

Breakout Session 6:

Scaling Up Literature Annotations with Cloud Computing in PubTator 3.0

Dr. Zhiyong Lu (Moderator),
*Deputy Director for Literature Search, NCBI; Senior Investigator,
NIH/NLM*

Scaling up literature annotations with cloud computing in PubTator 3.0

Zhiyong Lu, PhD FACMI FIAHSI

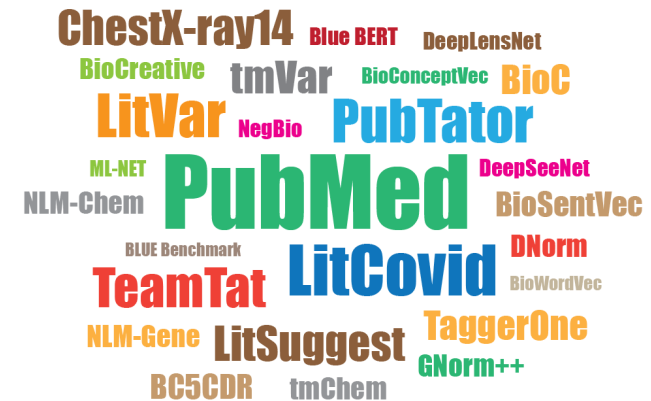
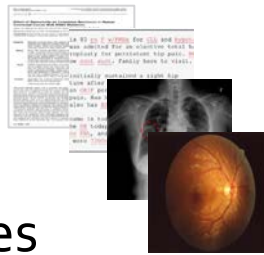
NCBI, NLM, NIH

2024 NIH/ODSSCloud Program PI Meeting

January 18, 2024

Our research at NLM IRP

- Research Areas
 - ✦ Machine Learning; Deep Learning
 - ✦ Natural Language Processing (NLP)
 - ✦ Medical Image Analysis
- Text & Image data
 - ✦ Biomedical Literature
 - ✦ Clinical notes, EMRs
 - ✦ CT, CXR & retinal images
- Application areas:
 - ✦ Literature Retrieval (e.g., PubMed Search; LitCovid)
 - ✦ Information Extraction/Curation (e.g., LitVar, PubTator, LitSuggest)
 - ✦ AI in Healthcare (e.g., machine diagnosis and prognosis)



Gene-disease-variant relations for genomic medicine

[J Alzheimers Dis.](#) 2012;32(2) **Disease** **A** **Gene** **Variant**

Highly pathogenic **Alzheimer's disease** **presenilin 1** **P117R** mutation causes a specific increase in p53 and p21 protein levels and cell cycle dysregulation in human lymphocytes.

[Bialopiotrowicz E](#)¹, [Szybinska A](#), [Kuzniewska B](#), [Buizza L](#), [Uberti D](#), [Kuznicki J](#), [Wojda U](#).



Rank	Gene or Protein ID	Gene SYM	WTAA	MTAA	POS	Disease	PMIDs
1	Q13131	PRKAA1	Q	R	16	Breast cancer	16959974
2	P31749	AKT1	E	K	17	Breast cancer	17611497 18954143 19713527 21793738
3	P10275	AR	H	Y	874	Prostate cancer	17591767

Automatic entity recognition is non-trivial

- Term variation
 - c.77A>C; c.77A->C; A77C;
- Synonymy
 - TP53; tumor protein p53; p53; BCC7; LFS1
- Ambiguity
 - Activated protein C (APC)
 - Antigen-presenting cell (APC)
 - Amino acid-polyamine-organocation (APC)
 - Anaphase-promoting complex (APC)
 - Argon plasma coagulation (APC)

Challenges in relation extraction

Multiple Diseases

Multiple Genes

Multiple Variants

TITLE:

Factors associated with oxidative stress and cancer risk in the Breast and Prostate Cancer Cohort Consortium.

