An Efficient Method for Indexing Temporal Gene Expression Datasets

Shawn Oliai, Casimir Tokarski, Guenter Tusch, PhD
Health Informatics and Bioinformatics Graduate Program, School of Computing and Information Systems
Grand Valley State University, Grand Rapids, MI, USA

Our Project

Purpose: Identifying Temporal High Throughput Gene Expression Data Sets for Comparative Transcriptome Analysis.

BACKGROUND: Comparative transcriptome analysis of high throughput gene expression temporal experiments helps with understanding of the complexities of living organisms with great value for diagnosis, treatment, and prevention of human diseases.

PROCEDURES: This study explores the feasibility of searching for temporal patterns based on knowledge-based temporal abstractions. Those imply conversion of expression values into an interval-based qualitative representation expressing amount of change over time in terms of trends, as "increasing", "decreasing", "constant" etc. We searched 20,000 articles between 2008-2018, and mined for keywords such as *'high throughput'*, *'gene expression'*, *'transcriptome analysis'*, *'temporal*', *'time series'*, *'longitudinal*, *'chronological'* and other synonym terms.

METHODS AND MATERIALS: To identify those temporal patterns, these studies need to be indexed appropriately since much of the publicly available high throughput expression data is of non-temporal nature. The index needs to be based on the MIAME standard as used for example for abstracting NCBI's Gene Expression Omnibus (GEO) datasets. A simple keyword search of abstracts from NCBI GEO will only result in a large number of false positives.

RESULTS: We essentially see this research as a text mining process to find the correct set of pertinent articles for the topic at hand. Using random samples of keyword search results, we repeatedly refined the search query to obtain better indexing of the datasets from temporal studies. We utilized a series of appropriate words for the algorithm to find in the search and thus were able to significantly improve on false positive and false negative search results.

Conclusion

CONCLUSIONS: After finding 9,694 articles with our selected terms, the were 212 articles including all of the terms and 'temporal' words. Of those 41 had one or more related GSE files for our final consideration.

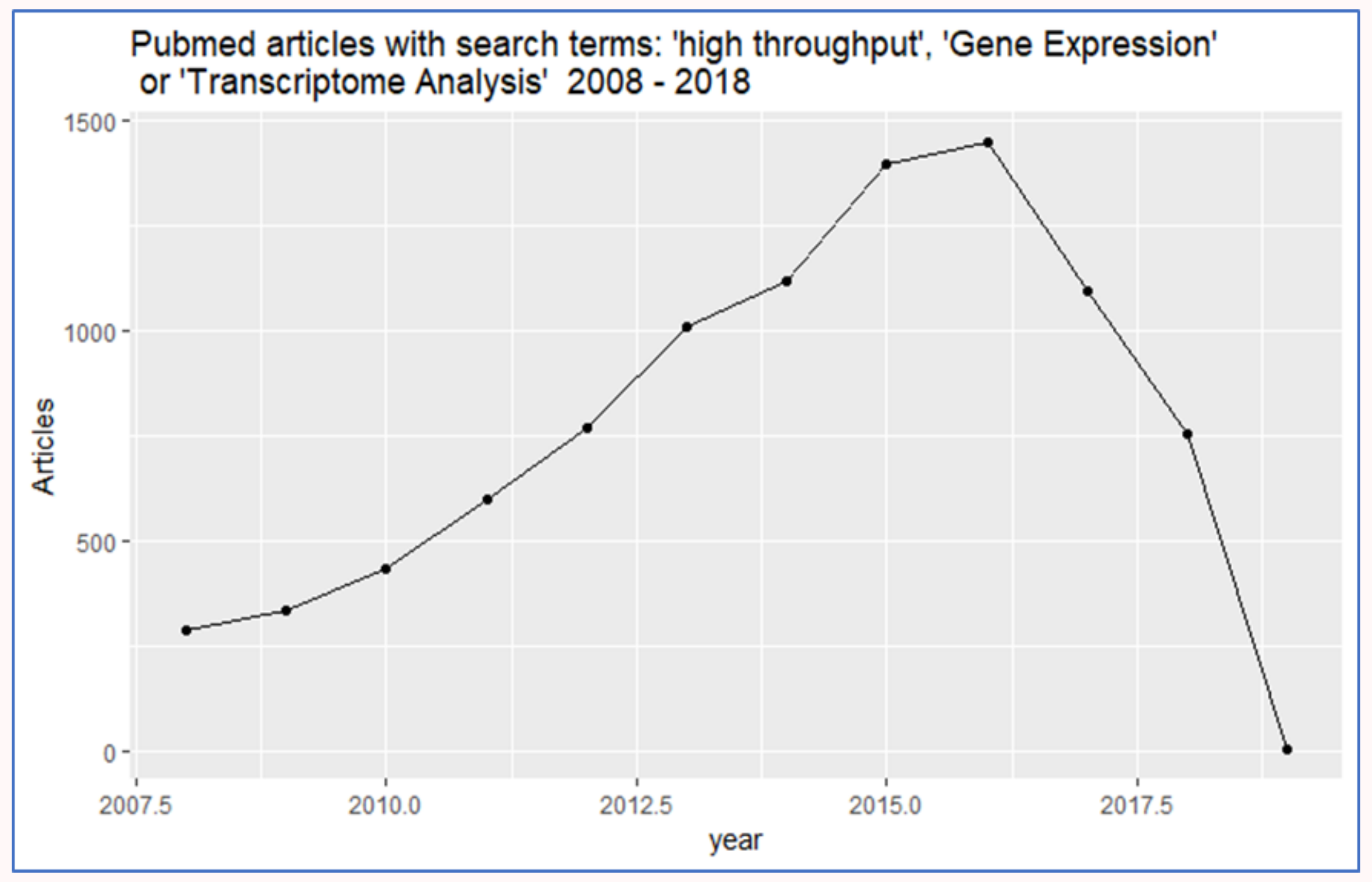
Our promising model can assist researchers in bioinformatics, genetics, medicine and scientific investigations, to find more appropriate selection of scientific data online, while saving various resources (time, capital, talent). The refined indexing algorithm can be used in future studies to compare patterns of gene expression in bioinformatics.

