Meta expression analysis of regulatory T cell experiments for gene regulatory network reconstruction

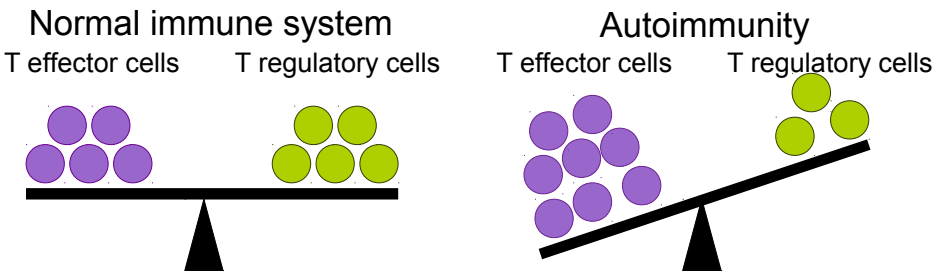
Stefan Kröger*, Melanie Venzke+, Ria Baumgrass+ and Ulf Leser*

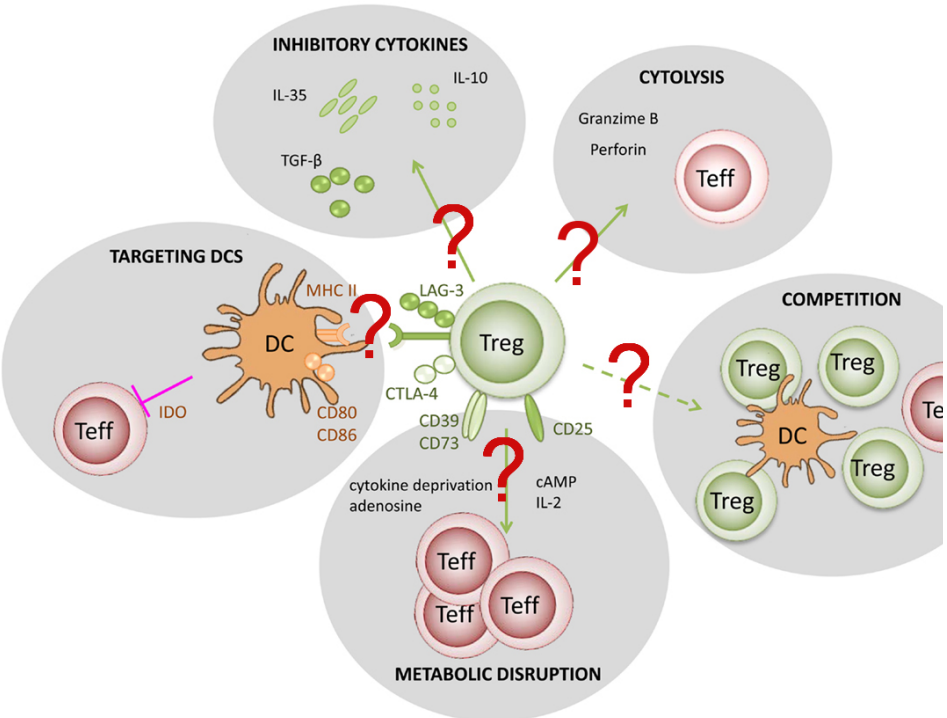
*Humboldt Universität zu Berlin, Institute for Computer Science, Berlin
+Deutsches Rheuma-Forschungszentrum, a Leibniz Institute, Berlin

