

Data Acquisition

The Big Data Era in Biology and Data Integration



Wish List



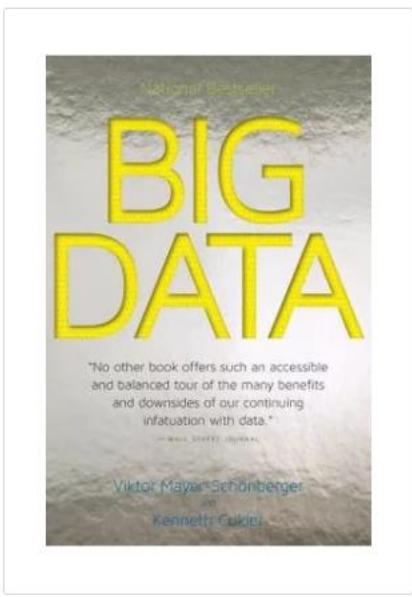
Sign in



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

All Categories ▾ About Us Contact Us

/ All Categories / Business / Information Management / Big Data: A Revolution That Will Transform How We Live, Work, and 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

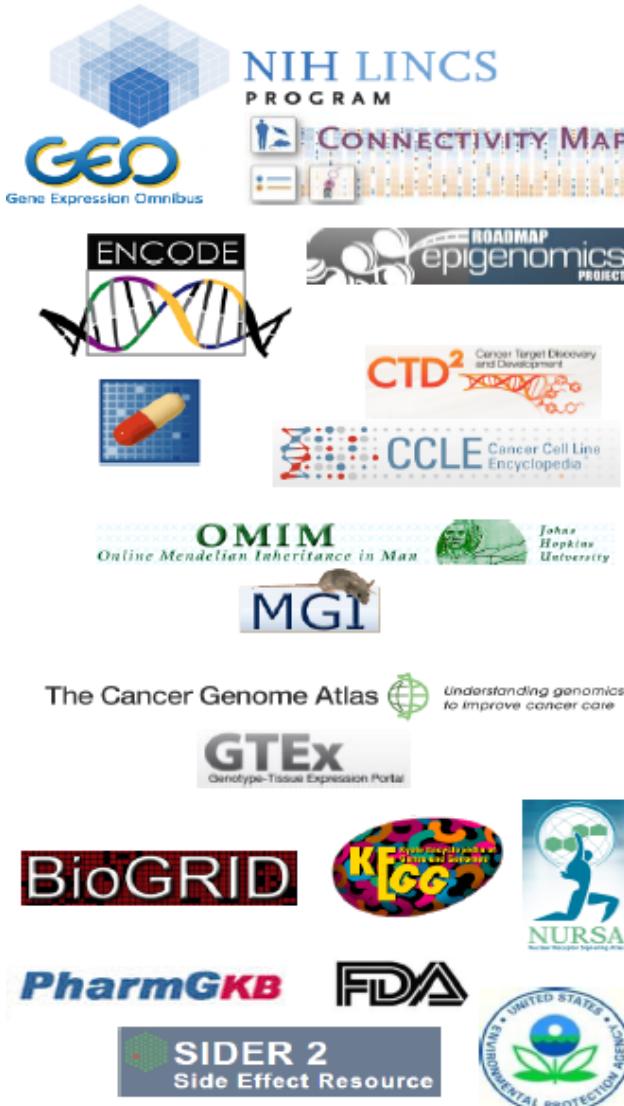
Good

\$3.88

New

\$15.95

BIG Data



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

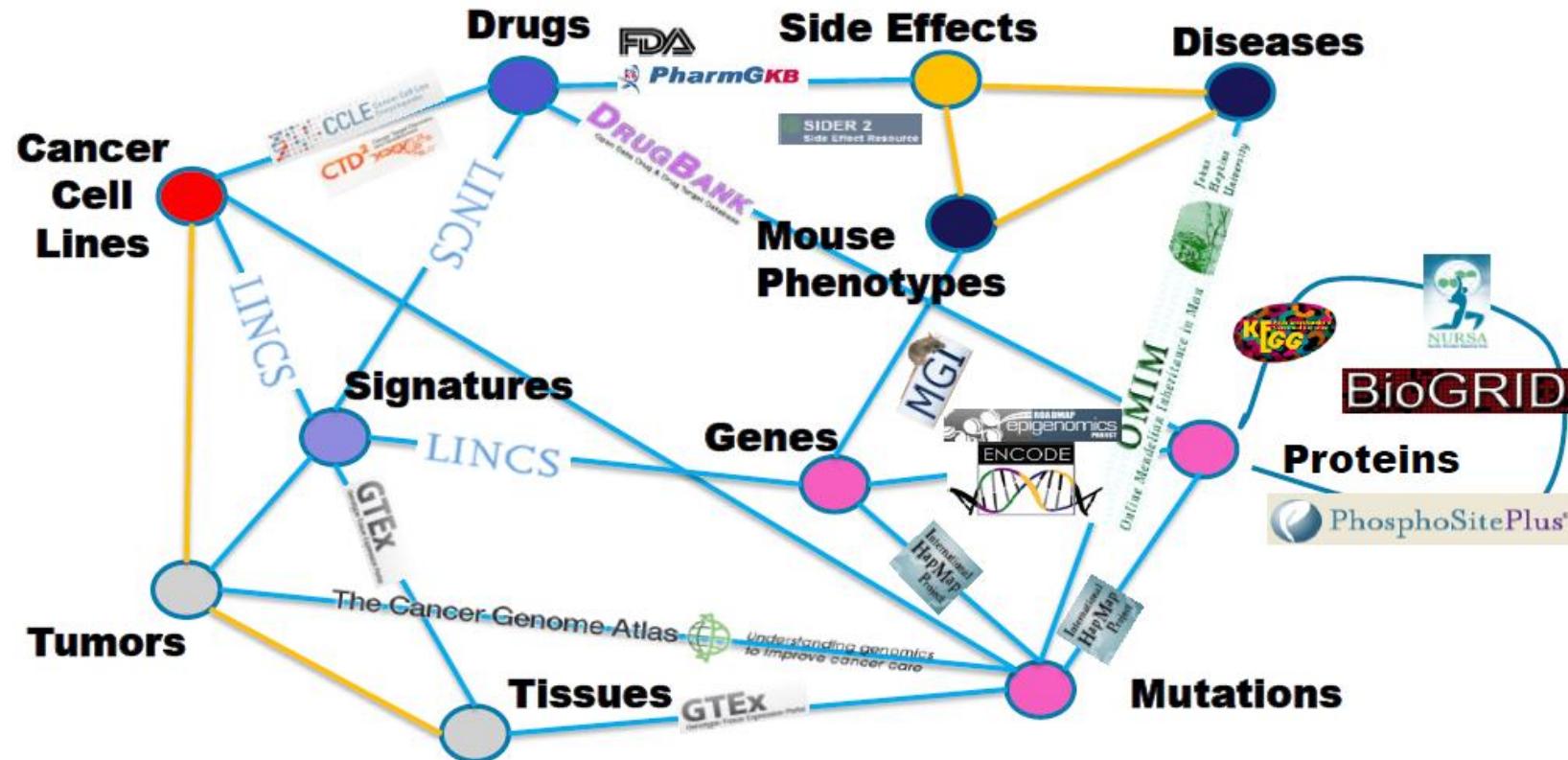
Drugs and Toxic Chemicals that Cause Adverse Events

