Data Acquisition

The Big Data Era in Biology and Data Integration

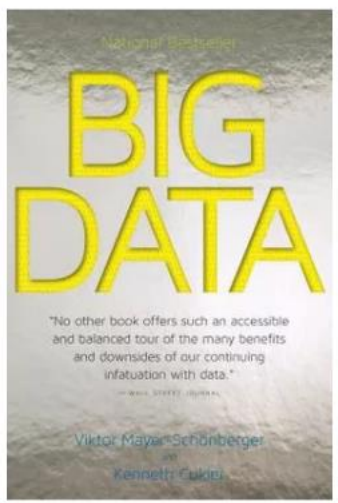
SecondSale

0 Wish List Sign in

Enter ISBN(s),title or keyword(s)

All Categories About Us Contact Us

/ All Categories / Business / Information Management / Big Data: A Revolution That Will Transfor Think



Big Data: A Revolution That Will Transform How We Live, Work, and Think

by **Viktor MayerSchonberger** and Kenneth Cukier

Select Format

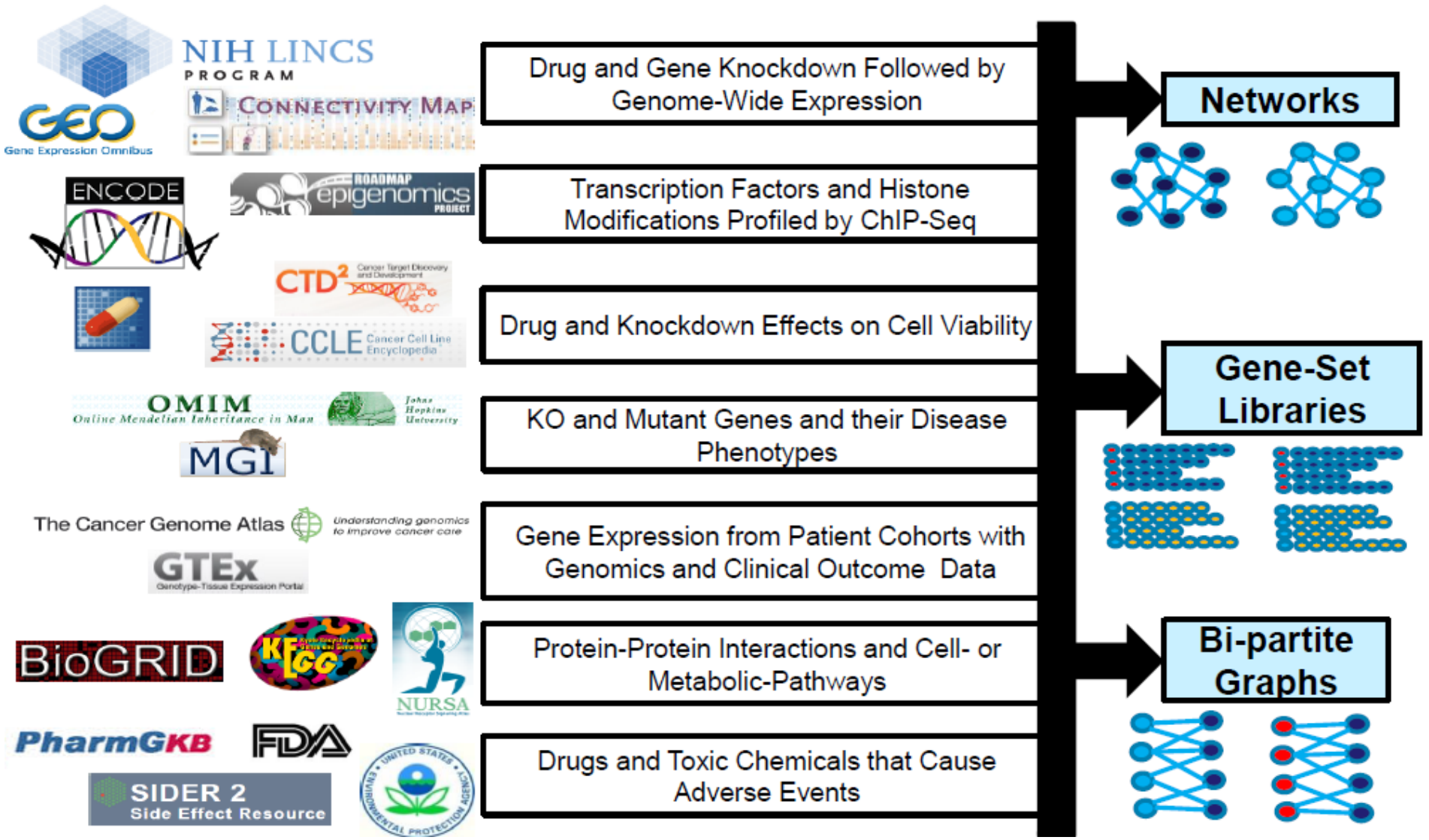
Paperback
$3.88 - $15.95

Select Conditions

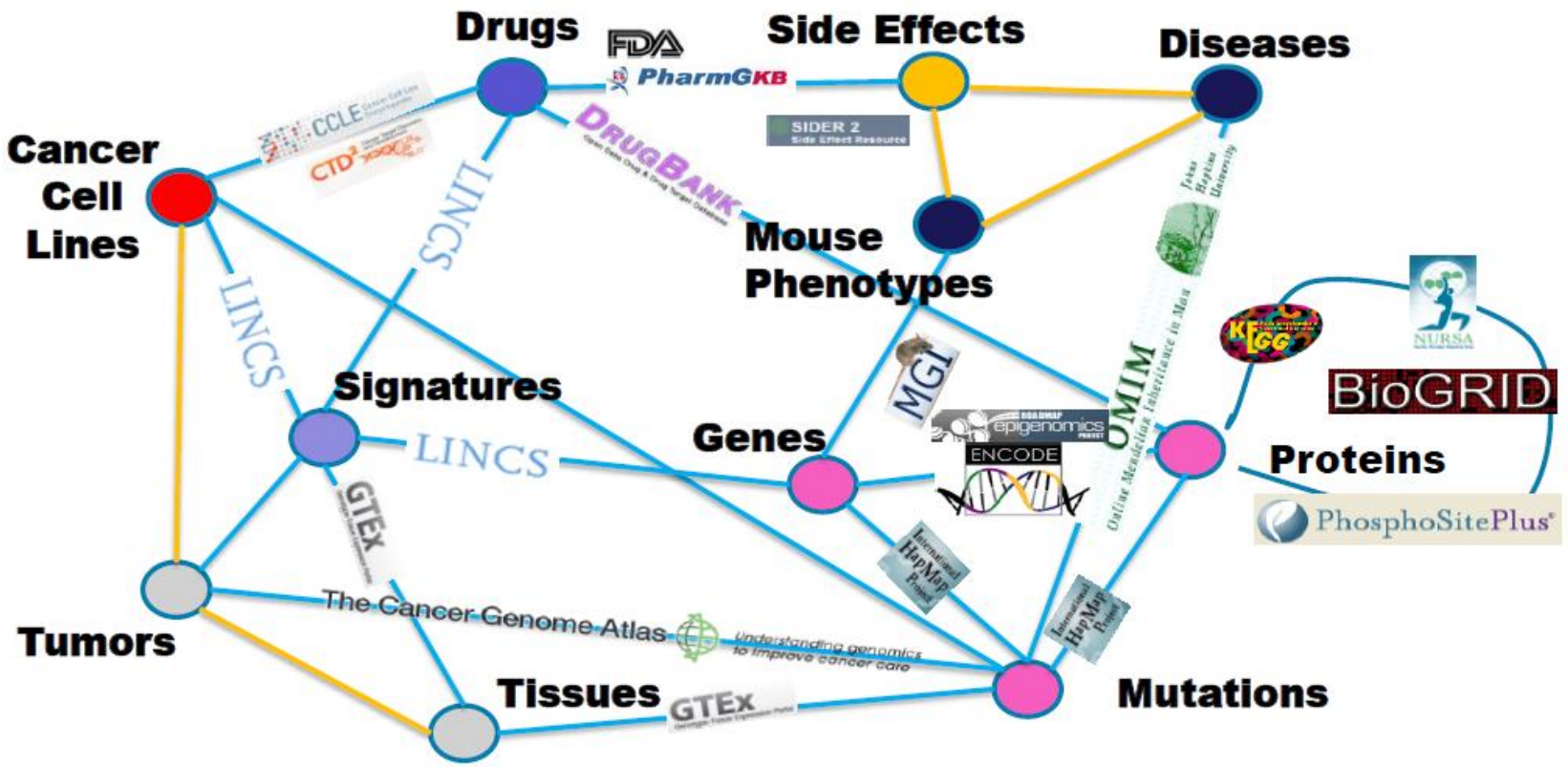
Good
$3.88

New
$15.95

BIG Data


NIH LINCS PROGRAM
CONNECTIVITY MAP
GEO
Gene Expression Omnibus
ENCODE
ROADMAP epigenomics PROJECT
CTD2
CCLE Cancer Cell Line Encyclopedia
OMIM
Online Mendelian Inheritance in Man
Johns Hopkins University
MGI
The Cancer Genome Atlas
understanding genomics to improve cancer care
GTEx
BioGRID
KEGG
NURSA
PharmGKB
FDA
SIDER 2
Side Effect Resource
Drug and Gene Knockdown Followed by Genome-Wide Expression
Transcription Factors and Histone Modifications Profiled by ChIP-Seq
Drug and Knockdown Effects on Cell Viability
KO and Mutant Genes and their Disease Phenotypes
Gene Expression from Patient Cohorts with Genomics and Clinical Outcome Data
Protein-Protein Interactions and Cell- or Metabolic-Pathways
Drugs and Toxic Chemicals that Cause Adverse Events
Networks
Gene-Set Libraries
Bi-partite Graphs

Global Picture of the Puzzle



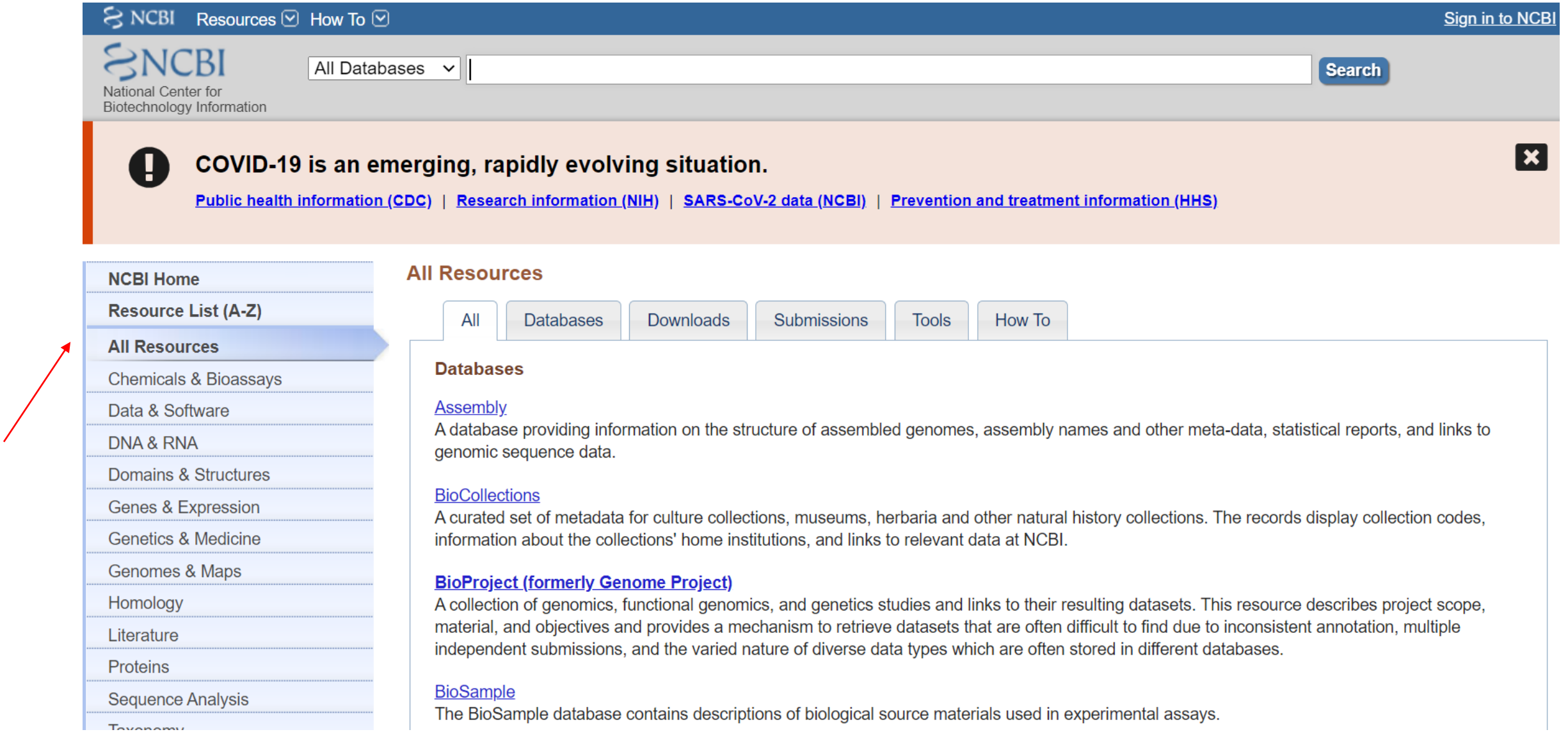
Avi Ma’ayan lab

Topics of this lecture

- Proteins
- Genes
- Drugs
- Health data
- Clinical data
- Economical Data, etc…

Census data, NHANES, Gene expression omnibus (GEO), Genomic data commons, dbGap, ENCODE, Kaggle, SEER, etc. , Ethical issues and responsibilities , Challenges of handling big-data

NCBI biological database (search for BBX)



Big Data Resources that can be Converted to Gene Networks, Gene Set Libraries, and Bipartite Graphs

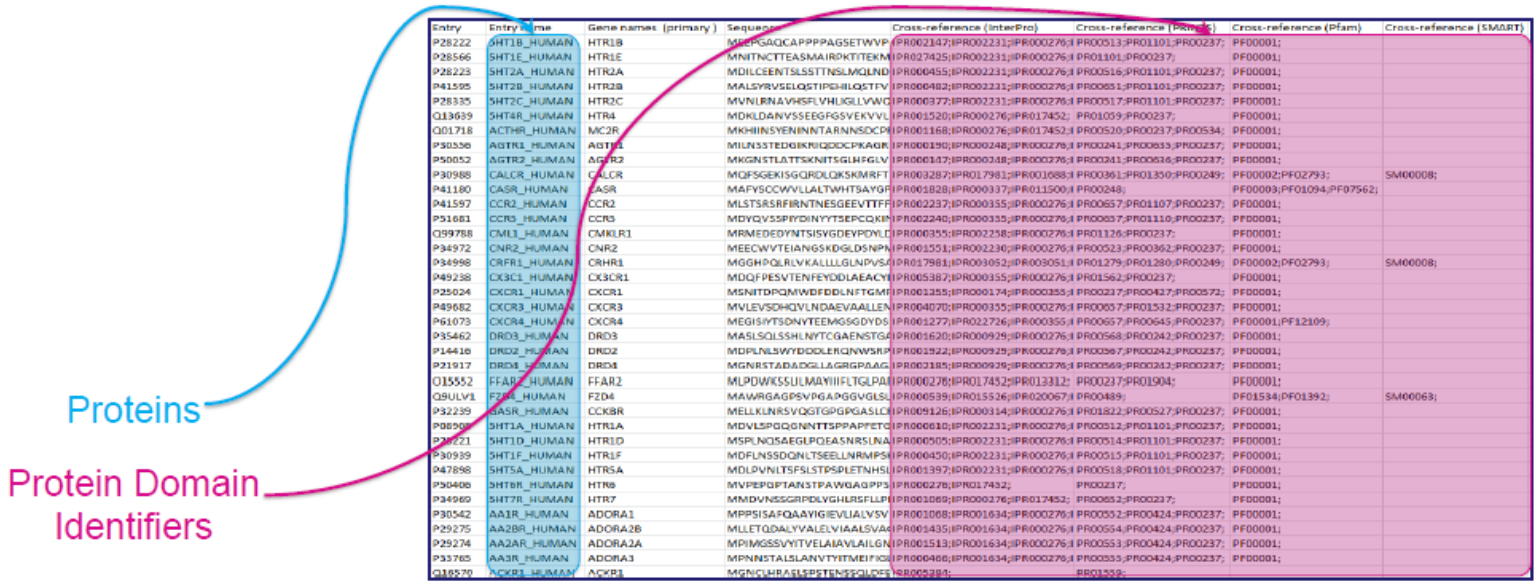
Universal Protein Resource (UniProt)

- UniProt (The Uniprot Consortium, Nucleic Acids Research, 2014) is a database containing diverse types of information about proteins
- UniProt hosts protein domain annotations from other resources, such as Pfam (Punta, Nucleic Acids Research, 2014) and Simple Modular Architecture Tool (SMART) (Letunic, 2012, Nucleic Acids Research)

The UniProt database can be mined for protein domain annotations and many other curated structural and functional annotations of proteins

Protein information

Universal Protein Resource (UniProt)