kroeger@informatik.hu-berlin.de

BioNetVisA workshop on September 7th, 2014

T regulatory cell

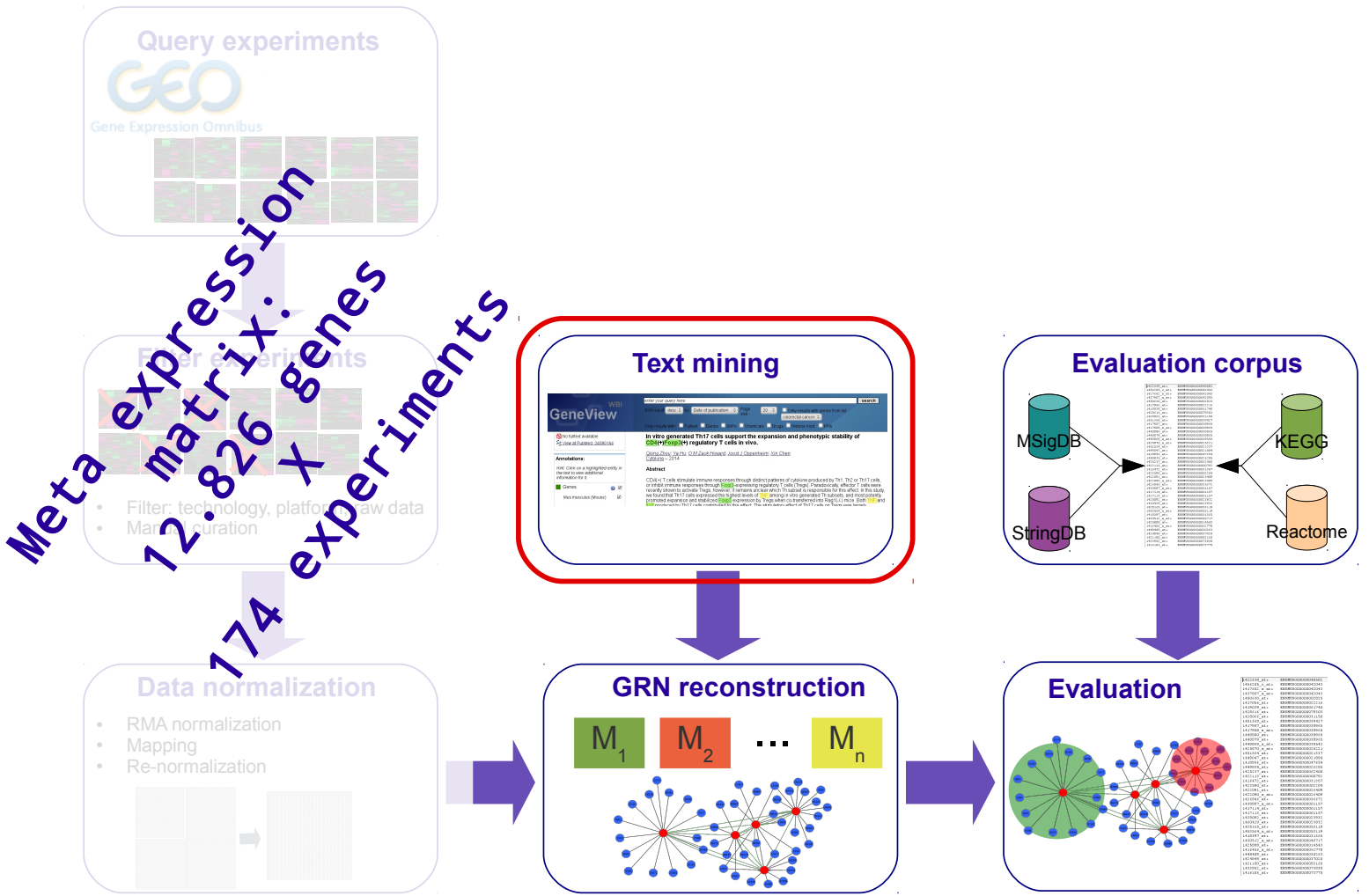




Source: Caridade2013, doi: 10.3389/fimmu.2013.00378

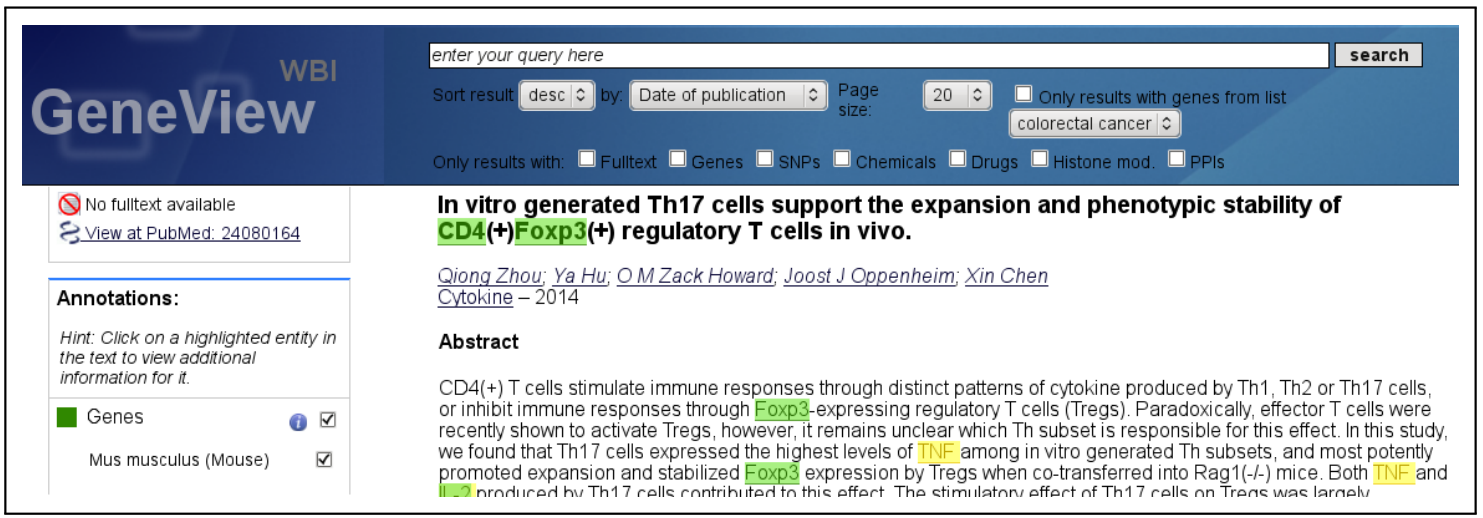
Thus, given that developmental or functional anomalies of nTreg cells are causative of autoimmune and other immunological diseases, it