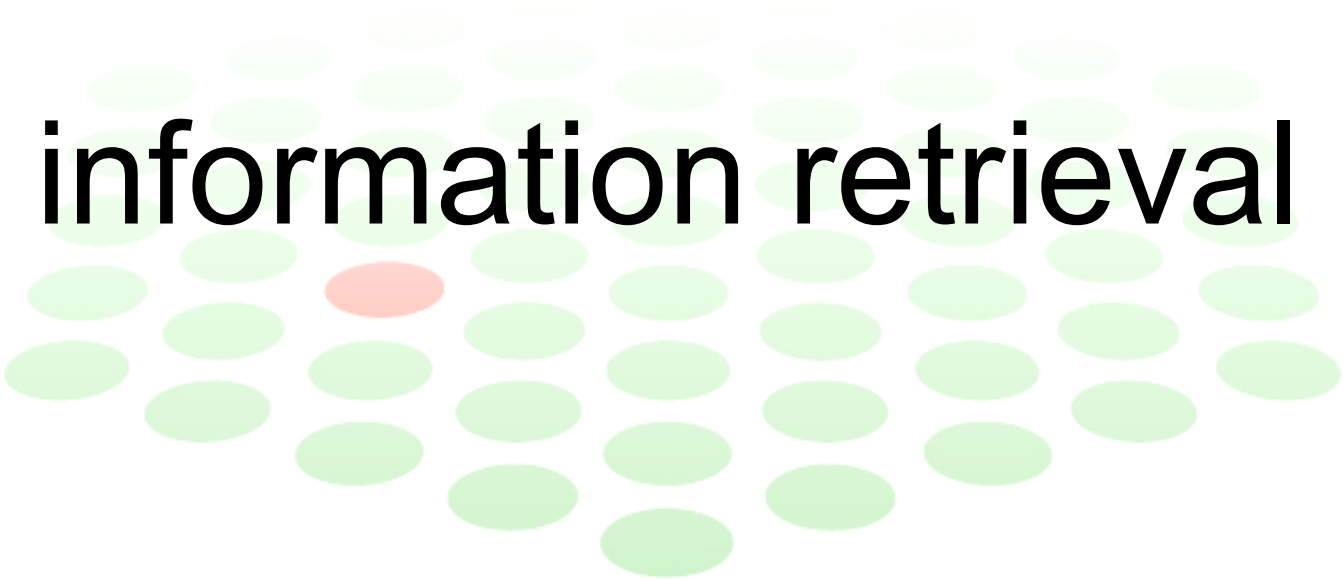
# why biomedicine?

# why literature mining?

# MEDLINE




Jensen et al., Nature Reviews Genetics, 2006


information retrieval

# finding the papers

NCBI

PubMed

National Library of Medicine NLM

My NCBI
[Sign In] [Register]

All Databases | PubMed | Nucleotide | Protein | Genome | Structure | OMIM | PMC | Journals | Books

Search PubMed for Go Clear

Limits | Preview/Index | History | Clipboard | Details


About Entrez

Text Version

Entrez PubMed
Overview
Help | FAQ
Tutorial
New/Noteworthy
E-Utilities

PubMed Services
Journals Database
MeSH Database
Single Citation Matcher
Batch Citation Matcher
Clinical Queries
Special Queries
LinkOut
My NCBI (Cubby)

Related Resources
Order Documents
NLM Mobile
NLM Catalog
NLM Gateway
TOXNET
Consumer Health
Clinical Alerts
ClinicalTrials.gov
PubMed Central


- To get started, enter one or more search terms.
- Search terms may be topics, authors or journals.


**News flash: deliver PubMed search results directly to your desktop with an RSS feed.**

To set up an RSS feed:
(1) Run your search in PubMed.
(2) Select *RSS Feed* from the *Send to* menu.
(3) Click *Create Feed* and copy the XML icon into your RSS Reader.

Read the PubMed Help to explore other options for automated e-mail updates using My NCBI.

PubMed is a service of the National Library of Medicine that includes over 15 million citations from MEDLINE and other life science journals for biomedical articles back to the 1950s. PubMed includes links to full text articles and other related resources.

Mitotic cyclin (Clb2)-bound Cdc28 (Cdk1 homolog) directly phosphorylated Swe1 and this modification served as a priming step to promote subsequent Cdc5-dependent Swe1 hyperphosphorylation and degradation

# entity recognition

# identifying the substance(s)

Mitotic cyclin (Clb2)-bound Cdc28 (Cdk1 homolog) directly phosphorylated Swe1 and this modification served as a priming step to promote subsequent Cdc5-dependent Swe1 hyperphosphorylation and degradation

iHOp

information hyperlinked over proteins

Search Gene

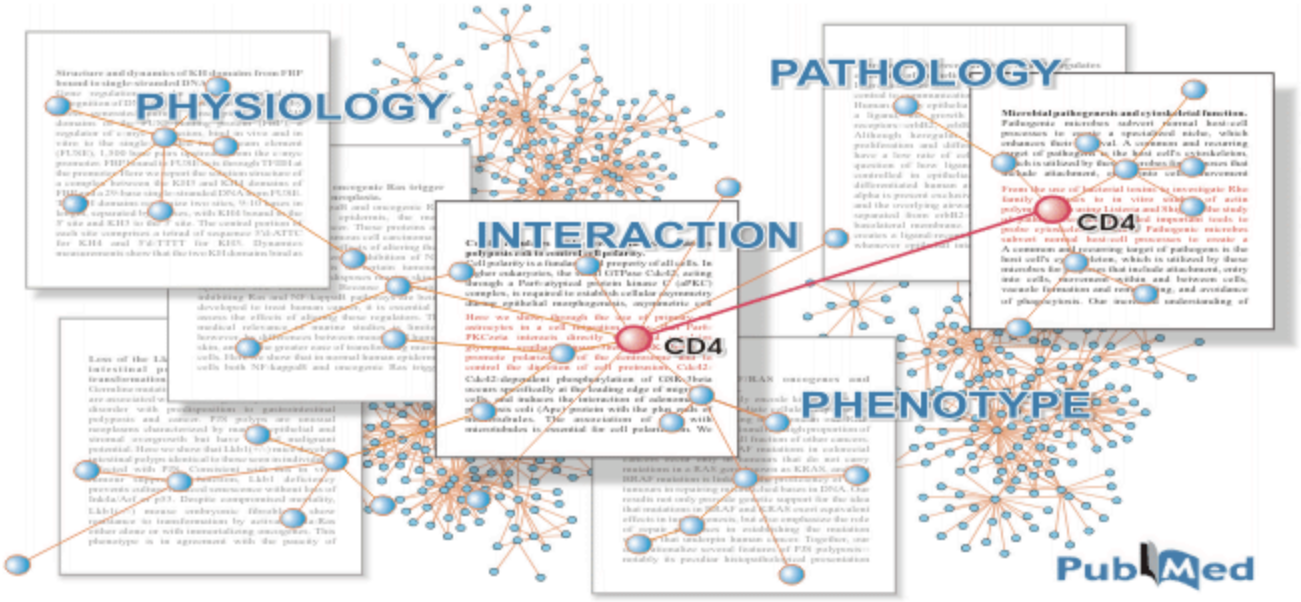
Gene Model
Developer's Zone new
Contact
Help

cocoon

Java

site powered by pdg

Concept & Implementation
by Robert Hoffmann




Hoffmann, R., Valencia, A. A Gene Network for Navigating the Literature. Nature Genetics 36, 664 (2004)