FUTURE EXPLORATION: There are several appropriate considerations on this project:

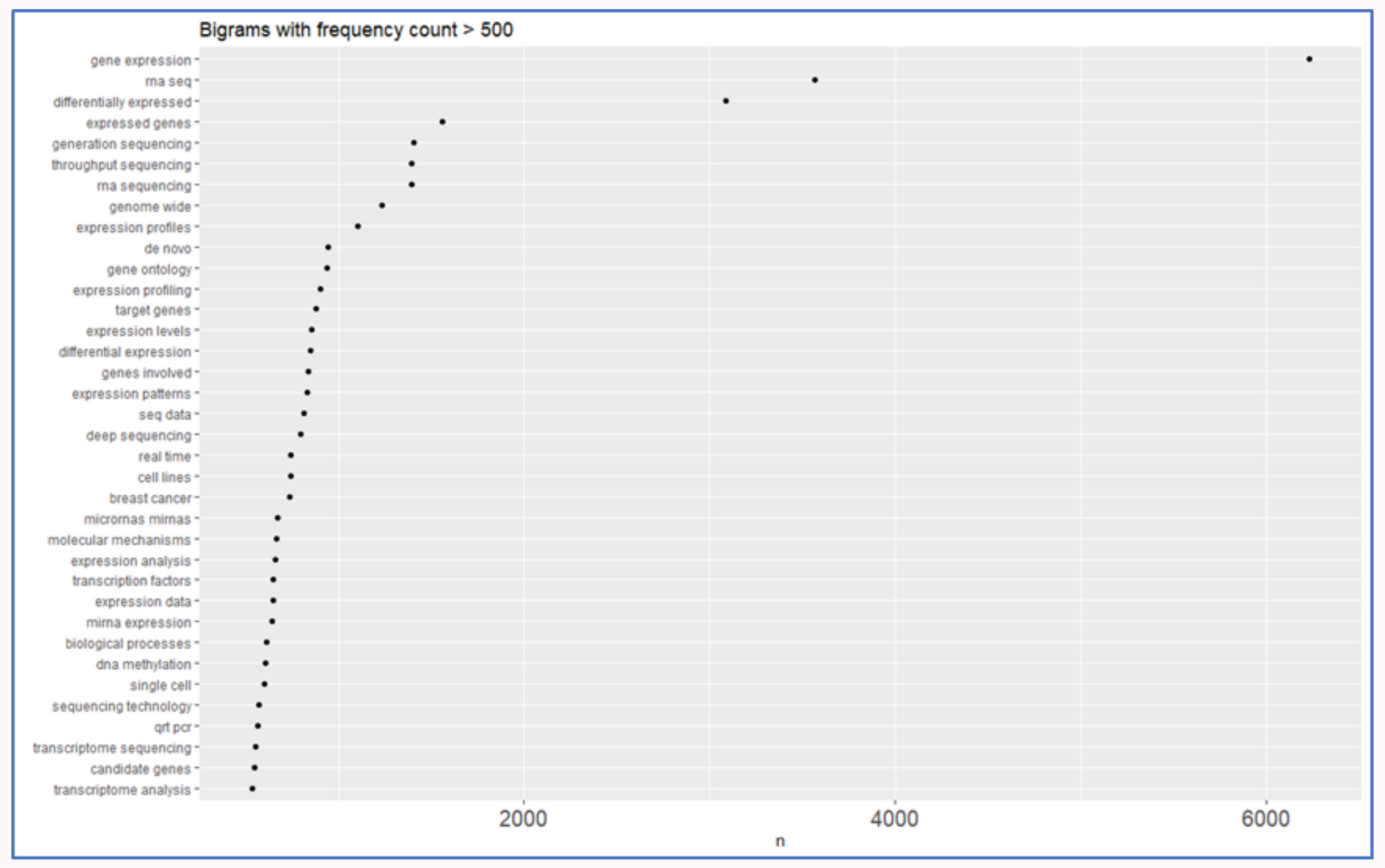
- Examine the methodology in other biological science fields and databases beyond PubMed, and compare the results
- Compare/contrast results with the PubMed's internal text mining tools
- Develop a Topic Network for further exploration, along with deeper insights into LDA modelling