remains a key issue to decipher at the molecular level [...] . Immunol Rev. 2014 May;259(1):192-205. doi: 10.1111/imr.12174

Workflow


Query experiments
Gene Expression Omnibus
Meta expression matrix: 12.826 genes X 174 experiments
Filter experiments
Filter: technology, platform, raw data
Manual curation
Data normalization
RMA normalization
Mapping
Re-normalization
Text mining
GeneView
Evaluation corpus
MSigDB
KEGG
StringDB
Reactome
GRN reconstruction
M1
M2
...
Mn
Evaluation

Gene set extraction using text mining

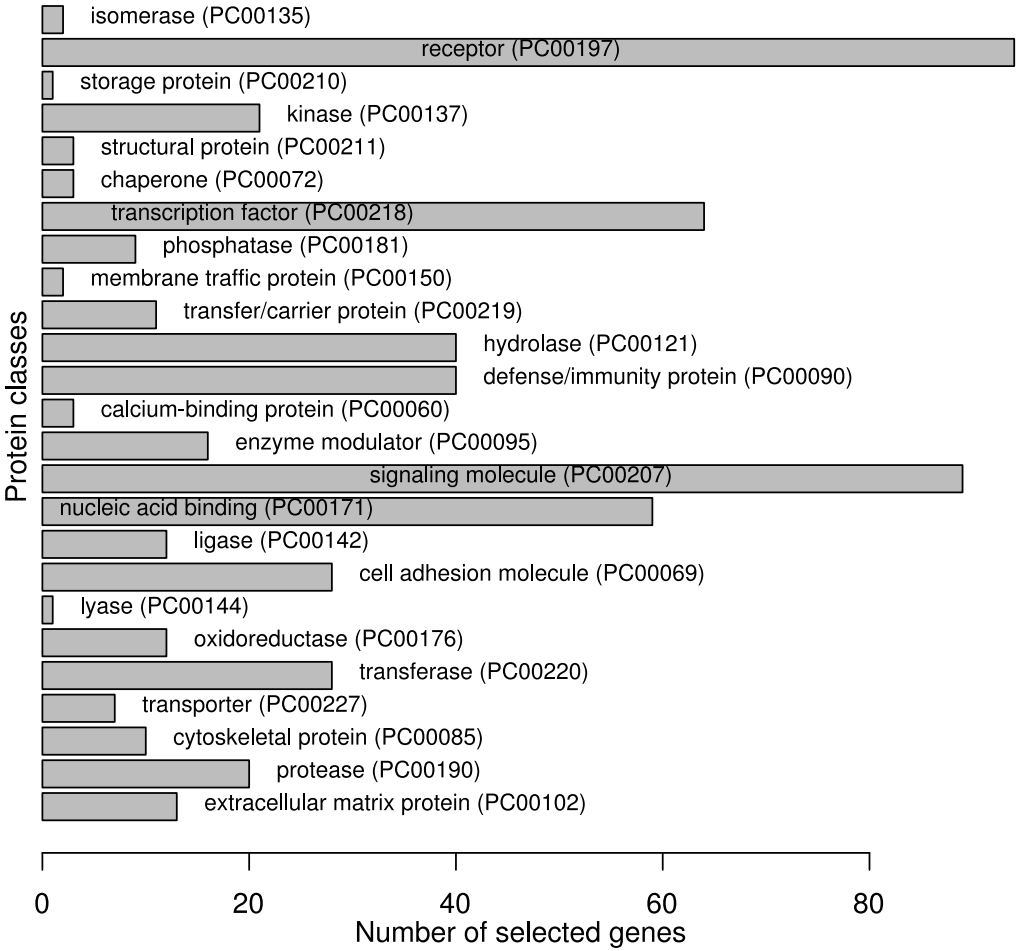
GeneView is an online text mining tool integrating approx. 22M abstracts and full texts. http://bc3.informatik.hu-berlin.de/, Thomas et al. 2012 [doi:10.1093/nar/gks563]