Networks

Gene-Set Libraries

Bi-partite Graphs

Global Picture of the Puzzle



Topics of this lecture

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

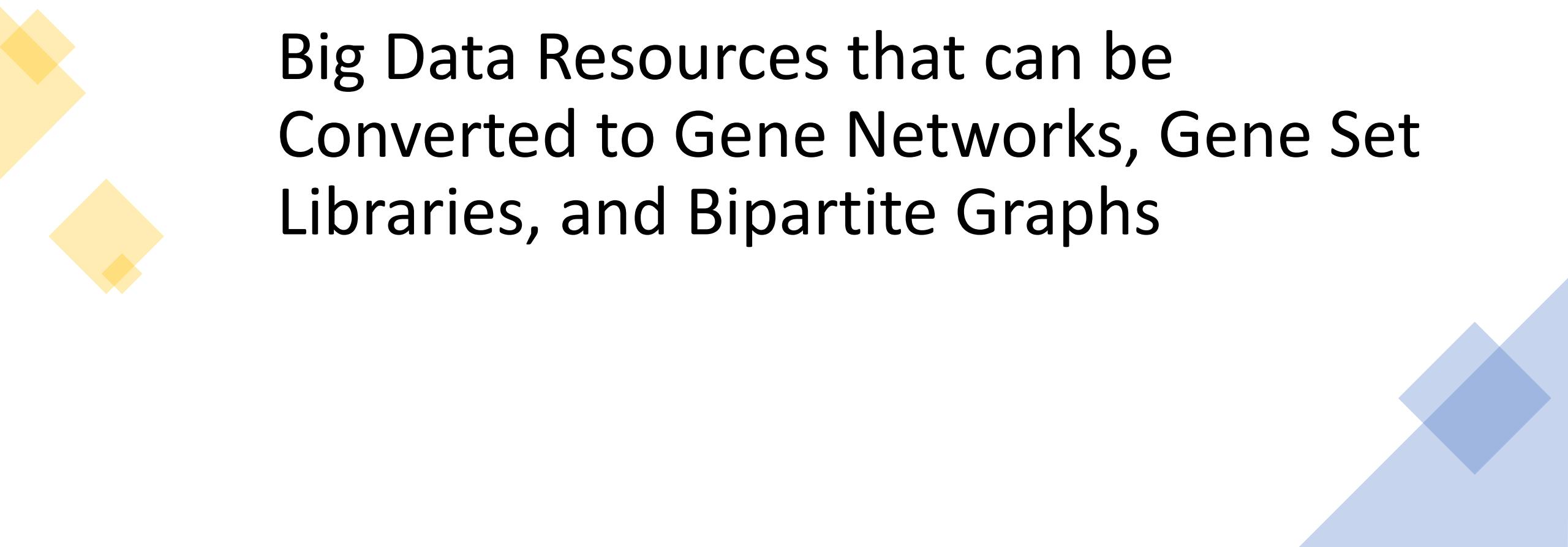
Census data, NHANES, Gene expression omnibus (GEO), Genomic data commons, dbGap, ENCODE, Kaggle, SEER, etc. , Ethical issues and responsibilities , Challenges of handling big-data

NCBI biological database (search for BBX)

The screenshot shows the NCBI All Resources page. On the left, there is a sidebar with a red arrow pointing to the "All Resources" link under the "Resource List (A-Z)" heading. The main content area has a title "All Resources" and a navigation bar with tabs: All, Databases, Downloads, Submissions, Tools, and How To. The "Databases" tab is selected. Below it, there are several database entries:

- Assembly**: A database providing information on the structure of assembled genomes, assembly names and other meta-data, statistical reports, and links to genomic sequence data.
- BioCollections**: A curated set of metadata for culture collections, museums, herbaria and other natural history collections. The records display collection codes, information about the collections' home institutions, and links to relevant data at NCBI.
- BioProject (formerly Genome Project)**: A collection of genomics, functional genomics, and genetics studies and links to their resulting datasets. This resource describes project scope, material, and objectives and provides a mechanism to retrieve datasets that are often difficult to find due to inconsistent annotation, multiple independent submissions, and the varied nature of diverse data types which are often stored in different databases.
- BioSample**: The BioSample database contains descriptions of biological source materials used in experimental assays.

At the top of the page, there is a banner about COVID-19 and links to CDC, NIH, NCBI, and HHS resources. The top navigation bar includes links for NCBI, Resources, How To, and Sign in to NCBI.



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

Protein information

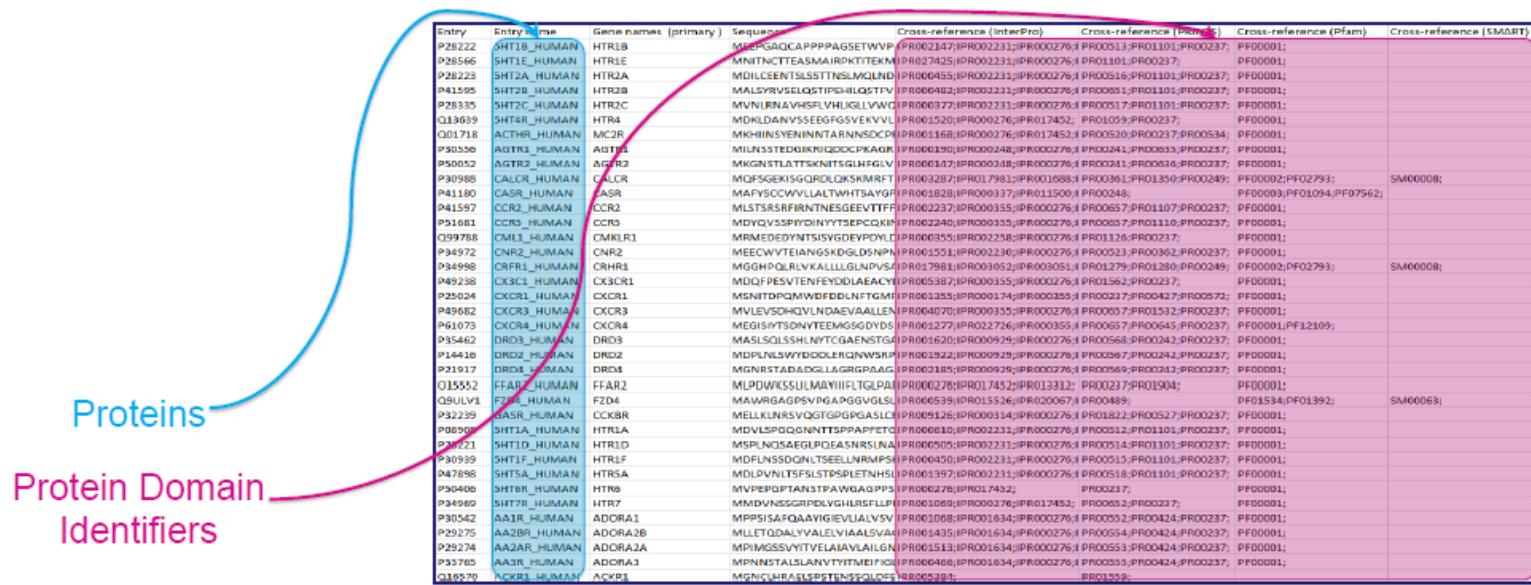
Universal Protein Resource (UniProt)