Acknowledgement:

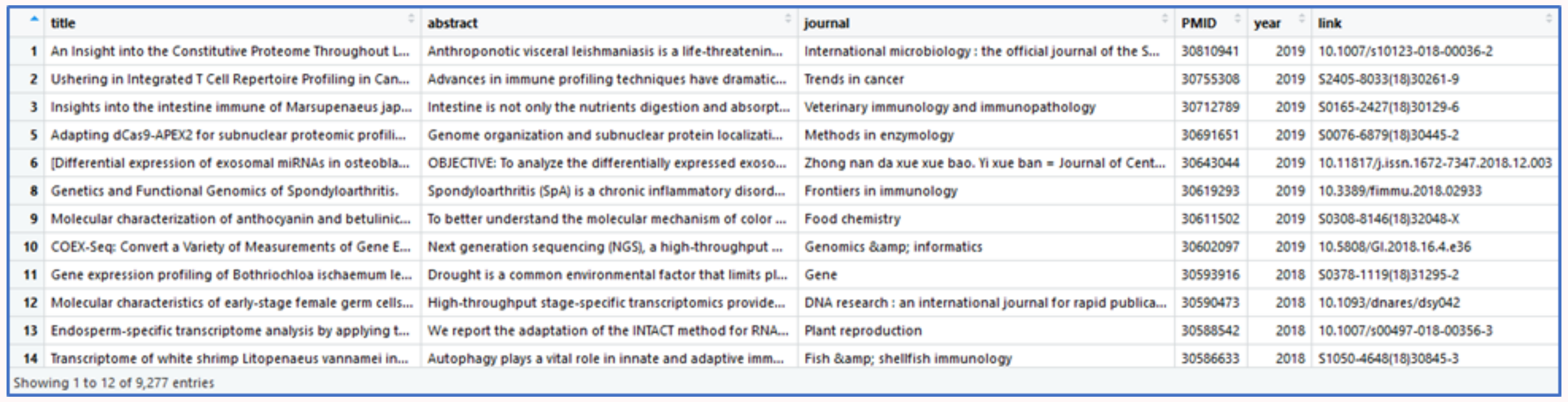
We would like to thank Professor G. Tusch, without whose support and guidance this study would not have been possible.