1. Named entity recognition (NER) of genes for each document
2. Extraction of context specific documents matching reg. expression
3. Calculation of a p-value for each gene, that depicts its over-representation in extracted documents using χ^2
4. Filter genes ($p \leq 0.05$)

Selected genes

- Extracted 399 genes
- 194 of selected genes have Immunome Knowledge Base[1](IKB) entries
- 64 transcription factors (TFs), including master TFs like Foxp3, Tbx21, Rorγ, Ifng, Runx, Gata3, Stat3,Cxcr3, Tnf
- 41 cytokines/chemokines including IL-6, IL-10, IL-12,

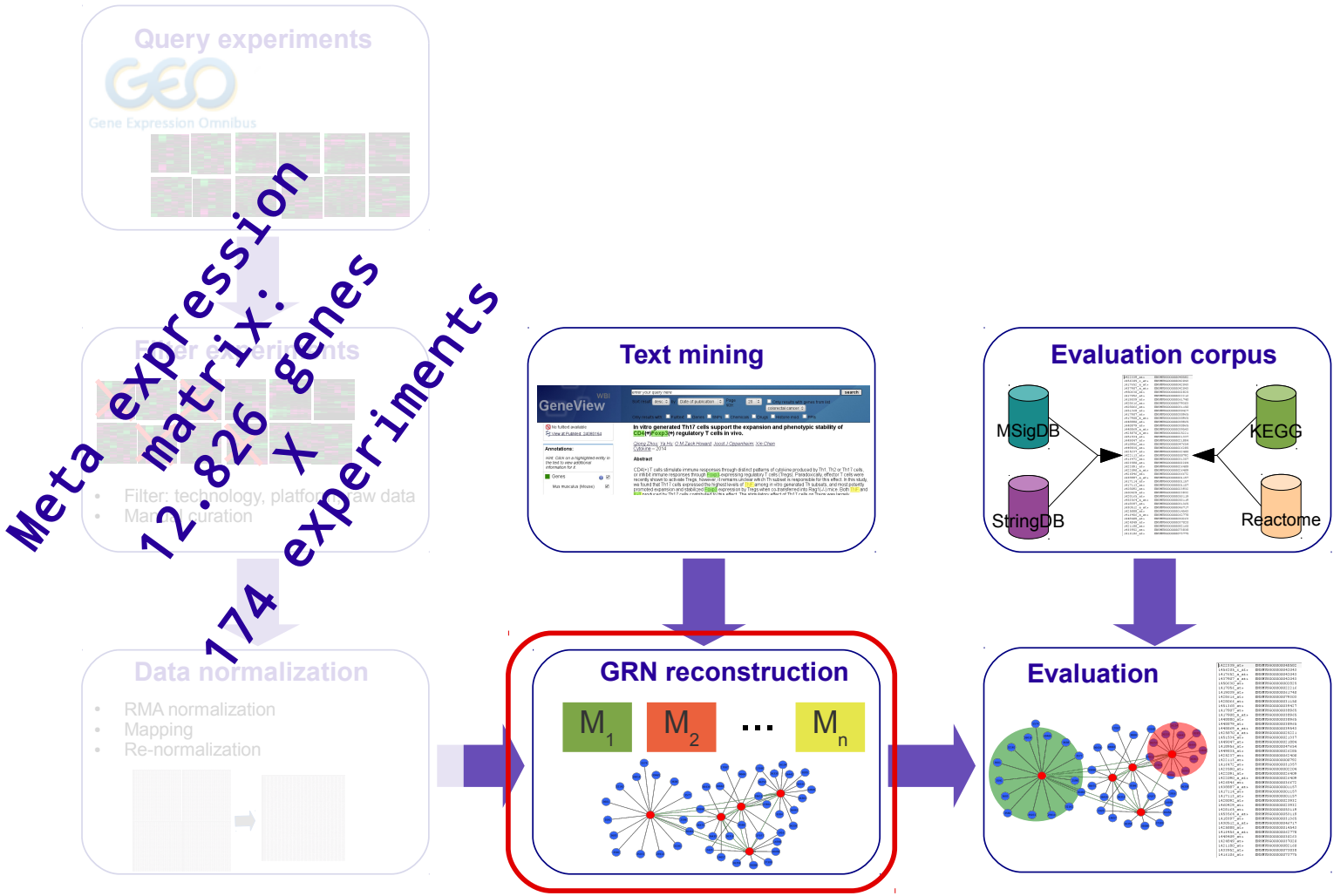


Created with: PANTER, Mi 2012, doi: 10.1093/nar/gks1118

Meta expression matrix of 364 extracted genes and 174 experiments

[1] Ortutay and Vihinen 2009,BMC Immunology 2009 [doi:10.1186/1471-2172-10-3]

Workflow


Query experiments
Gene Expression Omnibus
Meta expression matrix: 12.826 genes X 174 experiments
Filter experiments
Filter: technology, platform, raw data
Manual curation
Data normalization
RMA normalization
Mapping
Re-normalization
Text mining
GeneView
Evaluation corpus
MSigDB
KEGG
StringDB
Reactome
GRN reconstruction
M1
M2
...
Mn
Evaluation

Network reconstruction algorithms

Algorithm	Method	Performance	Reference
ARACNE	Relevance Network and DPI	40% precision on artificial data	Margolin2006
CLR	MI [2]+ Context Likelihood of Relatedness	36% precision on e.coli data	Faith2007
MRNET	MI [2]+ Maximum Relevance Minimum Redundancy	F-measure 0.1 to 0.45 on artificial data	Peng2005
Genie3	Decomposition into regression problems for each gene	$<$ 6% precision on e.coli data	Huynh-Thu2010
Coexpression	Spearman correlation		
Consensus	Combination of algorithms		