- 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, Nucleic Acids Research)

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



Universal Protein Resource (UniProt)

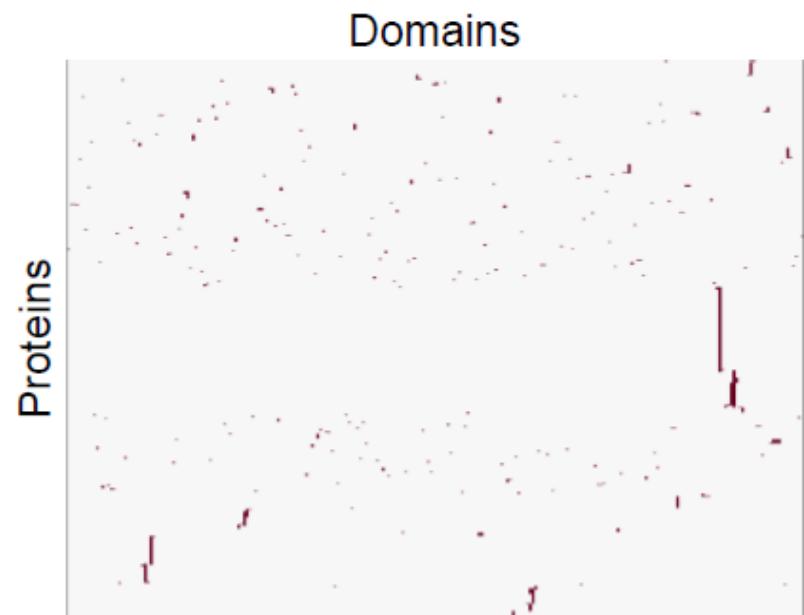


Entry	Entry name	Gene names (primary)	Sequence	Cross-reference (InterPro)	Cross-reference (Pfam)	Cross-reference (Pfam)	Cross-reference (SMART)
P28222	HTR1B_HUMAN	HTR1B	MPSAQCQAPPPAGSETWVP	[P]R002147;PR002231;PR000276;PR00513;PR01101;PR0237; PF00001;			
D28566	HTR1E_HUMAN	HTR1E	MNTINCTTEASMAIRKTTKEM	[P]R027425;PR002231;PR000276;PR01101;PR0237; PF00001;			
P28223	HTR2A_HUMAN	HTR2A	MDILCEENTSLSTTNSLMQLND	[P]R000455;PR002231;PR000276;PR00516;PR01101;PR0237; PF00001;			
P41595	HTR2B_HUMAN	HTR2B	MALSYRVSELSQTIPSHLQSTV	[P]R000482;PR002231;PR000276;PR00517;PR01101;PR0237; PF00001;			
P28335	HTR2C_HUMAN	HTR2C	MVNLRNAYVSEFVHUGLIVWNG	[P]R000377;PR002231;PR000276;PR00517;PR01101;PR0237; PF00001;			
Q13639	HTR4_HUMAN	HTR4	MDKLADANVYSSSEGFGSVEKYV	[P]R001520;PR000276;PR017452; PR01059;PR0237; PF00001;			
Q01718	ACTHR_HUMAN	MC2R	MKHINRSYENNTNARTARNNSQCPV	[P]R001168;PR000276;PR017452;PR00519;PR0237;PR00534; PF00001;			
P30358	A2TR2_HUMAN	A2TR2	MILNSTEDGKRNQDQCPKAQGV	[P]R000249;PR000276;PR002231;PR00535;PR0237; PF00001;			
P50652	A2TR2_HUMAN	A2TR2	MKGNSITLATTSSKNTSGIHLFGLV	[P]R000187;PR000238;PR002276;PR00341;PR00536;PR0237; PF00001;			
P30980	CALCR_HUMAN	CALCR	MQFSGEKISGQIROLQKSMKIFT	[P]R003287;PR017981;PR0010883;PR00361;PR01350;PR00249; PF00002;PF02793; SM00008;			
P41180	CASR_HUMAN	CASR	MAFVSCCWVLALWTHTSAYGF	[P]R001628;PR000337;PR011500;PR00248; PF00003;PF01094;PF07562;			
P41597	CCR2_HUMAN	CCR2	MLSTSRSRFRINTNTNEEGEEVTFF	[P]R002237;PR000355;PR002276;PR00657;PR01107;PR0237; PF00001;			
P51681	CCR_HUMAN	CCR	MDVQV3SPYDINYYTSEPCQQXP	[P]R002240;PR000355;PR002276;PR00837;PR01110;PR0237; PF00001;			
Q39788	CMKL1_HUMAN	CMKL1	MREMEDEDYNTISVGDEYEVPOYL	[P]R003555;PR002250;PR002761;PR011226;PR0237; PF00001;			
P34972	CNR2_HUMAN	CNR2	MEECWVTEIAANSKDGLDSNPW	[P]R0011551;PR002230;PR002276;PR00223;PR00362;PR0237; PF00001;			
P34998	CRFR1L_HUMAN	CRFR1L	MGGHPQAVLVKALLLGLDNPWS	[P]R001981;PR000052;PR003051;PR01279;PR01280;PR00249; PF00002;PF02793; SM00008;			
P49238	CX3CL1_HUMAN	CX3CR1	MDFQPSTVENFEYDQDAAEACY	[P]R005387;PR000355;PR002276;PR01562;PR0237; PF00001;			
P25024	CXCR1_HUMAN	CXCR1	MSNNTDPQMWDFFDLNTGTM	[P]R001355;PR000355;PR002276;PR0232;PR00427;PR00572; PF00001;			
P49682	CXCR3_HUMAN	CXCR3	MVLEVSDHQVNLDAEVAALEN	[P]R004070;PR000355;PR002276;PR00657;PR01352;PF00001;			
P61073	CXCR4_HUMAN	CXCR4	MEIGSIIYTSQNYTEEMASGQDYD5	[P]R001277;PR002276;PR00355;PR00657;PR00645;PR0237; PF00001;PF12109;			
P35462	DRD3_HUMAN	DRD3	MASLSQLSSHLYNTCAENNSTG	[P]R001620;PR000929;PR002276;PR00562;PR00242;PR0237; PF00001;			
P14410	DRD2_HUMAN	DRD2	MDPILNWLSDYDQDLERIGNWRSRP	[P]R001922;PR000929;PR002761;PR00397;PR00242;PR0237; PF00001;			
P21917	DRD4_HUMAN	DRD4	MGRN1STADADGILAGRGGPAAG	[P]R002185;PR000929;PR002276;PR00599;PR00242;PR0237; PF00001;			
Q15552	FFAR2_HUMAN	FFAR2	MDPDWESSLULMAYIII LTGLPAI	[P]R000276;PR017452;PR013312; PR00237;PR01904; PF00001;			
P91UL1	F73E_HUMAN	F7D4	MAWRGAGPSVVAAGPGVGVL	[P]R000593;PR015526;PR020067;PR00489; PF01534;PF01392; SM00068;			