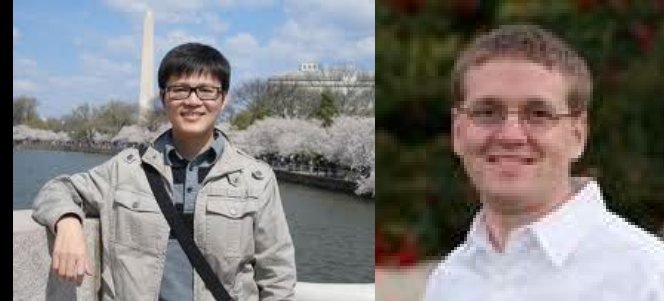
ABSTRACT:

Both endogenous factors (genomic variations) and exogenous factors (environmental exposures, lifestyle) impact the balance of reactive oxygen species (ROS). Variants of the ND3 (rs2853826, G10398A) gene of the mitochondrial genome, manganese superoxide dismutase (MnSOD; rs4880 Val16Ala) and glutathione peroxidase (GPX-1; rs1050450 Pro198Leu), are purported to have functional effects on regulation of ROS balance. In this study, we examined associations of breast and prostate cancer risks and survival with these variants, and interactions between rs4880-rs1050450, and alcohol consumption-rs2853826. Nested case-control studies were conducted in the Breast and Prostate Cancer Cohort Consortium (BPC3), consisting of nine cohorts. The analyses included over 10726 post-menopausal breast and 7532 prostate cancer cases with matched controls. Logistic regression models were used to evaluate associations with risk, and proportional hazard models were used for survival outcomes. We did not observe significant interactions between polymorphisms in MnSOD and GPX-1, or between mitochondrial polymorphisms and alcohol intake and risk of either breast (p-interaction of 0.34 and 0.98, respectively) or prostate cancer (p-interaction of 0.49 and 0.50, respectively). We observed a weak inverse association between prostate cancer risk and GPX-1 Leu198Leu carriers (OR 0.87, 95% CI 0.79-0.97, p = 0.01). Overall survival among women with breast cancer was inversely associated with G10398 carriers who consumed alcohol (HR 0.66 95% CI 0.49-0.88). Given the high power in our study, it is unlikely that interactions tested have more than moderate effects on breast or prostate cancer risk. Observed associations need both further epidemiological and biological confirmation.

Positive Assertion

Negative Finding

Our Entity Taggers



Disease

DNorm (2013) – 80.9%



Mutation

tmVar (2013) – 91.4%
tmVar 3.0 (2022) – 94.3%



Gene/Protein

GNormPlus – 86.7% (2015)
GNorm2 – 89.4% (2023)

Species

SR4GN (2012) – 85.4%

Chemical/Drug

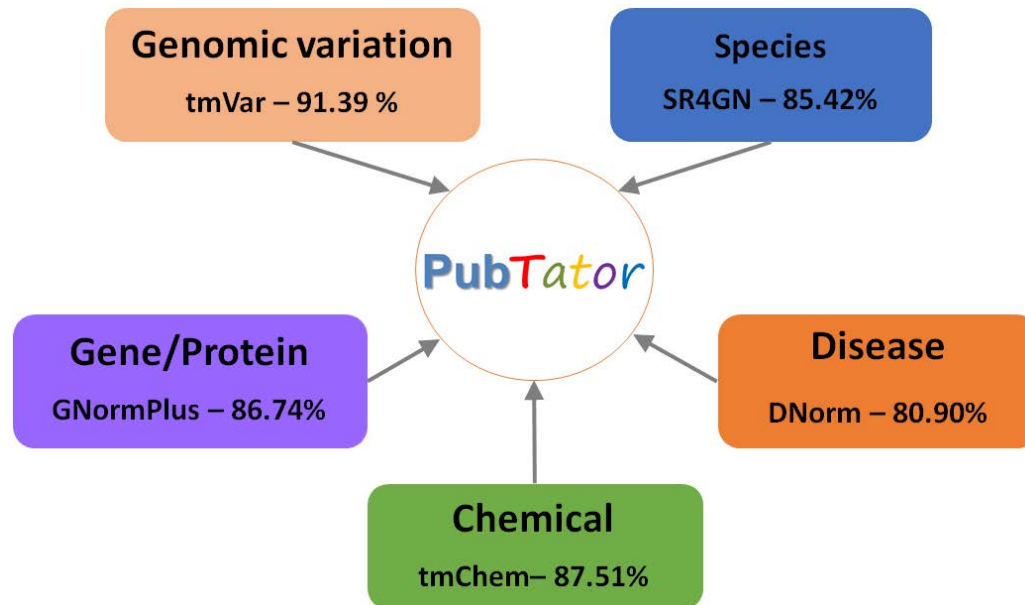
tmChem (2015) – 88.3%



All-in-one Tagger
AIONER (2023)