PubMed Articles with Selected Key Terms



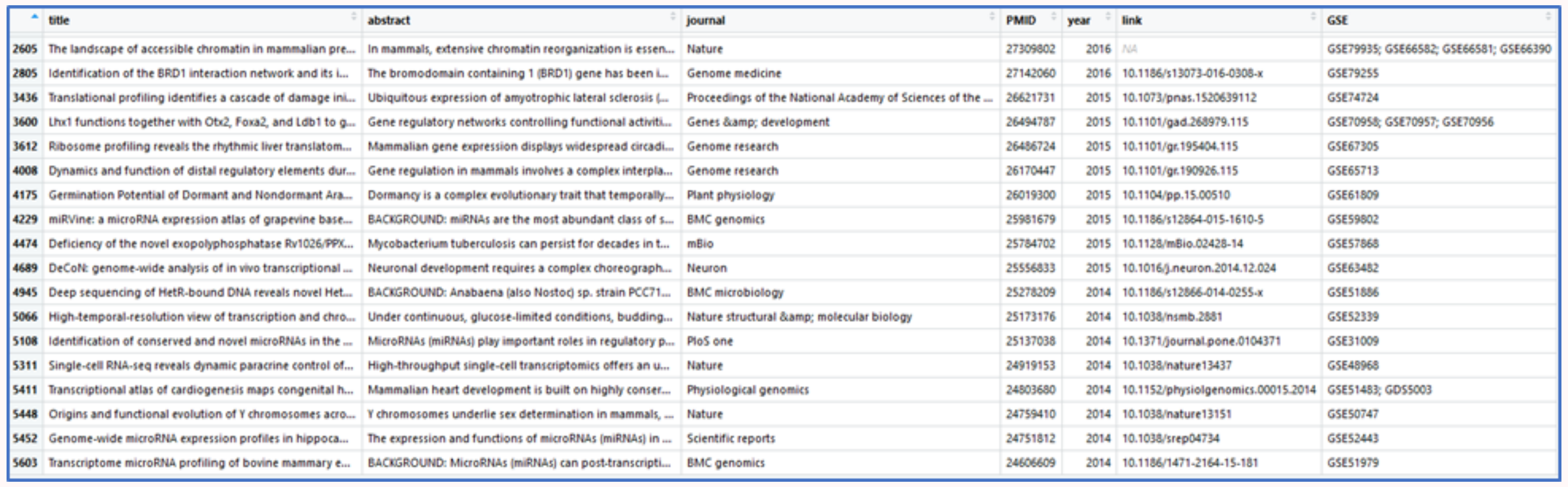
Bigram for high throughput



	title	abstract	journal	PMID	year	link
1	An Insight into the Constitutive Proteome Throughout L...	Anthroponotic visceral leishmaniasis is a life-threatenin...	International microbiology : the official journal of the S...	30810941	2019	10.1007/s10123-018-00036-2
2	Ushering in Integrated T Cell Repertoire Profiling in Can...	Advances in immune profiling techniques have dramatic...	Trends in cancer	30755308	2019	S2405-8033(18)30261-9
3	Insights into the intestine immune of Marsupenaeus jap...	Intestine is not only the nutrients digestion and absorpt...	Veterinary immunology and immunopathology	30712789	2019	S0165-2427(18)30129-6
5	Adapting dCas9-APEX2 for subnuclear proteomic profili...	Genome organization and subnuclear protein localizati...	Methods in enzymology	30691651	2019	S0076-6879(18)30445-2
6	[Differential expression of exosomal miRNAs in osteobla...	OBJECTIVE: To analyze the differentially expressed exoso...	Zhong nan da xue xue bao. Yi xue ban = Journal of Cent...	30643044	2019	10.11817/j.issn.1672-7347.2018.12.003
8	Genetics and Functional Genomics of Spondyloarthritis.	Spondyloarthritis (SpA) is a chronic inflammatory disord...	Frontiers in immunology	30619293	2019	10.3389/fimmu.2018.02933
9	Molecular characterization of anthocyanin and betulinic...	To better understand the molecular mechanism of color ...	Food chemistry	30611502	2019	S0308-8146(18)32048-X
10	COEX-Seq: Convert a Variety of Measurements of Gene E...	Next generation sequencing (NGS), a high-throughput ...	Genomics &amp; informatics	30602097	2019	10.5808/GI.2018.16.4.e36
11	Gene expression profiling of Bothriochloa ischaemum le...	Drought is a common environmental factor that limits pl...	Gene	30593916	2018	S0378-1119(18)31295-2
12	Molecular characteristics of early-stage female germ cells...	High-throughput stage-specific transcriptomics provide...	DNA research : an international journal for rapid publica...	30590473	2018	10.1093/dnares/dsy042
13	Endosperm-specific transcriptome analysis by applying t...	We report the adaptation of the INTACT method for RNA...	Plant reproduction	30588542	2018	10.1007/s00497-018-00356-3
14	Transcriptome of white shrimp Litopenaeus vannamei in...	Autophagy plays a vital role in innate and adaptive imm...	Fish &amp; shellfish immunology	30586633	2018	S1050-4648(18)30845-3