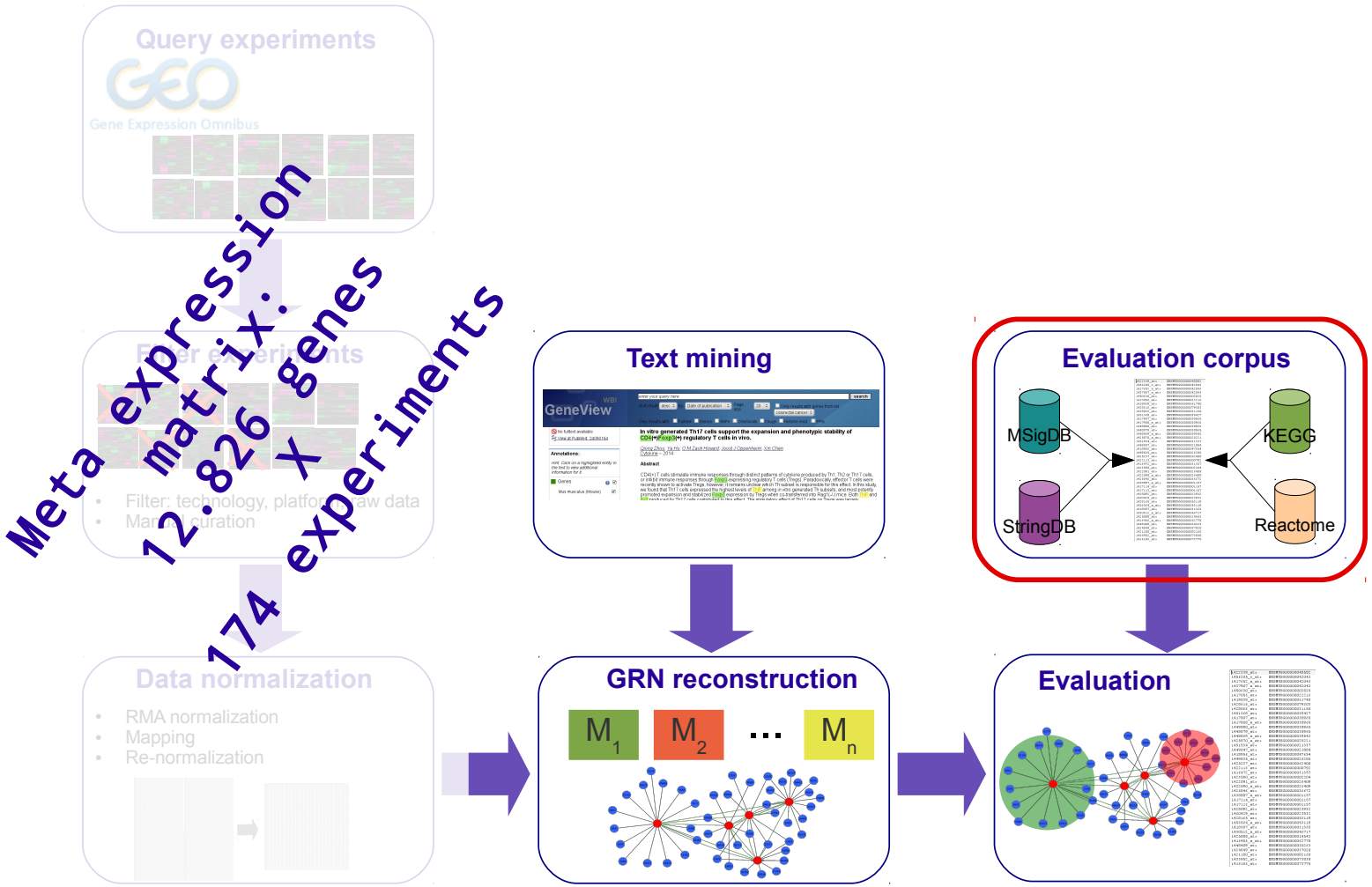
- Consensus: aggregated edge score of integrated algorithms + additional score α each time an edge is repeatedly inferred

$$\sum_{e}^{\textit{edges}} \sum_{a}^{\textit{algos}} \textit{norm}(\textit{rank}_a(e)) + \alpha \ , \ 0 \leq \alpha \leq 1$$

- Edge weights translated to relative rank-based score for comparison of inferred networks

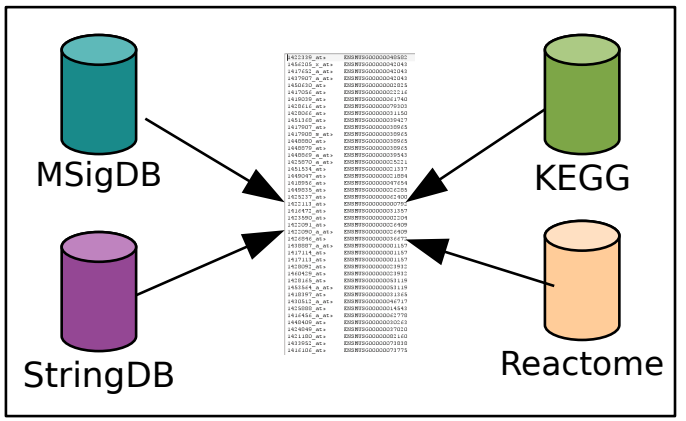
[2] MI - Mutual Information

Workflow


Query experiments
Gene Expression Omnibus
Meta expression matrix: 12.826 genes X 174 experiments
Filter experiments
Filter: technology, platform, raw data
Manual curation
Data normalization
RMA normalization
Mapping
Re-normalization
Text mining
GeneView
Evaluation corpus
MSigDB
KEGG
StringDB
Reactome
GRN reconstruction
M1
M2
...
Mn
Evaluation