- State-of-the-art Performance
 - CHEMPROT-Elsevier Challenge Award, 2017
 - GNormPlus: 1st in BioCreative V CHEMDNER-GPRO Task
 - tmChem: 1st in BioCreative V CHEMDNER-CEMP Task
 - DNorm: 1st in ShARe/CLEF Disease Normalization Task
 - tmChem: 1st in BioCreative IV CHEMDNER-CEM Task
- Hybrid approach (mostly machine learning)
- Freely available and open source
- Scalable and interoperable
- Large user base

PubTator (2012 -): integrating text-mined results at PubMed scale



Curatable
 Not Curatable
 TBD
 PubTator
 Disease
 Species
 Mutation
 Chemical
 Gene

PMID:26022131 Selection of a novel DNA thioaptamer against HER2 structure.

Publication: Clinical _ translational oncology : official publication of the Federation of Spanish Oncology Societies and of the National Cancer Institute of Mexico; 2015 May 29 [Full text links]

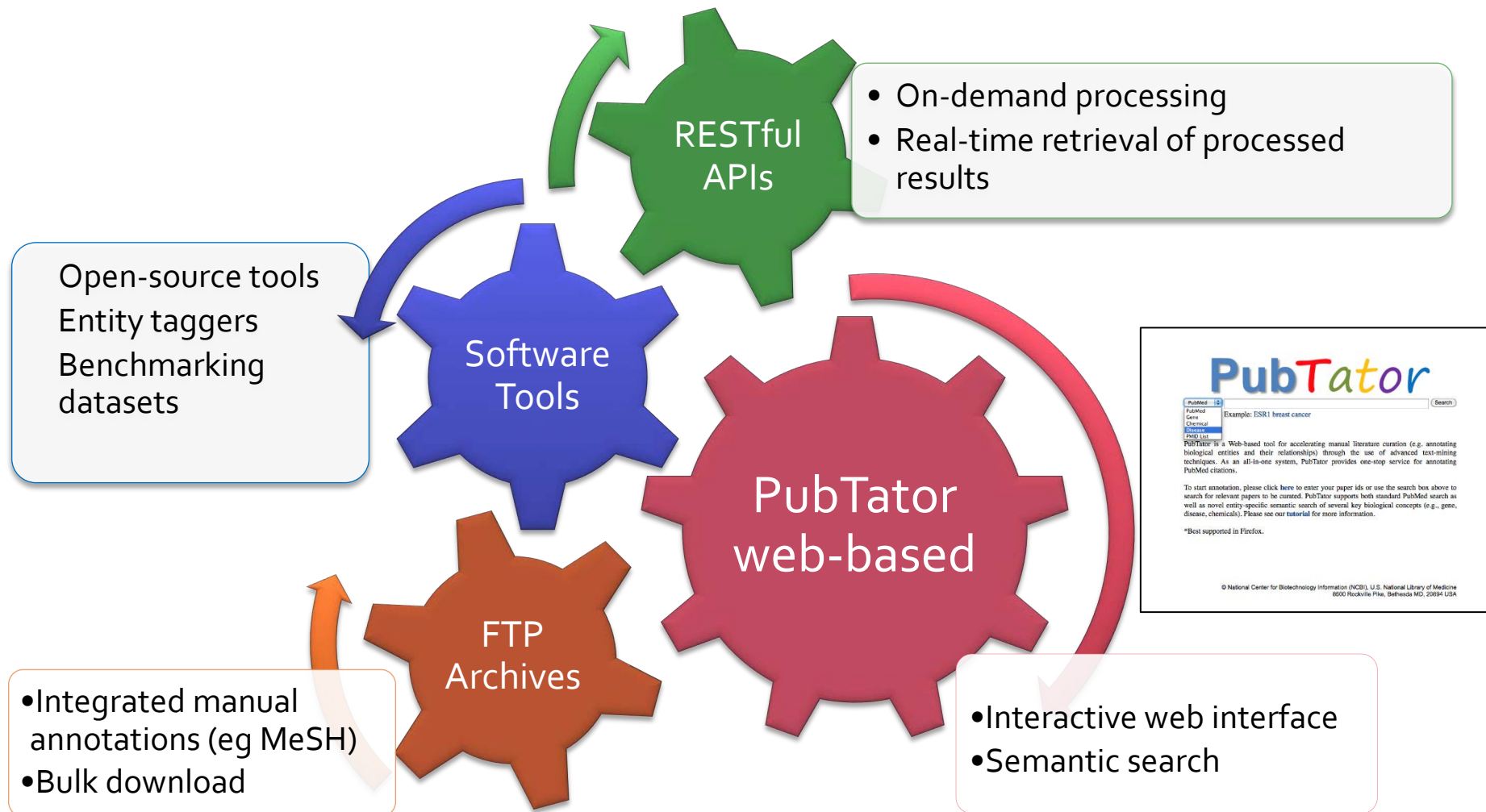
TITLE: Selection of a novel DNA thioaptamer against HER2 structure.
ABSTRACT:

