Text Mining in der biomedizinischen Forschung

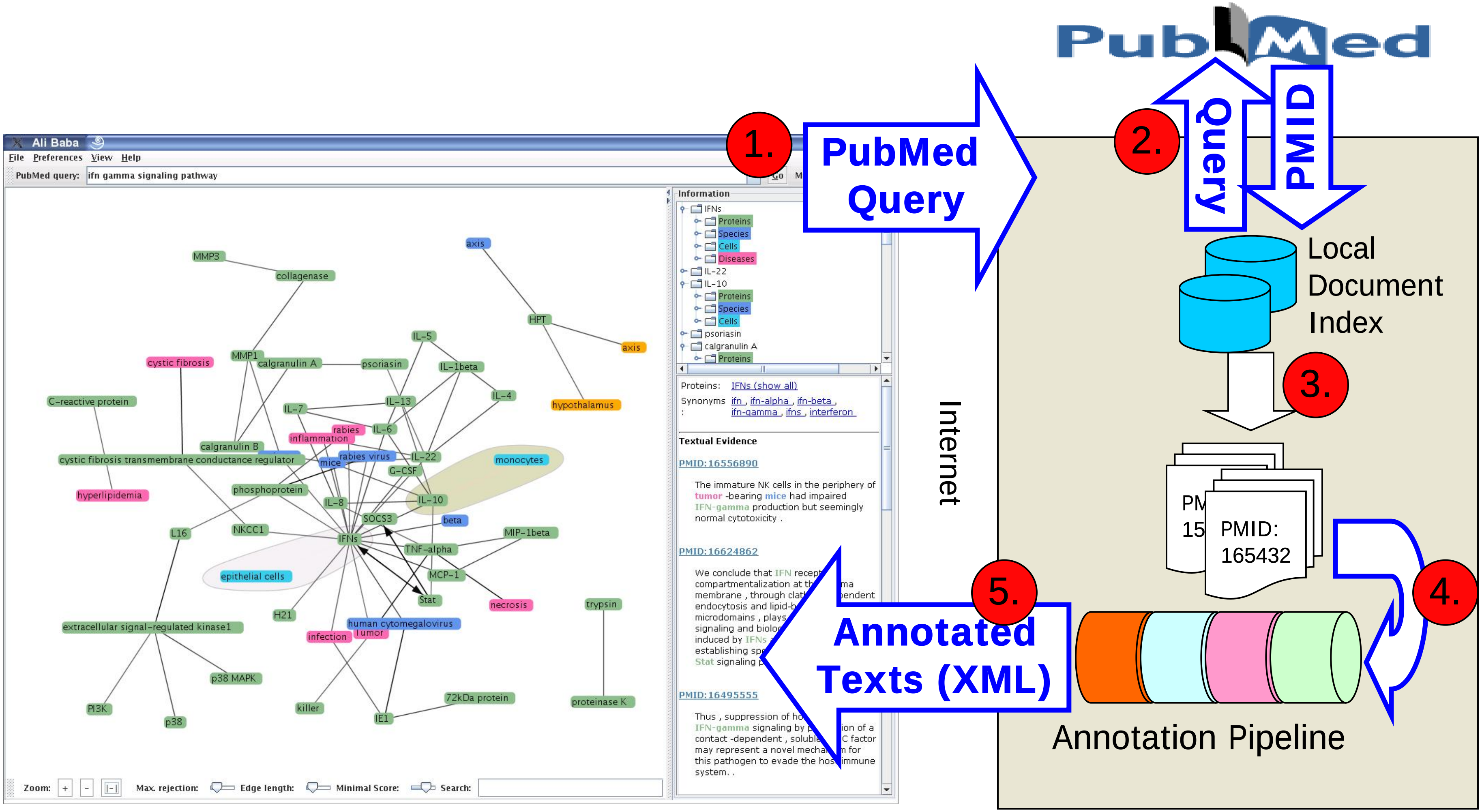
Ulf Leser, Humboldt-Universität zu Berlin

Case Report

- Patient with pneumonia and cough
- Normal dosage of codeine
- Patient not responding any more at day 4
- What‘s going on?
  - PubMed „Codeine intoxication“ -> 70 abstracts
  - Aren’t there better ways?

Case report from Univ. Hospital Geneva, thanks to Christian Meisel, Roche

AliBaba (Plake et al. 2006, Hakenberg et al. 2010)



PubMed query: ifn gamma signaling pathway

Proteins: IFNs (show all)

Synonyms: ifn , ifn-alpha , ifn-beta , ifn-gamma , ifns , interferon

Textual Evidence

PMID:16556890

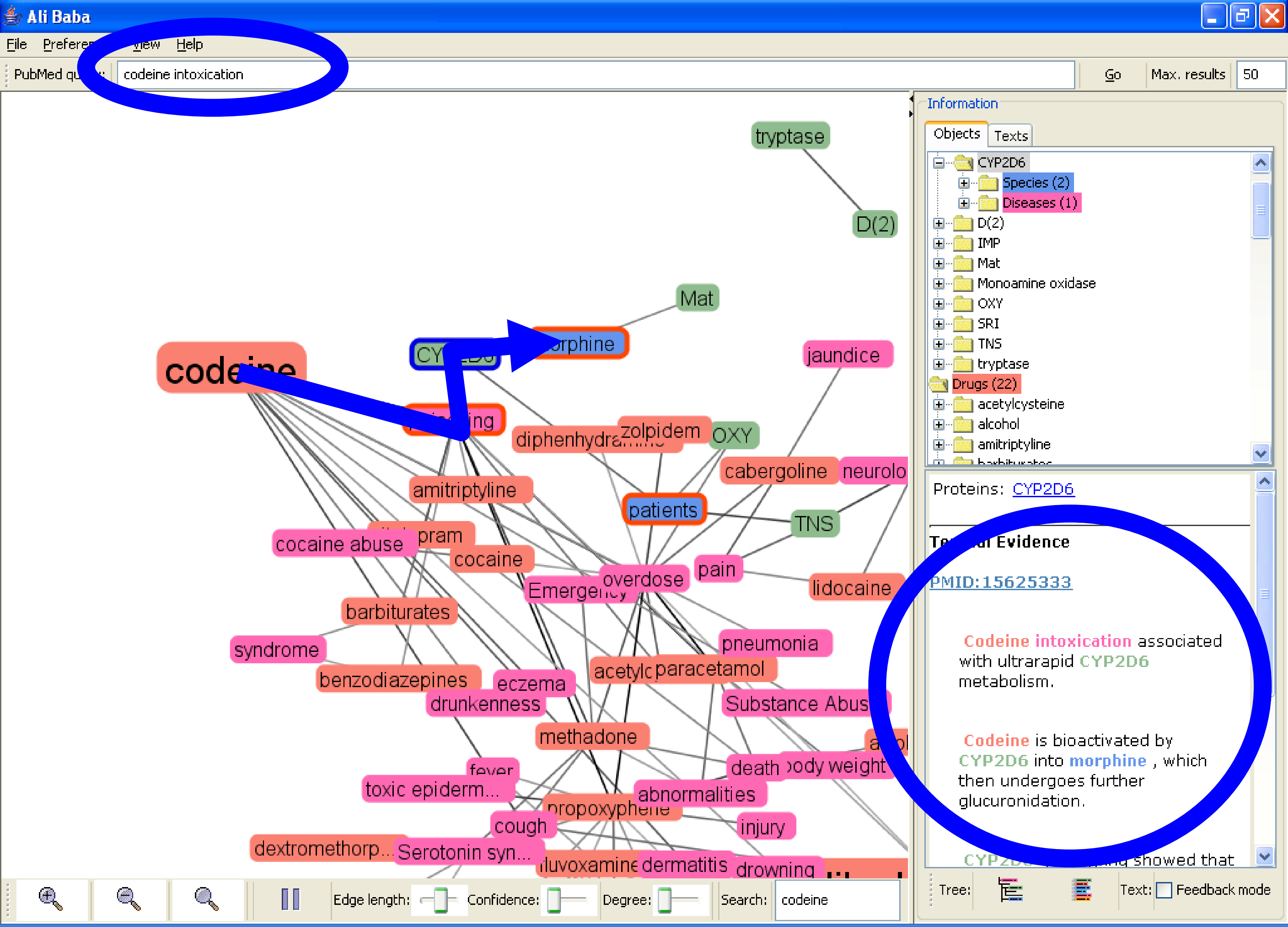
The immature NK cells in the periphery of tumor -bearing mice had impaired IFN-gamma production but seemingly normal cytotoxicity .