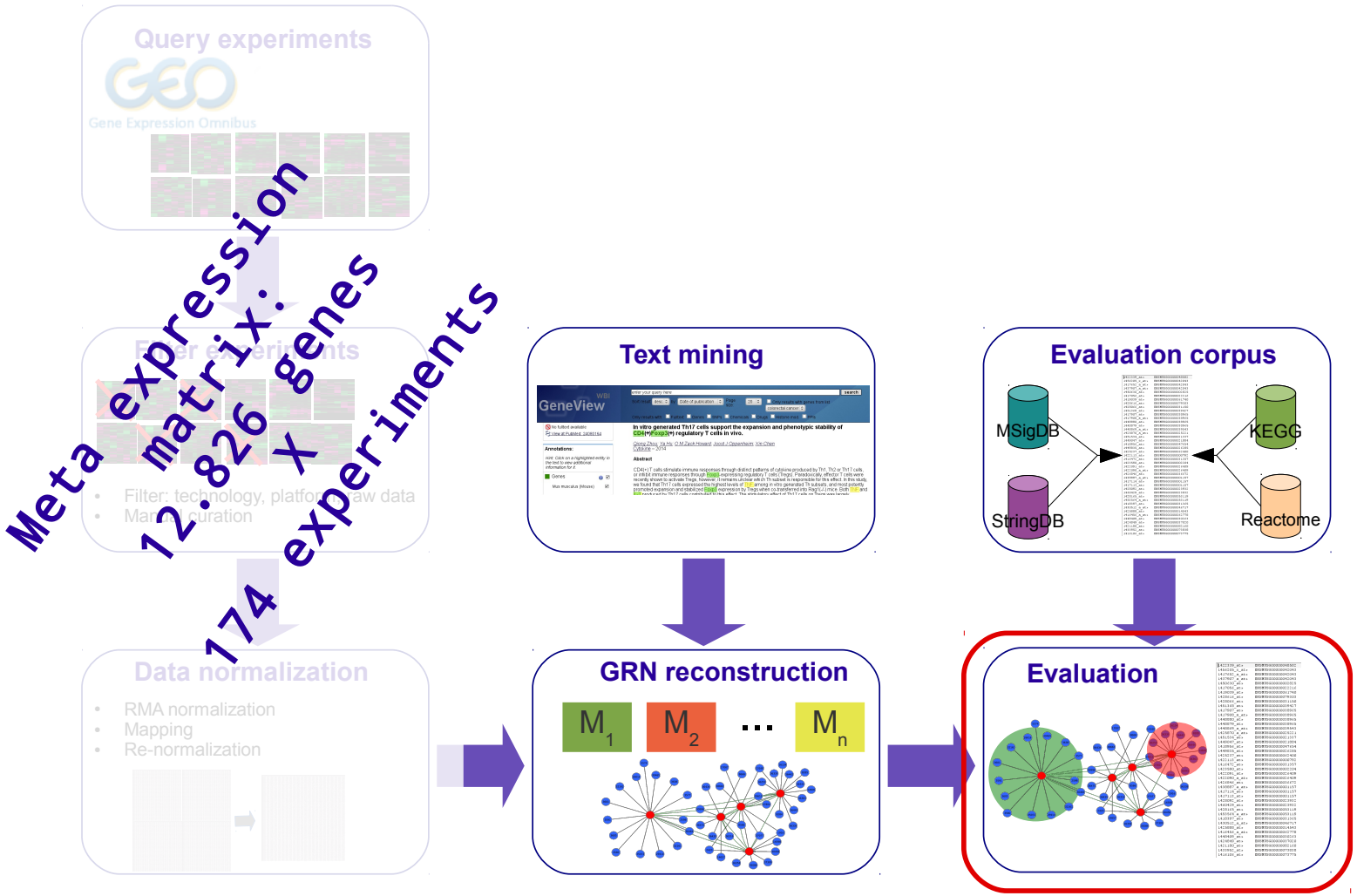
Evaluation

- Evaluation using public databases
  - MSigDB: a binding site of a node in the sequence of an adjacent one
  - StringDB: two nodes share an interaction
  - KEGG / Reactome: two nodes share a common pathway



- Database information are integrated into a single list of edges, here used as "gold standard" (GS)
- Networks
  - Text mining gene set: 364 genes and 20945 edges in GS
  - Set of randomly selected genes: 364 genes and 1507 edges in GS