**Search for a gene *synonym* or *accession number*...** (Click here for an example: SNF1)

all fields in all organism

[SEARCH]

| Symbol | Name | Synonyms | Organism |
|---|---|---|---|
| **CDC28** | Cell division control protein 28 | CDK1, HSL5, SRM5, YBR1211, YBR160W | Saccharomyces cerevisiae |

| | |
|---|---|
| UniProt | P00546 |
| IntAct | P00546 |
| NCBI Gene | 852457 |
| **NCBI RefSeq** | NP_009718 |
| NCBI Accession | CAA25065, CAA56509, CAA85119 |

**Homologues of CDC28 ...** new

**Definitions for CDC28** ...

**Enhanced PubMed/Google query ...** new

WARNING: Please keep in mind that gene detection is done automatically and can exhibit a certain error. Read more.

Find in this Page

Furthermore, SW14 associates with CLB2 protein and is a substrate for the CLB2-**associated** CDC28 kinase in vitro.

Furthermore, the Cks1 protein was shown to be physically **associated** with active forms of the Cdc28 protein kinase.

The cyclin-dependent kinase Cdc28p **associates** with the cyclin Clb2p to induce mitosis in the yeast Saccharomyces cerevisiae.

We find that G1 arrest in the cdc37-1 mutant is accompanied by a decrease in the Cdc28 activity **associated** with the G1 cyclin Cln2.

We found that Hct1 was **phosphorylated** in vivo at multiple CDK consensus sites during cell cycle stages when activity of the cyclin-dependent kinase Cdc28 is high and APC activity is low.

It is likely, therefore, that Cks1 mediates a more specialized **function** of the Cdc28 kinase such as its ability to form specific multimeric complexes or to localize properly in cellular compartments.

Cdc37 **promotes** the stability of protein kinases Cdc28 and Cak1.

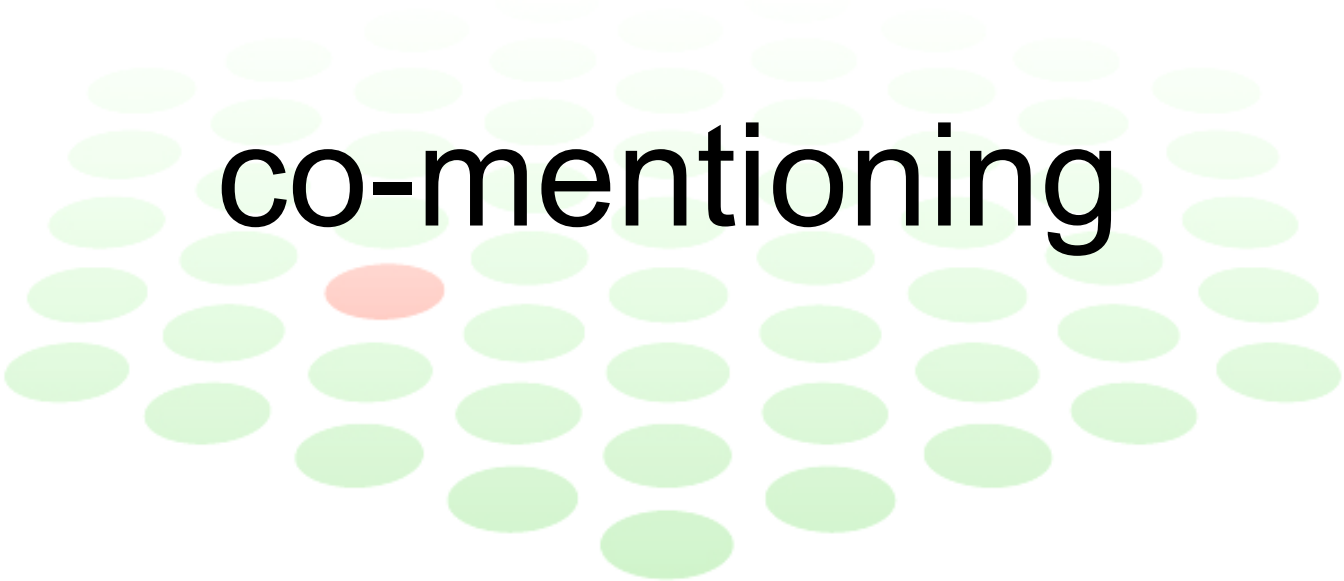
In addition, Cdc37 **promotes** the production of Cak1, but not that of Cdc28, when coexpressed in insect cells.

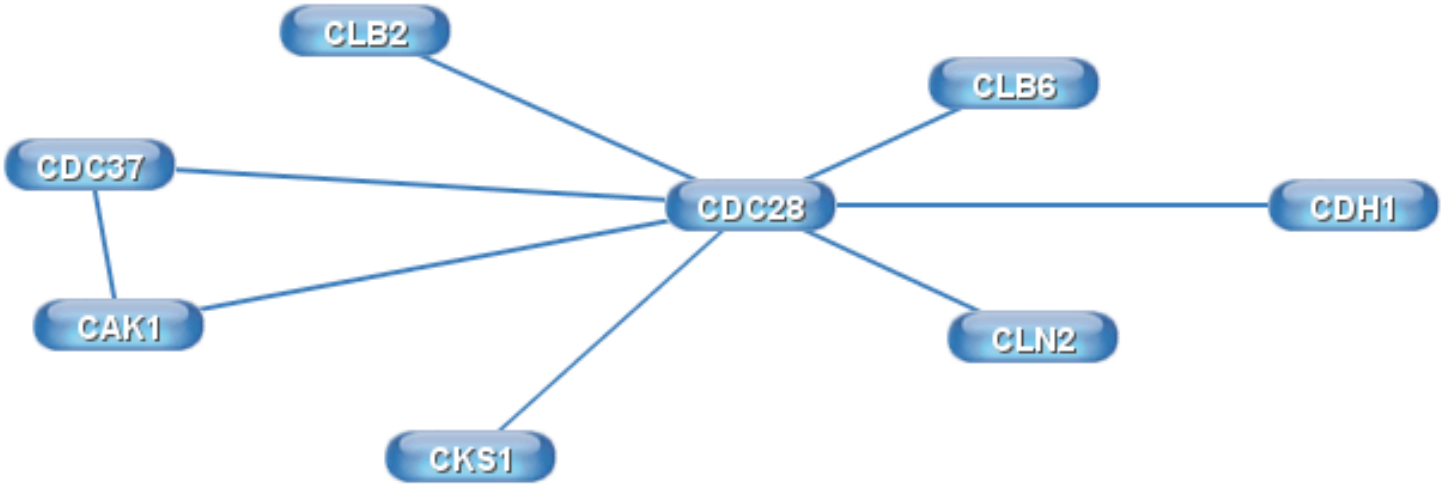
The B-type cyclins Clb5 and Clb6 are the primary activators of the S phase **function** of the budding yeast CDK Cdc28.