PMID:16624862

We conclude that IFN recept compartmentalization at th ma membrane , through clat pendent endocytosis and lipid-b microdomains , plays signaling and biolo induced by IFNs establishing sp Stat signaling p

PMID:16495555

Thus , suppression of ho IFN-gamma signaling by p ion of a contact -dependent , soluble C factor may represent a novel mecha n for this pathogen to evade the hos immune system. .



PubMed query: codeine intoxication

Max. results 50

Information

Objects | Texts

- CYP2D6
  - Species (2)
  - Diseases (1)
- D(2)
- IMP
- Mat
- Monoamine oxidase
- OXY
- SRI
- TNS
- tryptase
- Drugs (22)
- acetylcysteine
- alcohol
- amitriptyline
- barbiturates

Proteins: CYP2D6

Te al Evidence

PMID:15625333

Codeine intoxication associated with ultrarapid CYP2D6 metabolism.

Codeine is bioactivated by CYP2D6 into morphine , which then undergoes further glucuronidation.

CYP2D showed that

Edge length: Confidence: Degree: Search: codeine

Text: □ Feedback mode

What we Need to do

Z-100 is an arabinomannan extracted from Mycobacterium tuberculosis that has various immunomodulatory activities, such as the induction of interleukin 12, interferon gamma (IFN-gamma) and beta-chemokines. The effects of Z-100 on human immunodeficiency virus type 1 (HIV-1) replication in human monocyte-derived macrophages (MDMs) are investigated in this paper. In MDMs, Z-100 markedly suppressed the replication of not only macrophage-tropic (M-tropic) HIV-1 strain (HIV-1JR-CSF), but also HIV-1 pseudotypes that possessed amphotropic Moloney murine leukemia virus or vesicular stomatitis virus G envelopes. Z-100 was found to inhibit HIV-1 expression, even when added 24 h after infection. In addition, it substantially inhibited the expression of the pNL43lucDeltaenv vector (in which the env gene is defective and the nef gene is replaced with the firefly luciferase gene) when this vector was transfected directly into MDMs. These findings suggest that Z-100 inhibits virus replication, mainly at HIV-1 transcription. However, Z-100 also downregulated expression of the cell surface receptors CD4 and CCR5 in MDMs, suggesting some inhibitory effect on HIV-1 entry. Further experiments revealed that Z-100 induced IFN-beta production in these cells, resulting in induction of the 16-kDa CCAAT/enhancer binding protein (C/EBP) beta transcription factor that represses HIV-1 long terminal repeat transcription. These effects were alleviated by SB 203580, a specific inhibitor of p38 mitogen-activated protein kinases (MAPK), indicating that the p38 MAPK signalling pathway was involved in Z-100-induced repression of HIV-1 replication in MDMs. These findings suggest that Z-100 might be a useful immunomodulator for control of HIV-1 infection.

Find Entities

Z-100 is an ***arabinomannan*** extracted from Mycobacterium tuberculosis that has various immunomodulatory activities, such as the induction of **interleukin 12**, **interferon gamma** (**IFN-gamma**) and beta-chemokines. The effects of ***Z-100*** on human immunodeficiency virus type 1 (HIV-1) replication in human **monocyte-derived macrophages** (**MDMs**) are investigated in this paper. In **MDMs**, ***Z-100*** markedly suppressed the replication of not only macrophage-tropic (M-tropic) HIV-1 strain (HIV-1JR-CSF), but also HIV-1 pseudotypes that possessed amphotropic Moloney murine leukemia virus or vesicular stomatitis virus G envelopes. ***Z-100*** was found to inhibit HIV-1 expression, even when added 24 h after infection. In addition, it substantially inhibited the expression of the pNL43lucDeltaenv vector (in which the ***env*** gene is defective and the ***nef*** gene is replaced with the ***firefly luciferase*** gene) when this vector was transfected directly into **MDMs**. These findings suggest that ***Z-100*** inhibits virus replication, mainly at HIV-1 transcription. However, ***Z-100*** also downregulated expression of the cell surface receptors **CD4** and **CCR5** in **MDMs**, suggesting some inhibitory effect on HIV-1 entry. Further experiments revealed that ***Z-100*** induced **IFN-beta** production in these cells, resulting in induction of the 16-kDa **CCAAT/enhancer binding protein** (**C/EBP**) **beta transcription factor** that represses HIV-1 long terminal repeat transcription. These effects were alleviated by SB 203580, a specific inhibitor of **p38 mitogen-activated protein kinases** (**MAPK**), indicating that the **p38 MAPK** signalling pathway was involved in ***Z-100***-induced repression of HIV-1 replication in **MDMs**. These findings suggest that ***Z-100*** might be a useful immunomodulator for control of HIV-1 infection.