PURPOSE: Human epithelial growth factor receptor 2 (HER2) is over-expressed in several malignancies and represents an important therapeutic target. Aptamers are oligonucleotides that may potentially serve as tumor-homing ligand with excellent affinity and specificity for targeted cancer therapy. However, aptamers need to have nuclease resistance in order to function in vivo. The aim of this study was to generate a novel HER2 thioaptamer with enhanced nuclease resistance. **METHODS:** The HER2 thioaptamer is selected in an evolutionary process called systematic evolution of ligands by exponential enrichment. **RESULTS:** The thioaptamer could bind to the extracellular domain of HER2 with a K d of 172 nM and had minimal cross reactivity to trypsin or IgG. Moreover, the thioaptamer was found capable of binding with the HER2-positive breast cancer cells SK-BR-3 and MDA-MB-453, but not the HER2-negative cells MDA-MB-231. Notably, the thioaptamer HY6 largely maintained its structural integrity facing the nucleases in serum, while regular DNA aptamers were mostly digested. Additionally, the thioaptamer retained the capability of binding with the HER2-positive cells in the presence of serum, whereas non-thionated HER2 aptamer lost the binding function. **CONCLUSION:** The results indicated that the selected thioaptamer was more resistant to nuclease than regular DNA aptamers and might potentially function as a HER2-targeting ligand in complicated environment.