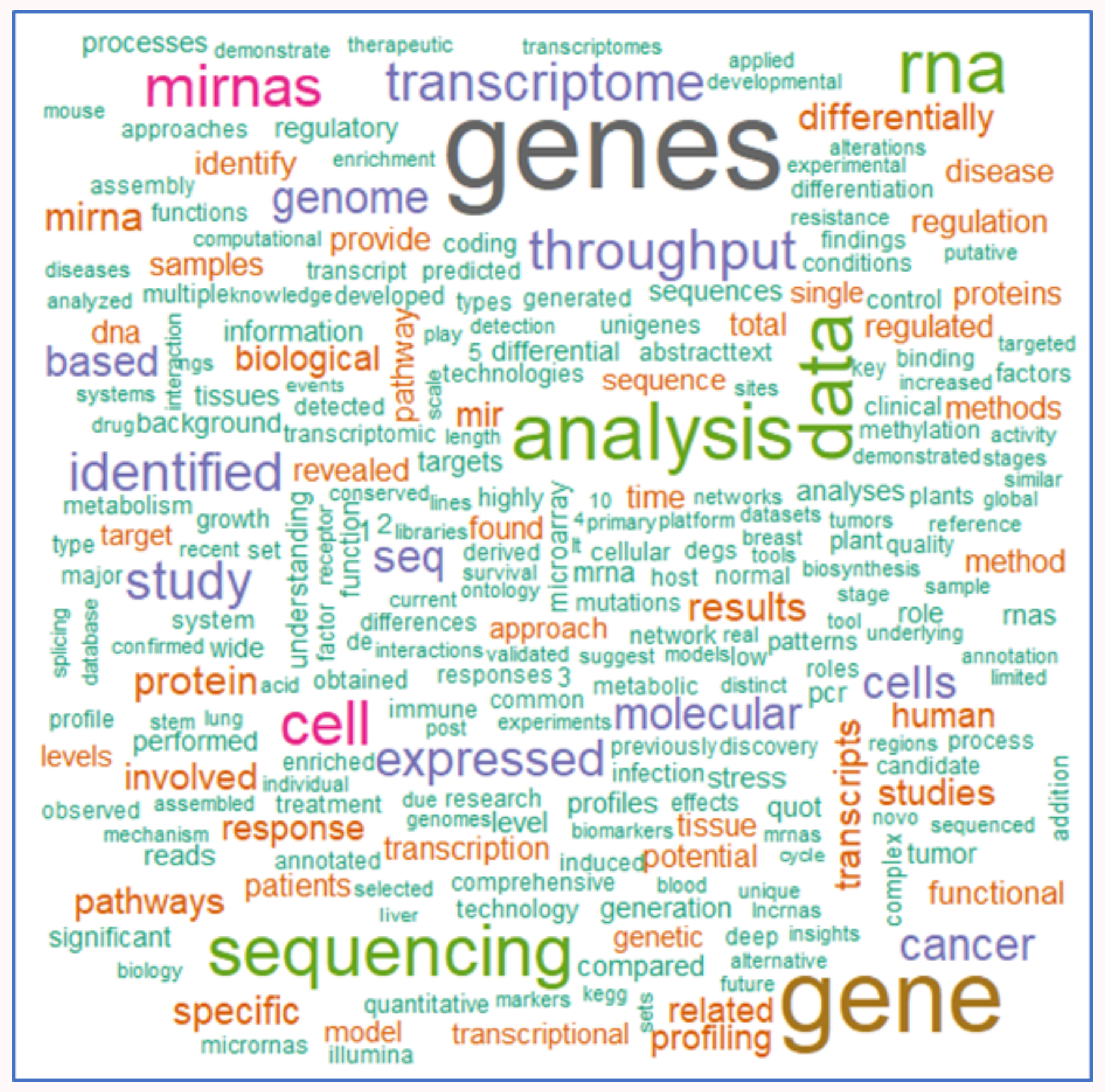
Showing 1 to 12 of 9,277 entries

Articles based on Any of the Keyword Term Sets



	title	abstract	journal	PMID	year	link	GSE
2605	The landscape of accessible chromatin in mammalian pre...	In mammals, extensive chromatin reorganization is essen...	Nature	27309802	2016	NA	GSE79935; GSE66582; GSE66581; GSE66390
2805	Identification of the BRD1 interaction network and its i...	The bromodomain containing 1 (BRD1) gene has been i...	Genome medicine	27142060	2016	10.1186/s13073-016-0308-x	GSE79255
3436	Translational profiling identifies a cascade of damage ini...	Ubiquitous expression of amyotrophic lateral sclerosis (...	Proceedings of the National Academy of Sciences of the ...	26621731	2015	10.1073/pnas.1520639112	GSE74724
3600	Lhx1 functions together with Otx2, Foxa2, and Ldb1 to g...	Gene regulatory networks controlling functional activiti...	Genes &amp; development	26494787	2015	10.1101/gad.268979.115	GSE70958; GSE70957; GSE70956
3612	Ribosome profiling reveals the rhythmic liver translatom...	Mammalian gene expression displays widespread circadi...	Genome research	26486724	2015	10.1101/gr.195404.115	GSE67305
4008	Dynamics and function of distal regulatory elements dur...	Gene regulation in mammals involves a complex interpla...	Genome research	26170447	2015	10.1101/gr.190926.115	GSE65713
4175	Germination Potential of Dormant and Nondormant Ara...	Dormancy is a complex evolutionary trait that temporally...	Plant physiology	26019300	2015	10.1104/pp.15.00510	GSE61809
4229	miRVine: a microRNA expression atlas of grapevine base...	BACKGROUND: miRNAs are the most abundant class of s...	BMC genomics	25981679	2015	10.1186/s12864-015-1610-5	GSE59802