Normalize Entities

Tax: 1773

Entrez: 3458

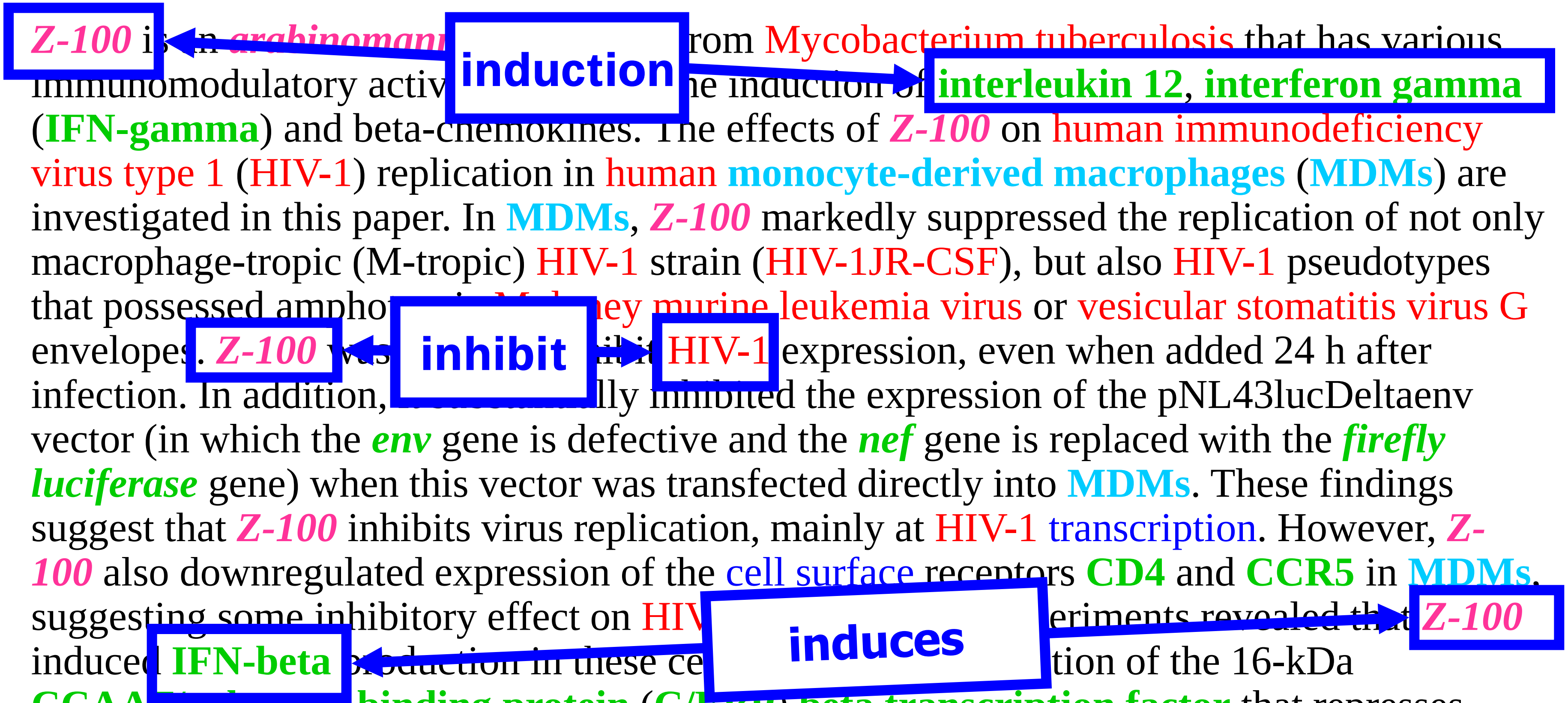
Tax: 9606

UMLS: C0001175

GO:0009986

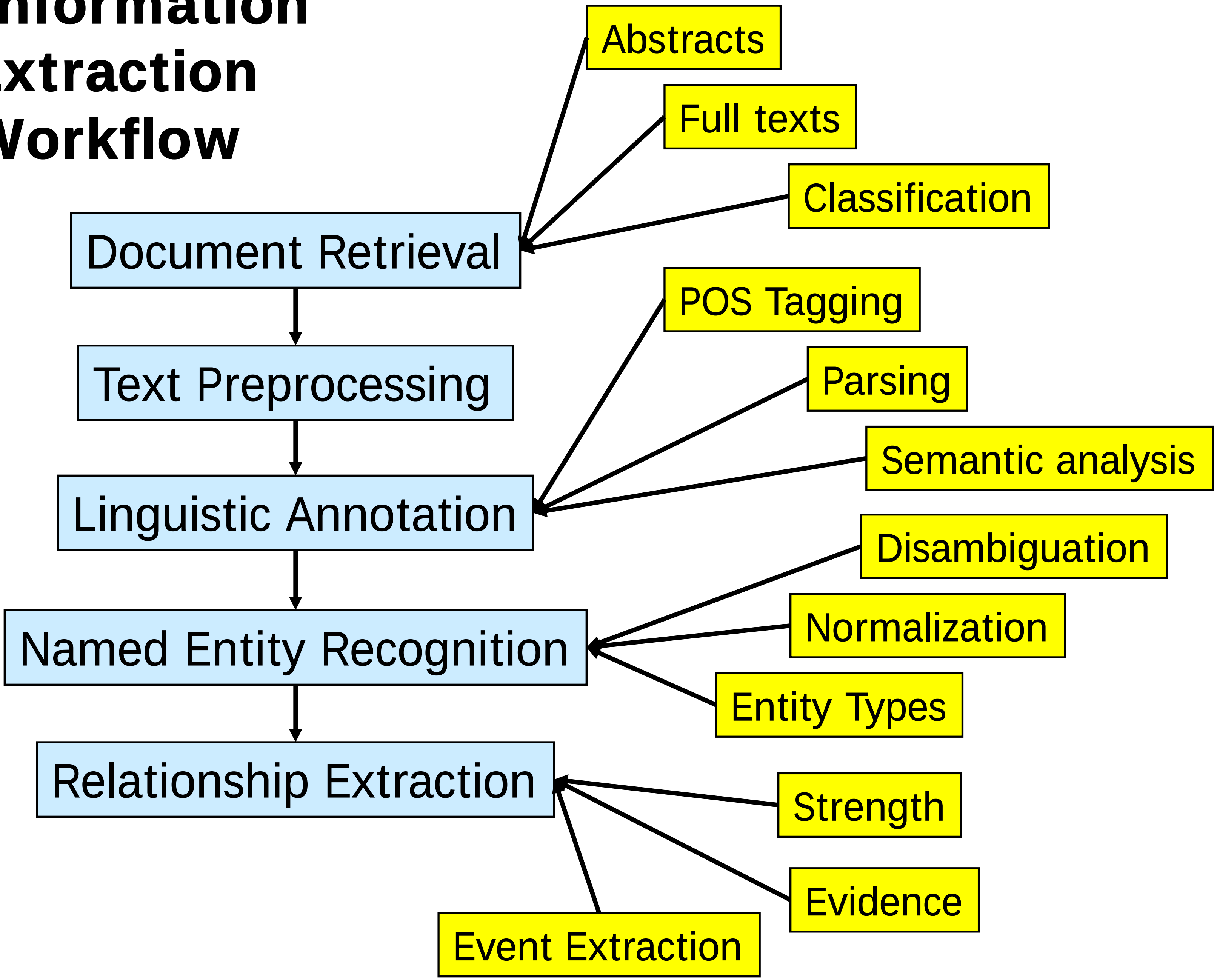
Z-100 is an ***arabinomannan*** extracted from Mycobacterium tuberculosis th has various immunomodulatory activities, ction of **interleukin 12**, **interferon gamma** (**IFN-gamma**) and beta-chemokin The effects of *Z-100* on y virus type 1 (HIV-1) replication in human **monocyte-derived** are investigated in this paper. In **MDMs**, *Z-100* markedly suppressed the plication of not only macrophage-tropic (M-tropic) HIV-1 strain (HIV-1JR-CSF), but also HIV-1 pseudotypes that possessed amphotropic Moloney murine leukemia virus or vesicular stomatitis virus G envelopes. *Z-100* was found to inhibit HIV-1 expression, even when added 24 h after infection. In addition, it substantially inhibited the expression of the pNL43lucDeltaenv vector (in which the ***env*** gene is defectiv aced with the ***firefly luciferase*** gene) when this vector was tr **DMs**. These findings suggest that *Z-100* inhibits virus replication, ma y at HIV-1 transcription. However, *Z-100* also downregulated expression of the cell surface receptors **CD4** and **CCR5** in **MDMs**, suggesting some inhibitory effect on HIV-1 entry. Further experiments revealed that *Z-100* induced **IFN-beta** production in these cells, resulting in induction of the 16-kDa **CCAAT/enhancer binding protein** (**C/EBP**) **beta transcription factor** that represses HIV-1 long terminal repeat transcription. These effects were alleviated by SB 203580, a specific inhibitor of **p38 mitogen-activated protein kinases** (**MAPK**), indicating that the **p38 MAPK** signalling pathway was involved in *Z-100*-induced repression of HIV-1 replication in **MDMs**. These findings suggest that *Z-100* might be a useful immunomodulator for control of HIV-1 infection.

Find Relationships



Z-100 is an *arabinomannan* extracted from Mycobacterium tuberculosis that has various immunomodulatory activities, such as the induction of **interleukin 12**, **interferon gamma** (**IFN-gamma**) and beta-chemokines. The effects of *Z-100* on human immunodeficiency virus type 1 (HIV-1) replication in human **monocyte-derived macrophages** (**MDMs**) are investigated in this paper. In **MDMs**, *Z-100* markedly suppressed the replication of not only macrophage-tropic (M-tropic) HIV-1 strain (HIV-1JR-CSF), but also HIV-1 pseudotypes that possessed amphotropic Moloney murine leukemia virus or vesicular stomatitis virus G envelopes. *Z-100* was found to inhibit HIV-1 expression, even when added 24 h after infection. In addition, it substantially inhibited the expression of the pNL43lucDeltaenv vector (in which the ***env*** gene is defective and the ***nef*** gene is replaced with the ***firefly luciferase*** gene) when this vector was transfected directly into **MDMs**. These findings suggest that *Z-100* inhibits virus replication, mainly at HIV-1 transcription. However, *Z-100* also downregulated expression of the cell surface receptors **CD4** and **CCR5** in **MDMs**, suggesting some inhibitory effect on HIV-1 entry. Further experiments revealed that *Z-100* induced **IFN-beta** production in these cells, resulting in induction of the 16-kDa **CCAAT/enhancer binding protein** (**C/EBP**) **beta transcription factor** that represses HIV-1 long terminal repeat transcription. These effects were alleviated by SB 203580, a specific inhibitor of **p38 mitogen-activated protein kinases** (**MAPK**), indicating that the **p38 MAPK** signalling pathway was involved in *Z-100*-induced repression of HIV-1 replication in **MDMs**. These findings suggest that *Z-100* might be a useful immunomodulator for control of HIV-1 infection.

Information Extraction Workflow



Document Retrieval

Abstracts

Full texts

Classification

Text Preprocessing

Linguistic Annotation

POS Tagging

Parsing

Semantic analysis

Named Entity Recognition

Disambiguation

Normalization

Entity Types

Relationship Extraction

Strength

Evidence

Event Extraction

Why Text-Mine?