All three cak1 mutants displayed significant synthetic interactions with loss-of-**function** mutations in CDC28 and KIN28.

# information extraction

# formalizing the facts


co-mentioning




[Redo graph layout]

The B-type cyclins Clb5 [CLB5] and Clb6 [CLB6] are the primary activators of the S phase **function** of the budding yeast CDK Cdc28 [CDC28].

Furthermore, SW14 associates with CLB2 [CLB2] protein and is a substrate for the CLB2-**associated** CDC28 [CDC28] kinase in vitro.

The cyclin-dependent kinase Cdc28 [CDC28]p **associates** with the cyclin Clb2 [CLB2]p to induce mitosis in the yeast Saccharomyces cerevisiae.

In addition, Cdc37 [CDC37] **promotes** the production of Cak1 [CAK1], but not that of Cdc28 [CDC28], when coexpressed in insect cells.

We found that Hct1 [CDH1] was **phosphorylated** in vivo at multiple CDK consensus sites during cell cycle stages when activity of the cyclin-dependent kinase Cdc28 [CDC28] is high and APC activity is low.

All three cak1 [CAK1] mutants displayed significant synthetic interactions with loss-of-**function** mutations in CDC28 [CDC28] and KIN28 [KIN28].

Cdc37 [CDC37] **promotes** the stability of protein kinases Cdc28 [CDC28] and Cak1 [CAK1].

Furthermore, the Cks1 [CKS1] protein was shown to be physically **associated** with active forms of the Cdc28 [CDC28] protein kinase.

We find that G1 arrest in the cdc37-1 mutant is accompanied by a decrease in the Cdc28 [CDC28] activity **associated** with the G1 cyclin Cln2 [CLN2].

It is likely, therefore, that Cks1 [CKS1] mediates a more specialized **function** of the Cdc28 [CDC28] kinase such as its ability to form specific multimeric complexes or to localize properly in cellular compartments.

# NLP

## Natural Language Processing

Mitotic cyclin (Clb2)-bound Cdc28 (Cdk1 homolog) directly phosphorylated Swe1 and this modification served as a priming step to promote subsequent Cdc5-dependent Swe1 hyperphosphorylation and degradation

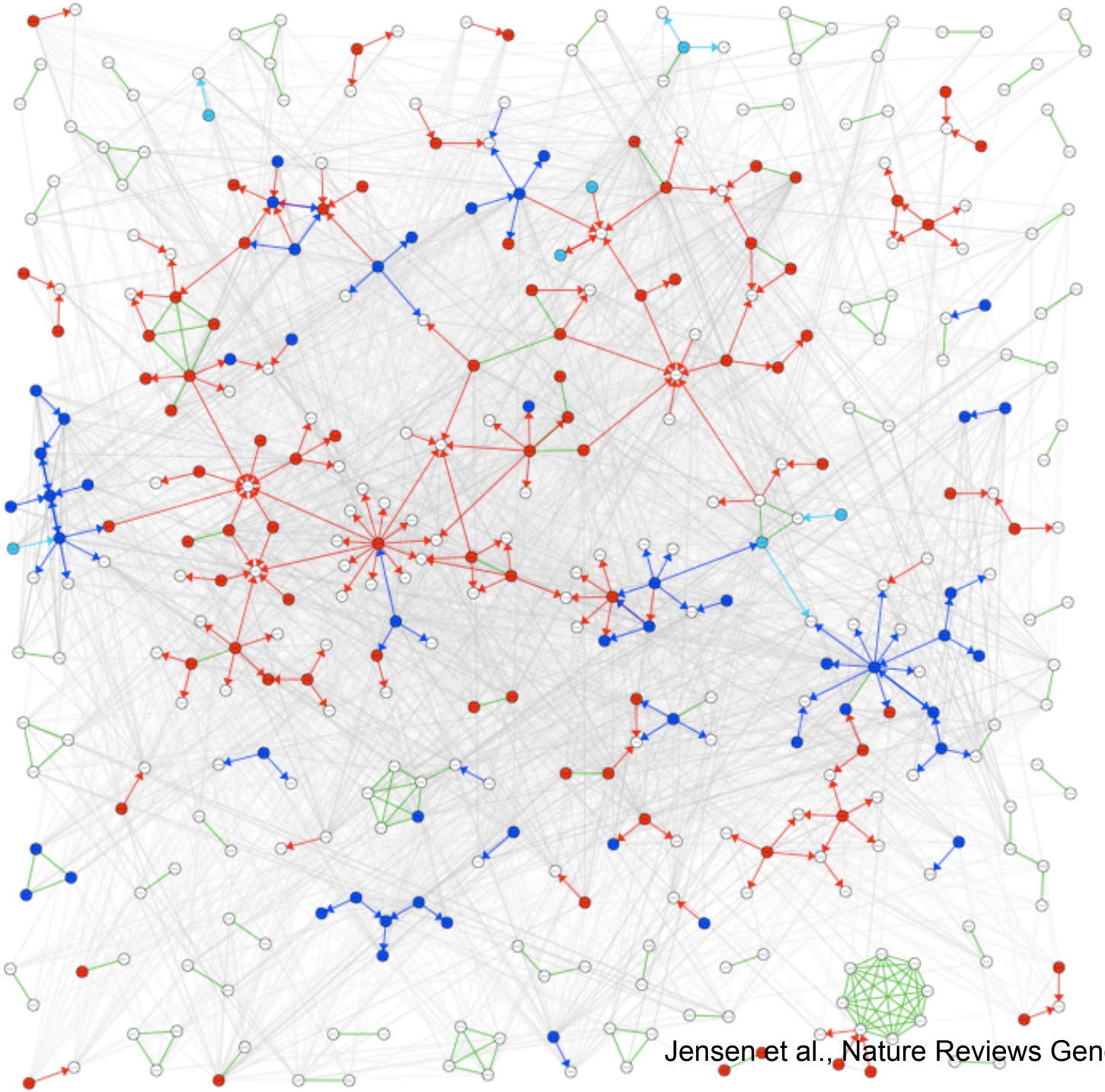
Gene and protein names

Cue words for entity recognition

Verbs for relation extraction

[nxgene The GAL4 gene]