4474	Deficiency of the novel exopolyphosphatase Rv1026/PPX...	Mycobacterium tuberculosis can persist for decades in t...	mBio	25784702	2015	10.1128/mBio.02428-14	GSE57868
4689	DeCoN: genome-wide analysis of in vivo transcriptional ...	Neuronal development requires a complex choreograph...	Neuron	25556833	2015	10.1016/j.neuron.2014.12.024	GSE63482
4945	Deep sequencing of HetR-bound DNA reveals novel Het...	BACKGROUND: Anabaena (also Nostoc) sp. strain PCC71...	BMC microbiology	25278209	2014	10.1186/s12866-014-0255-x	GSE51886
5066	High-temporal-resolution view of transcription and chro...	Under continuous, glucose-limited conditions, budding...	Nature structural &amp; molecular biology	25173176	2014	10.1038/nsmb.2881	GSE52339
5108	Identification of conserved and novel microRNAs in the ...	MicroRNAs (miRNAs) play important roles in regulatory p...	PloS one	25137038	2014	10.1371/journal.pone.0104371	GSE31009
5311	Single-cell RNA-seq reveals dynamic paracrine control of...	High-throughput single-cell transcriptomics offers an u...	Nature	24919153	2014	10.1038/nature13437	GSE48968
5411	Transcriptional atlas of cardiogenesis maps congenital h...	Mammalian heart development is built on highly conser...	Physiological genomics	24803680	2014	10.1152/physiolgenomics.00015.2014	GSE51483; GDS5003
5448	Origins and functional evolution of Y chromosomes acro...	Y chromosomes underlie sex determination in mammals, ...	Nature	24759410	2014	10.1038/nature13151	GSE50747
5452	Genome-wide microRNA expression profiles in hippoca...	The expression and functions of microRNAs (miRNAs) in ...	Scientific reports	24751812	2014	10.1038/srep04734	GSE52443
5603	Transcriptome microRNA profiling of bovine mammary e...	BACKGROUND: MicroRNAs (miRNAs) can post-transcripti...	BMC genomics	24606609	2014	10.1186/1471-2164-15-181	GSE51979