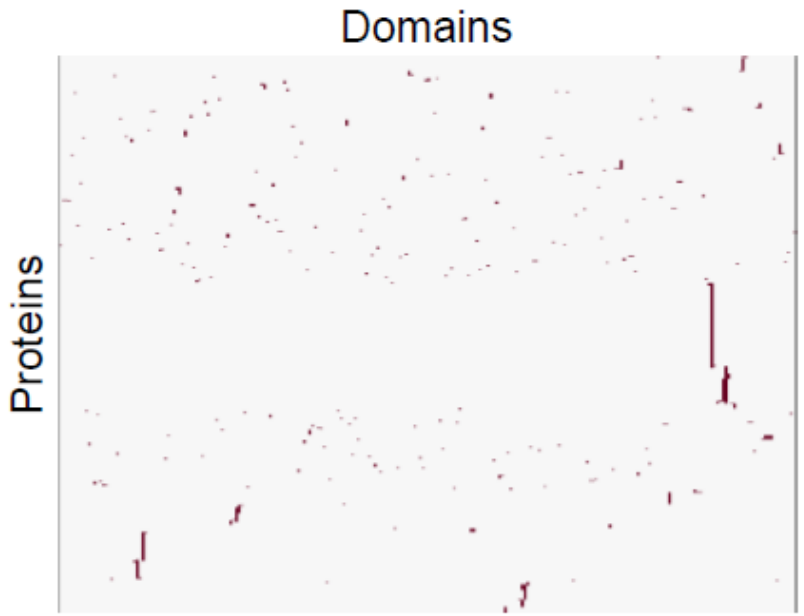


Entry	Entry name	Gene names (primary)	Sequence	Cross-reference (InterPro)	Cross-reference (PRINTS)	Cross-reference (Pfam)	Cross-reference (SMART)
P28222	5HT1B_HUMAN	HTR1B	MLEPGAQCAPPPPAGSETWVP	IPR002147;IPR002231;IPR000276;	PR00513;PR01101;PR00237;	PF00001;	
P28566	5HT1E_HUMAN	HTR1E	MNITNCTTEASMAIRDKTITEKM	IPR027425;IPR002231;IPR000276;	PR01101;PR00237;	PF00001;	
P28223	5HT2A_HUMAN	HTR2A	MDILCEENTSLSSTTNSLMQLND	IPR000455;IPR002231;IPR000276;	PR00516;PR01101;PR00237;	PF00001;	
P41595	5HT2B_HUMAN	HTR2B	MALSYRVSELQSTIPEHILQSTFV	IPR000482;IPR002231;IPR000276;	PR00651;PR01101;PR00237;	PF00001;	
P28335	5HT2C_HUMAN	HTR2C	MVNLRNAVHSFLVHLIGLLVWQ	IPR000377;IPR002231;IPR000276;	PR00517;PR01101;PR00237;	PF00001;	
Q13639	5HT4R_HUMAN	HTR4	MDKLDANVSSEEGFGSVEKVVL	IPR001520;IPR000276;IPR017452;	PR01059;PR00237;	PF00001;	
Q01718	ACTHR_HUMAN	MC2R	MKHIINSYENENNTARNNSDCPR	IPR001168;IPR000276;IPR017452;	PR00520;PR00237;PR00534;	PF00001;	
P30556	AGTR1_HUMAN	AGTR1	MILNSSTEDGIKRIQDDCPKAGR	IPR000190;IPR000248;IPR000276;	PR00241;PR00633;PR00237;	PF00001;	
P50052	AGTR2_HUMAN	AGTR2	MKGNSTLATTSKNITSGLIHFGLV	IPR000147;IPR000248;IPR000276;	PR00241;PR00636;PR00237;	PF00001;	
P30988	CALCR_HUMAN	CALCR	MQFSGEKISGCRDLQKSKMRFT	IPR003287;IPR017981;IPR001688;	PR00361;PR01350;PR00249;	PF00002;PF02793;	SM00008;
P41180	CASR_HUMAN	CASR	MAFVSCCWVLLALTWHTSAYGP	IPR001828;IPR000337;IPR011500;	PR00248;	PF00003;PF01094;PF07562;	
P41597	CCR2_HUMAN	CCR2	MLSTSRSRFIRNTNESGEEVTTFF	IPR002237;IPR000355;IPR000276;	PR00657;PR01107;PR00237;	PF00001;	
P51681	CCR3_HUMAN	CCR3	MDYQVSSPIYDINYYTSEPCQKIN	IPR002240;IPR000355;IPR000276;	PR00657;PR01110;PR00237;	PF00001;	
Q99788	CML1_HUMAN	CMKLR1	MRMEDEDYNTSISVGDEVPDYLD	IPR000355;IPR002258;IPR000276;	PR01126;PR00237;	PF00001;	
P34972	CNR2_HUMAN	CNR2	MEECWVTEIANGSKDGLDSNPM	IPR001551;IPR002230;IPR000276;	PR00523;PR00362;PR00237;	PF00001;	
P34998	CRFR1_HUMAN	CRHR1	MGGHPQLRLVKALLLLGLNPVSA	IPR017981;IPR003052;IPR003051;	PR01279;PR01280;PR00249;	PF00002;PF02793;	SM00008;
P49238	CX3C1_HUMAN	CX3CR1	MDQFPESVTENFEYDDLAEACYI	IPR005387;IPR000355;IPR000276;	PR01562;PR00237;	PF00001;	
P25024	CXCR1_HUMAN	CXCR1	MSNITDPQMWDFDDLNFTGMP	IPR001235;IPR000174;IPR000355;	PR00237;PR00427;PR00572;	PF00001;	
P49682	CXCR3_HUMAN	CXCR3	MVLEVSDHQVLNDAEVAALLEN	IPR004070;IPR000355;IPR000276;	PR00657;PR01532;PR00237;	PF00001;	
P61073	CXCR4_HUMAN	CXCR4	MEGISIYTSDNYTEEMGSGDYDS	IPR001277;IPR022726;IPR000355;	PR00657;PR00645;PR00237;	PF00001;PF12109;	
P35462	DRD3_HUMAN	DRD3	MASLSQLSSHLNYTCGAENSTGA	IPR001620;IPR000929;IPR000276;	PR00568;PR00242;PR00237;	PF00001;	
P14416	DRD2_HUMAN	DRD2	MDPLNLSWYDDDLERQNWSRP	IPR001922;IPR000929;IPR000276;	PR00567;PR00242;PR00237;	PF00001;	
P21917	DRD4_HUMAN	DRD4	MGNRSTADADGLLAGRGPAAG	IPR003185;IPR000929;IPR000276;	PR00569;PR00243;PR00237;	PF00001;	
O15552	FFAR2_HUMAN	FFAR2	MLPDWKSSLNLSAWYIIIFLTGLPAN	IPR000276;IPR017452;IPR013312;	PR00237;PR01904;	PF00001;	
Q9ULV1	FZD4_HUMAN	FZD4	MAWRGAGPSVPGAPGGVGLSL	IPR000539;IPR015526;IPR020067;	PR00489;	PF01534;PF01392;	SM00063;
P32239	GASR_HUMAN	CCK8R	MELLKLNRSVQGTGPGPGASLCH	IPR009126;IPR000314;IPR000276;	PR01822;PR00527;PR00237;	PF00001;	
P08908	5HT1A_HUMAN	HTR1A	MDVLSPGQGNNTTSPPAPFETG	IPR000610;IPR002231;IPR000276;	PR00512;PR01101;PR00237;	PF00001;	
P28221	5HT1D_HUMAN	HTR1D	MSPLNQSAEGLPQEASNRSLNA	IPR000505;IPR002231;IPR000276;	PR00514;PR01101;PR00237;	PF00001;	
P30939	5HT1F_HUMAN	HTR1F	MDFLNSSDQNLTSEELLNRMPSH	IPR000450;IPR002231;IPR000276;	PR00515;PR01101;PR00237;	PF00001;	
P47898	5HT5A_HUMAN	HTR5A	MDLDVNLTSFSLSTPSPLETNHSL	IPR001397;IPR002231;IPR000276;	PR00518;PR01101;PR00237;	PF00001;	
P50406	5HT6R_HUMAN	HTR6	MVPEPGPTANSTPAWGAQPPS	IPR000278;IPR017432;	PR00237;	PF00001;	
P34969	5HT7R_HUMAN	HTR7	MMDVNSSGRPDLVGHLRSFLLPR	IPR001089;IPR000276;IPR017452;	PR00652;PR00237;	PF00001;	
P30542	AA1R_HUMAN	ADORA1	MPPSISAFQAAYIGIEVLIALVSV	IPR001068;IPR001634;IPR000276;	PR00552;PR00424;PR00237;	PF00001;	
P29275	AA2BR_HUMAN	ADORA2B	MLLETQDALYVALELVIAALSVAG	IPR001435;IPR001634;IPR000276;	PR00554;PR00424;PR00237;	PF00001;	
P29274	AA2AR_HUMAN	ADORA2A	MPIMGSSVYITVELAIAVLAILGN	IPR001513;IPR001634;IPR000276;	PR00553;PR00424;PR00237;	PF00001;	
P33765	AA3R_HUMAN	ADORA3	MPNNSTALSLANVTYITMEIFIG	IPR000466;IPR001634;IPR000276;	PR00555;PR00424;PR00237;	PF00001;	
Q16570	ACKR3_HUMAN	ACKR3	MGNCLHPALVSPSTSNSSQLDFS	IPR000384;	PR01359;		

UniProt protein domain data are provided as a list of proteins paired with sets of known domains

Protein information

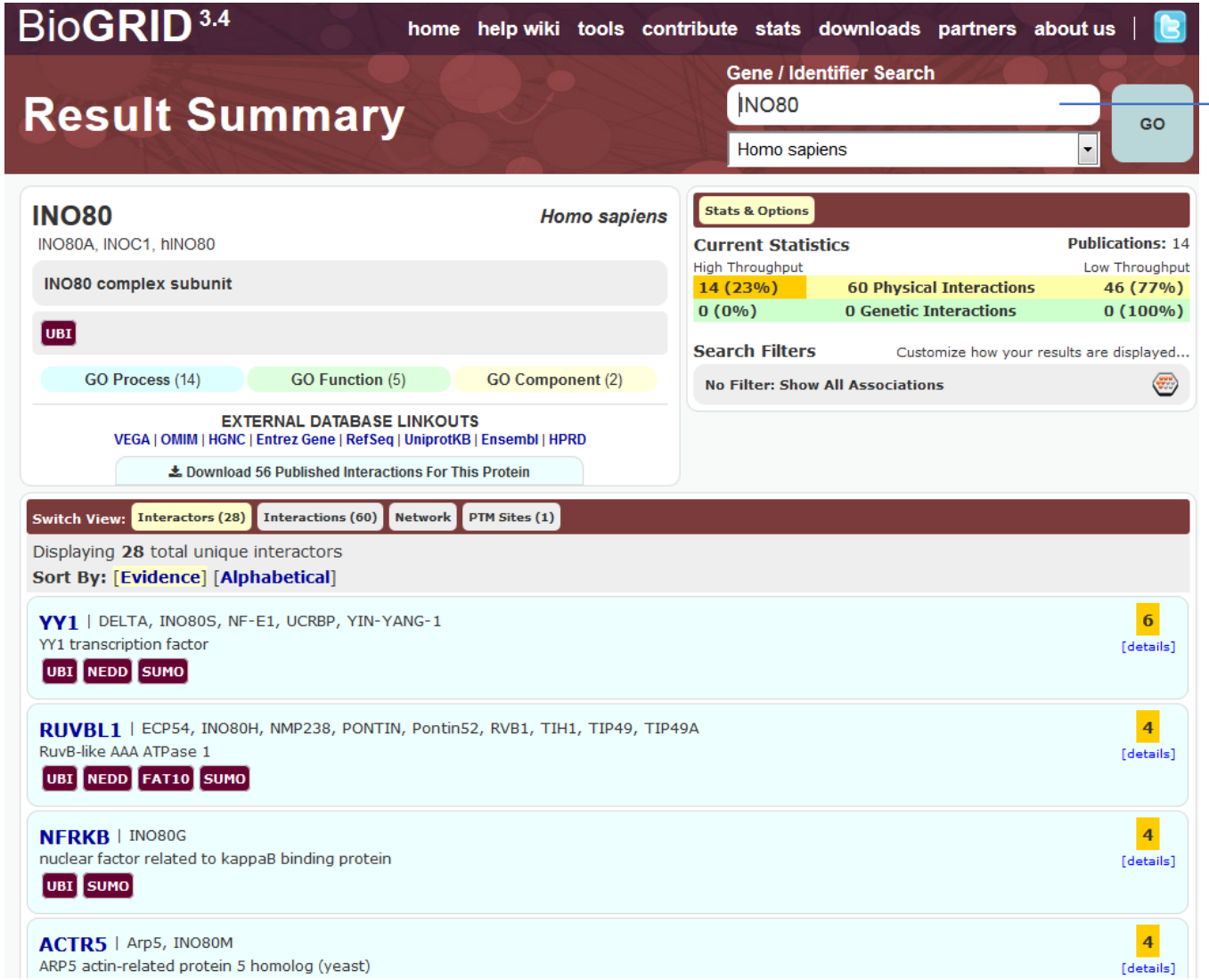
Universal Protein Resource (UniProt)



UniProt protein domain data organized as matrix/graph connecting proteins to protein domains

Protein information

Biological General Repository for Interaction Datasets (BioGRID)



Search for BRMS1L protein interactions

Protein interactions

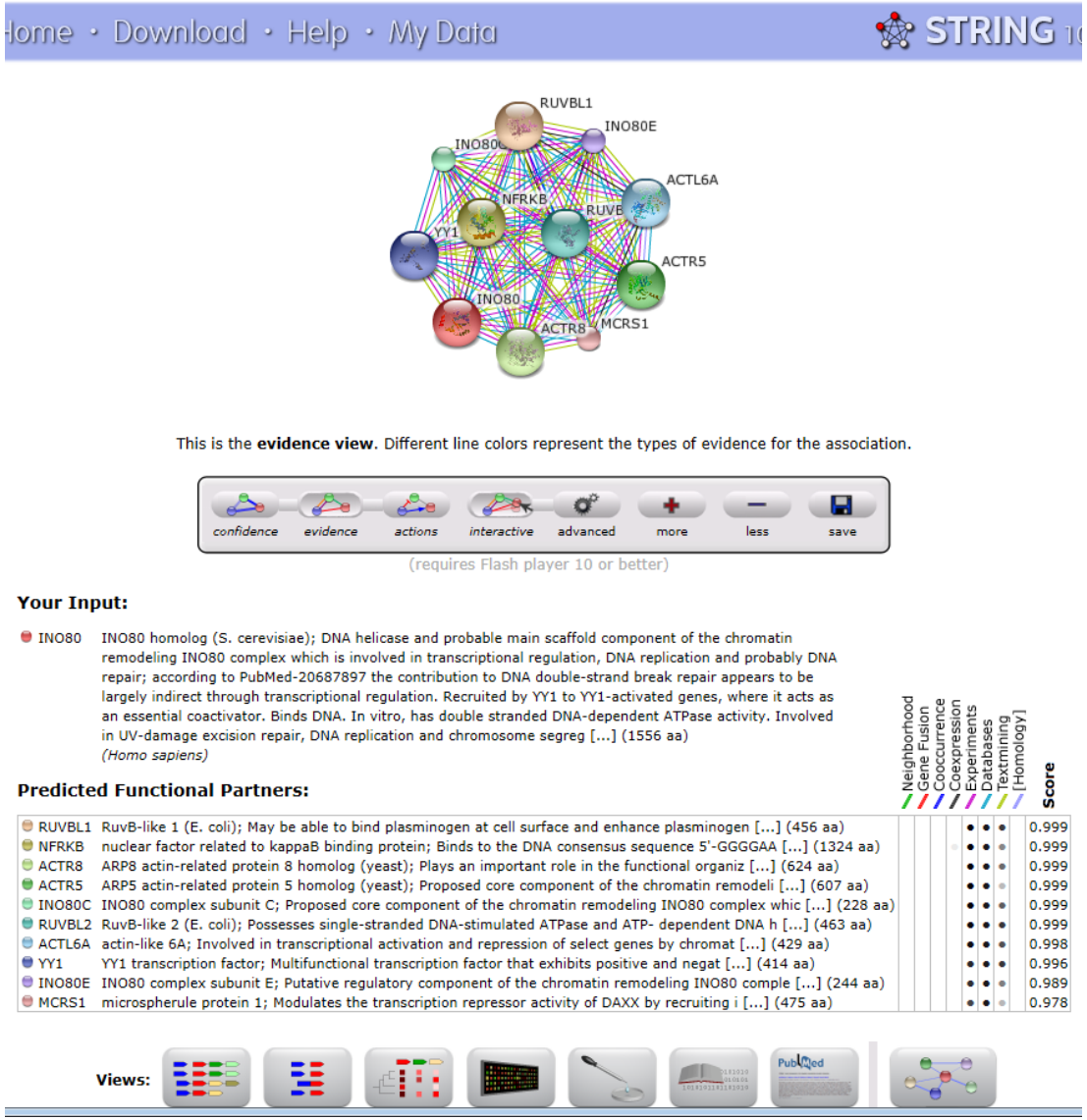
Biogrid.org

Biological General Repository for Interaction Datasets (BioGRID)

BioGRID data are provided as a list of protein-protein interactions with additional metadata.

#BioGRID	Entrez Ge	Entrez Ge	BioGRID II	BioGRID II	Systemati	Systemati	Official Sy	Official Sy	Synonyms	Synonyms	Experime	Experime	Author	Pubmed II	Organism	Organism	Throughp	Score
103	6416	2318	112315	108607	-	-	MAP2K4	FLNC	SAPKK1|JI	ABPL|MPI	Two-hybri	physical	Marti A (1	9006895	9606	9606	Low Throu	-
117	84665	88	124185	106603	-	-	MYPN	ACTN2	MYOP	CMD1AA	Two-hybri	physical	Bang ML (:	11309420	9606	9606	Low Throu	-
183	90	2339	106605	108625	-	-	ACVR1	FNTA	ACVRLK2|	PTAR2|FP	Two-hybri	physical	Wang T (1	8599089	9606	9606	Low Throu	-
278	2624	5371	108894	111384	-	-	GATA2	PML	NFE1B|DC	RNF71|TR	Two-hybri	physical	Tsuzuki S (	10938104	9606	9606	Low Throu	-
418	6118	6774	112038	112651	RP4-547C9	-	RPA2	STAT3	REPA2|RP	APRF|HIE	Two-hybri	physical	Kim J (200	10875894	9606	9606	Low Throu	-
586	375	23163	106870	116775	-	-	ARF1	GGA3	-	-	Two-hybri	physical	Dell'Ange	10747089	9606	9606	Low Throu	-
612	377	23647	106872	117174	-	-	ARF3	ARFIP2	-	POR1	Two-hybri	physical	Kanoh H (:	9038142	9606	9606	Low Throu	-
617	377	27236	106872	118084	-	-	ARF3	ARFIP1	-	HSU52521	Two-hybri	physical	Kanoh H (:	9038142	9606	9606	Low Throu	-
663	54464	226	119970	106728	-	-	XRN1	ALDOA	1-Sep	GSD12|AL	Two-hybri	physical	Lehner B (	15231747	9606	9606	High Throu	-
866	351	10513	106848	115769	-	-	APP	APPBP2	PN-II|ABF	HS.84084|	Two-hybri	physical	Zheng P (1	9843960	9606	9606	Low Throu	-
917	333	1600	106830	107970	-	RP6-239D2	APLP1	DAB1	APLP	-	Two-hybri	physical	Homayoui	10460257	9606	9606	Low Throu	-
1156	10370	7020	115649	112878	-	RP1-290I1	CITED2	TFAP2A	P35SRJ|M	BOFS|AP2	Two-hybri	physical	Braganca J	12586840	9606	9606	Low Throu	-
1161	2033	7020	108347	112878	RP1-85F18	RP1-290I1	EP300	TFAP2A	RSTS2|p3(	BOFS|AP2	Two-hybri	physical	Braganca J	12586840	9606	9606	Low Throu	-
1206	338	4547	106835	110641	-	-	APOB	MTTP	FLDB|LDL(	ABL|MTP	Two-hybri	physical	Bradbury I	9915855	9606	9606	Low Throu	-
1427	409	5900	106902	111836	-	RP11-326L	ARRB2	RALGDS	BARR2|AF	RalGEF|R(	Two-hybri	physical	Bhattacha	12105416	9606	9606	Low Throu	-
1611	1436	2885	107823	109142	-	-	CSF1R	GRB2	FMS|FIM2	EGFRBP-G	Two-hybri	physical	Mancini A	9380408	9606	9606	Low Throu	-
1695	7916	2885	113646	109142	DADB-70P	-	PRRC2A	GRB2	D6S51E|B/	EGFRBP-G	Two-hybri	physical	Lehner B (	14667819	9606	9606	Low Throu	-
1783	27257	4677	118104	110758	-	-	LSM1	NARS	CASM|YJL	NARS1|AS	Two-hybri	physical	Lehner B (	15231747	9606	9606	High Throu	-
1888	6521	22950	112412	116605	-	HLC3	SLC4A1	SLC4A1AP	BND3|WD	-	Two-hybri	physical	Chen J (19	9422766	9606	9606	Low Throu	-
1958	602	580	107074	107056	-	-	BCL3	BARD1	BCL4|D19	-	Two-hybri	physical	Dechend F	10362352	9606	9606	Low Throu	-
2006	153	10755	106662	115978	-	-	ADRB1	GIPC1	B1AR|ADF	NIP|SEMC	Two-hybri	physical	Hu LA (200	12724327	9606	9606	Low Throu	-
2368	672	466	107140	106956	-	-	BRCA1	ATF1	PPP1R53|	FUS/ATF-1	Two-hybri	physical	Houvras Y	10945975	9606	9606	Low Throu	-
2398	672	4436	107140	110573	-	-	BRCA1	MSH2	PPP1R53|	HNPCC1|H	Two-hybri	physical	Wang Q (2	11498787	9606	9606	Low Throu	-
2411	672	580	107140	107056	-	-	BRCA1	BARD1	PPP1R53|	-	Two-hybri	physical	Wu LC (19	8944023	9606	9606	Low Throu	-
2424	672	2956	107140	109211	-	-	BRCA1	MSH6	PPP1R53|	HNPCC5|(	Two-hybri	physical	Wang Q (2	11498787	9606	9606	Low Throu	-
2466	421	1013	106914	107448	-	-	ARVCF	CDH15	-	CDH3|MR	Two-hybri	physical	Kaufmann	11058098	9606	9606	Low Throu	-
2721	5092	775	111125	107229	-	-	PCBD1	CACNA1C	PCBD|DC(	CACNL1A1	Two-hybri	physical	Waters PJ	11461190	9606	9606	Low Throu	-
2765	5664	823	111643	107273	-	PIG30	PSEN2	CAPN1	AD3L|AD4	CANPL1|C	Two-hybri	physical	Shinozaki	9852298	9606	9606	Low Throu	-
2785	825	7273	107275	113124	-	-	CAPN3	TTN	nCL-1|p94	CMH9|MY	Two-hybri	physical	Ono Y (199	9642272	9606	9606	Low Throu	-
2827	3708	767	109913	107222	-	-	ITPR1	CA8	Insp3r1|S(	CALS|CA-1	Two-hybri	physical	Hirota J (2	12611586	9606	9606	Low Throu	-
3024	9223	1499	114655	107880	-	OK/SW-cl.	MAGI1	CTNNB1	AIP-3|MA	CTNNB	Two-hybri	physical	Dobrosots	10772923	9606	9606	Low Throu	-
3189	5925	1523	111860	107903	RP11-174I	-	RB1	CUX1	pp110|OS	p75|Clox	Two-hybri	physical	Gupta S (2	12891711	9606	9606	Low Throu	-
3224	7251	1026	113102	107460	-	-	TSG101	CDKN1A	VPS23|TS(	MDA-6|CA	Two-hybri	physical	Oh H (200	11943869	9606	9606	Low Throu	-
3312	4998	4171	111040	110339	-	-	ORC1	MCM2	ORC1L|PA	MITOTIN|	Two-hybri	physical	Kneissl M	12614612	9606	9606	High Throu	-
3313	5000	4171	111042	110339	-	-	ORC4	MCM2	ORC4L|OF	MITOTIN|	Two-hybri	physical	Kneissl M	12614612	9606	9606	High Throu	-
3315	4174	4171	110342	110339	RP5-824I1	-	MCM5	MCM2	P1-CDC46	MITOTIN|	Two-hybri	physical	Kneissl M	12614612	9606	9606	High Throu	-
3317	8317	4171	113914	110339	-	-	CDC7	MCM2	Hsk1|huC	MITOTIN|	Two-hybri	physical	Kneissl M	12614612	9606	9606	High Throu	-