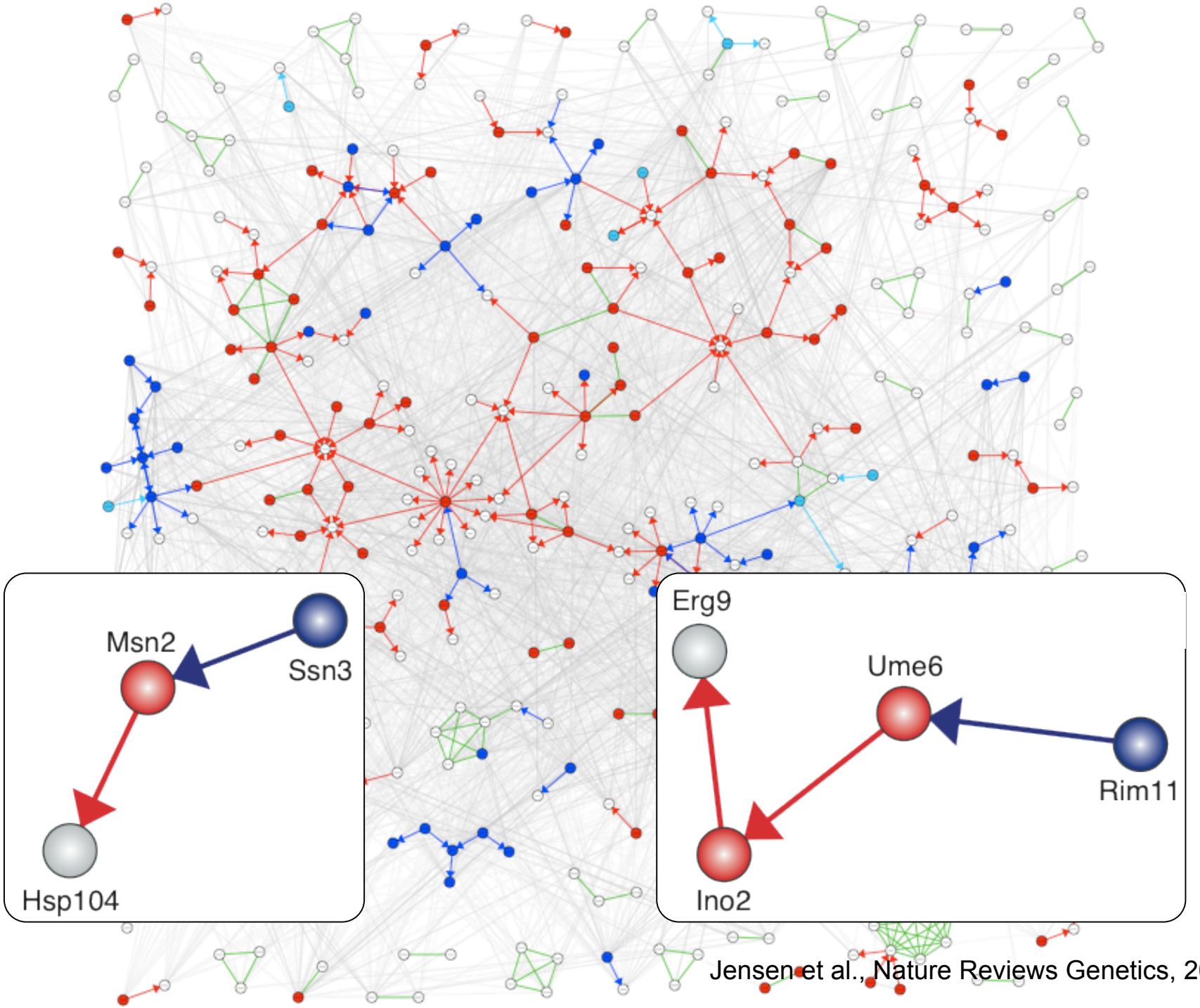
[nxexpr The expression of
[nxgene the cytochrome genes
[nxpg CYC1 and CYC7]]]
is controlled by
[nxpg HAP1]



Jensen et al., Nature Reviews Genetics, 2006


new discoveries

# text mining




Jensen et al., Nature Reviews Genetics, 2006

# integration of text and data

# open access databases

Home · Download · Help/Info STRING

## STRING - Search Tool for the Retrieval of Interacting Genes/Proteins

**Enter your gene/protein of interest ...**

identifier: e.g. 'trpB', 'ANP1_YEAST', ... you may also upload a list

alternatively, paste an amino-acid sequence:

interactors wanted: GO ! Reset COGs Proteins

**What it does ...**

STRING is a database of known and predicted protein-protein interactions.
The interactions include direct (physical) and indirect (functional) associations; they are derived from four sources:

Genomic Context | High-throughput Experiments | (Conserved) Coexpression | Previous Knowledge


STRING quantitatively integrates interaction data from these sources for a large number of organisms, and transfers information between these organisms where applicable. The database currently contains 736429 proteins in 179 species.