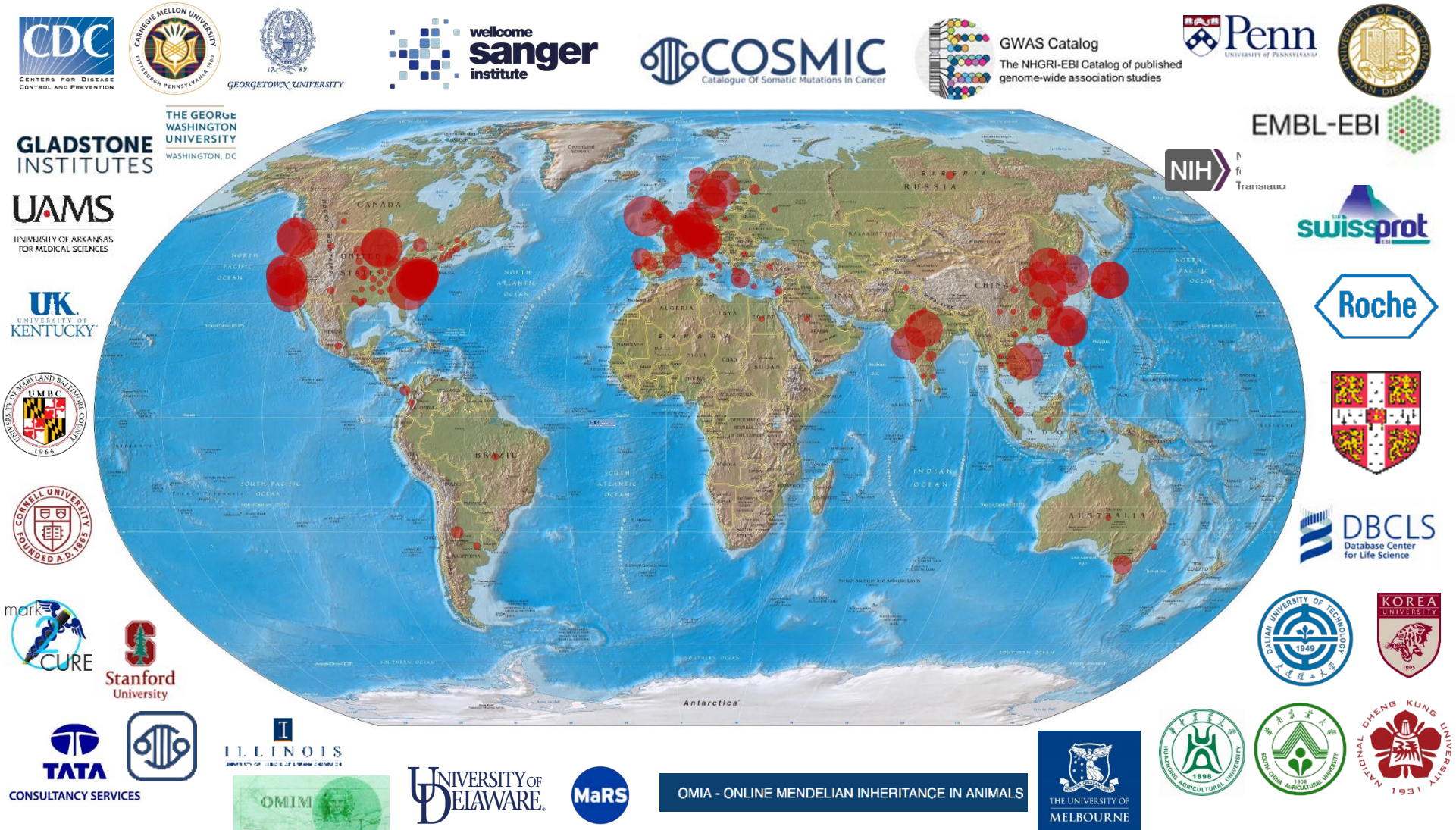
Concept View
 Mention View
 [Add bio-relation annotation to the table below.](#)

Entity type	Entity mention	Concept ID	Nomenclature	Delete
Disease	breast cancer	D001943	MEDIC	Delete
Disease	cancer tumor	D009369	MEDIC	Delete
Gene	HER2 Human epithelial growth factor receptor 2	2064	NCBI Gene	Delete

Data Sharing and Reuse



PubTator API usage: over 1 billion requests since April 2015



PubTator milestones (2012 -)

2019

Deep Learning for improved results



2016

Bulk download available via FTP



2018

Full-text PMC articles added (PubTator 2.0)



2023

Relation annotations (PubTator 3.0)



2015

APIs released



2012

PubTator first launched at BioCreative IV

PubTator interface showing a search result for a novel DNA thioaptamer against HER2 structure. The interface includes filters for Curatable, Not Curatable, TBD, Disease, Species, Mutation, Chemical, and Gene. The search results show a table of entities with their types, mentions, concept IDs, nomenclatures, and delete options.

Entity type	Entity mention	Concept ID	Nomenclature	Delete
Disease	breast cancer	0001943	MEDIC	Delete
Disease	cancer	0009999	MEDIC	Delete
Gene	HER2	3094	NCBI Gene	Delete
	Human epithelial growth factor receptor 2			

PubTator 3.0 beta is now available

The screenshot shows the top navigation bar of the PubTator 3.0 beta website. On the left, there are links for 'Home', 'Saved', and 'Playlists'. The main header area contains the text 'Search entities & relations in 35+ million biomedical publications.' and a search input field with the placeholder text 'Ex: Remdesivir'. To the right of the search bar is a magnifying glass icon. Below the search bar, there is a suggestion: 'Try: N-dimethylnitrosamine and Metformin COVID-19 and PON1'. The top right corner of the header has links for 'FTP', 'API', and 'FAQ'. The NIH and NLM logos are also present in the top right area.