Protein interactions



Search for BRMS1L protein interactions

Protein interactions

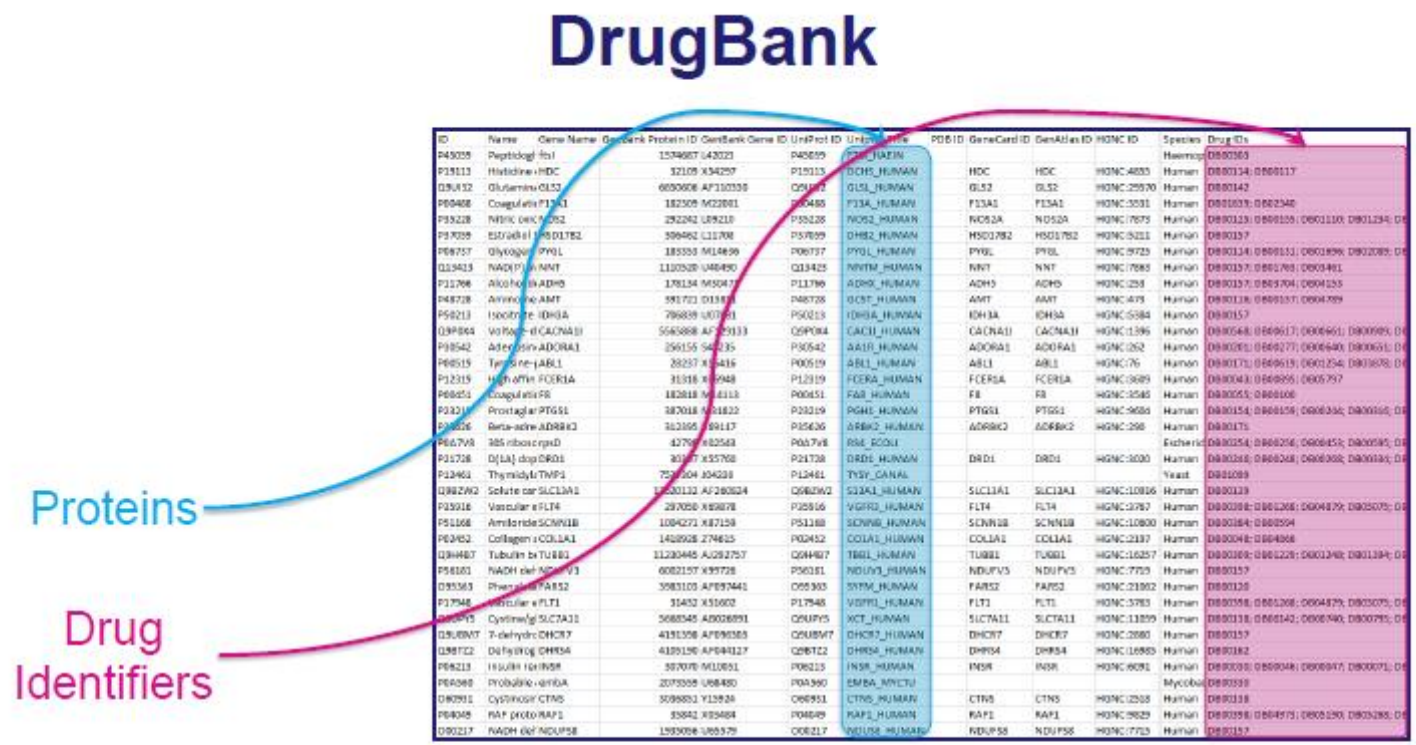
http://string-db.org/

DrugBank

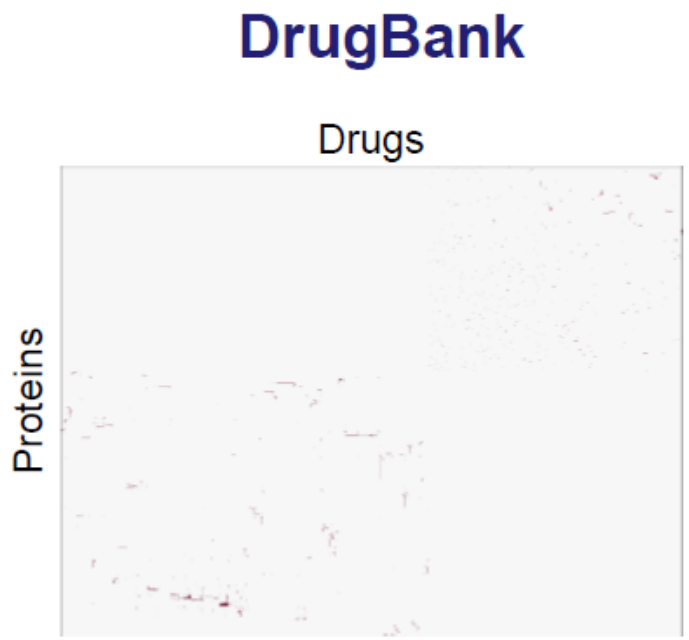
- DrugBank is a manually curated database of information about drugs and their targets (Knox, Nucleic Acids Research, 2011)
- Drug information: chemical composition, molecular structure, pharmacokinetic and pharmacodynamic properties, indications, mechanism of action…
- Target information: sequence, reactions, pathways, protein domain function, cellular location…

The DrugBank database can be mined for known targets of drugs and many other manually curated structural and functional annotations of drugs

Drug information



Protein-Drug relationship

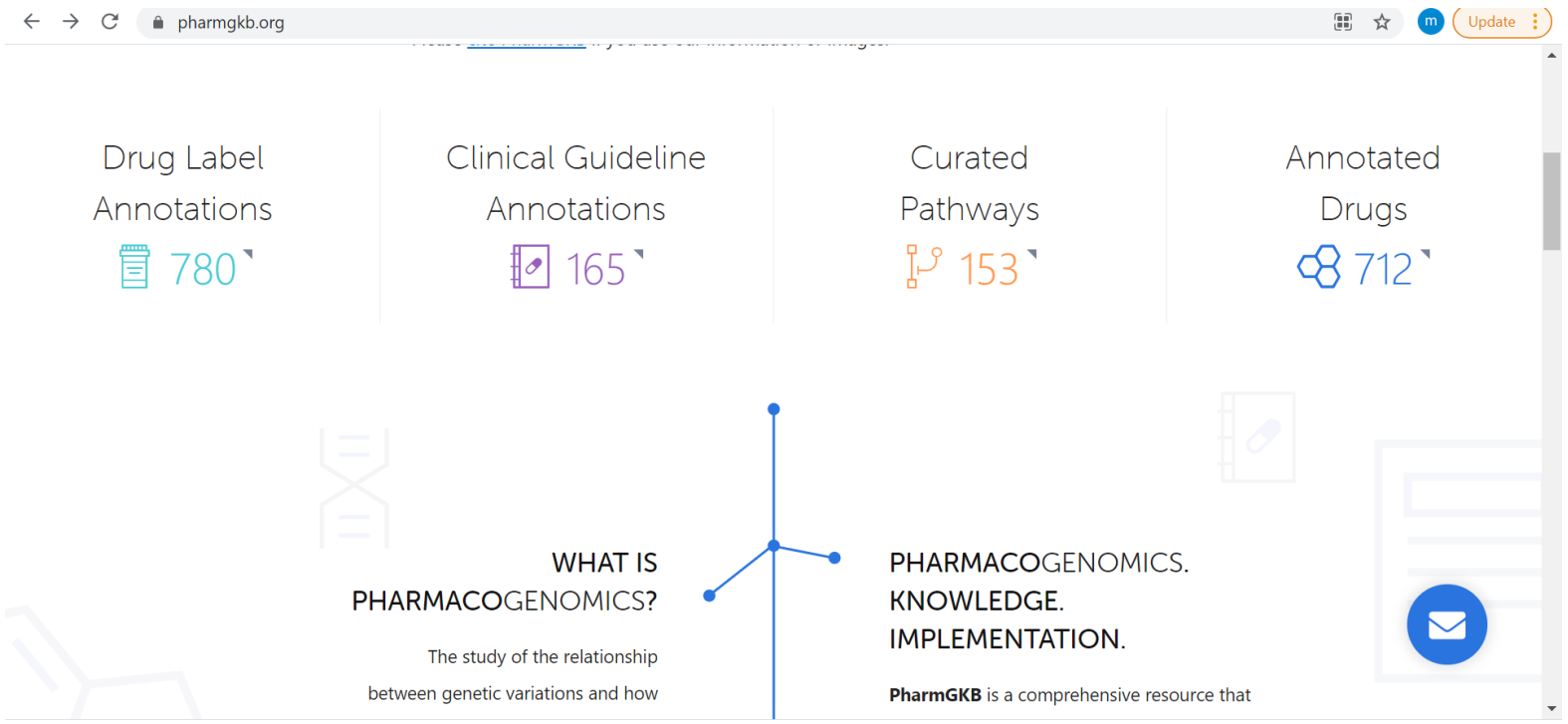


DrugBank data organized as matrix/graph connecting proteins to drugs known to target the proteins

Protein-Drug relationship

- Search for Acetazolamide

Clinical database

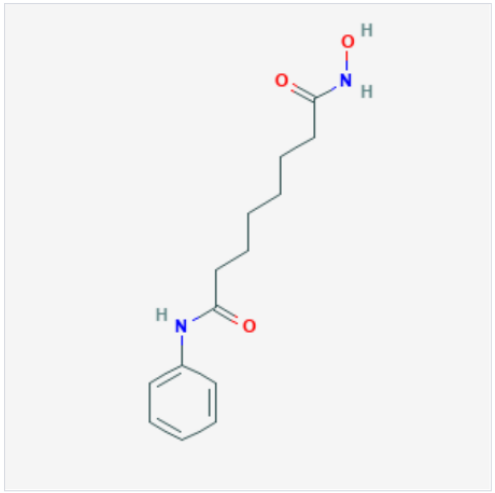


https://www.pharmgkb.org/

pharmgkb.org/chemical/PA164748224

- Overview
- Prescribing Info
- Drug Label Annotations
- Clinical Annotations
- Variant Annotations
- Literature
- Pathways

Structure



large version

3D version

source: PubChem

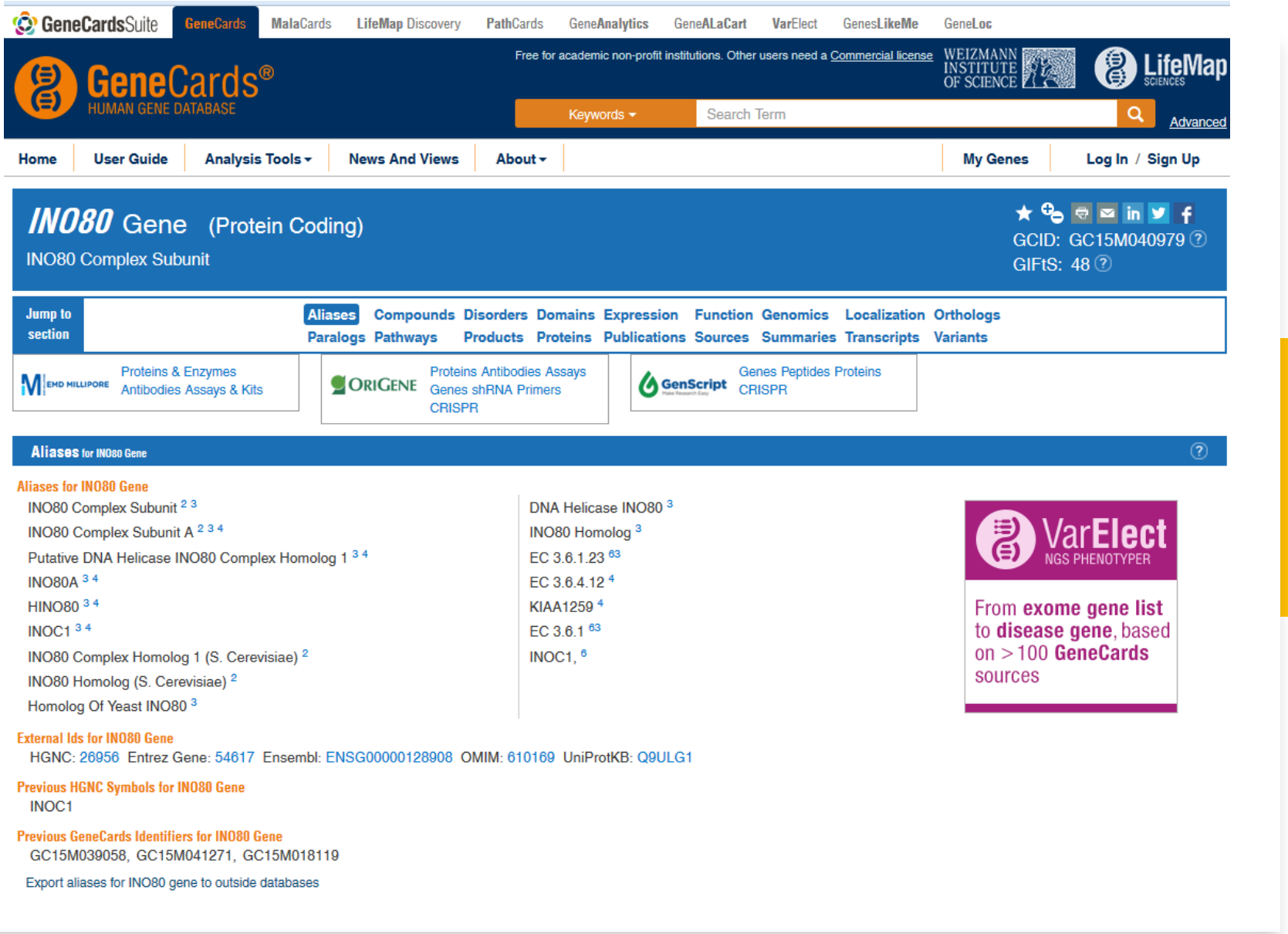
Type : Drug

Synonyms

MK0683, N-hydroxy-n'-phenyloctanediamide, N-hyrdroxy-n'-phenyloctanediamide, SAHA, SHH, Suberanilohydroxamic acid, suberoylanilide hydroxamic acid, vorinostat, Zolinza