- Curation
  Support construction of high quality knowledge bases
- Search
  Let users find specific information faster
- Biomedical Research
  Provide background information for specific types of biomedical analysis (network / systems biology)

Some WBI Projects

- EUMed: Genotype-phenotype relations
  - Mutations, genes, diseases, drugs, mutation-disease, DDI, ...
- CellFinder: Genes characteristic for a given cell
  - Genes, cells, cell lines, transcription, location, function
- OncoPath: Regulatory relationships between TF
  - Genes, species, methods, regulation
- Virtual Liver: Metabolic reactions
  - Chemicals, quantities, units, reactions, methods
- ColoNet: Mammalian clock genes and the clock network

Topics Today

- Named Entity Recognition
- Applications in Curation
- Application in Search

Detecting Gene Names

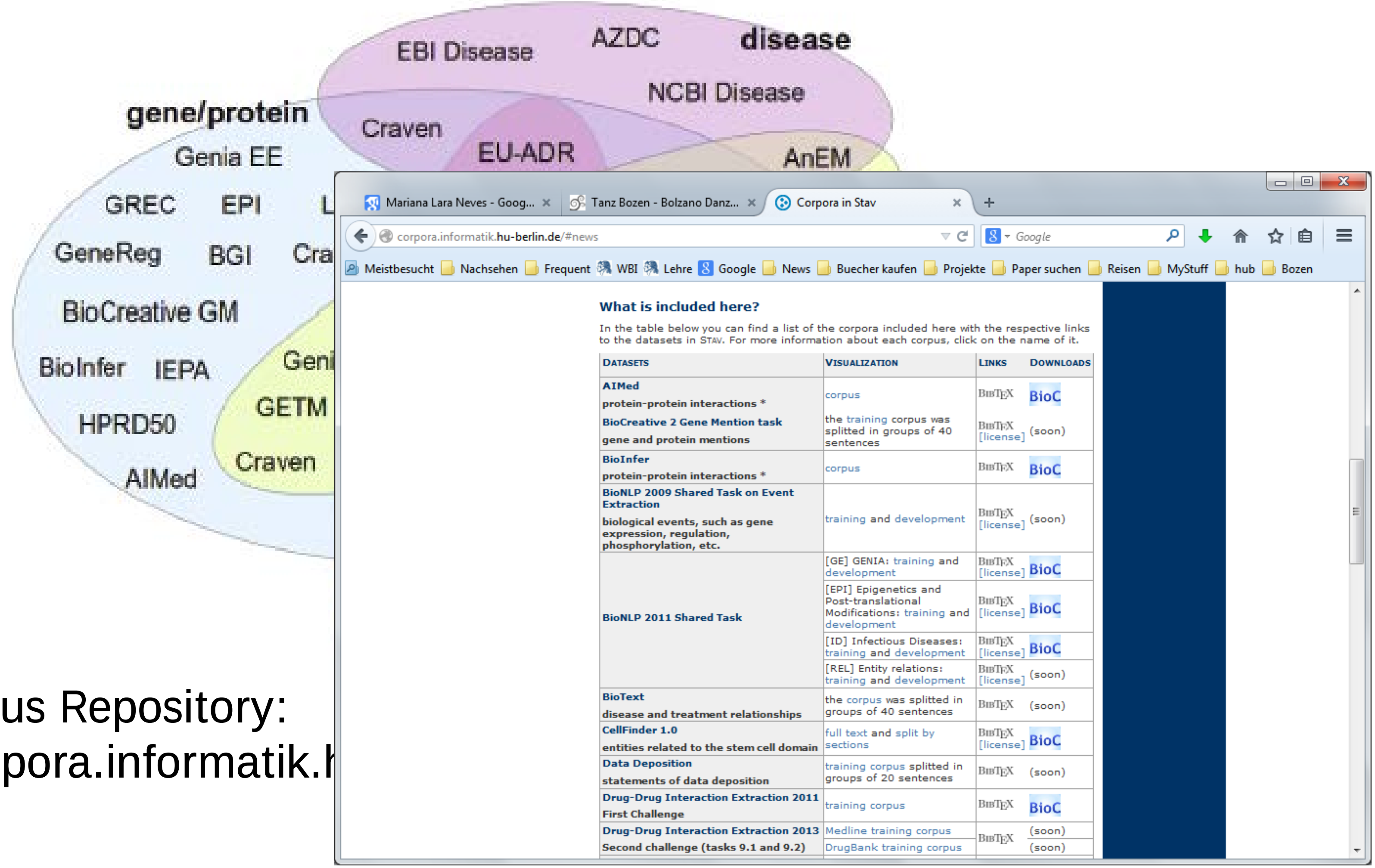
The human T cell leukemia lymphotropic virus type 1 Tax protein represses MyoD-dependent transcription by inhibiting MyoD-binding to the KIX domain of p300.

The human T cell leukemia lymphotropic virus type 1 Tax protein represses MyoD-dependent transcription by inhibiting MyoD-binding to the KIX domain of p300.

State-of-the-Art Solutions

- Example: GNAT
  - Large dictionaries, 2nd order CRF
  - Species disambiguation using Linnaeus (NER again)
  - Background texts describing genes and their function
  - Single sense per discourse: Further appearances in the same text
  - No resolution of type and splice variants
- Evaluated performance: ~85% F1-measure
  - Only measured on abstracts, full texts are different
  - Correctness is in the eye of the beholder
  - Different results on different corpora
  - Different quality per species

What Bio-TM People Study (Neves 2014)



What is included here?

In the table below you can find a list of the corpora included here with the respective links to the datasets in STAV. For more information about each corpus, click on the name of it.

DATASETS	VISUALIZATION	LINKS	DOWNLOADS
AIMed **protein-protein interactions ***	corpus	BibTeX	BioC
BioCreative 2 Gene Mention task **gene and protein mentions**	the training corpus was splitted in groups of 40 sentences	BibTeX [license]	(soon)
BioInfer **protein-protein interactions ***	corpus	BibTeX	BioC
BioNLP 2009 Shared Task on Event Extraction **biological events, such as gene expression, regulation, phosphorylation, etc.**	training and development	BibTeX [license]	(soon)
BioNLP 2011 Shared Task	[GE] GENIA: training and development	BibTeX [license]	BioC
	[EPI] Epigenetics and Post-translational Modifications: training and development	BibTeX [license]	BioC
	[ID] Infectious Diseases: training and development	BibTeX [license]	BioC
	[REL] Entity relations: training and development	BibTeX [license]	(soon)
BioText **disease and treatment relationships**	the corpus was splitted in groups of 40 sentences	BibTeX	(soon)
CellFinder 1.0 **entities related to the stem cell domain**	full text and split by sections	BibTeX [license]	BioC
Data Deposition **statements of data deposition**	training corpus splitted in groups of 20 sentences	BibTeX	(soon)
Drug-Drug Interaction Extraction 2011 **First Challenge**	training corpus	BibTeX	BioC
Drug-Drug Interaction Extraction 2013 **Second challenge (tasks 9.1 and 9.2)**	Medline training corpus	BibTeX	(soon)
	DrugBank training corpus		(soon)

WBI Corpus Repository:
http://corpora.informatik.h

Experiences: NER+NEN

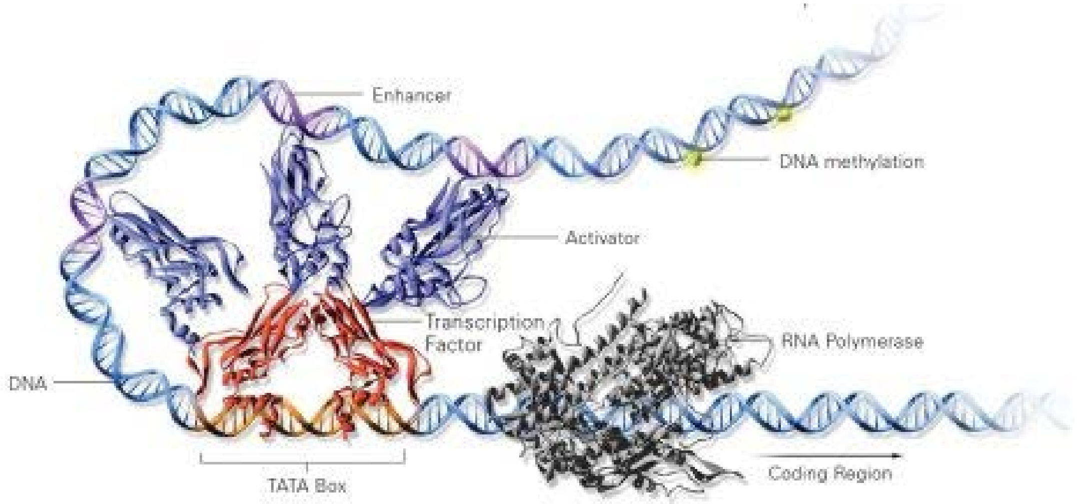
- No true applications without normalization
  - ChemSpot2 includes OPSIIN
- Every entity type requires a mid-size project
  - Training data, annotator recruitment, domain-specific features, etc.
- In general: The more concrete, the easier
  - Species names versus gene function
- Evaluation usually is a nightmare
  - “The human FGD-3’ protein is soluble in water at ...”
  - w/o “protein”, w/o splice variants, w/o taxa, w/o water, ...