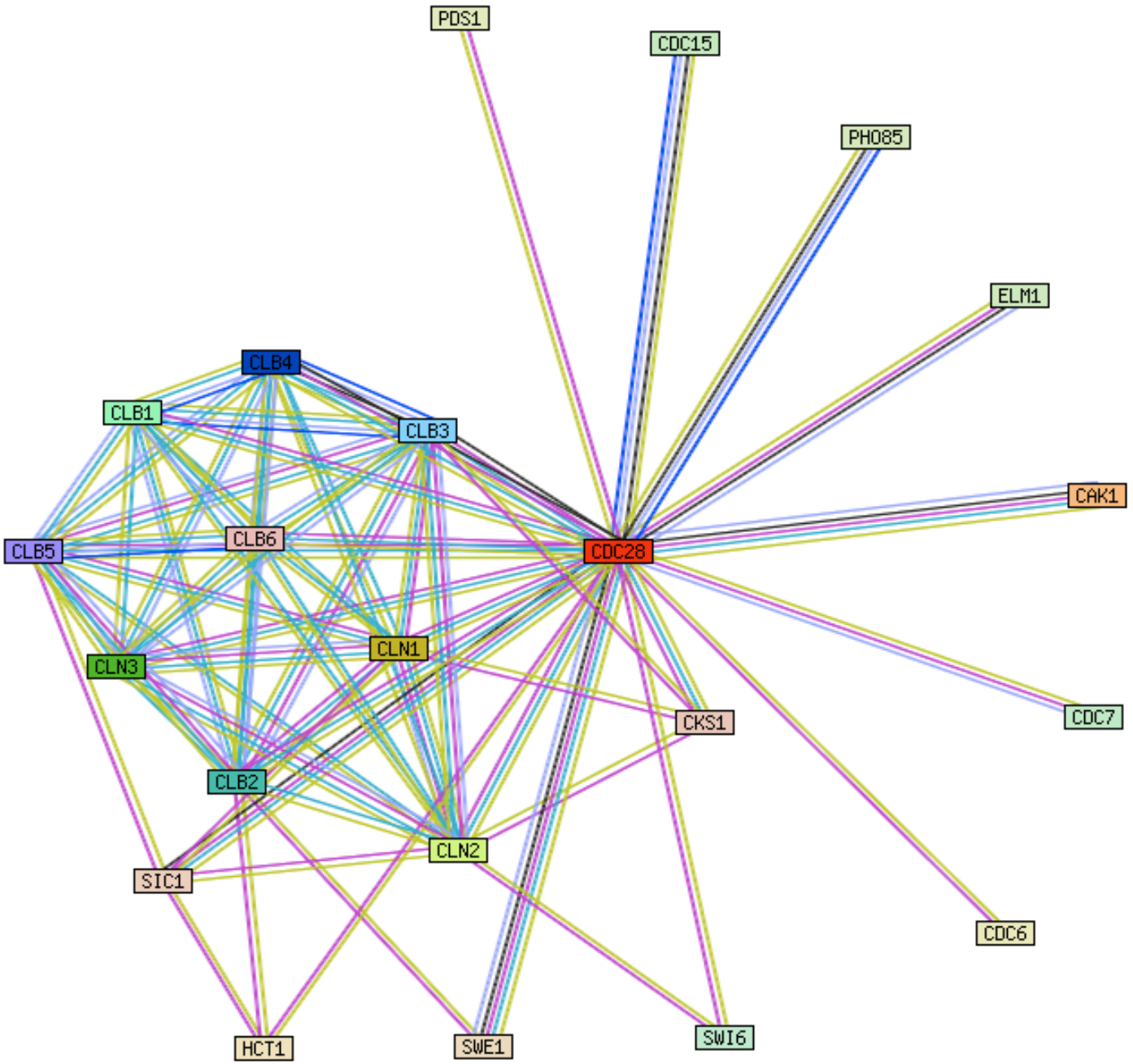
**References / Info ...**

STRING uses orthology information from the excellent COG database (Ref).
Up-to-date genomes and proteins are maintained at SWISSPROT and ENSEMBL
STRING references: von Mering et.al. 2005 / von Mering et.al. 2003 / Snel et.al. 2000.
Miscellaneous: Access Statistics, Robot Access Guide, Supported Browsers.

EMBL
• BIOCOMPUTING
• BORK GROUP

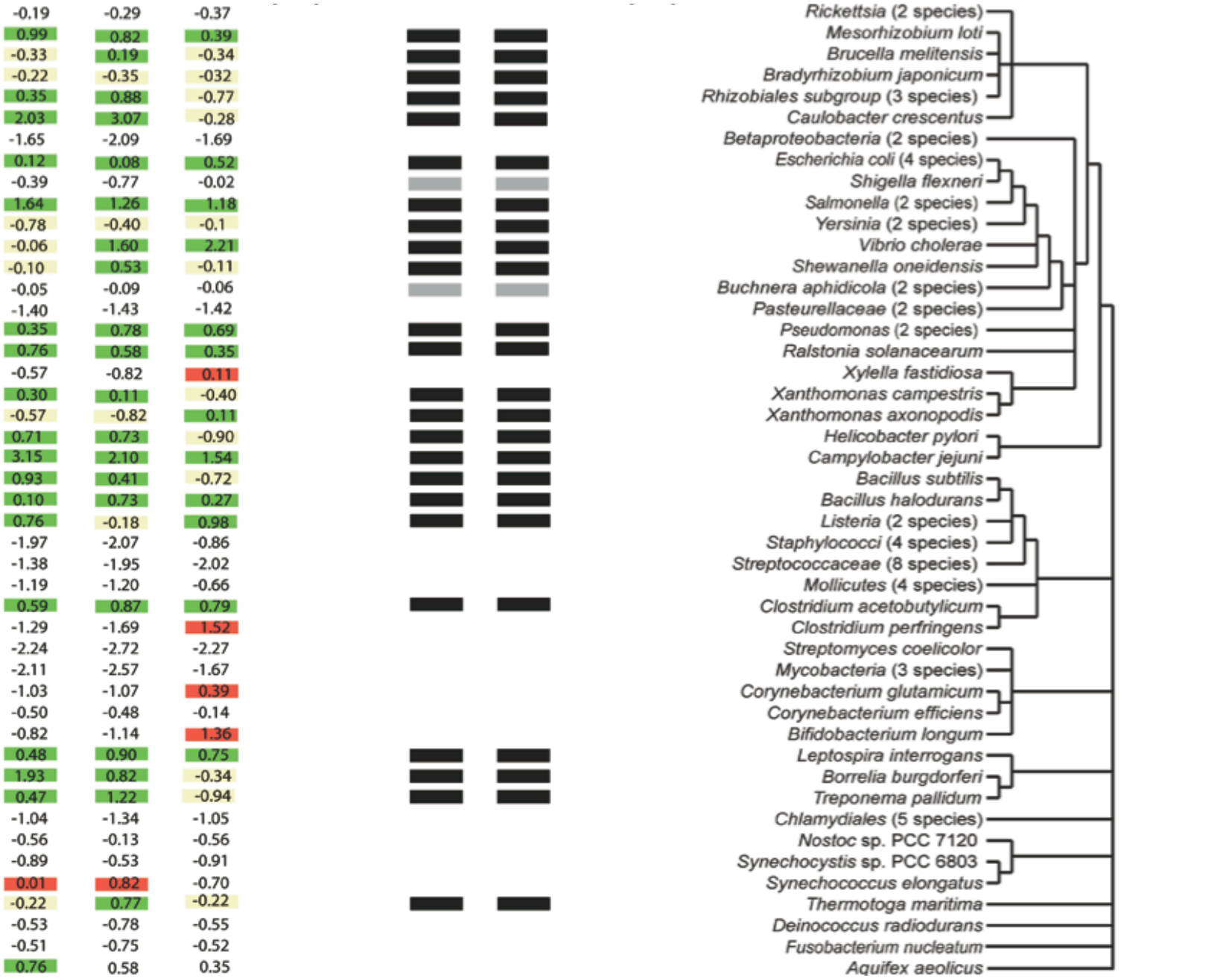
**What's New?** You are looking at release 6.2 of STRING - latest additions are the 'HPRD' and 'Reactome' databases.
**Previous Releases:** Trying to reproduce an earlier finding? Confused? Try our old releases: version 6.0, version 5.1