P32239	GASR_HUMAN	CCKBR	MELLKLNR5VOGTGQGPQGSACL	[P]R009126;PR000314;PR000276;PR01822;PR00527;PR0237; PF00001;			
P08500	HHT1A_HUMAN	HTR1A	MDSLSPCGQNNTTSPPAFPETG	[P]R000610;PR002231;PR000276;PR00312;PR01101;PR0237; PF00001;			
P28221	HHT1D_HUMAN	HTR1D	MSPLNQSAEGLPQEASNRSLNA	[P]R000505;PR002231;PR000276;PR00314;PR01101;PR0237; PF00001;			
P09399	HHT1F_HUMAN	HTR1F	MDFLNSSDQNLTESELLNRMPS	[P]R000450;PR002231;PR000276;PR00515;PR01101;PR0237; PF00001;			
P47896	HHT5A_HUMAN	HTR5A	MIDLNVNTSFSLTPSPLENTNHSL	[P]R001397;PR002231;PR000276;PR00518;PR01101;PR0237; PF00001;			
P50400	HHT6_HUMAN	HTR6	MVPEPGPTANSTFAWGAGPSP	[P]R000276;PR017452; PF00237; PF00001;			
P34966	HHT7_HUMAN	HTR7	MMDVNNSGGRPDLYGHURSLP	[P]R001069;PR000276;PR017452; PR00852;PR0237; PF00001;			
P30542	AA1R_HUMAN	ADORA1	MPPSIQFOAAAYIGEVILAVLVSY	[P]R001068;PR001034;PR000276;PR00552;PR00424;PR0237; PF00001;			
P29275	AA2B1_HUMAN	ADORA2B	MILETQDALYYALEVLIAVAILSV	[P]R001435;PR001634;PR000276;PR00554;PR00424;PR0237; PF00001;			
P29274	AA2A2R_HUMAN	ADORA2A	MPIMGSSVYTTEVLAIALVILGN	[P]R001513;PR001634;PR000276;PR00553;PR00424;PR0237; PF00001;			
P33760	AA3R_HUMAN	ADORA3	MPPINSTALSLANVTVITMEIICL	[P]R000646;PR001634;PR000276;PR00333;PR00424;PR0237; PF00001;			
P316570	NCKR1_HUMAN	ACKR1	MGNICLHARAEELSPTEENPQQLP	[P]R005281; PR01539;			

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)

The screenshot shows the BioGRID 3.4 Result Summary page for the gene INO80 in Homo sapiens. The search bar at the top contains 'INO80' and 'Homo sapiens'. Below the search bar, the gene name 'INO80' is displayed along with its aliases: INO80A, INOC1, hINO80. A 'Stats & Options' section provides current statistics: Publications: 14, High Throughput: 14 (23%), Low Throughput: 46 (77%), and interaction types: 60 Physical Interactions, 0 Genetic Interactions. The 'Search Filters' section allows customization of results, with 'No Filter: Show All Associations' selected. The main content area displays 28 unique interactors, sorted by evidence and alphabetically. Each interactor entry includes the protein name, aliases, description, and a count of interactions (e.g., YY1 has 6 interactions, RUVBL1 has 4, NFRKB has 4, ACTR5 has 4). Each entry also includes a 'details' link.

Result Summary

INO80 *Homo sapiens*

INO80A, INOC1, hINO80

INO80 complex subunit

UBI

GO Process (14) GO Function (5) GO Component (2)

EXTERNAL DATABASE LINKOUTS
VEGA | OMIM | HGNC | Entrez Gene | RefSeq | UniprotKB | Ensembl | HPRD
Download 56 Published Interactions For This Protein

Switch View: Interactors (28) Interactions (60) Network PTM Sites (1)

Displaying 28 total unique interactors

Sort By: [Evidence] [Alphabetical]

YY1 | DELTA, INO80S, NF-E1, UCRBP, YIN-YANG-1
YY1 transcription factor
UBI NEDD SUMO

RUVBL1 | ECP54, INO80H, NMP238, PONTIN, Pontin52, RVB1, TIH1, TIP49, TIP49A
RuvB-like AAA ATPase 1
UBI NEDD FAT10 SUMO

NFRKB | INO80G
nuclear factor related to kappaB binding protein
UBI SUMO

ACTR5 | Arp5, INO80M
ARP5 actin-related protein 5 homolog (yeast)

Search for BRMS1L protein interactions

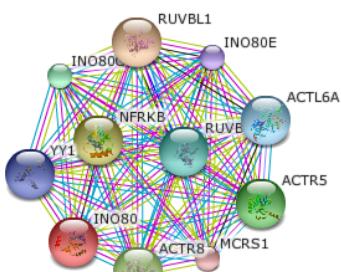
Protein interactions
Biogrid.org

Biological General Repository for Interaction Datasets (BioGRID)