Topics Today

- Named Entity Recognition
- Applications in Curation
- Application in Search

Gene-Regulatory Relationships

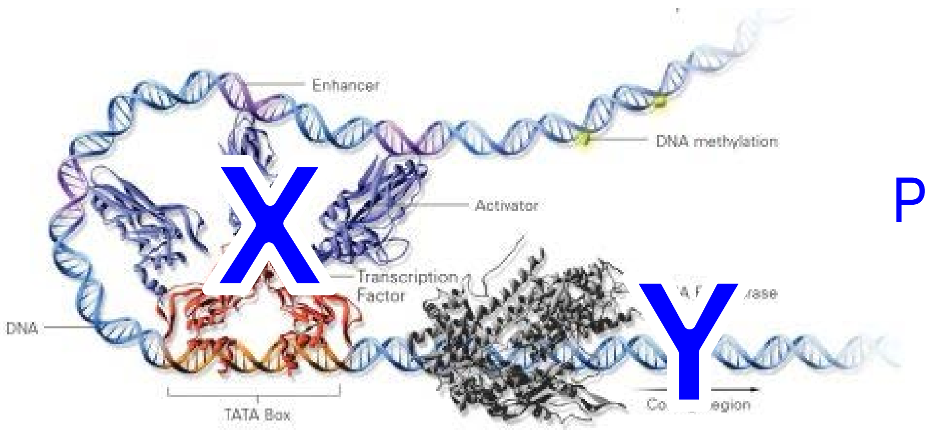
- Many published “PPI” actually are derived from co-expression, in-silico predictions, phylogenetic inference, ...
- Essential information: Strength of experimental evidence?
- Example: Gene regulation and transcription factors



Source: http://apt.bea.ki.se

Showing that a TF X directly regulates gene Y

TF X must bind to the
promoter region of gene Y

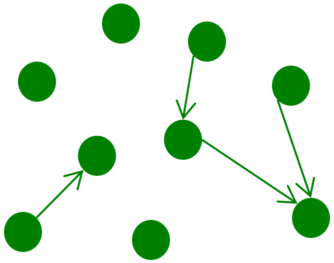


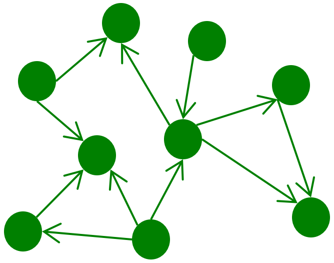
Presence of TF X
must change
expression of
gene Y

Destroying the promoter
where X binds must change
expression of gene Y

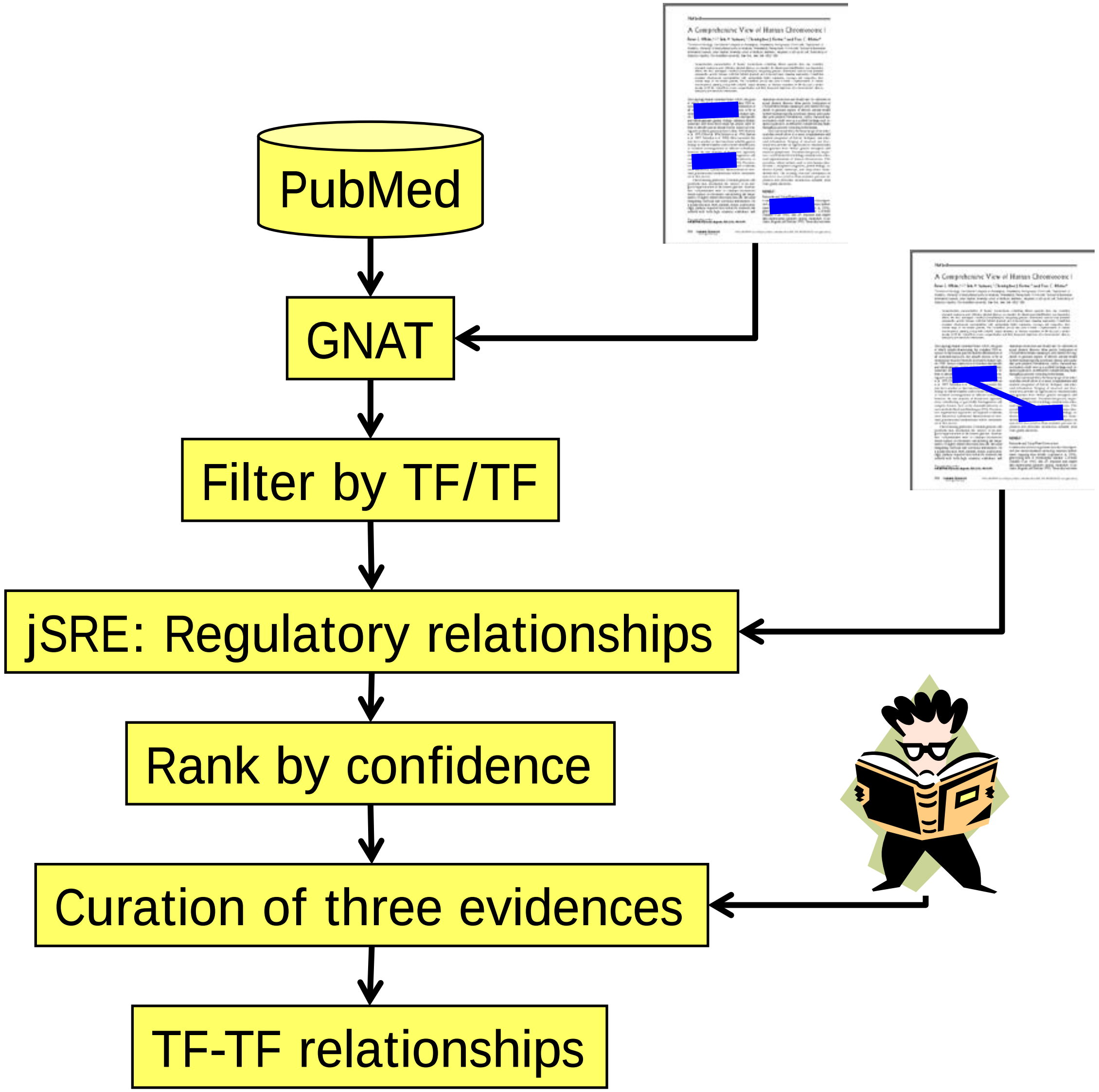
Towards a Human TF Network (Thomas et al. 2015)

- Human: ~1200 TFs
- Only ~500 TF-TF relationships are publically available
  - Transfac, ORegAnno, TRRD
- TM-REG: Enhance human TF core network
  - Identification of potential TF-TF relationships using text mining
  - Manual verification by biological experts
  - Distinguish hypothetical from proven relationships