von Mering et al., Nucleic Acids Research, 2005

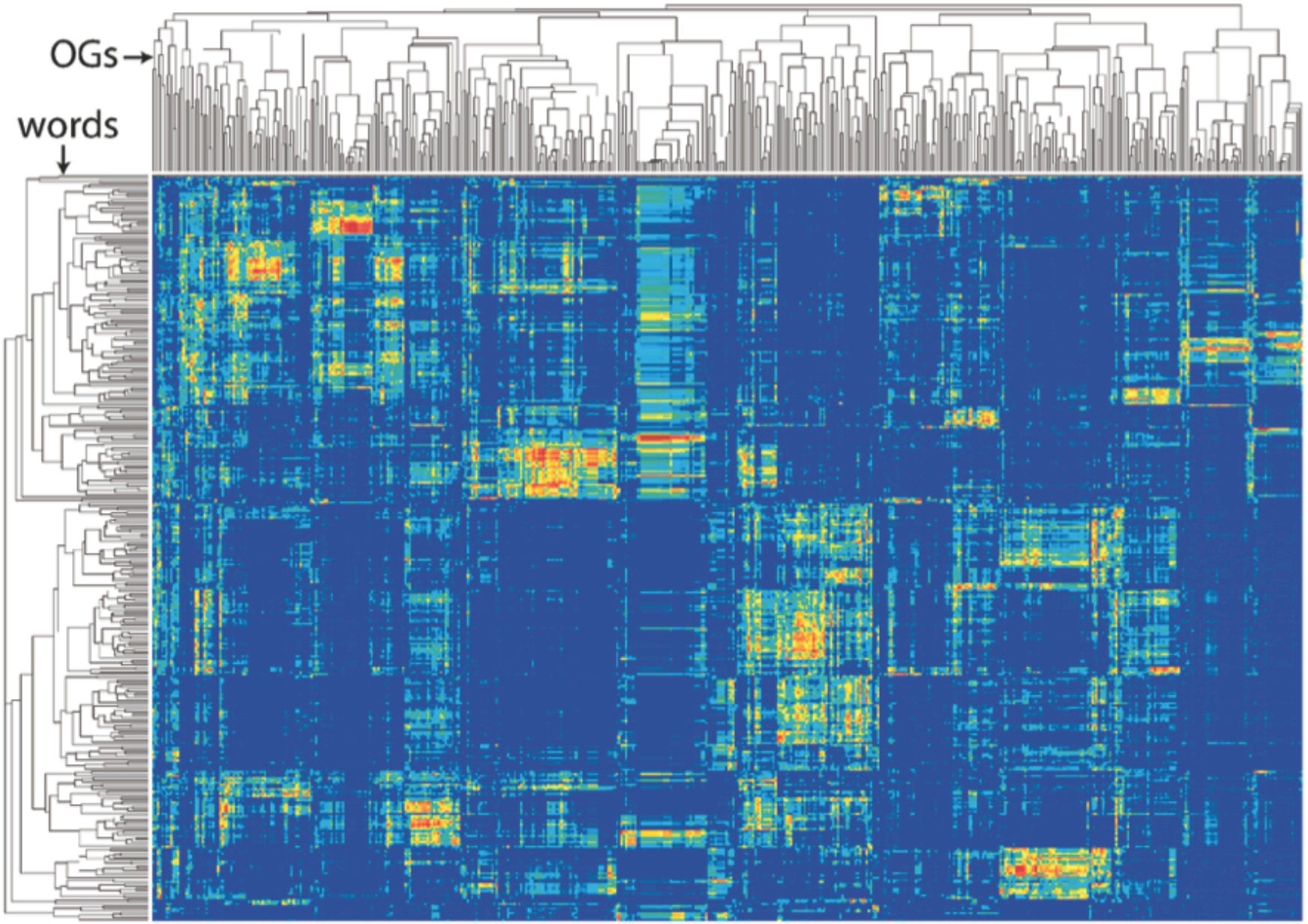

PDS1
CDC15
PHO85
ELM1
CLB4
CLB1
CLB3
CAK1
CLB6
CLB5
CDC28
CLN1
CLN3
CDC7
CKS1
CLB2
CLN2
SIC1
CDC6
HCT1
SWE1
SWI6

# network mining

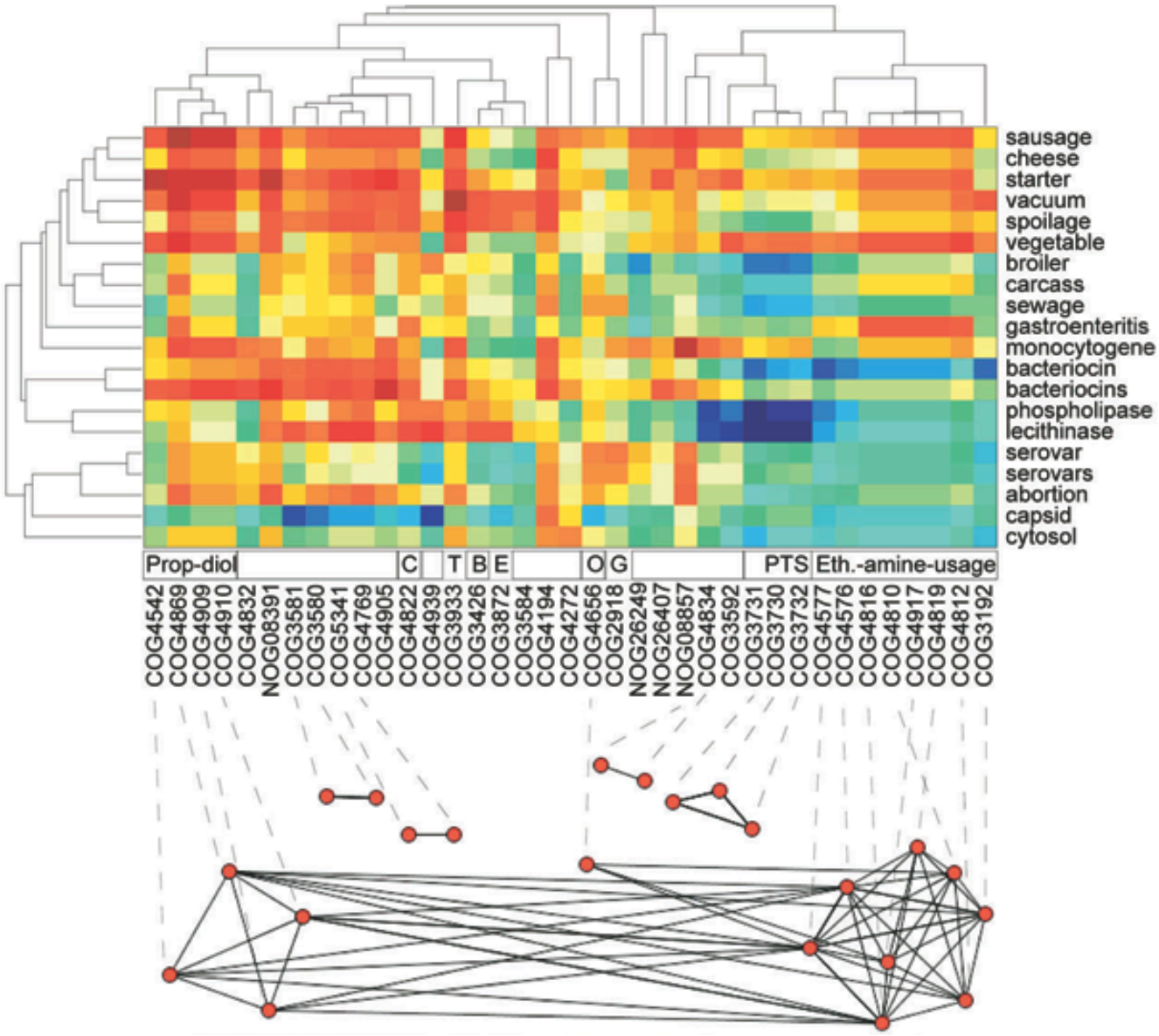
# genotype to phenotype




Korbel et al., PLoS Biology, 2005