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

#BioGRID	Entrez Gei	Entrez Gei	BioGRID ID	BioGRID ID	Systemati	Systemati	Official Sy	Official Sy	Synonyms	Synonyms	Experime	Experime	Author	Pubmed ID	Organism	Organism	Throughput	Score
103	6416	2318	112315	108607	-	-	MAP2K4	FLNC	SAPKK1 J1 ABPL MPI	Two-hybrid physical	Marti A (1	9006895	9606	9606	Low Throu-			
117	84665	88	124185	106603	-	-	MYPN	ACTN2	MYOP	CMD1AA	Two-hybrid physical	Bang ML (11309420	9606	9606	Low Throu-		
183	90	2339	106605	108625	-	-	ACVR1	FNTA	ACVRLK2 PTAR2 FNTA	Two-hybrid physical	Wang T (1	8599089	9606	9606	Low Throu-			
278	2624	5371	108894	111384	-	-	GATA2	PML	NFE1B DCRNF71 TR	Two-hybrid physical	Tsuzuki S (10938104	9606	9606	Low Throu-			
418	6118	6774	112038	112651	RP4-547C5	-	RPA2	STAT3	REPA2 RPAPRF HIE	Two-hybrid physical	Kim J (20	10875894	9606	9606	Low Throu-			
586	375	23163	106870	116775	-	-	ARF1	GGA3	-	-	Two-hybrid physical	Dell'Ange	10747089	9606	9606	Low Throu-		
612	377	23647	106872	117174	-	-	ARF3	ARFIP2	-	POR1	Two-hybrid physical	Kanoh H (9038142	9606	9606	Low Throu-		
617	377	27236	106872	118084	-	-	ARF3	ARFIP1	-	HSU5251	Two-hybrid physical	Kanoh H (9038142	9606	9606	Low Throu-		
663	54464	226	119970	106728	-	-	XRN1	ALDOA	1-Sep GSD12 AL	Two-hybrid physical	Lehner B (15231747	9606	9606	High Throu-			
866	351	10513	106848	115769	-	-	APP	APPBP2	PN-II ABP HS.84084	Two-hybrid physical	Zheng P (1	9843960	9606	9606	Low Throu-			
917	333	1600	106830	107970	-	RP6-239D APLP1	DAB1	APLP	-	Two-hybrid physical	Homayoun	10460257	9606	9606	Low Throu-			
1156	10370	7020	115649	112878	RP1-2901I	CITED2	TFAP2A	P35SRJ M BOFS AP2	Two-hybrid physical	Braganca J	12586840	9606	9606	Low Throu-				
1161	2033	7020	108347	112878	RP1-85F1E RP1-2901I	EP300	TFAP2A	RSTS2 p30 BOFS AP2	Two-hybrid physical	Braganca J	12586840	9606	9606	Low Throu-				
1206	338	4547	106835	110641	-	-	APOB	MTTP	FLDB LDL ABL MTP	Two-hybrid physical	Bradbury I	9915855	9606	9606	Low Throu-			
1427	409	5900	106902	111836	-	RP11-326L ARRB2	RALGDS	BARR2 AFRaGEF R	Two-hybrid physical	Bhattachar	12105416	9606	9606	Low Throu-				
1611	1436	2885	107823	109142	-	-	CSF1R	GRB2	FMS FIM2EGFRBP-G	Two-hybrid physical	Mancini A	9380408	9606	9606	Low Throu-			
1695	7916	2885	113646	109142	DADB-70P-	-	PRRC2A	GRB2	D6551E B EGFRBP-G	Two-hybrid physical	Lehner B (14667819	9606	9606	Low Throu-			
1783	27257	4677	118104	110758	-	-	LSM1	NARS	CASM YJL NARS1 A	Two-hybrid physical	Lehner B (15231747	9606	9606	High Throu-			
1888	6521	22950	112412	116605	-	HLC3	SLC4A1	SLC4A1AP BND3 WD-	-	Two-hybrid physical	Chen J (19	9422766	9606	9606	Low Throu-			
1958	602	580	107074	107056	-	-	BCL3	BARD1	BCL4 D19 -	-	Two-hybrid physical	Dechend I	10362352	9606	9606	Low Throu-		
2006	153	10755	106662	115978	-	-	ADRB1	GIPC1	B1AR ADNFIP1 SEMCT	Two-hybrid physical	Hu LA (20	12724327	9606	9606	Low Throu-			
2368	672	466	107140	106956	-	-	BRCA1	ATF1	PPP1R53 FUS ATF1	Two-hybrid physical	Houvaras Y	10945975	9606	9606	Low Throu-			
2398	672	4436	107140	110573	-	-	BRCA1	MSH2	PPP1R53 HNPPCC1	Two-hybrid physical	Wang Q (2	11498787	9606	9606	Low Throu-			
2411	672	580	107140	107056	-	-	BRCA1	BARD1	PPP1R53 -	-	Two-hybrid physical	Wu LC (15	8944023	9606	9606	Low Throu-		
2424	672	2956	107140	109211	-	-	BRCA1	MSH6	PPP1R53 HNPPCC5	Two-hybrid physical	Wang Q (2	11498787	9606	9606	Low Throu-			
2466	421	1013	106914	107448	-	-	ARVCF	CDH15	-	CDH3 MR	Two-hybrid physical	Kaufmann	11058098	9606	9606	Low Throu-		
2721	5092	775	111125	107229	-	-	PCBD1	CACNA1C	PCBD DC CACNL1A1	Two-hybrid physical	Waters PJ	11461190	9606	9606	Low Throu-			
2765	5664	823	111643	107273	-	PIG30	PSEN2	CAPN1	AD3L AD4CANPL1 C	Two-hybrid physical	Shinozaki	9852298	9606	9606	Low Throu-			
2785	825	7273	107275	113124	-	-	CAPN3	TTN	nCL-1 p94 CMH9 MY	Two-hybrid physical	Ono Y (19	9642272	9606	9606	Low Throu-			
2827	3708	767	109913	107222	-	-	ITPR1	CA8	Insp3r1 S CALS CA-	Two-hybrid physical	Hirota J (12611586	9606	9606	Low Throu-			
3024	9223	1499	114655	107880	-	OK/SW-cl.	MAGI1	CTNNB1	AIP-3 MA CTNNB	Two-hybrid physical	Dobrosots	10772923	9606	9606	Low Throu-			
3189	5925	1523	111860	107903	RP11-174I	-	RB1	CUX1	pp110 OS p75 Clos	Two-hybrid physical	Gupta S (2	12891711	9606	9606	Low Throu-			
3224	7251	1026	113102	107460	-	-	TSG101	CDKN1A	VPS23 TS MDA-6 C	Two-hybrid physical	Oh H (20	11943869	9606	9606	Low Throu-			
3312	4998	4171	111040	110339	-	-	ORC1	MCM2	ORC1 PAMITOTIN	Two-hybrid physical	Kneissl M	12614612	9606	9606	High Throu-			
3313	5000	4171	111042	110339	-	-	ORC4	MCM2	ORC4 OF MITOTIN	Two-hybrid physical	Kneissl M	12614612	9606	9606	High Throu-			
3315	4174	4171	110342	110339	RP5-824I1	-	MCM5	MCM2	P1-CDC46 MITOTIN	Two-hybrid physical	Kneissl M	12614612	9606	9606	High Throu-			
3317	8317	4171	113914	110339	-	-	CDC7	MCM2	Hsk1 huC MITOTIN	Two-hybrid physical	Kneissl M	12614612	9606	9606	High Throu-			

Protein
interactions



This is the **evidence view**. Different line colors represent the types of evidence for the association.