Workflow


Query experiments
Gene Expression Omnibus
Meta expression matrix: 12.826 X genes 174 experiments
Filter experiments
Filter: technology, platform, raw data
Manual curation
Data normalization
RMA normalization
Mapping
Re-normalization
Text mining
GeneView
Evaluation corpus
MSigDB
KEGG
StringDB
Reactome
GRN reconstruction
M1
M2
...
Mn
Evaluation

Results - on different gene sets

- *Consensus* := $f(CLR, Genie3, Coexpression)$
- $>$ 10% of the edges are exclusively predicted by a single method (except Consensus)
- Consensus: known edges are high ranked, but false positive rate increases at lower ranks
- Recall for all algorithms $< 3\%$

Comparison of algorithms on top 270 edges

	Accuracy		#TP edges	
	text mining	random	text mining	random
ARACNE	0,2529	0,0075	109	4
CLR	0,2414	0,0405	105	21
Coexpression	0,2706	0,0169	115	9
Genie3	0,2245	**0,0445**	99	**23**
MRNET	0,2385	0,0227	104	12
Consensus (α=0)	0,2558	0,0385	110	20
Consensus (α=1)	**0,2736**	0,0385	**116**	20

Comparison of algorithms on all inferred edges

	Precision		Network size	
	text mining	random	text mining	random
ARACNE	0,4	0,028	280	6088
CLR	0,376	0,024	740	42380
Coexpression	**0,422**	**0,032**	275	14944
Genie3	0,363	0,023	1035	66066
MRNET	0,365	0,024	706	42049
Consensus	0,363	0,023	1035	66066

Conclusion

Conclusion

- Text mining can be used to identify Treg-specific gene sets for further analysis
- Consensus approach aggregates good results of integrated algorithms for top ranked edges.

Open tasks

- Include more GRN reconstruction algorithms
- Integrate further prior knowledge
- Evaluate high ranked “unknown” edges

Literature

IKB Ortutay and Vihinen. Immunome Knowledge Base (IKB): An integrated service for immunome research., 2009,BMC Immunology 2009 [doi:10.1186/1471-2172-10-3]

MSigDB Subramanian A et al., Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles, PNAS 2005

StringDB Jensen et al. , STRING 8–a global view on proteins and their functional interactions in 630 organisms., Nucleic Acids Res. 2009, 37(Database issue):D412-6

KEGG Kanehisa, M and Goto, S; KEGG Kyoto Encyclopedia of Genes and Genomes. NAR. 2000.

REACTOME D'Eustachio, P, Pathway databases: making chemical and biological sense of the genomic data flood. Chem Biol. 2013

NCBI GEO Barrett T et al. NCBI GEO: archive for high-throughput functional genomic data. NAR. 2009.

GeneView Thomas P et al., GeneView: a comprehensive semantic search engine for PubMed. NAR. 2012.

Panther Huaiyu Mi, Anushya Muruganujan and Paul D. Thomas. PANTHER in 2013: modeling the evolution of gene function, and other gene attributes, in the context of phylogenetic trees. , Nucl. Acids Res. (2012) doi: 10.1093/nar/gks1118

ARACNE Margolin, A. A; Nemenman, I.; Basso, K.; Wiggins, C.; Stolovitzky, G.; Favera, R. D. & Califano, A. ARACNE: an algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context. BMC Bioinformatics, Department of Biomedical Informatics, Columbia University, New York, NY 10032, USA. adam@dbmi.columbia.edu, 2006, 7 Suppl 1, S7

CLR Faith, J. J.; Hayete, B.; Thaden, J. T.; Mogno, I.; Wierzbowski, J.; Cottarel, G.; Kasif, S.; Collins, J. J. & Gardner, T. S. Large-Scale Mapping and Validation of Escherichia coli , Transcriptional Regulation from a Compendium of Expression Profiles PLoS Biol, Public Library of Science, 2007, 5, e8

Genie3 Huynh-Thu, V. A.; Irrthum, A.; Wehenkel, L. & Geurts, P. Inferring regulatory networks from expression data using tree-based methods. PLoS One, Department of Electrical Engineering and Computer Science, Systems and Modeling, University of Liége, Liége, Belgium., 2010, 5

MRNET H. Peng, F.long and C.Ding. Feature selection based on mutual information: Criteria of max-dependency, max relevance and min redundancy. IEEE transaction on Pattern Analysis and Machine Intelligence, 2005.