Workflow



PubMed

GNAT

Filter by TF/TF

jSRE: Regulatory relationships

Rank by confidence

Curation of three evidences

TF-TF relationships

Top-Scores

- Ranking is crucial to use human time effectively

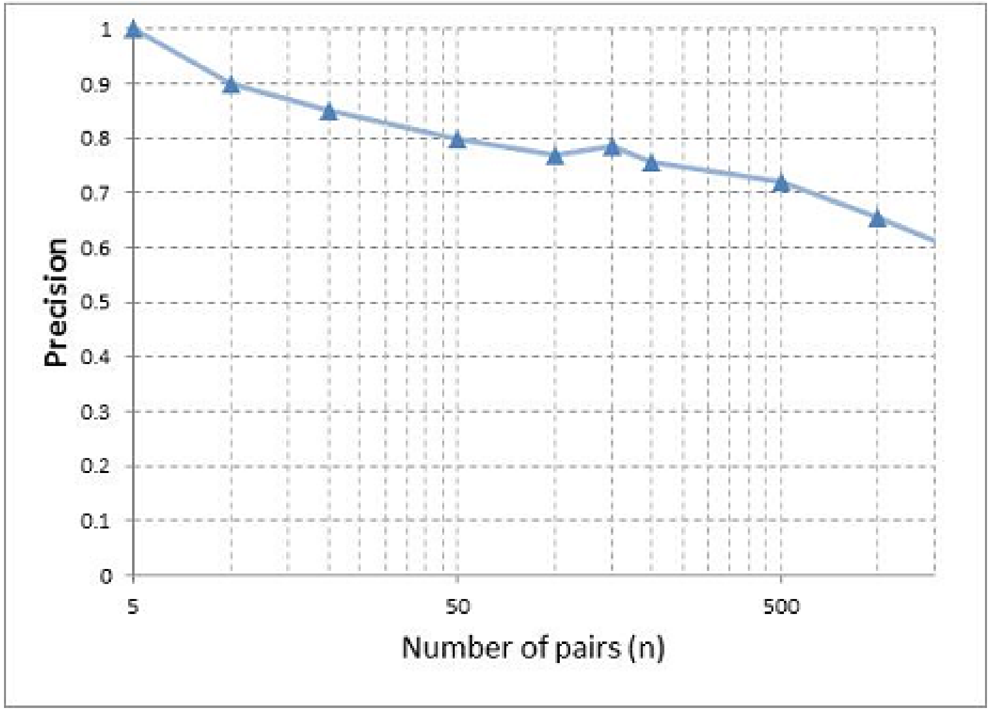
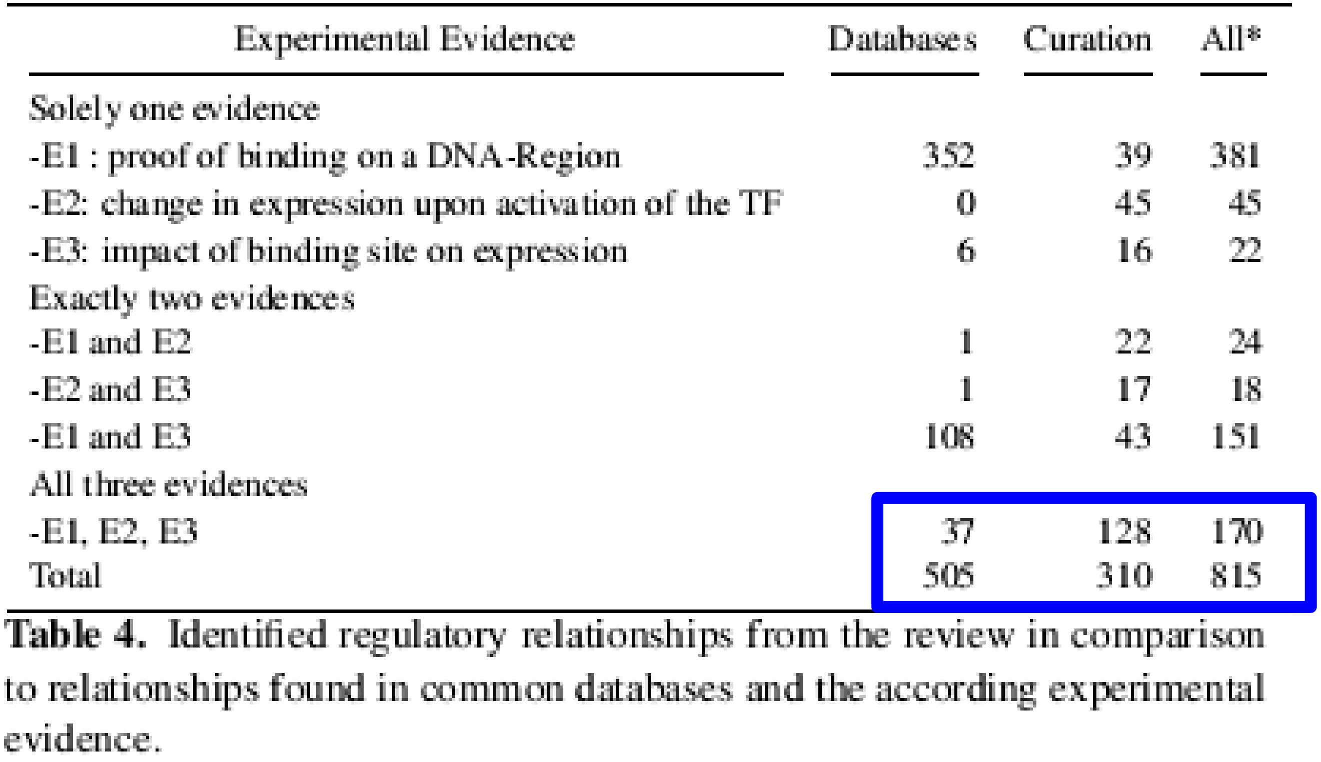


Fig. 4. Precision of our workflow for the n most confidently classified and manually curated sentences. Pairs already contained in a regulatory database are ignored (see Table 3).

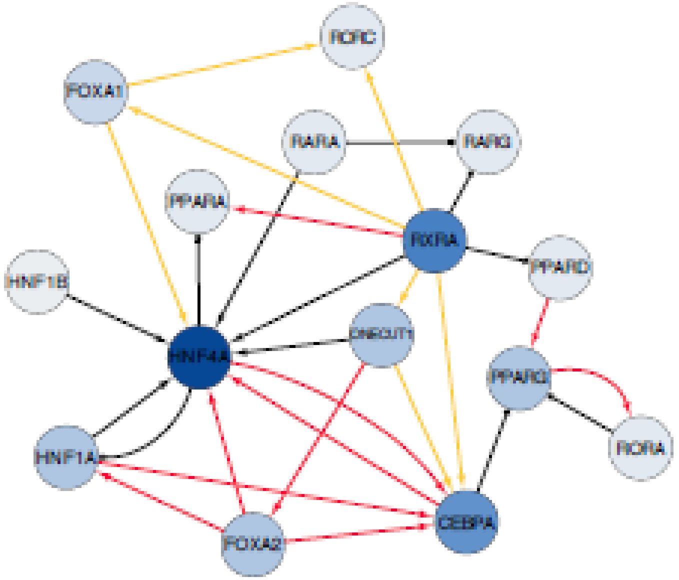
Results
(Top 2000 sentences, known relationships filtered)



Experimental Evidence	Databases	Curation	All*
Solely one evidence			
-E1 : proof of binding on a DNA-Region	352	39	381
-E2: change in expression upon activation of the TF	0	45	45
-E3: impact of binding site on expression	6	16	22
Exactly two evidences			
-E1 and E2	1	22	24
-E2 and E3	1	17	18
-E1 and E3	108	43	151
All three evidences			
-E1, E2, E3	37	128	170
Total	505	310	815

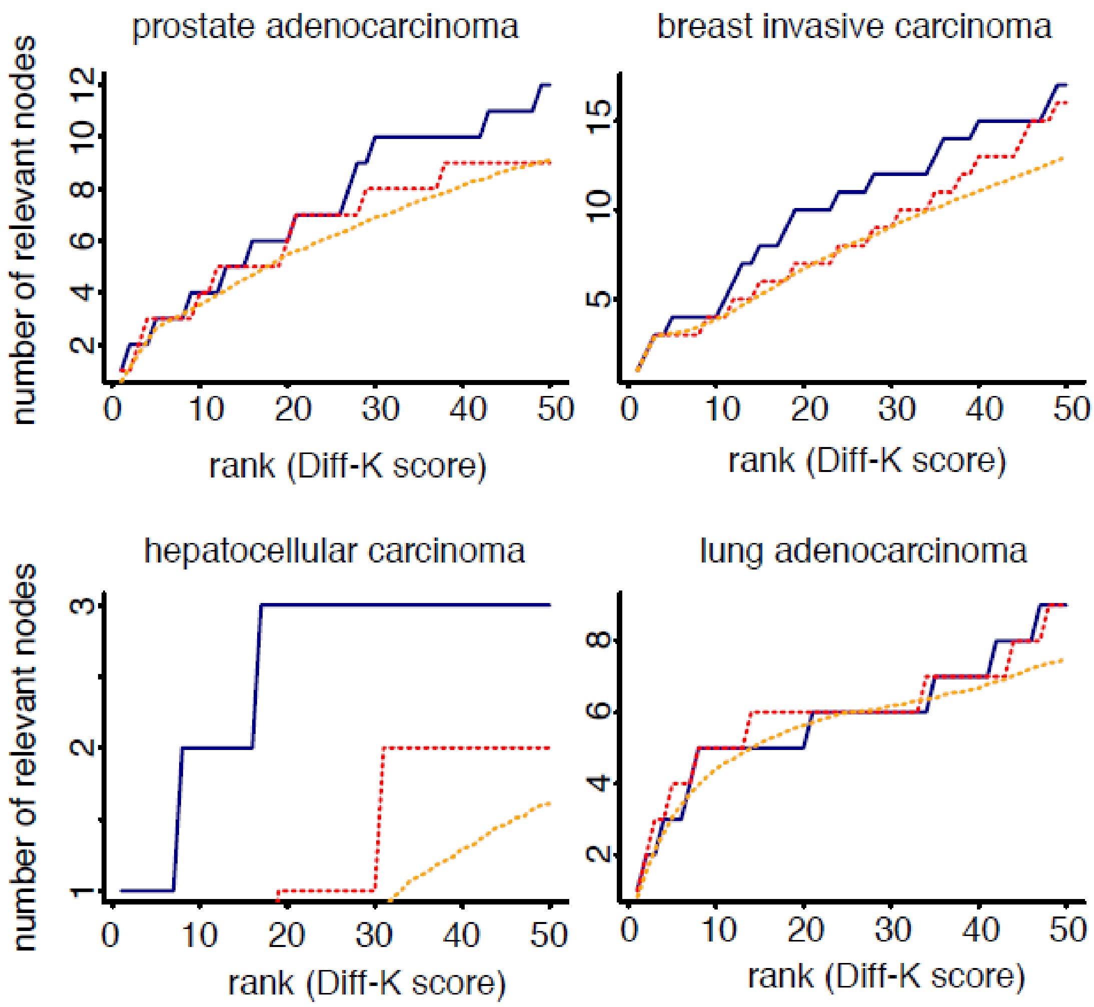
Table 4. Identified regulatory relationships from the review in comparison to relationships found in common databases and the according experimental evidence.