(requires Flash player 10 or better)

Your Input:

INO80 INO80 homolog (*S. cerevisiae*); DNA helicase and probable main scaffold component of the chromatin remodeling INO80 complex which is involved in transcriptional regulation, DNA replication and probably DNA repair; according to PubMed-20687897 the contribution to DNA double-strand break repair appears to be largely indirect through transcriptional regulation. Recruited by YY1 to YY1-activated genes, where it acts as an essential coactivator. Binds DNA. In vitro, has double stranded DNA-dependent ATPase activity. Involved in UV-damage excision repair, DNA replication and chromosome segreg [...] (1556 aa) (*Homo sapiens*)

Predicted Functional Partners:

		Neighborhood	Gene Fusion	Coccur	Coexpression	Experiments	Databases	Textmining	[Homology]	Score
INO80	RuvB-like 1 (<i>E. coli</i>); May be able to bind plasminogen at cell surface and enhance plasminogen [...] (456 aa)	● ● ●	● ● ●	● ● ●	● ● ●	● ● ●	● ● ●	● ● ●	● ● ●	0.999
NFRKB	nuclear factor related to kappaB binding protein; Binds to the DNA consensus sequence 5'-GGGGAA [...] (1324 aa)	● ● ●	● ● ●	● ● ●	● ● ●	● ● ●	● ● ●	● ● ●	● ● ●	0.999
ACTR8	ARP8 actin-related protein 8 homolog (<i>yeast</i>); Plays an important role in the functional organiz [...] (624 aa)	● ● ●	● ● ●	● ● ●	● ● ●	● ● ●	● ● ●	● ● ●	● ● ●	0.999
ACTR5	ARP5 actin-related protein 5 homolog (<i>yeast</i>); Proposed core component of the chromatin remodeli [...] (607 aa)	● ● ●	● ● ●	● ● ●	● ● ●	● ● ●	● ● ●	● ● ●	● ● ●	0.999
INO80C	INO80 complex subunit C; Proposed core component of the chromatin remodeling INO80 complex whic [...] (228 aa)	● ● ●	● ● ●	● ● ●	● ● ●	● ● ●	● ● ●	● ● ●	● ● ●	0.999
RUVBL2	RuvB-like 2 (<i>E. coli</i>); Possesses single-stranded DNA-stimulated ATPase and ATP- dependent DNA h [...] (463 aa)	● ● ●	● ● ●	● ● ●	● ● ●	● ● ●	● ● ●	● ● ●	● ● ●	0.998
ACTL6A	actin-like 6A; Involved in transcriptional activation and repression of select genes by chromat [...] (429 aa)	● ● ●	● ● ●	● ● ●	● ● ●	● ● ●	● ● ●	● ● ●	● ● ●	0.996
YY1	YY1 transcription factor; Multifunctional transcription factor that exhibits positive and negat [...] (414 aa)	● ● ●	● ● ●	● ● ●	● ● ●	● ● ●	● ● ●	● ● ●	● ● ●	0.989
INO80E	INO80 complex subunit E; Putative regulatory component of the chromatin remodeling INO80 comple [...] (244 aa)	● ● ●	● ● ●	● ● ●	● ● ●	● ● ●	● ● ●	● ● ●	● ● ●	0.978
MCRS1	microspherule protein 1; Modulates the transcription repressor activity of DAXX by recruiting i [...] (475 aa)	● ● ●	● ● ●	● ● ●	● ● ●	● ● ●	● ● ●	● ● ●	● ● ●	

Views:



Search for BRMS1L protein interactions



Protein interactions

<http://string-db.org/>

Drug information

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 screenshot shows the homepage of DrugBank Version 4.1. At the top, there's a navigation bar with links for Browse, Search, Downloads, About, Help, Tools, and Contact Us. Below the header, the DrugBank logo is displayed, followed by a brief description of the database: "DrugBank is a unique bioinformatics and cheminformatics resource that combines detailed drug target (i.e. sequence, structure, and pathway) information with comprehensive drug (i.e. chemical, pharmacological and pharmaceutical) data." It highlights that the database contains 7740 drug entries, 1584 FDA-approved small molecule drugs, 157 FDA-approved biotech (protein/peptide) drugs, 89 nutraceuticals, and over 6000 experimental drugs. The "Drug of the day" section features Ciclopirox, showing its chemical structure and a brief description: "Ciclopirox olamine (used in preparations called Buratain, Lopex, Mycostat, Penlac and Staprox) is a synthetic antifungal agent for topical dermatologic treatment of superficial mycoses. It is most useful against *Tinea versicolor*. [NIHpedia]".