PharmGKB ID : PA164748224

SEARCH FOR SAHA



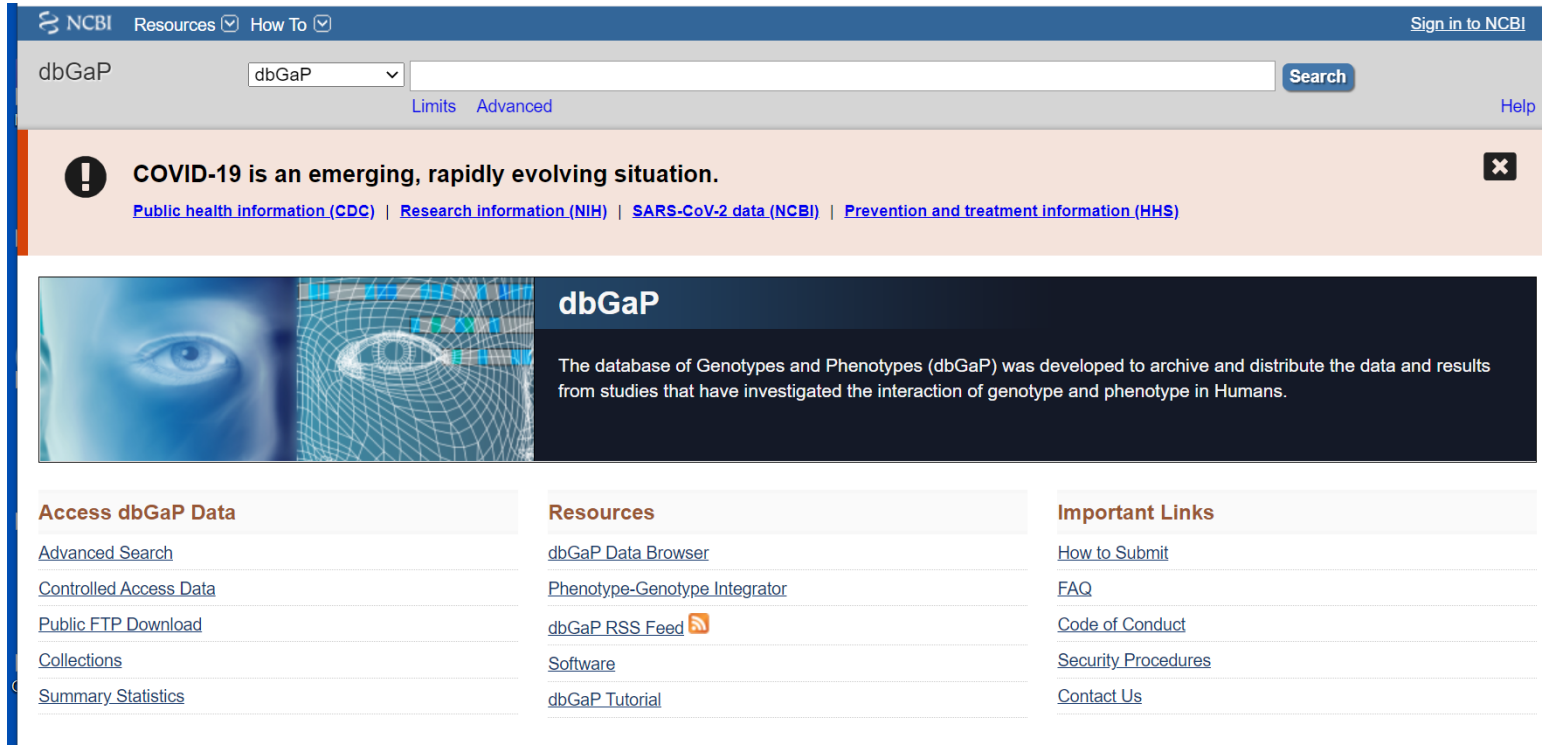
Gene/Protein catalog

- Search and learn about WDR76 gene

Genecards.org

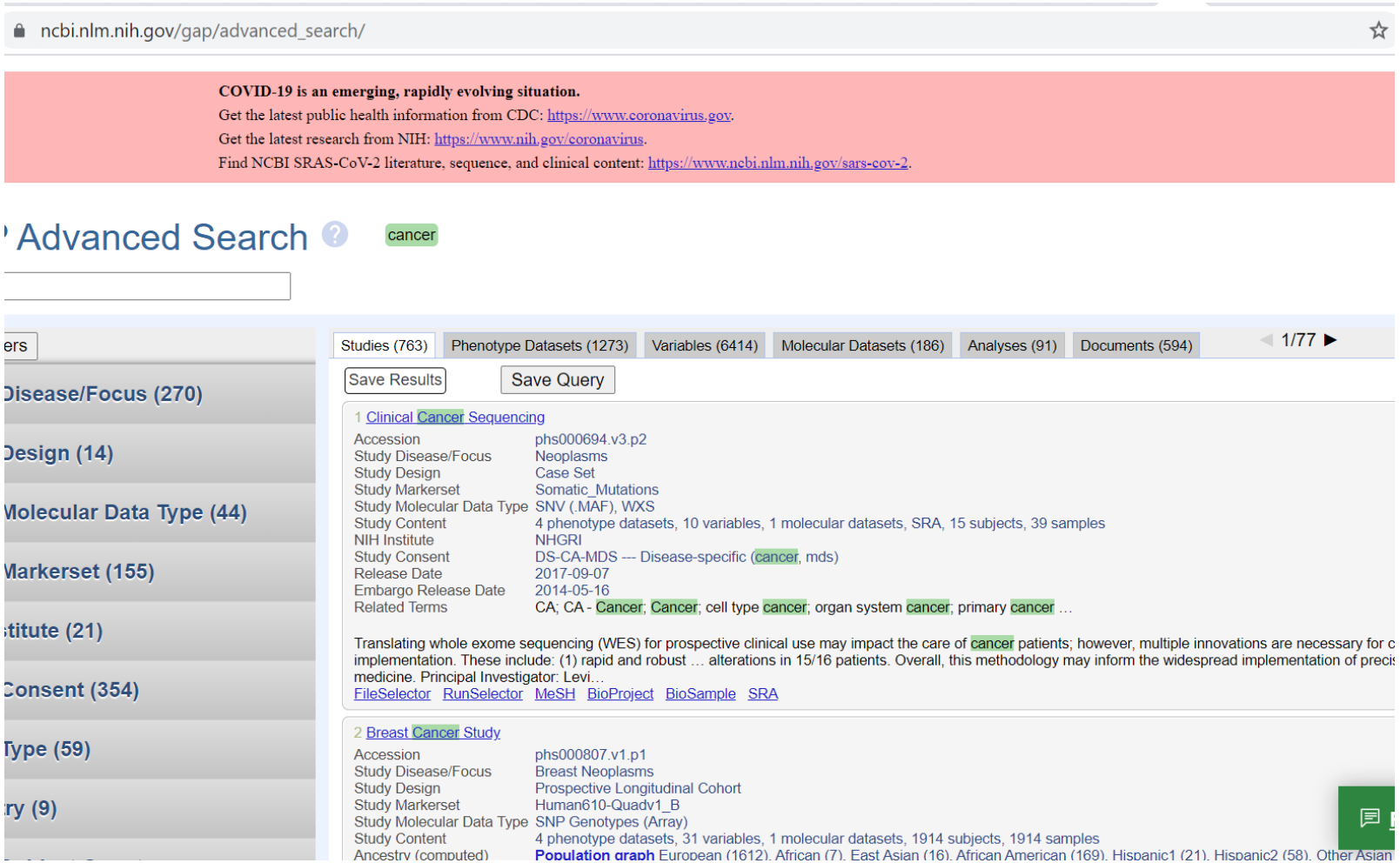
- https://www.ncbi.nlm.nih.gov/geo/info/datasets.html
- You can analyze and visualize existing gene expression data

GEO (Gene Expression Omnibus)



Genotype and Phenotype in Humans

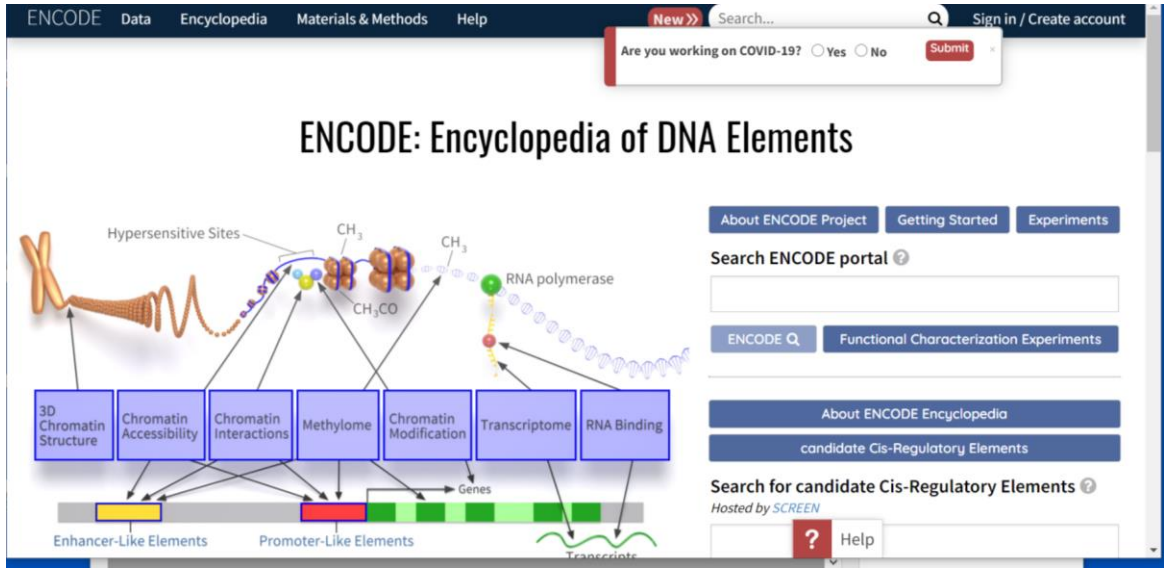
https://www.ncbi.nlm.nih.gov/gap/



Genotype and Phenotype in Humans

Encyclopedia of DNA Elements (ENCODE)

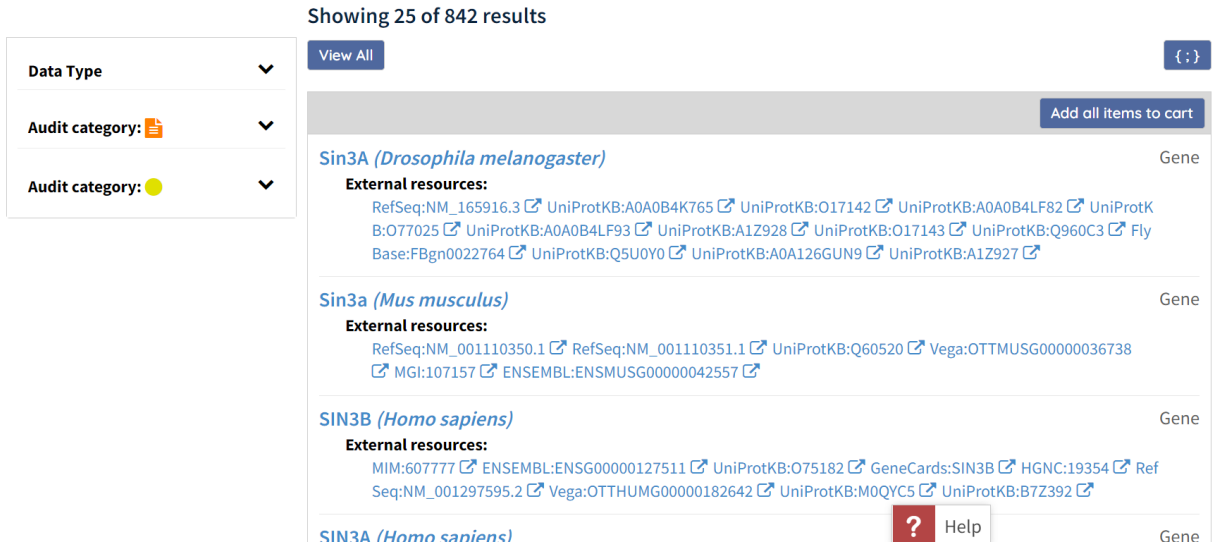
- ENCODE is a genome mapping project that seeks to annotate the human genome with information about genes and elements that regulate gene transcription, such as transcription factor binding sites (ENCODE Consortium, Science, 2004)
- Chip- Enrichment Analysis is a related resource that has accumulated transcription factor binding site data generated separately from ENC

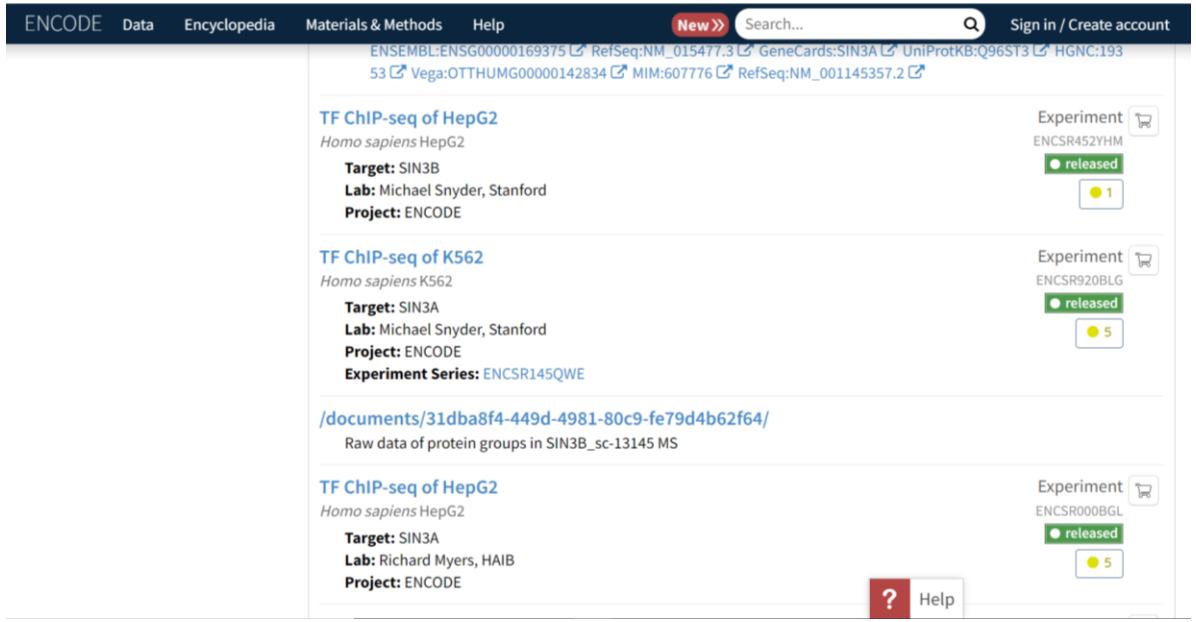


The ENCODE dataset can be mined for putative transcription factors regulating expression of a gene

Search for the SIN3 gene

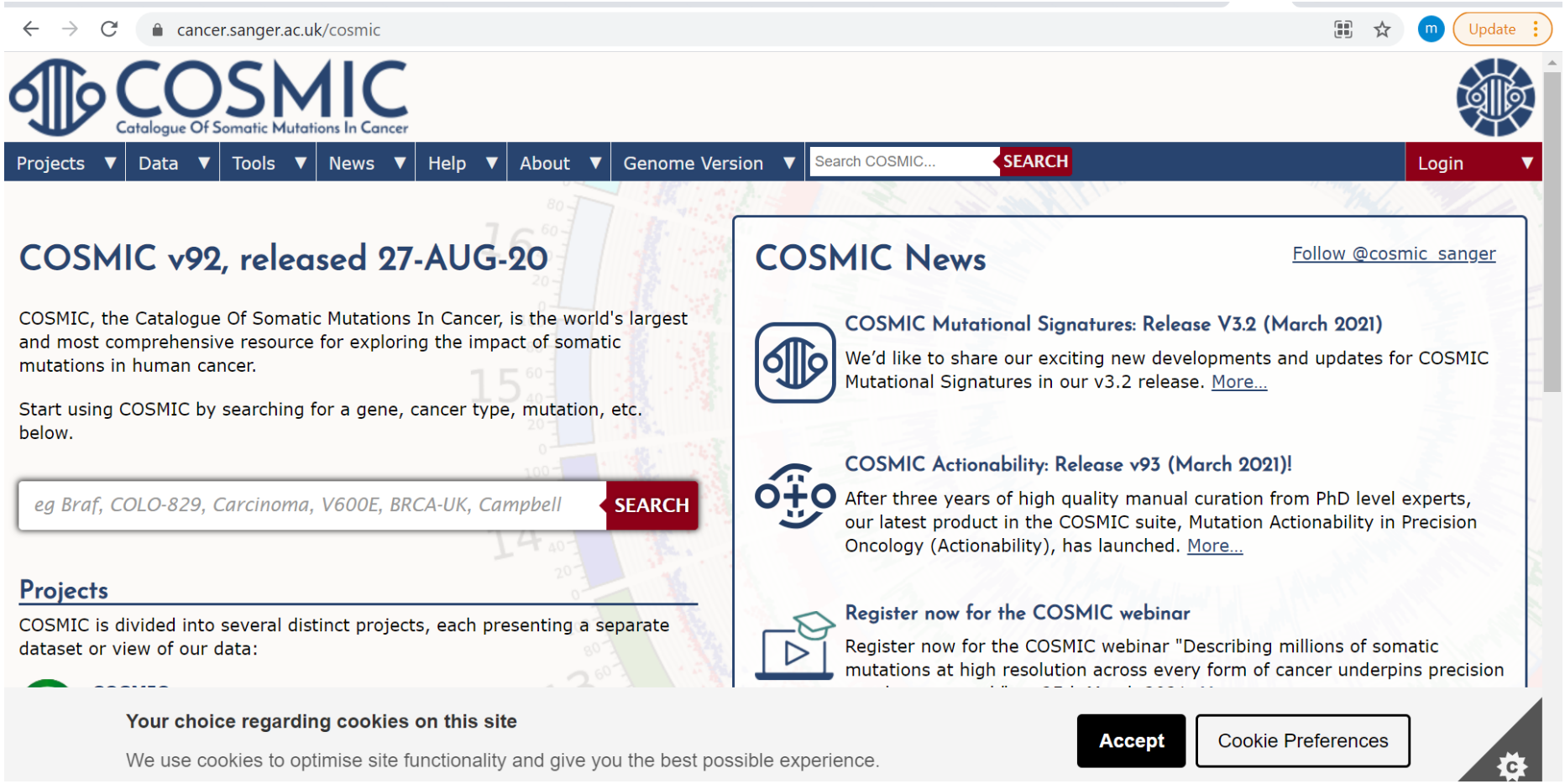
- https://www.encodeproject.org/search/?searchTerm=SIN3