The diagram shows a search input field with the text 'Can'. Below the input field, a dashed line connects to a box labeled 'Neoplasms'. To the right of the 'Neoplasms' box is a small icon of a person with a stethoscope. Below the diagram, there are icons of a test tube, a pipette, and a DNA double helix.

Entities Autocomplete

PubTator3 uses a high-performance entities search engine, to normalize different forms of the same entity into a unique standardized name to returned all matching articles.

The diagram shows an open book with two pages. The left page is titled 'PubTator3 : an improved search engine' and the right page is titled 'PubTator Central: search engine for biomedical in PubMed and PMC'. Below the book, there is a DNA double helix icon.

Full Text Search

PubTator3 provides unified access to the entire 35+ million abstracts in PubMed and nearly 6 million full-text articles in the PMC Text Mining subset.

The diagram shows a person with a stethoscope, a DNA double helix, and a test tube. Dashed lines connect the person to the DNA, and the DNA to the test tube, illustrating a relationship between entities.

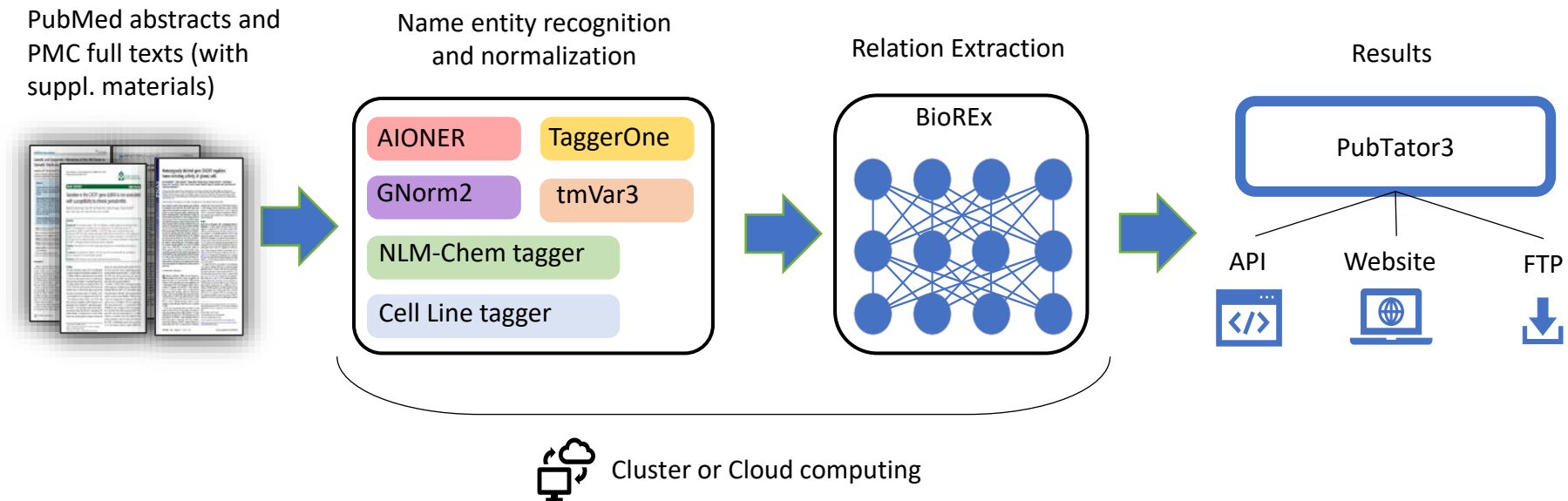
Relations

PubTator3 allows to filter results to only return publications containing specific relations between two entities, such as diseases, chemicals, genes or variants.

<https://www.ncbi.nlm.nih.gov/research/pubtator3/>

Supported by NIH ODSS Exploratory Cloud Program

Computational tasks need GPU access



- IE methods enhancement
- (Re-)process the entire PubMed/PMC datasets
- Annotate new articles each day
 - 5k-100K PubMed abstracts / 1-5K full texts (daily)

Available computing resources

CPU/GPU	#	Resource
Nvidia A100	0-1	Biowulf
Nvidia V100	0-1	Biowulf
Nvidia Tesla K80	0-5	Biowulf
CPU	0-400	Biowulf
CPU	0-300	NCBI computer cluster



CPU/GPU	#	Resource
Nvidia A100	2	Google cloud
CPU	24	Google cloud

- Most of our IE tools are developed using deep learning, which rely on GPUs.
- The availability of NCBI/Biowulf resources is unstable and often occupied by other users.