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

Proteins
Drug
Identifiers

DrugBank



ID	Name	Gene Name	UniProt ID	GenBank Protein ID	GenBank Gene ID	UniProt ID	Uniprot	Species	PDB ID	GeneCard ID	GeneBank ID	HGNC ID	Species	Drug IDs
P45029	Peptidoglycan		15746007 (43022)	P45029				Homo sapiens						DB00163
P21113	Histidine-HDC		32109 X34297	P15113	DCHS_HUMAN	HDC	HDC	HGNC:4803	Human	DB00134; DB00117				
DB0132	Glutamin-GLU		6050008 AF310338	P01312	GLSL_HUMAN	GLS2	GLS2	HGNC:29570	Human	DB00142				
P04048	Coagulation F12A1		182509 M22801	P04048	F12A1_HUMAN	F12A1	F12A1	HGNC:3331	Human	DB00133; DB00340				
P30228	NFKB1_DNA		292242 U94218	P30228	NFKB2_HUMAN	NFKB2A	NFKB2A	HGNC:7673	Human	DB00133; DB00128; DB01110; DB01128; D				
P37039	Striatal 11031782		399462 L11708	P37039	DHB2_HUMAN	DHB2	DHB2	HGNC:5211	Human	DB00137				
P8757	Oxygenase_P450		183555 ML4699	P8757	PTG1_HUMAN	PTG1	PTG1	HGNC:5729	Human	DB00134; DB00111; DB00189; DB02089; D				
DB13413	NAD(P)H:quinone		1103520 U46490	P13413	QD3425	NDHM_HUMAN	NDHM	HGNC:7683	Human	DB00137; DB00141				
P17986	Alcohol dehydrogenase		178134 MS0477	P17986	ADHD_HUMAN	ADH3	ADH3	HGNC:259	Human	DB00137; DB001704; DB00123				
P48728	Aminopeptidase		591772 Q13871	P48728	CD36_HUMAN	AMT	AMT	HGNC:479	Human	DB00138; DB00137; DB00479				
P52113	Prostaglandin D2		796839 U07181	P52113	DRHA_HUMAN	DRHA	DRHA	HGNC:5384	Human	DB00157				
DB48644	Vitamin-B6-dependent		50563888 AF29133	DB48644	GACI_HUMAN	GACNAJ1	GACNAJ1	HGNC:1396	Human	DB00140; DB00167; DB00160; DB00095; D				
P28542	Adenosine A2A receptor		256155 54-235	P28542	AD2AR_HUMAN	ADORA2A	ADORA2A	HGNC:262	Human	DB00102; DB00177; DB001440; DB00065; D				
P66519	Tyrosine-tRNA		28237 1-416	P66519	ABLI_HUMAN	ABLI	ABLI	HGNC:76	Human	DB00170; DB00165; DB00124; DB00198; D				
P23115	High affinity FcR		131518 K19498	P23115	FCER2A_HUMAN	FCERA	FCERA	HGNC:369	Human	DB00040; DB00055; DB00079				
P68654	Coagulation Factor		182818 M-EL113	P68654	F4F_HUMAN	F4	F4	HGNC:3548	Human	DB00055; DB00106				
P23210	Prostaglandin PTG1		182018 M-EL122	P23210	PGM1_HUMAN	PTG1	PTG1	HGNC:9688	Human	DB00154; DB00159; DB00098; D				
P61226	Beta-actin ADRA2A2		162168 49117	P61226	ADRB2_HUMAN	ADRB2	ADRB2	HGNC:2992	Human	DB00175				
P61788	28S ribosomal RNA		42795 42583	P61788	RPS_ECOLI									
P21708	D(LA)-dop CR001		20117 K57670	P21708	DRD1_HUMAN	DRD1	DRD1	HGNC:3020	Human	DB00128; DB00124; DB00098; DB00098; D				
P23461	ThymidylylTMP1		752104 104238	P23461	TSY_CANAL									DB00189
P288292	Solute carrier SCLCJA1		1202132 AF280324	P288292	S32A1_HUMAN	SLC13A1	SLC13A1	HGNC:10816	Human	DB00113				
P39318	Vascular eNOS		297058 K59878	P39318	VGRD_HUMAN	FL14	FL14	HGNC:2387	Human	DB00098; DB01206; DB004879; DB002075; D				
P51168	Amiloride SCNN1B		1084271 XI7159	P51168	SCNN1B_HUMAN	SCNN1B	SCNN1B	HGNC:10800	Human	DB00184; DB00294				
P61632	Collagen I COL1A1		1428928 274615	P61632	COL1A1_HUMAN	COL1A1	COL1A1	HGNC:2337	Human	DB00044; DB00488				
DB48487	Tubulin beta TUBB		11230485 AU28757	DB48487	TBLB_HUMAN	TUBB	TUBB	HGNC:16257	Human	DB00106; DB01229; DB01194; DB01194; D				
P58101	NADH dehydrogenase		0052157 K59723	P58101	NDU1_HUMAN	NDUFV3	NDUFV3	HGNC:7723	Human	DB00137				
P35543	Phenylalanyl tRNA		3983103 AF897441	P35543	SYTH_HUMAN	PAR52	PAR52	HGNC:21902	Human	DB00125				
P27948	Protein F11T1		31432 1-31602	P27948	VSP1_HUMAN	FLT1	FLT1	HGNC:3783	Human	DB00358; DB001206; DB004879; DB002075; D				
P61615	Cysteineglutathione		5668545 AF003691	P61615	XCT_HUMAN	SCLCA11	SCLCA11	HGNC:13899	Human	DB00118; DB00140; DB00140; DB00075; D				
P09897	7-dehydro-DHCR7		4181598 AF095363	P09897	DHCR7_HUMAN	DHCR7	DHCR7	HGNC:2980	Human	DB00157				
P68722	Di-hydroxy DHRS4		4181596 AF041127	P68722	DHRS4_HUMAN	DHRS4	DHRS4	HGNC:16895	Human	DB00182				
P68213	Hydroxy DHRS4		3970706 M10851	P68213	DSR_HUMAN	DSR	DSR	HGNC:6091	Human	DB00090; DB00048; DB00047; DB00071; D				
P68480	Proline-rich 40kDa		2073558 U58480	P68480	EM8A_MYTU									Mycobacterium
DB02915	Cysteine-rich CTNS		3098803 Y12828	DB02915	CTNS_HUMAN	CTNS	CTNS	HGNC:2538	Human	DB00118				
P68489	Protein MAP1		33842 R33484	P68489	KAP1_HUMAN	KAP1	KAP1	HGNC:2952	Human	DB00358; DB004973; DB00290; DB03285; D				
P60217	NADH dehydrogenase		1593508 U65379	P60217	NDUFS2_HUMAN	NDUF58	NDUF58	HGNC:7723	Human	DB00157				

Protein-
Drug
relationship

DrugBank data are provided as a list of proteins paired with sets of drugs known to target the proteins

DrugBank



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

|| Protein-
Drug
relationship

- [Search for Acetazolamide](#)

Clinical database

The screenshot shows the homepage of the PharmGKB website (<https://www.pharmgkb.org/>). The page features a navigation bar at the top with links for Home, About, Contact, and Log In. Below the navigation, there are four main data points: "Drug Label Annotations" (780), "Clinical Guideline Annotations" (165), "Curated Pathways" (153), and "Annotated Drugs" (712). A central graphic illustrates the concept of pharmacogenomics as a network of genetic variations and their relationship to drug response. The text "WHAT IS PHARMACOGENOMICS?" is displayed above the network diagram, which is described as "The study of the relationship between genetic variations and how". To the right, a vertical bar chart shows two bars reaching the top of the scale. The PharmGKB logo and a brief description of the resource are also present.