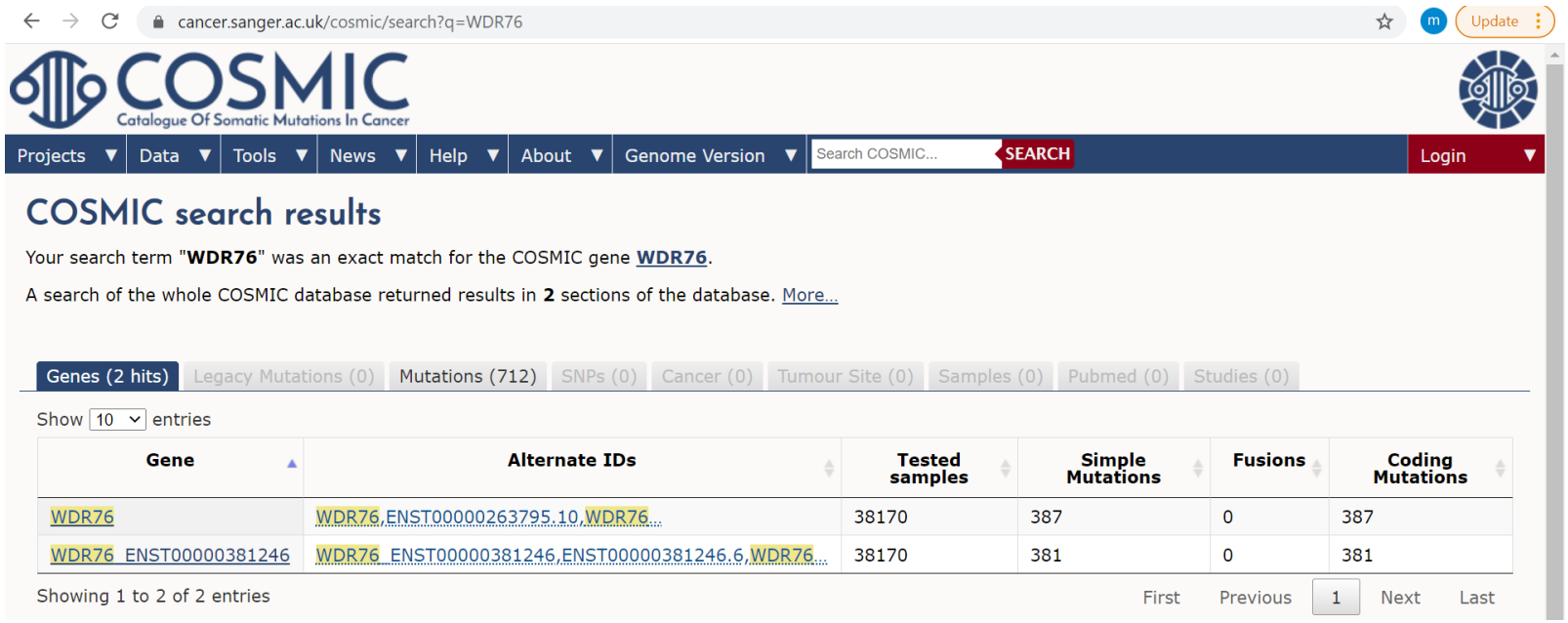




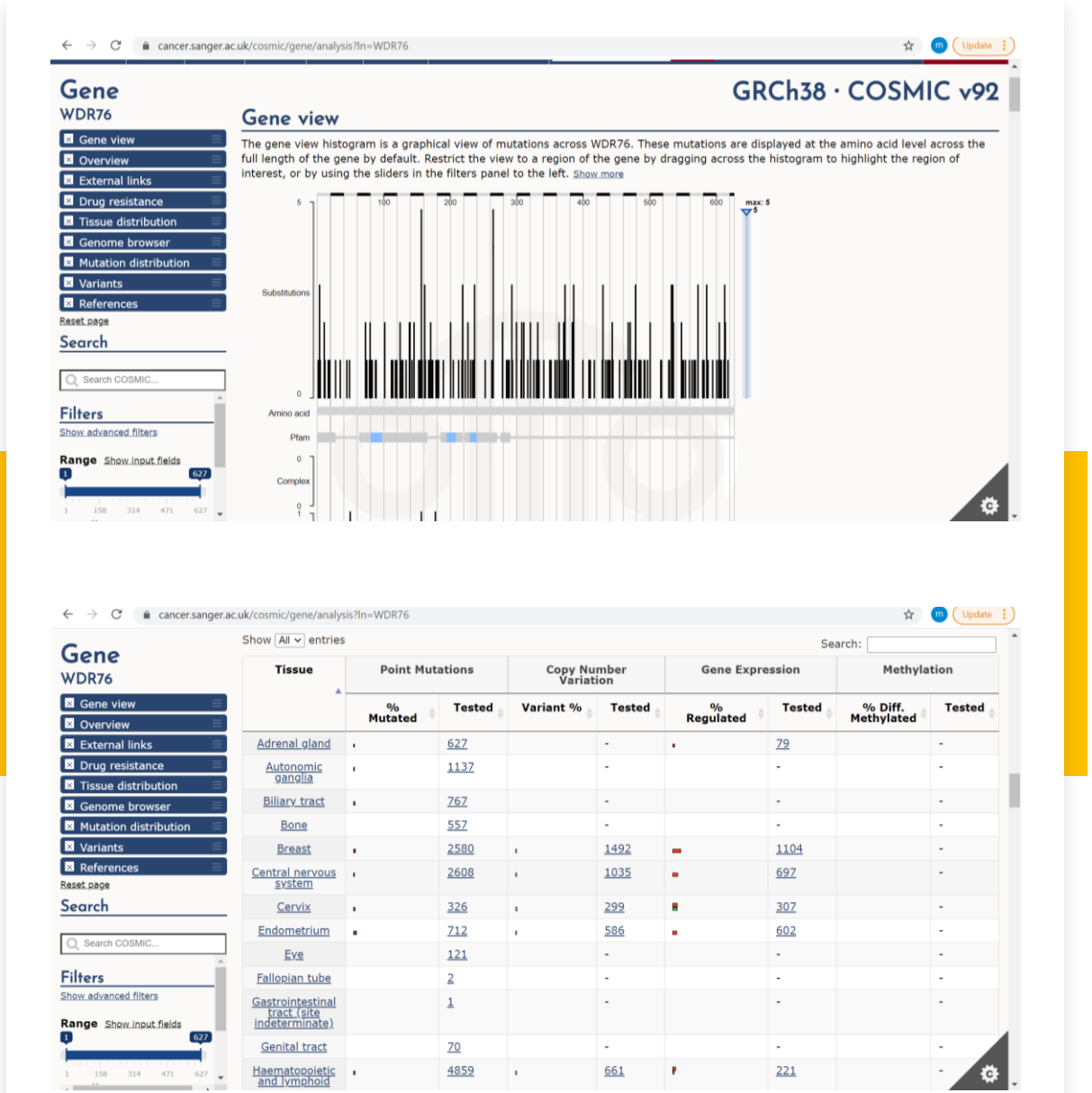
COSMIC (Catalog of Somatic Mutations in Cancer)

- https://cancer.sanger.ac.uk/cosmic



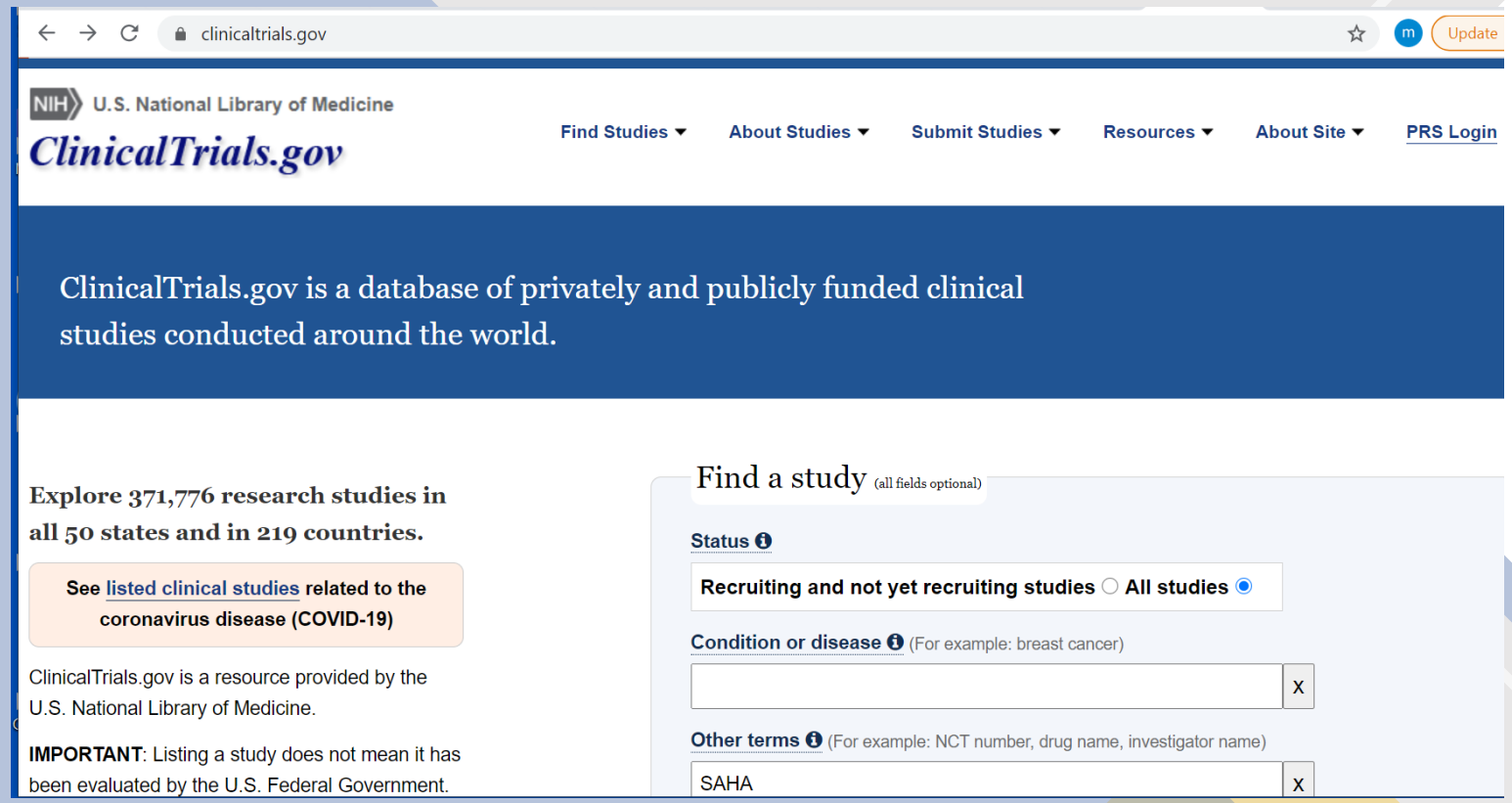


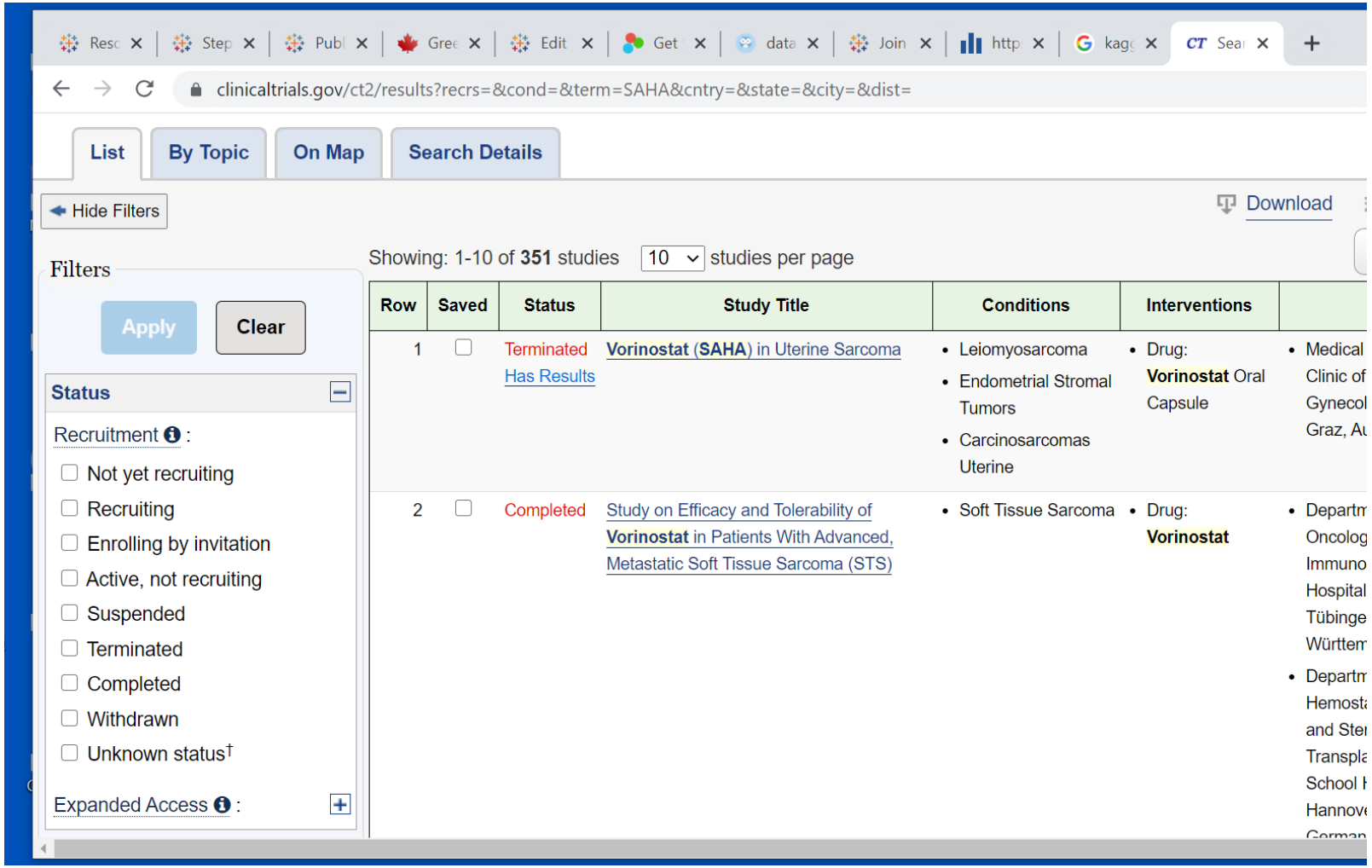
Search for WDR76



Search for WDR76

Clinical Trials





Search for SAHA drug

Table 1.

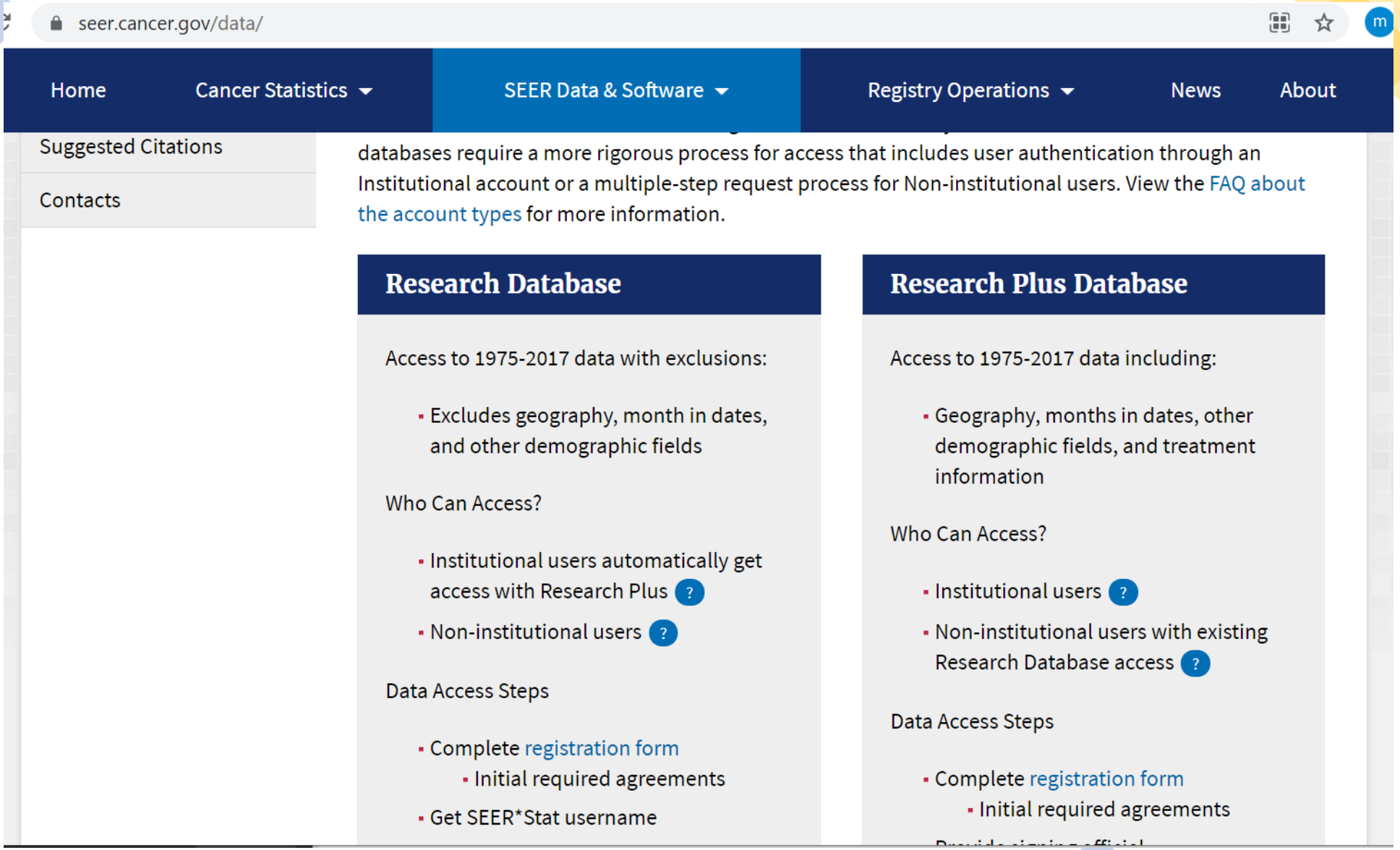
List of multi-omics data repositories.

Data repository	Web link	Disease	Types of multi-omics data available
The Cancer Genome Atlas (TCGA)	https://cancergenome.nih.gov/	Cancer	RNA-Seq, DNA-Seq, miRNA-Seq, SNV, CNV, DNA methylation, and RPPA
Clinical Proteomic Tumor Analysis Consortium (CPTAC)	https://cptac-data-portal.georgetown.edu/cptacPublic/	Cancer	Proteomics data corresponding to TCGA cohorts
International Cancer Genomics Consortium (ICGC)	https://icgc.org/	Cancer	Whole genome sequencing, genomic variations data (somatic and germline mutation)
Cancer Cell Line Encyclopedia (CCLE)	https://portals.broadinstitute.org/ccle	Cancer cell line	Gene expression, copy number, and sequencing data; pharmacological profiles of 24 anticancer drugs
Molecular Taxonomy of Breast Cancer International Consortium (METABRIC)	http://molonc.bccrc.ca/aparicio-lab/research/metabric/	Breast cancer	Clinical traits, gene expression, SNP, and CNV
TARGET	https://ocg.cancer.gov/programs/target	Pediatric cancers	Gene expression, miRNA expression, copy number, and sequencing data
Omics Discovery Index	https://www.omicsdi.org	Consolidated data sets from 11 repositories in a uniform framework	Genomics, transcriptomics, proteomics, and metabolomics

List of multi-omics data repositories.