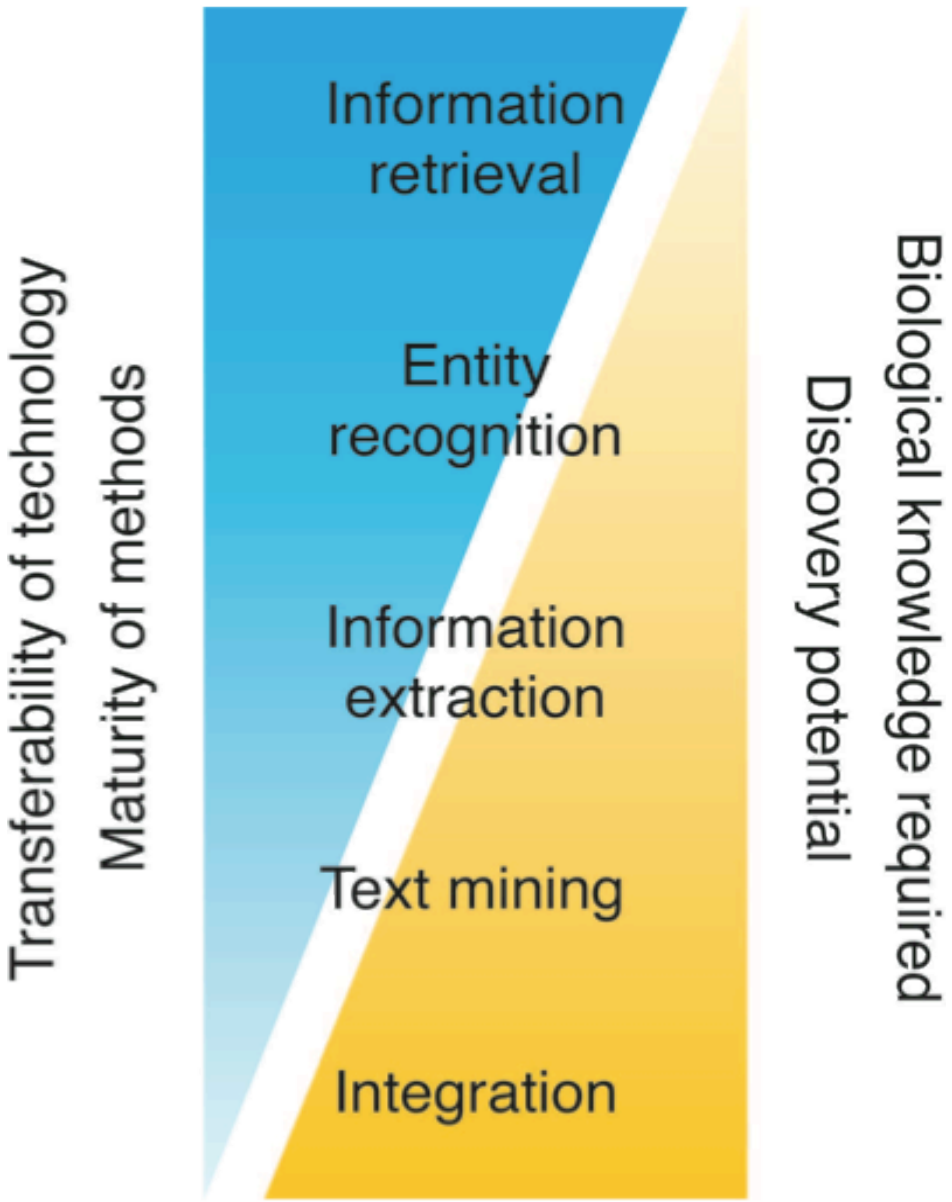



Korbel et al., PLoS Biology, 2005

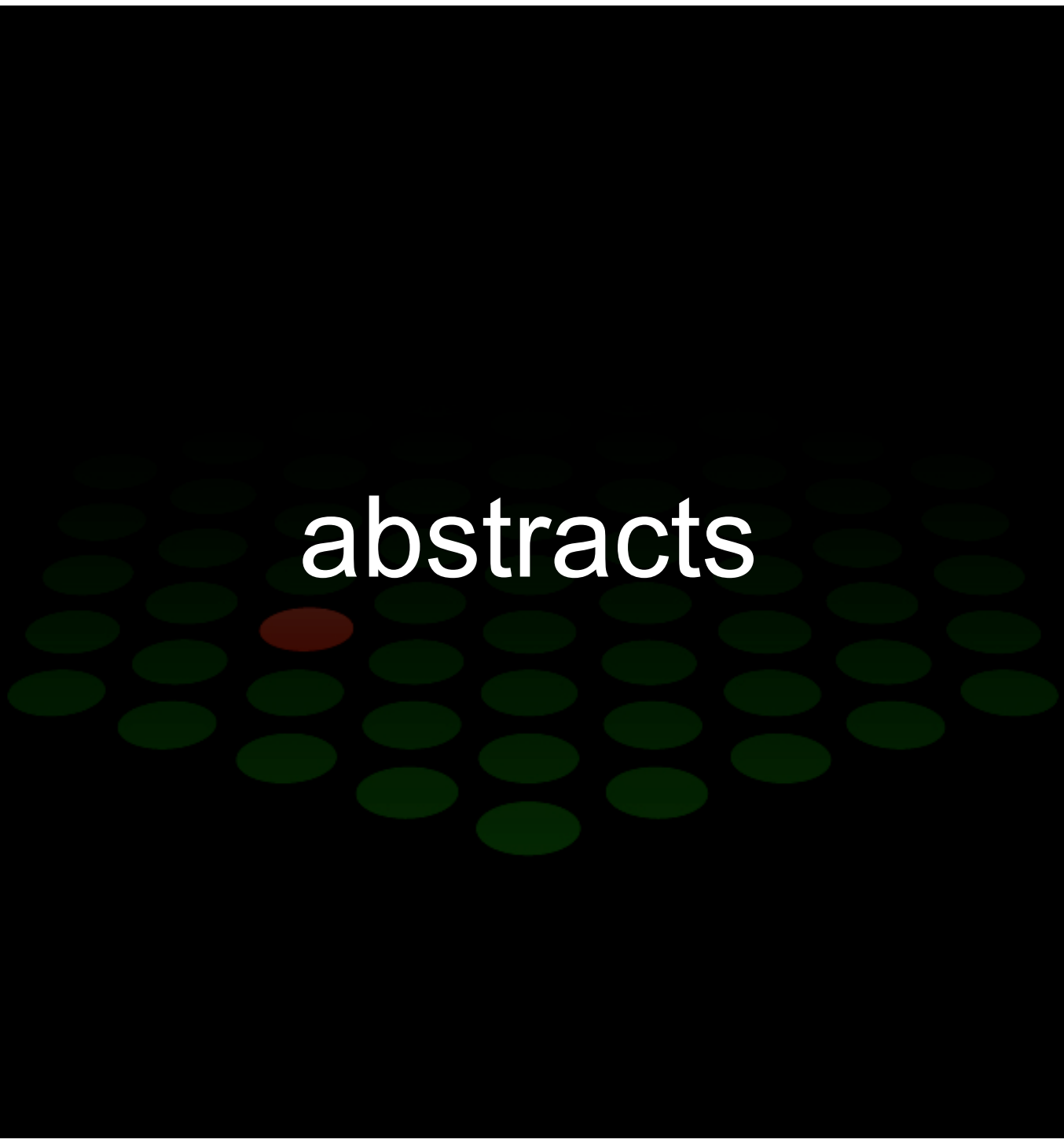


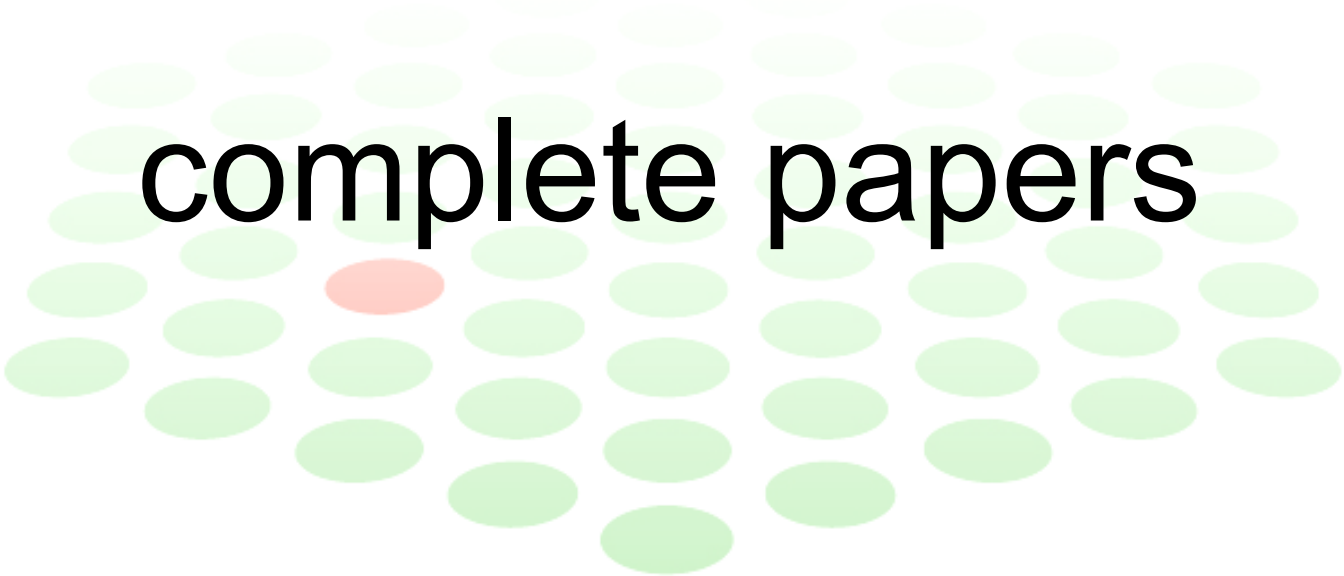

Korbel et al., PLoS Biology, 2005

# where are we now?




Jensen et al., Nature Reviews Genetics, 2006

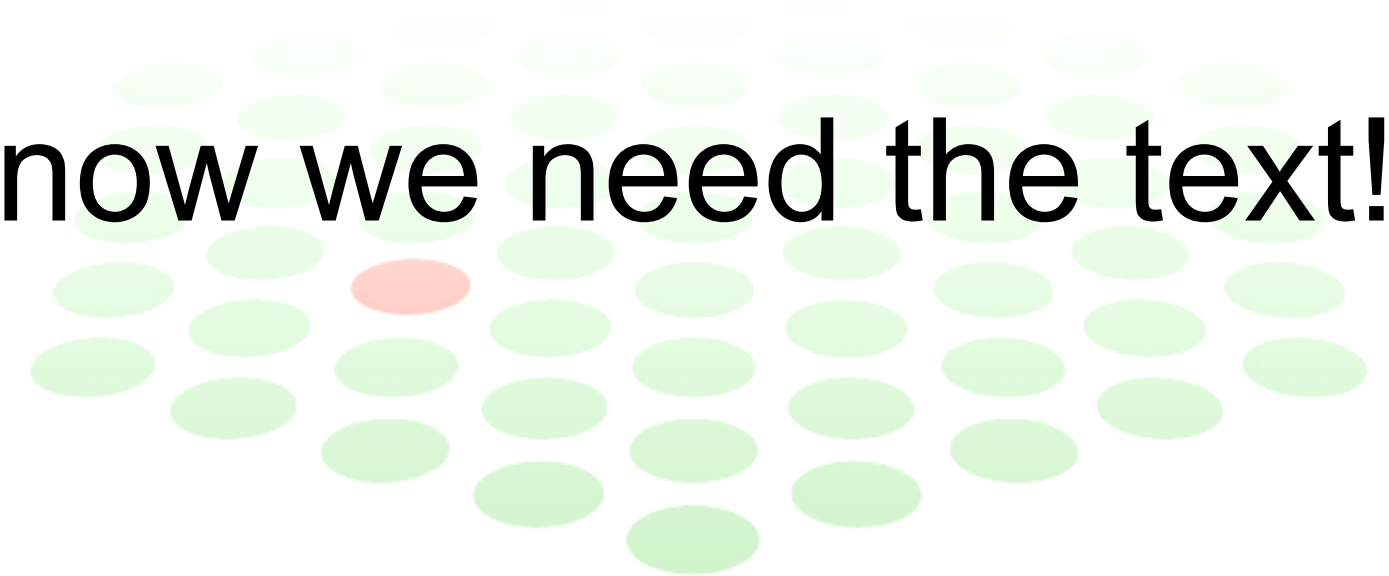

abstracts


complete papers


restricted access


open access


the tools are there


now we need the text!

# Acknowledgments


Work done in collaboration with

Peer Bork

Jasmin Saric

Rossitza Ouzounova

Michael Kuhn

Isabel Rojas

Presentation style stolen from

Lawrence Lessig

Dick Clarence Hardt

# Thank you!