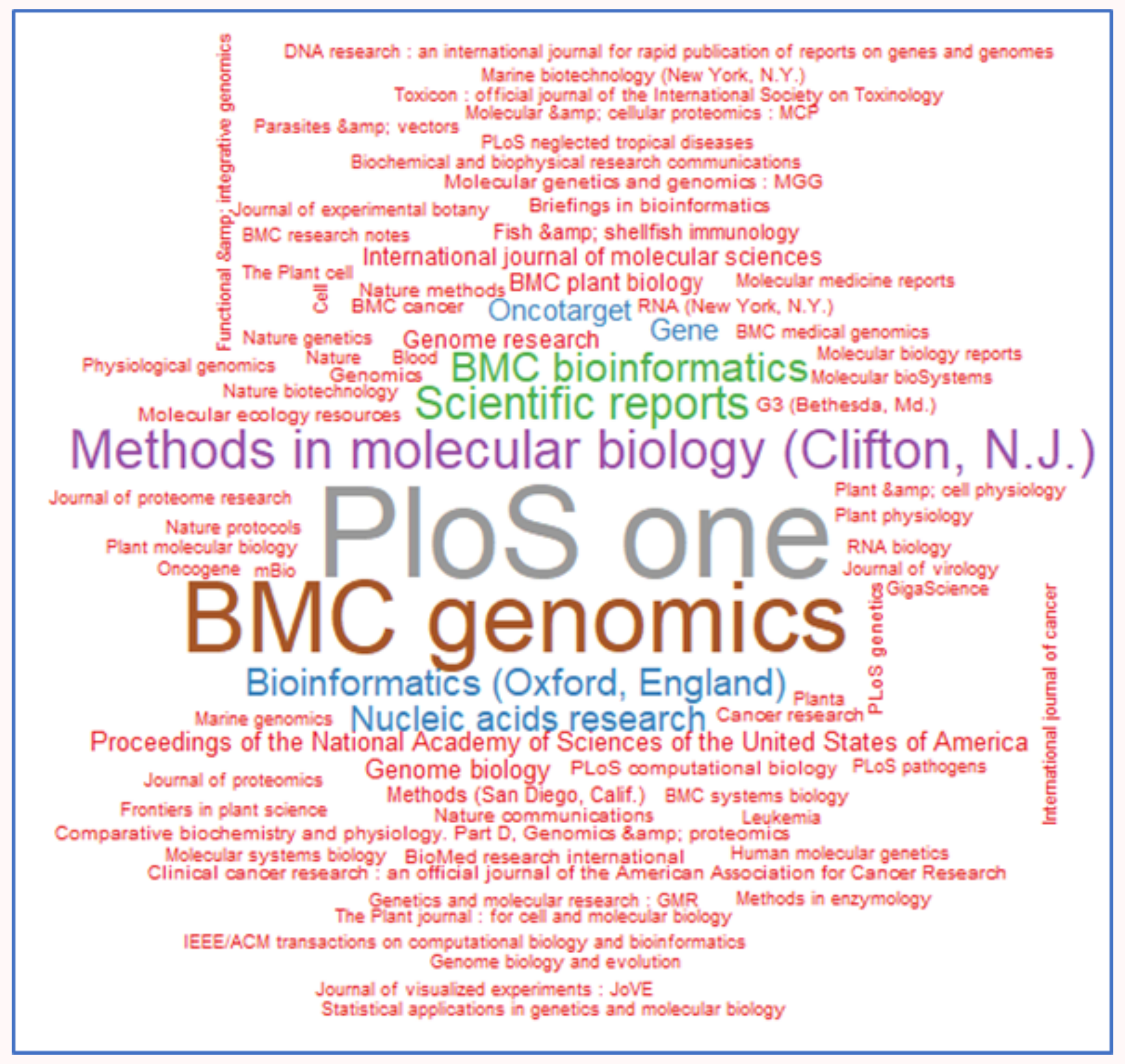
Final Results of Appropriate Articles with GSEs



Word Cloud for top 1500, repeated at least 750 times

gene expression
generation sequencing
throughput screening
differentially expressed
mass spectrometry
rna sequencing
real time
single cell
genome wide
breast cancer
sequencing data
bacterial community

Bigram of words with n>500 high throughput



Journals with at least 10 articles on the topic words

	2008-2018	March 2018
	Entire Set	Sample Set
Run Time	384 s	42 s
Raw Data	9694	1077
Clean (!na)	9277	919
'gene expression'	3520	324
'high throughput'	729	55
'temporal'	212	20
'time series'	36	1
'longitud'*	21	0
'chronol'*	6	0
Article + GSE	40	3