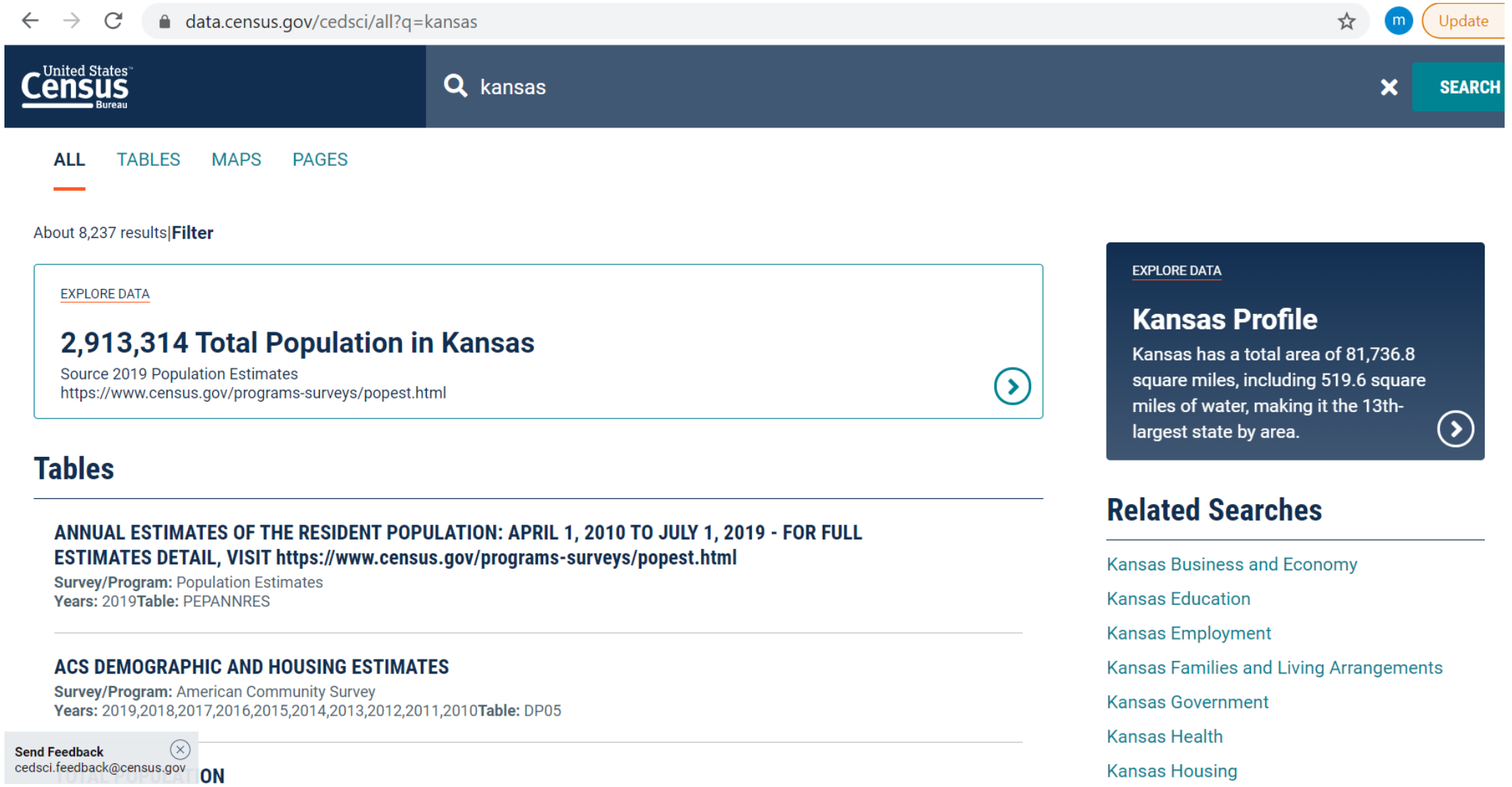
SEER CANCER- Demographic database

- https://seer.cancer.gov/data/

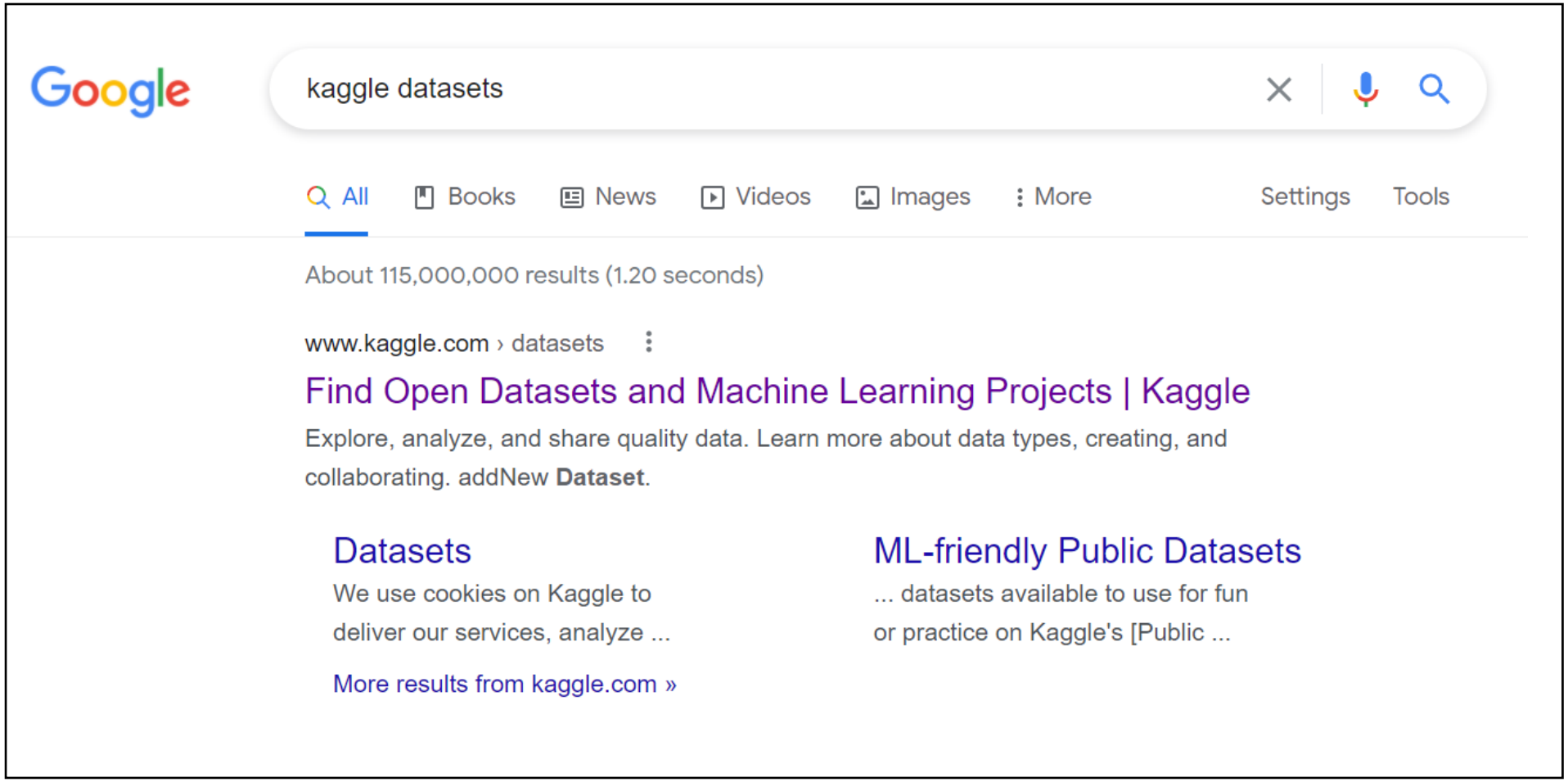

seer.cancer.gov/data/
Home
Cancer Statistics
SEER Data & Software
Registry Operations
News
About
Suggested Citations
Contacts
databases require a more rigorous process for access that includes user authentication through an Institutional account or a multiple-step request process for Non-institutional users. View the FAQ about the account types for more information.
Research Database
Access to 1975-2017 data with exclusions:
Excludes geography, month in dates, and other demographic fields
Who Can Access?
Institutional users automatically get access with Research Plus
Non-institutional users
Data Access Steps
Complete registration form
Initial required agreements
Get SEER*Stat username
Research Plus Database
Access to 1975-2017 data including:
Geography, months in dates, other demographic fields, and treatment information
Who Can Access?
Institutional users
Non-institutional users with existing Research Database access
Data Access Steps
Complete registration form
Initial required agreements

Census database-Demographic data

- https://www.census.gov/data.html



Kaggle databases for Machine Learning



Datasets

Datasets ×

76,684 Datasets

Hottest

Reddit Vaccine Myths
Gabriel Preda · Updated 6 hours ago
Usability **10.0** · 1 File (CSV) · 217 KB · 1 Task

102
Bronze

A Large Scale Fish Dataset
Oğuzhan Ulucan · Updated a month ago
Usability **10.0** · 18006 Files (other) · 3 GB

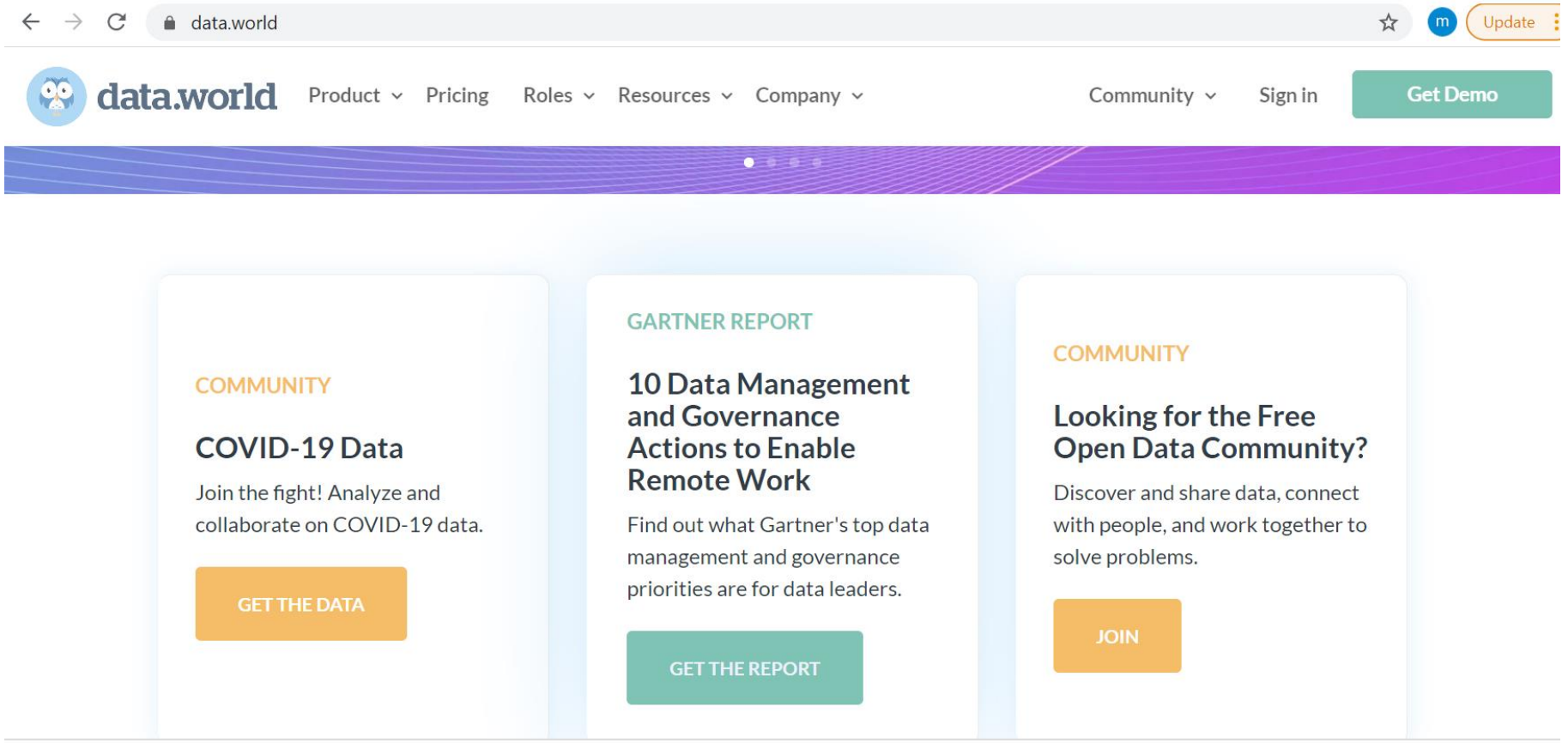
84
Bronze

Wikibooks Dataset
Dhruvil Dave · Updated a month ago
Usability **10.0** · 7 Files (CSV) · 1 GB

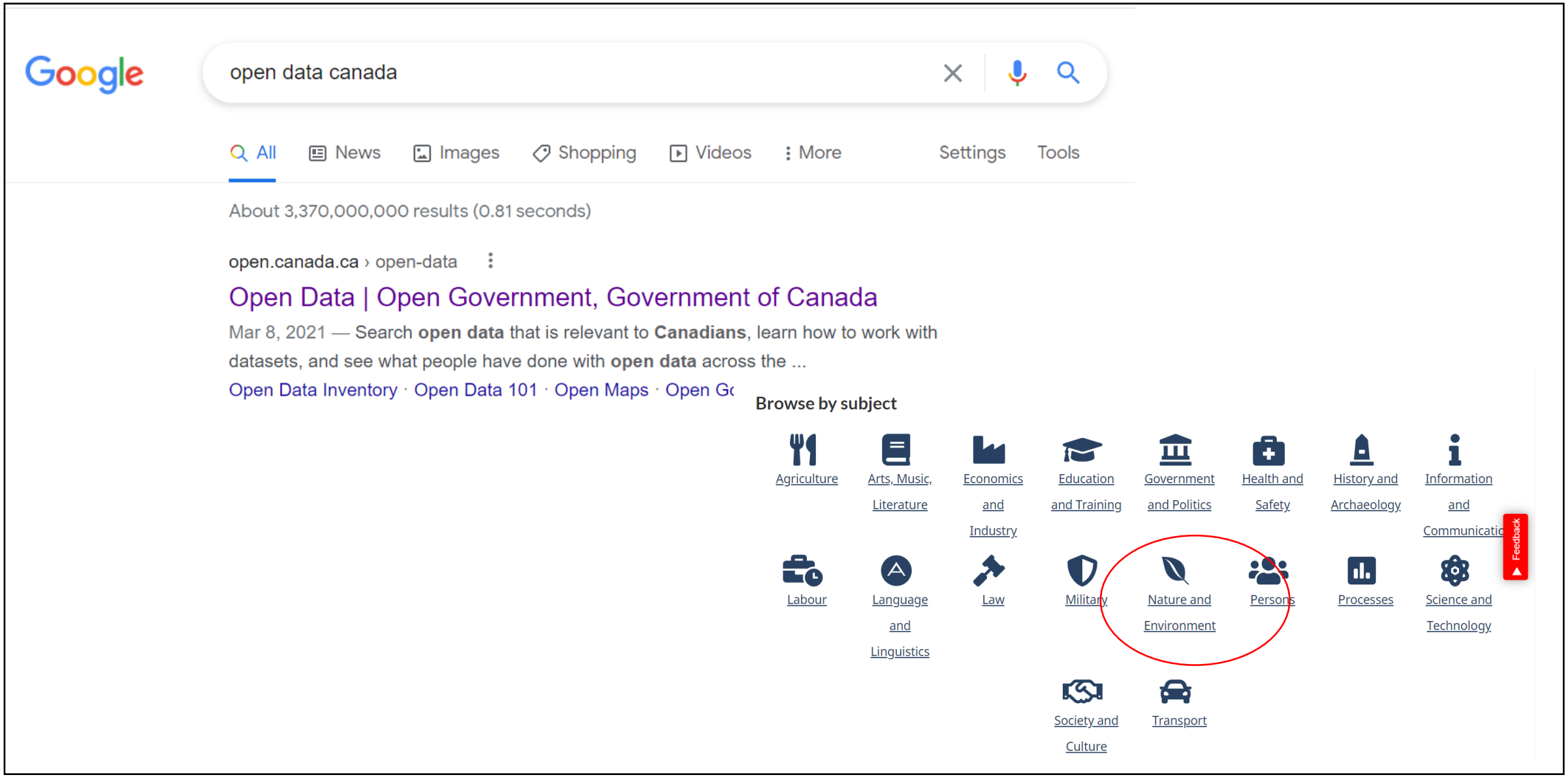
80
Silver

Data.world

- https://data.world/



Open DATA


Google
open data canada
All
News
Images
Shopping
Videos
More
Settings
Tools
About 3,370,000,000 results (0.81 seconds)
open.canada.ca › open-data
Open Data | Open Government, Government of Canada
Mar 8, 2021 — Search open data that is relevant to Canadians, learn how to work with datasets, and see what people have done with open data across the ...
Open Data Inventory · Open Data 101 · Open Maps · Open G
Browse by subject
Agriculture
Arts, Music, Literature
Economics and Industry
Education and Training
Government and Politics
Health and Safety
History and Archaeology
Information and Communicatio
Labour
Language and Linguistics
Law
Military
Nature and Environment
Persons
Processes
Science and Technology
Society and Culture
Transport
Feedback

Open Government Portal

Found 9969 records

Clear all choices

Download Search Results

green house emission facilities Search

Order By Best match

CSV ✖

Open Data Portal Catalogue Dataset

Suggest a Dataset

▶ Organization

▶ Portal Type

Atlantic Colonies - Density Analysis

Federal

Data Sources: Banque informatisée des oiseaux de mer au Québec (BIOMQ: ECCC-CWS Quebec Region) Atlantic Colonial Waterbird Database (ACWD: ECCC-CWS Atlantic Region).. Both the BIOMQ and ACWD contain records of individual colony counts, by species, for known colonies located in Eastern Canada. Although some colonies are censused annually, most are visited much less frequently. Methods used to derive

Feedback

data.worldbank.org

Learn how the World Bank Group is helping countries with COVID-19 (coronavirus). Find Out >

THE WORLD BANK IBRD • IDA | Data

This page in: English Español Français العربية 中文

New to this site? Start Here

DataBank Microdata Data Catalog

World Bank Open Data

Free and open access to global development data

Search data e.g. GDP, population, Indonesia

Browse by Country or Indicator

MOST RECENT

A year into the COVID-19 pandemic: what rounds 4 and 5 of Uganda High-Frequency Phone Survey tell

Aziz Atamanov, Frédéric Cochinard, John Ilukor, Giulia Ponzini, Jun 10, 2021

WHAT YOU CAN LEARN WITH OPEN DATA

Poverty headcount ratio at $1.90 a day (2011 PPP) (% of population)

Help / Feedback

Downloading files using R environment

Get/set your working directory

- A basic component of working with data is knowing your working directory
- Two main commands are getwd() and setwd().
- Be aware of relative versus abosulute paths

- **Relative –** setwd (“./data”), setwd(“../”)
- **Absolute –** setwd(“users/Mihaela/data/”)

- Important difference in Windows setwd(“C:\\Users\\Mihaela\\Downloads”)

Checking for and creating directories

- file.exists (“directoryName”) will check to see if the directory exists
- dir.create(“directoryName”) will create a directory if it doesn’t exist