Comparison

Tasks	NCBI Cluster/BioWulf	Google Cloud
Re-process the entire PubMed/PMC	100-600 CPUs + 1 K80 GPU	2 A100 + 24 CPUs
	5+ months	1-2 months
Method improvement	GPU (not guaranteed)	2 A100
	Days to Weeks	Hours to Days
Regularly process new articles	100-300 CPUs	1 A100 + 24 CPUs
	1-5 days	Daily

Acknowledgments



Team members: Alexis Allot, Don Comeau, Qingyu Chen, Rezarta Dogan, Amr Elsayy, Aadit Kapoor, Won Kim, Qiao Jin, Robert Leaman, **Po-Ting Lai**, Ling Luo, Shubo Tian, **Chih-Hsuan Wei**, John Wilbur, Natalie Xie, Yifan Yang, Lana Yeganova, Qingqing Zhu

CC: Ron Summers, Le Lu, Xiaosong Wang, Ke Yan, etc.

NCATS: Tyler Beck, Christine Colvis, Noel Southall

NCI: Daniela Seminara, Travis Hyams, Brionna Hair

NHGRI: Vence Bonham, Larry Brody, Julia Byeon

NIAID: Morgan Similuk, Daniel Veltri, Sandhya Xirasagar, Rajarshi Ghosh, Andrew Oler

PubMed: Kathi Canese, Grisha Starchenko, etc.

dbSNP: Lon Phan, Juliana Feltz, Rama Maiti, Tim Hefferon

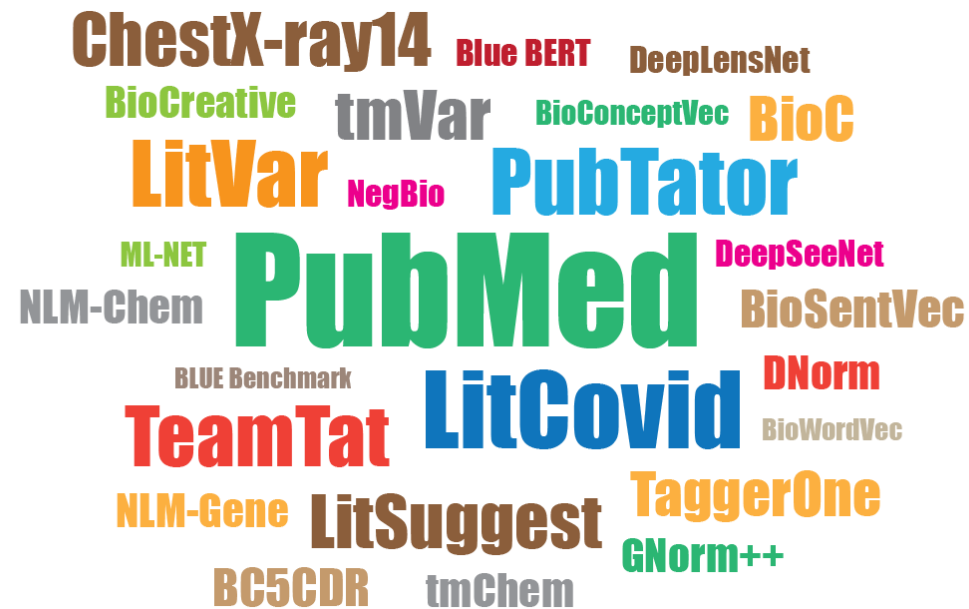
ClinVar: Melissa Landrum

UniProt: Cecilia Arighi, Alex Bateman, Alan Bridge, Livia Famiglietti, Michele Magrane, Sylvain Poux, Cathy Wu, Ioannis Xenarios, etc.

GWAS Catalog: Aoife McMahon, Jackie MacArthur, Fiona Cunningham, Helen Parkinson

And many others at NLM/NIH and beyond!!

Data Sharing and Open Science



zhiyong.lu@nih.gov