Specific Pathway: Regulation in Liver



- Black: regulations contained in existing RegDBs;
- Red: added by curation of the top 2,500 unspecific sentences;
- Orange: regulations found by manual curation of all 1,435 sentences with co-occurring liver specific TFs.

Biological Usefulness



- With TM_REG
- Current network
- Randomized

Experiences: Curation

- Key: Combine text mining and human wisdom
  - No way that current text mining techniques can check evidences at such detail automatically
- Confidence scores of classifiers do make sense
- High throughput is possible
  - Clear problem setting
  - High performance text mining
  - Appropriate curation tools
- Important difference
  - Fast curation: Get certain data from all of PubMed
  - Full curation: Get everything out of this paper

Manageable

Hard & costly

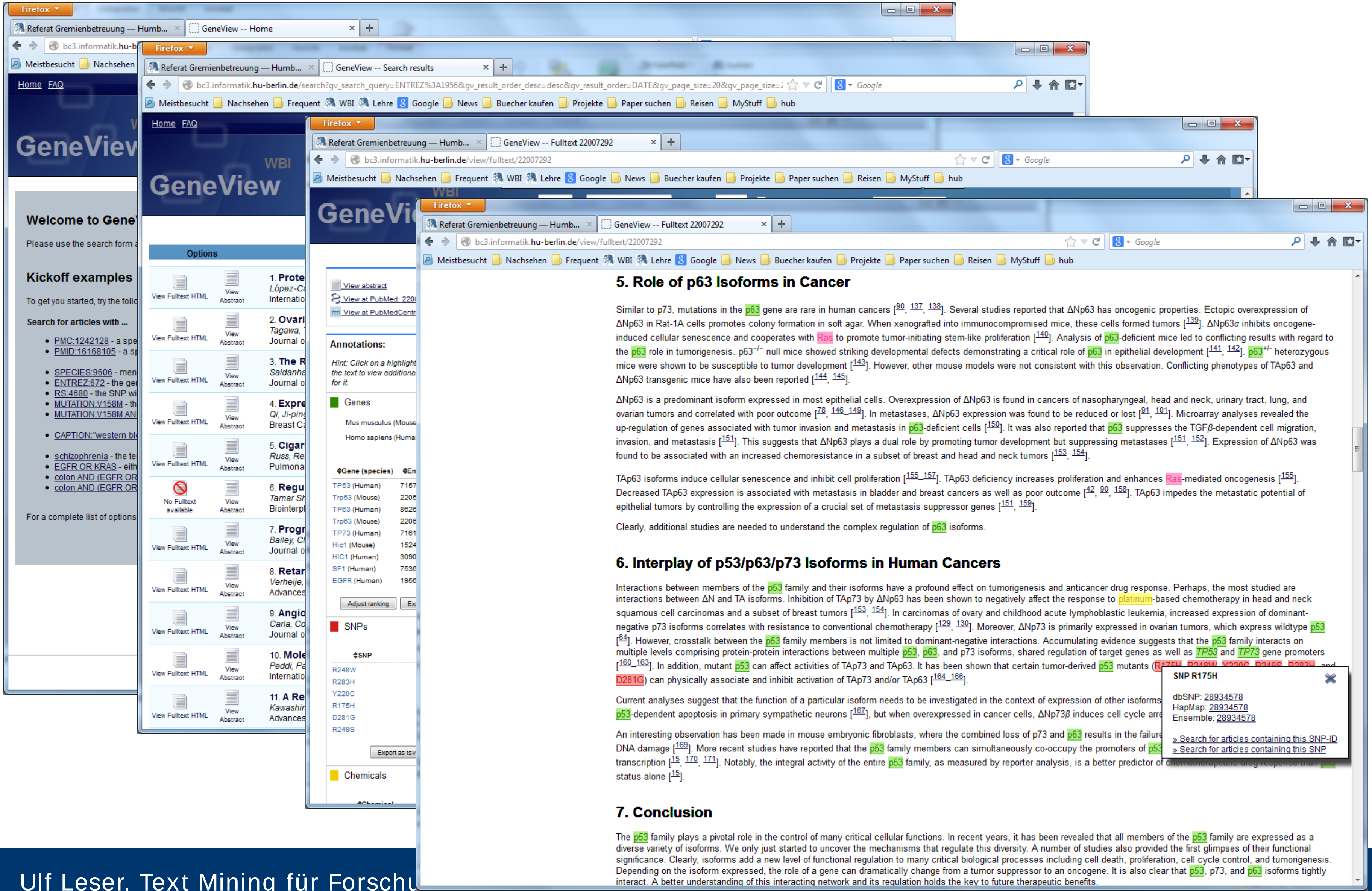
Topics Today

- Named Entity Recognition
- Application in Curation
- Application in Search

GeneView

- PubMed does not allow searching synonyms of entities
  - “Find mentionings of BRCA1 (170 synonyms)”
- PubMed does not resolve homonyms
  - “Find mentionings of human BRCA1”
- PubMed cannot rank by “semantic” content
  - “Find mutations in BRCA1”
- PubMed does not allow searching for relationships
  - “Find proteins interacting with BRCA1”
- Geneview does
  - Free web interface
  - Free database: Annotated PubMed

GeneView



Welcome to Gene

Please use the search form a

Kickoff examples

To get you started, try the follo

Search for articles with ...

- PMC:1242128 - a spe
- PMID:16168105 - a sp
- SPECIES:9606 - men
- ENTREZ:672 - the ge
- RS:4680 - the SNP wi
- MUTATION:V158M - th
- MUTATION:V158M AN
- CAPTION:"western bl
- schizophrenia - the te
- EGFR OR KRAS - eith
- colon AND (EGFR OR
- colon AND (EGFR OR

For a complete list of options

Annotations:

Hint: Click on a highligh the text to view additiona for it.

🟩 Genes

🟥 SNPs

🟨 Chemicals

5. Role of p63 Isoforms in Cancer

Similar to p73, mutations in the p63 gene are rare in human cancers [90, 137, 138]. Several studies reported that ΔNp63 has oncogenic properties. Ectopic overexpression of ΔNp63 in Rat-1A cells promotes colony formation in soft agar. When xenografted into immunocompromised mice, these cells formed tumors [139]. ΔNp63*α* inhibits oncogene-induced cellular senescence and cooperates with Ras to promote tumor-initiating stem-like proliferation [140]. Analysis of p63-deficient mice led to conflicting results with regard to the p63 role in tumorigenesis. $p63^{-/-}$ null mice showed striking developmental defects demonstrating a critical role of p63 in epithelial development [141, 142]. $p63^{+/-}$ heterozygous mice were shown to be susceptible to tumor development [143]. However, other mouse models were not consistent with this observation. Conflicting phenotypes of TAp63 and ΔNp63 transgenic mice have also been reported [144, 145].

ΔNp63 is a predominant isoform expressed in most epithelial cells. Overexpression of ΔNp63 is found in cancers of nasopharyngeal, head and neck, urinary tract, lung, and ovarian tumors and correlated with poor outcome [78, 146–149]. In metastases, ΔNp63 expression was found to be reduced or lost [91, 101]. Microarray analyses revealed the up-regulation of genes associated with tumor invasion and metastasis in p63-deficient cells [150]. It was also reported that p63 suppresses the TGF*β*-dependent cell migration, invasion, and metastasis [151]. This suggests that ΔNp63 plays a dual role by promoting tumor development but suppressing metastases [151, 152]. Expression of ΔNp63 was found to be associated with an increased chemoresistance in a subset of breast and head and neck tumors [153, 154].