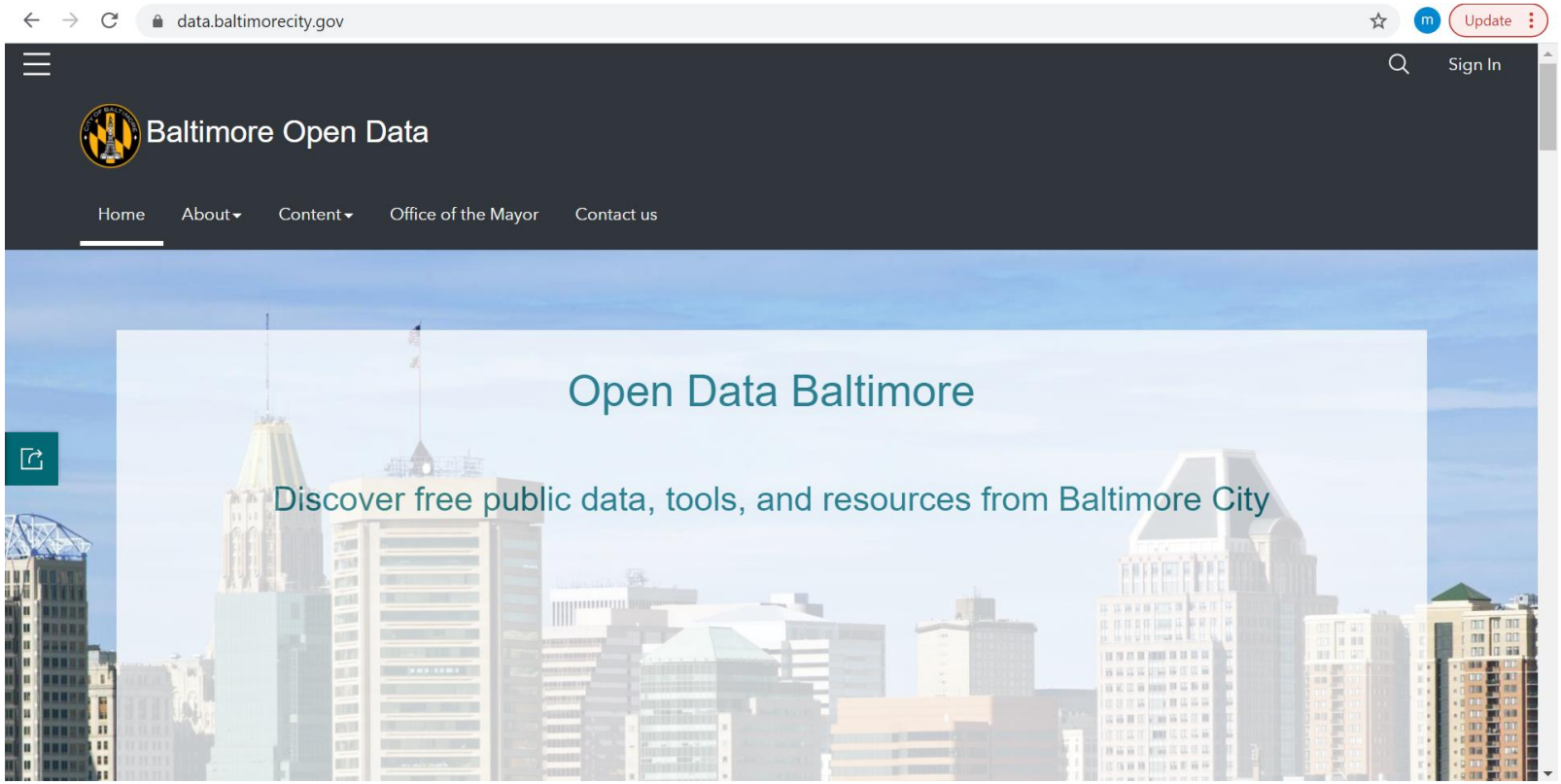
- Here is an example checking for a “data” directory and creating if it doesn’t exist

```
if(!file.exists(“data”)){
dir.create(“data”)
}
```

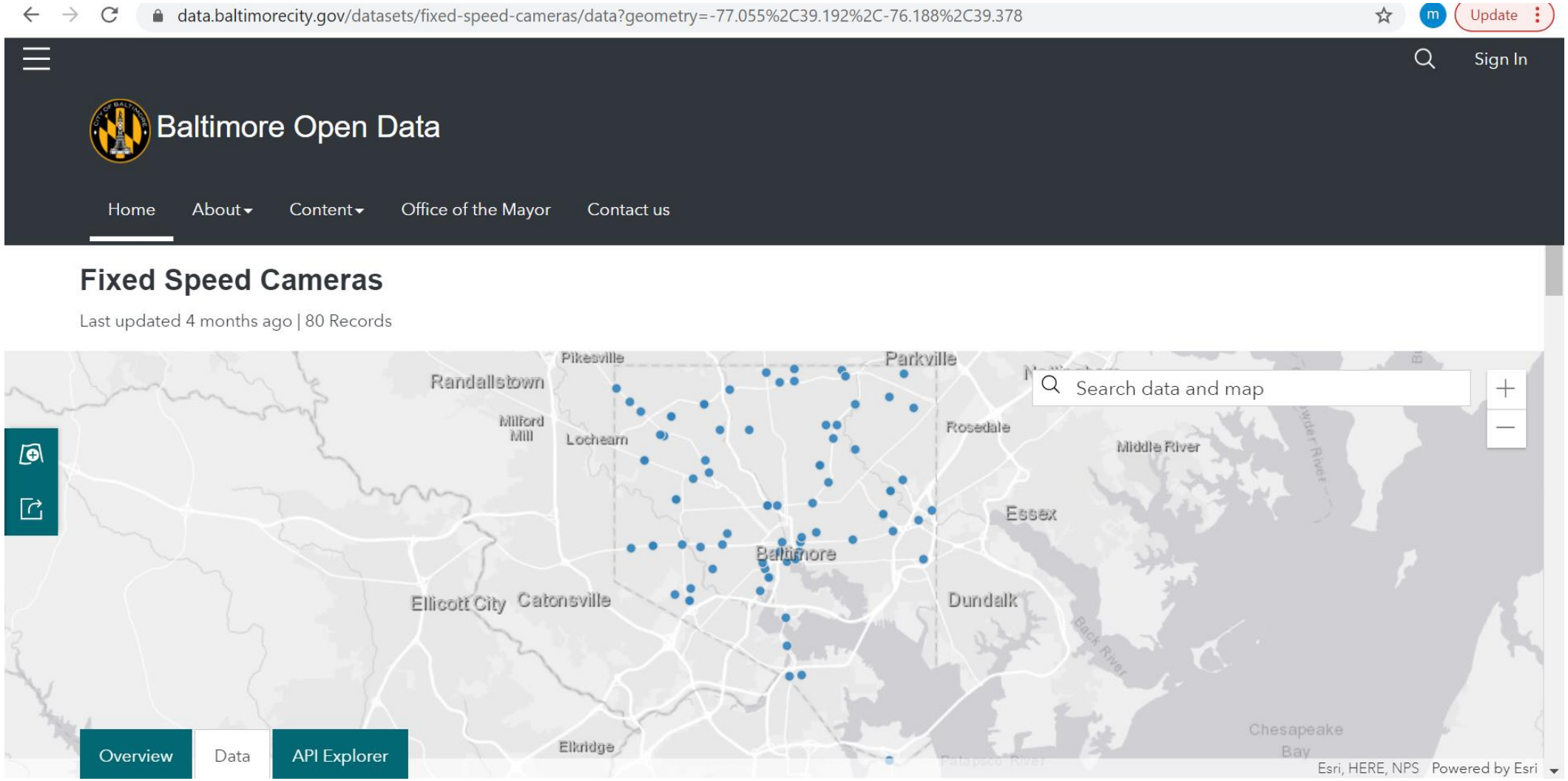
Getting data from the internet with download.file()

- Downloads a file from the internet
- Even if you could do this by hand, helps with reproducibility
- Important parameters are *url*, *destfile*, *method*
- Useful for downloading tab-delimited, csv, and other files

Example – Baltimore camera data



Example – Baltimore camera data



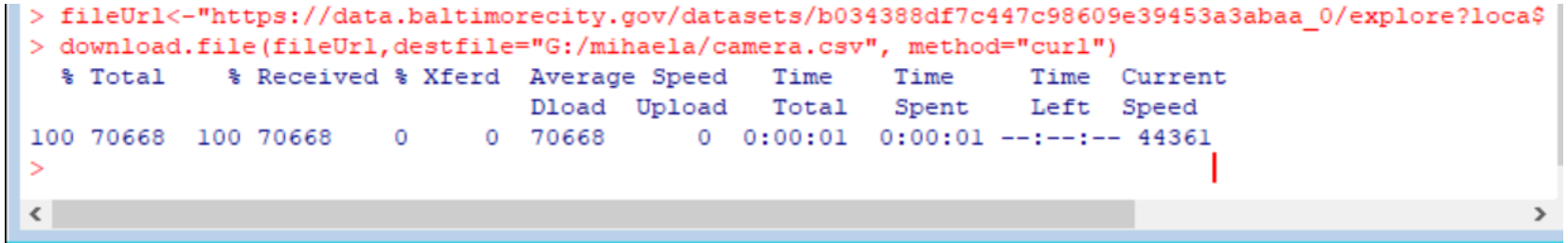
Download a file from the web

fileUrl<-”https://data.baltimorecity.gov/api/views/dz54-2aru/rows.csv?accessType=Download”

fileUrl<-
”https://data.baltimorecity.gov/datasets/b034388df7c447c98609e39453a3abaa_0/explore?location=40.661610%2C-76.593382%2C5.69”

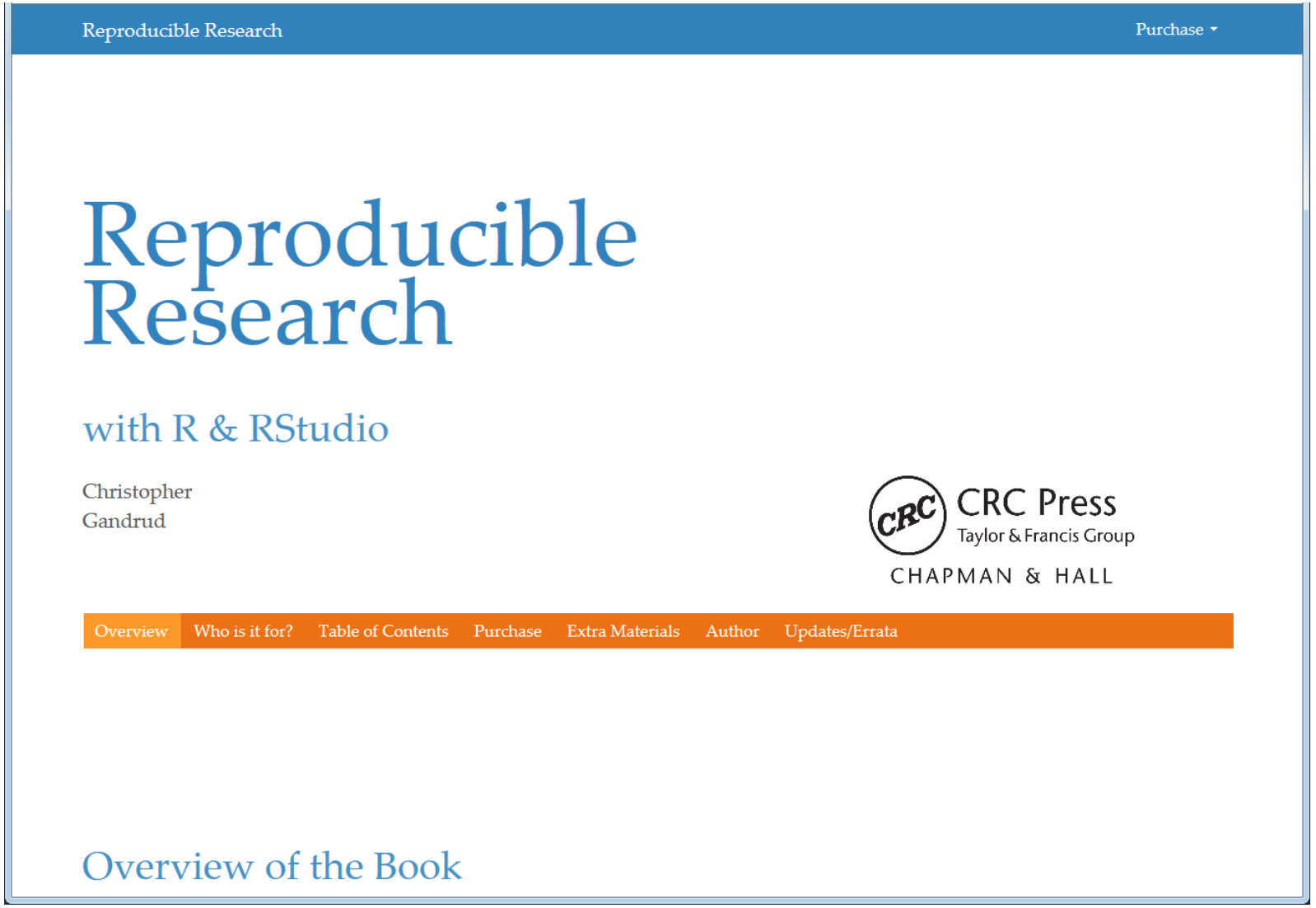
download.file(fileUrl,destfile="G:/mihaela/camera.csv", method="curl")

list.files("G:/mihaela")
[1] "camera.csv"



Some notes about download.file()

- If the url starts with *http* you can use download.file()
- If the url starts with *https* on Windows you may be ok
- If the url starts with *https* on Mac you may need to set *method="curl"*
- If the file is big, this might take a while
- Be sure to record when you downloaded.



Ethical issues and responsibilities

Why do we care?

Nature medicine

Genomic signatures to guide the use of chemotherapeutics

Anil Potti[1,2], Holly K Dressman[1,3], Andrea Bild[1,3], Richard F Riedel[1,2], Gina Chan[4], Robyn Sayer[4], Janiel Cragun[4], Hope Cottrill[4], Michael J Kelley[2], Rebecca Petersen[5], David Harpole[5], Jeffrey Marks[5], Andrew Berchuck[1,6], Geoffrey S Ginsburg[1,2], Phillip Febbo[1,2,3], Johnathan Lancaster[4] & Joseph R Nevins[1,2,3]

[1] Duke Institute for Genome Sciences and Policy, Duke University, Box 3382, Durham, North Carolina 27710, USA

[2] Department of Medicine, Duke University Medical Center, Box 31295, Durham, North Carolina 27710, USA

[3] Department of Molecular Genetics and Microbiology, Duke University Medical Center, Box 3054, Durham, North Carolina 27710, USA

[4] Division of Gynecologic Surgical Oncology, H. Lee Moffitt Cancer Center & Research Institute, University of South Florida, 12902 Magnolia Drive, Tampa, Florida 33612, USA

[5] Department of Surgery, Duke University Medical Center, Box 3627, Durham, North Carolina 27710, USA

[6] Department of Obstetrics and Gynecology, Duke University Medical Center, Box 3079, Durham, North Carolina 27710, USA

Correspondence should be addressed to Joseph R Nevins nevin001@mc.duke.edu

Using *in vitro* drug sensitivity data coupled with Affymetrix microarray data, we developed gene expression signatures that predict sensitivity to individual chemotherapeutic drugs. Each signature was validated with response data from an independent set of cell line studies. We further show that many of these signatures can accurately predict clinical response in individuals treated with these drugs. Notably, signatures developed to predict response to individual agents, when combined, could also predict response to multidrug regimens. Finally, we integrated the chemotherapy response signatures with signatures of oncogenic pathway deregulation to identify new therapeutic strategies that make use of all available drugs. The development of gene expression profiles that can predict response to commonly used cytotoxic agents provides opportunities to better use

The main conclusion is that we can use microarray data from cell lines (the NCI60) to define drug response "signatures", which can be used to predict whether patients will respond.

→ An exciting result

May/June 2009: clinical trials had begun.

Ann. Appl. Stat.
Volume 3, Number 4 (2009), 1309-1334.

Deriving chemosensitivity from cell lines: Forensic bioinformatics and reproducible research in high-throughput biology

Keith A. Baggerly and Kevin R. Coombes

Abstract

High-throughput biological assays such as microarrays let us ask very detailed questions about how diseases operate, and promise to let us personalize therapy. Data processing, however, is often not described well enough to allow for exact reproduction of the results, leading to exercises in "forensic bioinformatics" where aspects of raw data and reported results are used to infer what methods must have been employed. Unfortunately, poor documentation can shift from an inconvenience to an active danger when it obscures not just methods but errors. In this report we examine several related papers purporting to use microarray-based signatures of drug sensitivity derived from cell lines to predict patient response. Patients in clinical trials are currently being allocated to treatment arms on the basis of these results. However, we show in five case studies that the results incorporate several simple errors that may be putting patients at risk. One theme that emerges is that the most common errors are simple (e.g., row or column offsets); conversely, it is our experience that the most simple errors are common. We then discuss steps we are taking to avoid such errors in our own investigations.

1. We cannot reproduce their selection of cell lines. The sets of "sensitive" and "resistant" GI50 concentrations for pemetrexed overlap.

2. The lists of genes reported are wrong, due to an off-by-one indexing error.

3. Using their software, we can perfectly reproduce the published heatmaps for both cisplatin and pemetrexed. However, after correcting for the off-by-one error, we can only match the gene list exactly for pemetrexed.

4. For cisplatin, their software returns only 41 of the 45 reported probesets. The remaining 4 are "203719_at", "210158_at", "228131_at", and "231971_at" which correspond to ERCC1 (U133A), ERCC4 (U133A), ERCC1 (U133B), and FANCM (U133B) respectively. These probesets cannot be identified from the training data. Indeed, *the last two probesets are not physically present on the U133A arrays used in the training set.* ERCC1 and ERCC4 are the only genes named in the paper. Their paper also notes enrichment of the cisplatin signature for DNA repair genes. Their reported list contains 5 probesets with this function, 4 of which are those mentioned above.

5. In the case of pemetrexed, the sensitive/resistant labeling has been reversed. Using the model with this labeling suggests administering the drug only to the patients it would not benefit.

Sep-Oct: Story covered by The Cancer Letter, Duke starts internal investigation, suspends trials. So, what happened next?

Reply (Nat Med 13:1277-8)
Labels for Adria are correct – details on their web page.

They've gotten the approach to work again.

JOURNAL OF CLINICAL ON

Pharmacogenomic Strategies Provide a Rational Approach to the Treatment of Cisplatin-Resistant Patients With Advanced Cancer

David S. Hsu, Bala S. Balakumaran, Chaitanya R. Acharya, Vanja Vlahovic, Kelli S. Walters, Katherine Garman, Carey Anders, Richard F. Riedel, Johnathan Lancaster, David Harpole, Holly K. Dressman, Joseph R. Nevins, Phillip G. Febbo and Anil Potti

validation of gene signatures that predict the response of breast cancer to neoadjuvant chemotherapy: a substudy of the EORTC 10994/BIG 00-01 clinical trial

Hervé Bonnefoi [a], Anil Potti [a], Mauro Delorenzi [a], Louis Mauriac [a], Mario Campone [a], Michèle Tubiana-Hulin [a], Thierry Petit [a], Philippe Rouanet [a], Jacek Jassem [a], Emmanuel Blot [a], Véronique Becette [a], Pierre Farmer [a], Sylvie André [a], Chaitanya R Acharya [a], Sayan Mukherjee [a], David Cameron [a], Jonas Bergh [a], Joseph R Nevins [a], Richard D Iggo [a]

THE CANCER LETTER

PO Box 9905 Washington DC 20016 Telephone 202-362-1809

Duke In Process To Restart Three Trials Using Microarray Analysis Of Tumors

By Paul Goldberg

Duke University said it is in the process of restarting three clinical trials using microarray analysis of patient tumors to predict their response to chemotherapy.

Their investigation's results *"strengthen ... confidence in this evolving approach to personalized cancer treatment."*

THE CANCER LETTER

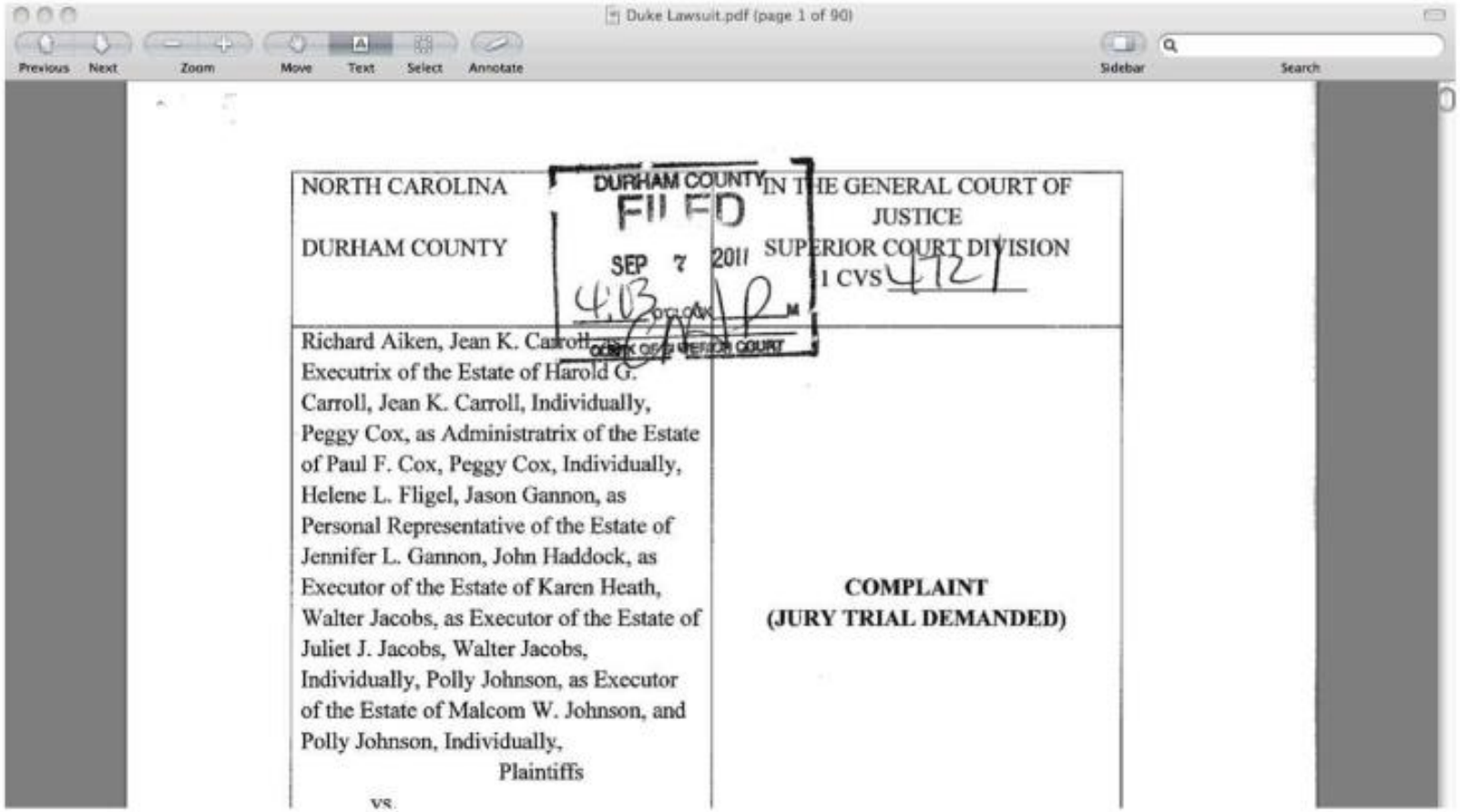
PO Box 9905 Washington DC 20016 Telephone 202-362-1809

Prominent Duke Scientist Claimed Prizes He Didn't Win, Including Rhodes Scholarship

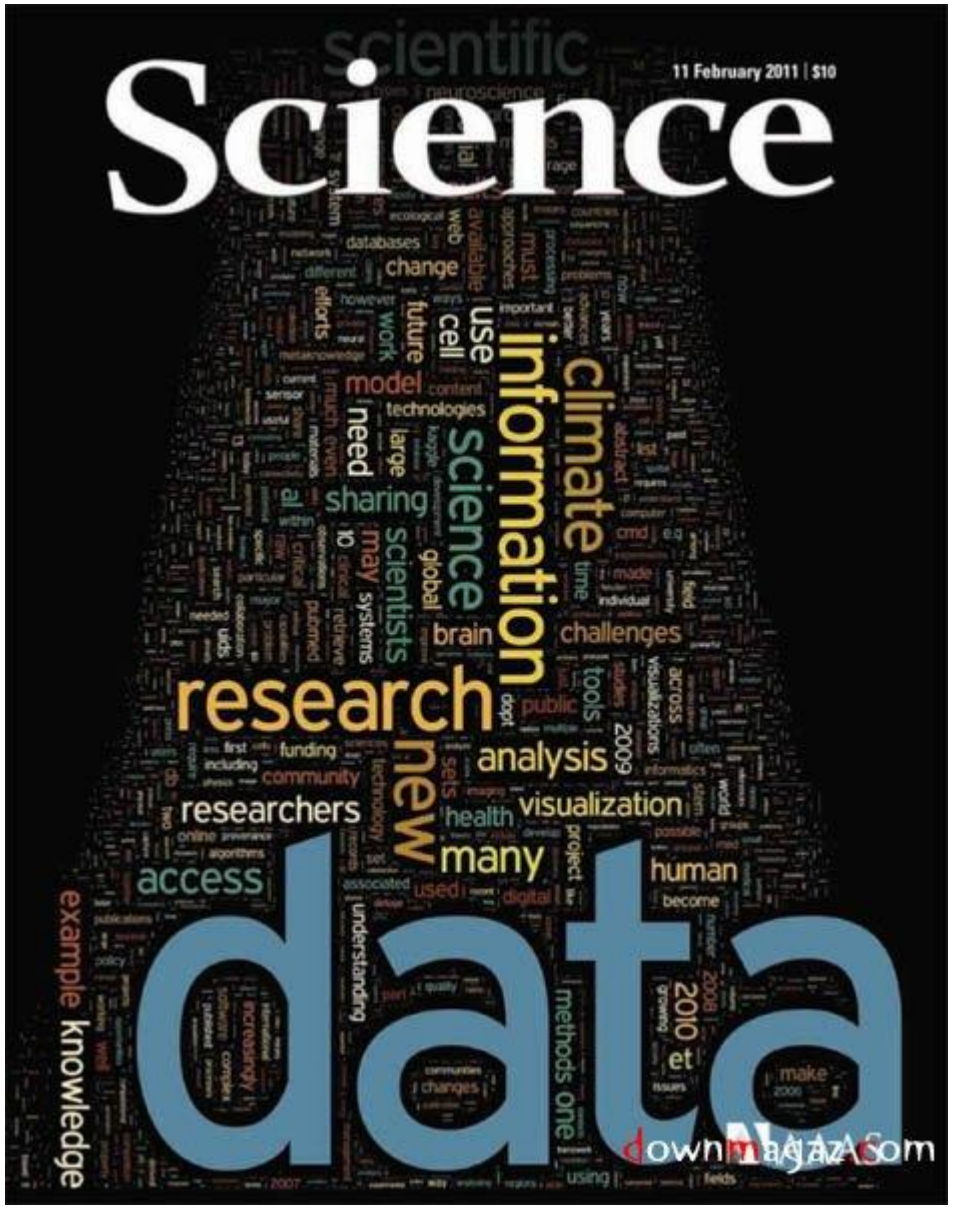
By Paul Goldberg

A high-profile cancer genomics researcher at Duke University claimed in multiple grant applications that he had been a Rhodes scholar, when, in fact, the Rhodes Trust states flatly that he was not.

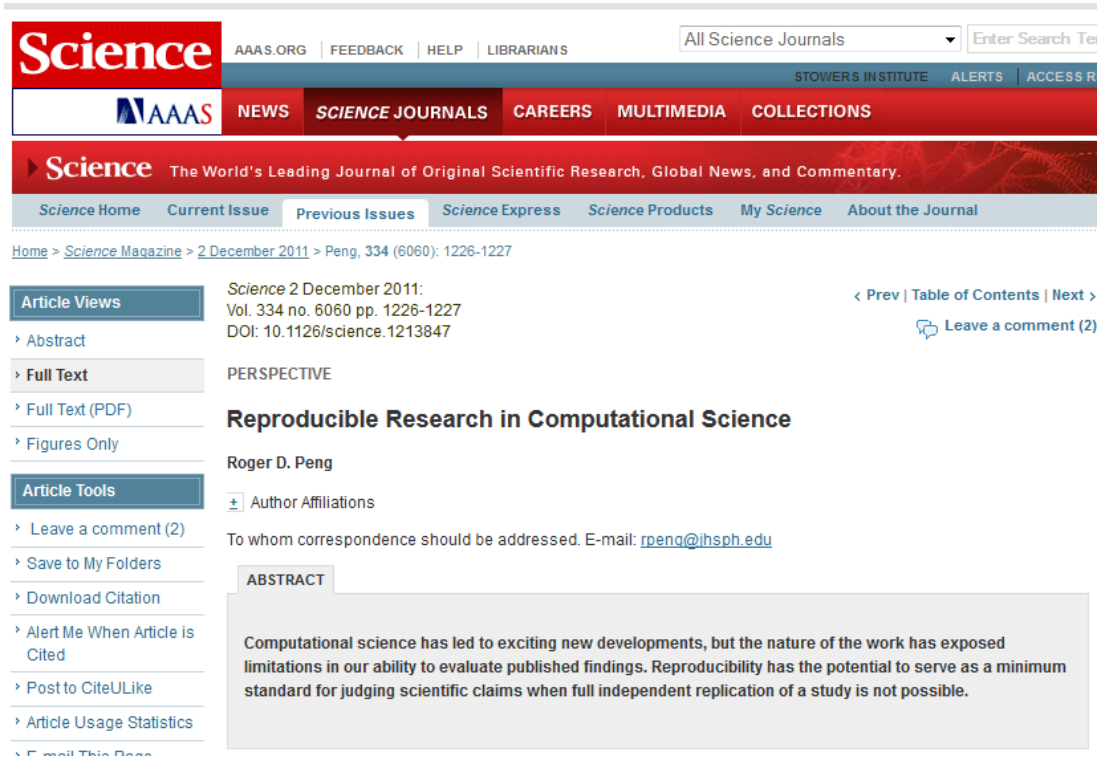
Why you should care - serious trouble



Recent Developments in Reproducible Research



Reproducible Research in Computational science



The journal *Biostatistics,*, has implemented a policy for encouraging authors of accepted papers to make their work reproducible by others. Authors can submit their code or data to the journal for posting as supporting online material and can additionally request a "reproducibility review," in which the associate editor for reproducibility runs the submitted code on the data and verifies that the code produces the results published in the article.

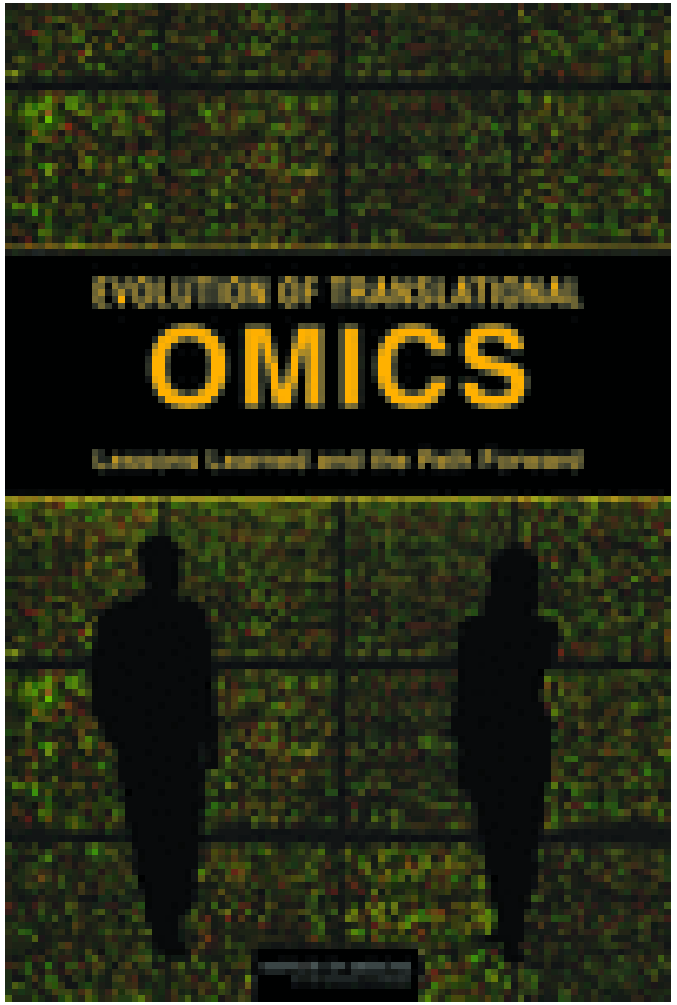
As of July 2011, 21 of 125 articles have been published with a kite-mark

Recent Developments in Reproducible Research

Evolution of Translational Omics

Lessons Learned and the Path Forward

Institute of Medicine
of the National Academies



<u>Recommended to be done to improve reproducibility</u>

Data/metadata used to develop test should be made publicly available

The computer code and fully specified computational procedures used for development of the candidate Omics-based test should be made sustainably available

"Ideally, the computer code that is released will encompass all of the steps of computational analysis including all data preprocessing steps, that have been described in this chapter. All aspects of the analysis need to be transparently reported"

The IOM Report

Support the IOM | Media Room | Directory | Videos | Pe

INSTITUTE OF MEDICINE
OF THE NATIONAL ACADEMIES
Advising the nation • Improving health

ABOUT THE IOM | REPORTS | ACTIVITIES | MEETINGS | Explore b

BROWSE HISTORY

News Release

IOM Report Recommends Evaluation and Validation Process to Prevent Problems Associated With Turning 'Omics' Research Into Clinical Tests

Released: 3/23/2012

Genomics, proteomics, and other branches of molecular bioscience offer the prospect of greater precision in medical care, but some clinical tests based on "omics" research have proved invalid and highlighted the challenges of dealing with complex data. To enhance the translation of omics-based discoveries to clinical use, a new report by the Institute of Medicine recommends a detailed process to evaluate whether the data and computational steps underlying such tests are sound and the tests are ready to be used in clinical trials. The proposed process defines responsibilities and best practices for the investigators, research institutions, funders, regulators, and journals involved in development and dissemination of clinical omics-based technologies.

The request for the IOM report stemmed in part from a series of events at Duke University in which researchers claimed that their genomics-based tests were reliable predictors of which chemotherapy would be most effective for specific cancer patients. Failure by many parties to detect or act on problems with key data and computational methods underlying the tests led to the inappropriate enrollment of patients in clinical trials, premature launch of companies, and retraction of dozens of research papers. Five years after they were first made public, the tests were acknowledged to be invalid.

Read More...

A recent investigation found that less than half of selected microarray experiments published in *Nature Genetics* could be reproduced. Issues that prevented reproduction included missing raw data, details in processing methods (especially computational ones), and software and hardware details.

To support reproducible computational research, the concept of **a Reproducible Research System (RRS)** has been proposed. An RRS provides an environment for performing and recording computational analyses and enabling the use or inclusion of these analyses when preparing documents for publications.

Bioconductor and Sweave, a 'literate programming' tool for documenting Bioconductor analyses, can be used to reproduce an analysis if a researcher has the original data, the Bioconductor scripts used in the analysis, and enough programming expertise to run the scripts.

Without programming or informatics expertise, scientists needing to use computational approaches are impeded by problems ranging from tool installation; to determining which parameter values to use; to efficiently combining multiple tools together in an analysis chain.

Literate (Statistical) Programming

- Knitr is a more recent package
- Brings together many features added on to Sweave to address limitations
- Knitr is a powerful tool for integrating code and text in a simple document form
- Knitr uses R as the programming language (although others are allowed) and variety of documentation languages LaTeX, Markdown, HTML
- Markdown is a text-to-HTML conversion tool for web writers. Markdown allows you to write using an easy-to-read, easy-to-write plain text format, then convert it to structurally valid XHTML (or HTML).

Example

RStudio

File Edit Code View Plots Session Build Debug Tools Help

PA1_template.Rmd | PA1_template.Rmd

Knit HTML | Run | Chunks

```
1  Loading and preprocessing the data
2  ==========================================================
3
4  ```{r}
5  if(!suppressMessages(require(ggplot2))){
6      print('trying to install ggplot2')
7      install.packages('ggplot2')
8      if(suppressMessagesrequire(ggplot2)){
9          print("ggplot2 installed and loaded")
10     } else {
11         stop("could not install ggplot2")
12     }
13 }
14
15 # Define some options for knitr
16 knitr::opts_chunk$set(tidy=FALSE, fig.path='figures/')
17
18
19 unzip("U:/mis/HOPKINS/repdata_data_activity.zip")
20 activity <- read.csv(file="U:/mis/HOPKINS/activity.csv",stringsAsFactors =
   FALSE)
21 activity$date <- as.Date(activity$date)
22
23 head(activity)
24
25 ##what is mean total number of steps taken per day?
26
27 daily_activity <-
28   aggregate(formula = steps~date, data = activity,
29             FUN = sum, na.rm=TRUE)
30
31
```

1:1 (Top Level) R Markdown

Console U:/mis/HOPKINS/

```
R version 3.1.1 (2014-07-10) -- "Sock it to Me"
Copyright (C) 2014 The R Foundation for Statistical Computing
Platform: x86_64-w64-mingw32/x64 (64-bit)
```

RStudio: Preview HTML

Preview: H:/RepData_PeerAssessment1/PA1_template.html | Log | Save As | Publish | Find

Loading and preprocessing the data

```
if(!suppressMessages(require(ggplot2))){
    print('trying to install ggplot2')
    install.packages('ggplot2')
    if(suppressMessagesrequire(ggplot2)){
        print("ggplot2 installed and loaded")
    } else {
        stop("could not install ggplot2")
    }
}

# Define some options for knitr
knitr::opts_chunk$set(tidy=FALSE, fig.path='figures/')


unzip("U:/mis/HOPKINS/repdata_data_activity.zip")
activity <- read.csv(file="U:/mis/HOPKINS/activity.csv",stringsAsFactors = FALSE)
activity$date <- as.Date(activity$date)

head(activity)
```

```
##   steps       date interval
## 1    NA 2012-10-01        0
## 2    NA 2012-10-01        5
## 3    NA 2012-10-01       10
## 4    NA 2012-10-01       15
## 5    NA 2012-10-01       20
## 6    NA 2012-10-01       25
```

```
##What is mean total number of steps taken per day?

daily_activity <-
  aggregate(formula = steps~date, data = activity,
            FUN = sum, na.rm=TRUE)


mean_steps <- round(mean(daily_activity$steps), 2)  # Mean
median_steps <- quantile(x = daily_activity$steps, probs = 0.5)  # Median, 50%Q
mean_steps
```

```
## [1] 10766
```

```
median_steps
```

```
##   50%
## 10765
```

```
histogram <-
qplot(x=date, y=steps,
      data=subset(activity, complete.cases(activity)),
      stat='summary', fun.y=sum, geom='bar') +
  labs(title='Figure 1: Number of steps taken daily\n',
       y='Total steps per day', x='Date')
plot(histogram)
```

Figure 1: Number of steps taken daily