TAp63 isoforms induce cellular senescence and inhibit cell proliferation [155–157]. TAp63 deficiency increases proliferation and enhances Ras-mediated oncogenesis [155]. Decreased TAp63 expression is associated with metastasis in bladder and breast cancers as well as poor outcome [42, 90, 158]. TAp63 impedes the metastatic potential of epithelial tumors by controlling the expression of a crucial set of metastasis suppressor genes [151, 159].

Clearly, additional studies are needed to understand the complex regulation of p63 isoforms.

6. Interplay of p53/p63/p73 Isoforms in Human Cancers

Interactions between members of the p53 family and their isoforms have a profound effect on tumorigenesis and anticancer drug response. Perhaps, the most studied are interactions between ΔN and TA isoforms. Inhibition of TAp73 by ΔNp63 has been shown to negatively affect the response to platinum-based chemotherapy in head and neck squamous cell carcinomas and a subset of breast tumors [153, 154]. In carcinomas of ovary and childhood acute lymphoblastic leukemia, increased expression of dominant-negative p73 isoforms correlates with resistance to conventional chemotherapy [129, 130]. Moreover, ΔNp73 is primarily expressed in ovarian tumors, which express wildtype p53 [64]. However, crosstalk between the p53 family members is not limited to dominant-negative interactions. Accumulating evidence suggests that the p53 family interacts on multiple levels comprising protein-protein interactions between multiple p53, p63, and p73 isoforms, shared regulation of target genes as well as *TP53* and *TP73* gene promoters [160–163]. In addition, mutant p53 can affect activities of TAp73 and TAp63. It has been shown that certain tumor-derived p53 mutants (R175H, R248W, Y220C, R249S, R283H, and D281G) can physically associate and inhibit activation of TAp73 and/or TAp63 [164–166].

Current analyses suggest that the function of a particular isoform needs to be investigated in the context of expression of other isoforms p53-dependent apoptosis in primary sympathetic neurons [167], but when overexpressed in cancer cells, ΔNp73*β* induces cell cycle arre

An interesting observation has been made in mouse embryonic fibroblasts, where the combined loss of p73 and p63 results in the failure DNA damage [169]. More recent studies have reported that the p53 family members can simultaneously co-occupy the promoters of p53 transcription [15, 170, 171]. Notably, the integral activity of the entire p53 family, as measured by reporter analysis, is a better predictor of chemotherapeutic drug response than p53 status alone [15].

SNP R175H

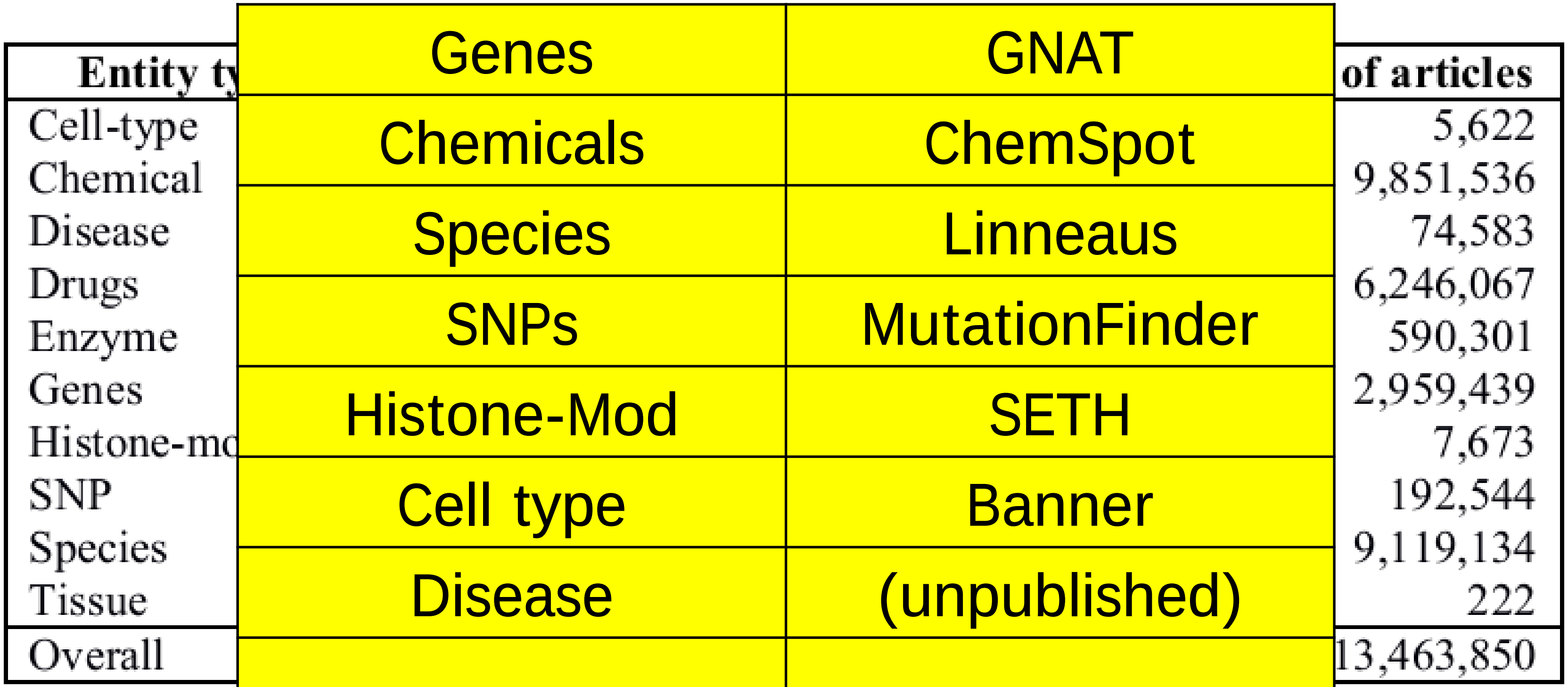
dbSNP: 28934578
HapMap: 28934578
Ensemble: 28934578

» Search for articles containing this SNP-ID
» Search for articles containing this SNP

7. Conclusion

The p53 family plays a pivotal role in the control of many critical cellular functions. In recent years, it has been revealed that all members of the p53 family are expressed as a diverse variety of isoforms. We only just started to uncover the mechanisms that regulate this diversity. A number of studies also provided the first glimpses of their functional significance. Clearly, isoforms add a new level of functional regulation to many critical biological processes including cell death, proliferation, cell cycle control, and tumorigenesis. Depending on the isoform expressed, the role of a gene can dramatically change from a tumor suppressor to an oncogene. It is also clear that p53, p73, and p63 isoforms tightly interact. A better understanding of this interacting network and its regulation holds the key to future therapeutic benefits.

Biomedical IE at Scale



Entity ty	of articles
Cell-type	5,622
Chemical	9,851,536
Disease	74,583
Drugs	6,246,067
Enzyme	590,301
Genes	2,959,439
Histone-mo	7,673
SNP	192,544
Species	9,119,134
Tissue	222
Overall	13,463,850

Genes	GNAT
Chemicals	ChemSpot
Species	Linneaus
SNPs	MutationFinder
Histone-Mod	SETH
Cell type	Banner
Disease	(unpublished)
…	…

Experiences: Search

- User interface more important than 2% more F-measure
- Errors are inevitable; sometimes accepted, sometimes not
- Keeping it up-to-date (tools & data) is a challenge
- Users use context you cannot capture (author, journal, ...)
- Much work, few papers
- Scalability is a real issue

Acknowledgements

- Philippe Thomas
- Mariana Lara Neves
- Tim Rocktäschel
- Torsten Huber
- Michael Weidlich

- Astrid Rheinländer
- Anja Kunkel
- Martin Beckmann
- Marc Bux
- Jörgen Brandt

- Christine Sers
- Nils Blüthgen
- Bertram KLinger
- Andreas Kurz
- Angela Relogio

HUMBOLDT-UNIVERSITÄT ZU BERLIN

Bayer HealthCare
Bayer Schering Pharma

DAAD
Deutscher Akademischer Austausch Dienst
German Academic Exchange Service

Bundesministerium
für Bildung
und Forschung

Bundesministerium
für Wirtschaft
und Technologie

Deutsche
Forschungsgemeinschaft
DFG

Alexander von Humboldt
Stiftung/Foundation

SEVENTH FRAMEWORK
PROGRAMME