Drug Label Annotations
780

Clinical Guideline Annotations
165

Curated Pathways
153

Annotated Drugs
712

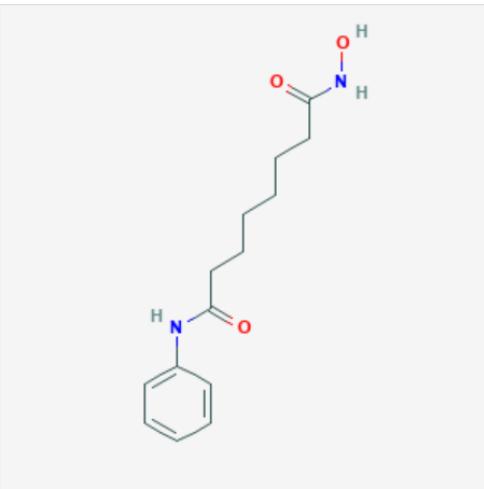
WHAT IS PHARMACOGENOMICS?
The study of the relationship between genetic variations and how

PHARMACOGENOMICS.
KNOWLEDGE.
IMPLEMENTATION.

PharmGKB is a comprehensive resource that

<https://www.pharmgkb.org/>

Structure



[large version](#)

[3D version](#)

source: [PubChem](#)

Overview >

Prescribing Info

Drug Label Annotations

Clinical Annotations

Variant Annotations

Literature

Pathways

Type : Drug

Synonyms

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

PharmGKB ID : PA164748224

SEARCH
FOR SAHA

GeneCardsSuite GeneCards MalaCards LifeMap Discovery PathCards GeneAnalytics GeneALaCart VarElect GenesLikeMe GeneLoc

Free for academic non-profit institutions. Other users need a [Commercial license](#). WEIZMANN INSTITUTE OF SCIENCE LifeMap SCIENCES

GeneCards® HUMAN GENE DATABASE

Keywords ▾ Search Term Advanced

Home User Guide Analysis Tools ▾ News And Views About ▾ My Genes Log In / Sign Up

INO80 Gene (Protein Coding)

INO80 Complex Subunit

Jump to section Aliases Compounds Disorders Domains Expression Function Genomics Localization Orthologs Paralogs Pathways Products Proteins Publications Sources Summaries Transcripts Variants

EMD MILLIPORE Proteins & Enzymes Antibodies Assays & Kits **ORIGENE** Proteins Antibodies Assays Genes shRNA Primers CRISPR **GenScript** Genes Peptides Proteins CRISPR

Aliases for INO80 Gene

Aliases for INO80 Gene

- INO80 Complex Subunit ^{2 3}
- INO80 Complex Subunit A ^{2 3 4}
- Putative DNA Helicase INO80 Complex Homolog 1 ^{3 4}
- INO80A ^{3 4}
- HINO80 ^{3 4}
- INOC1 ^{3 4}
- INO80 Complex Homolog 1 (S. Cerevisiae) ²
- INO80 Homolog (S. Cerevisiae) ²
- Homolog Of Yeast INO80 ³

DNA Helicase INO80 ³
INO80 Homolog ³
EC 3.6.1.23 ⁶³
EC 3.6.4.12 ⁴
KIAA1259 ⁴
EC 3.6.1 ⁶³
INOC1, ⁶

 **VarElect**
NGS PHENOTYPER
From exome gene list to disease gene, based on >100 GeneCards sources

External IDs for INO80 Gene

HGNC: 26956 Entrez Gene: 54617 Ensembl: ENSG00000128908 OMIM: 610169 UniProtKB: Q9ULG1

Previous HGNC Symbols for INO80 Gene

INOC1

Previous GeneCards Identifiers for INO80 Gene

GC15M039058, GC15M041271, GC15M018119

Export aliases for INO80 gene to outside databases

Gene/Protein catalog

- Search and learn about WDR76 gene

Genecards.org

The screenshot shows the NCBI GEO DataSets homepage. At the top, there's a blue header bar with the NCBI logo and the GEO logo. Below the header, a navigation bar includes links for GEO Publications, FAQ, MIAME, Email GEO, and Login. A breadcrumb trail shows the user is at NCBI > GEO > Info > About GEO DataSets. A prominent red banner at the top displays COVID-19 information from CDC and NIH, along with a link to NCBI SARS-CoV-2 content. The main content area has a light blue header titled "About GEO DataSets". Under this, there's a list of links: Background, GEO DataSets Results Page, GEO DataSet Record, GEO DataSet Analysis Tools (with sub-links for Find genes, Compare 2 sets of samples, Cluster heatmaps, and Experiment design and value distribution). Below this, a section titled "Background" provides a general overview of the database.

COVID-19 is an emerging, rapidly evolving situation.
Get the latest public health information from CDC: <https://www.coronavirus.gov>.
Get the latest research from NIH: <https://www.nih.gov/coronavirus>.
Find NCBI SARS-CoV-2 literature, sequence, and clinical content: <https://www.ncbi.nlm.nih.gov/sars-cov-2/>.

About GEO DataSets

- Background
- GEO DataSets Results Page
- GEO DataSet Record
- GEO DataSet Analysis Tools
 - Find genes
 - Compare 2 sets of samples
 - Cluster heatmaps
 - Experiment design and value distribution

Background

The GEO DataSets database stores original submitter-supplied records (Series, Samples and Platforms) as well as curated DataSets. See the [Overview](#) for information about these different records types and how they are related to each other.

Curated DataSets form the basis of GEO's advanced data display and analysis features, including tools to identify

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

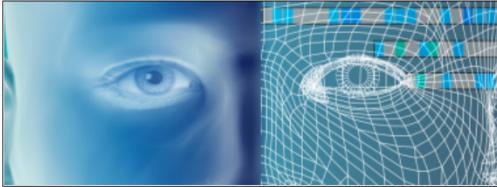
GEO (Gene Expression Omnibus)

NCBI Resources How To Sign in to NCBI

dbGaP dbGaP Search Limits Advanced Help

COVID-19 is an emerging, rapidly evolving situation. X

[Public health information \(CDC\)](#) | [Research information \(NIH\)](#) | [SARS-CoV-2 data \(NCBI\)](#) | [Prevention and treatment information \(HHS\)](#)



dbGaP

The database of Genotypes and Phenotypes (dbGaP) was developed to archive and distribute the data and results from studies that have investigated the interaction of genotype and phenotype in Humans.

Access dbGaP Data

[Advanced Search](#)
[Controlled Access Data](#)
[Public FTP Download](#)
[Collections](#)
[Summary Statistics](#)

Resources

[dbGaP Data Browser](#)
[Phenotype-Genotype Integrator](#)
[dbGaP RSS Feed](#)
[Software](#)
[dbGaP Tutorial](#)

Important Links

[How to Submit](#)
[FAQ](#)
[Code of Conduct](#)
[Security Procedures](#)
[Contact Us](#)

Genotype and Phenotype in Humans

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

COVID-19 is an emerging, rapidly evolving situation.

Get the latest public health information from CDC: <https://www.coronavirus.gov>.

Get the latest research from NIH: <https://www.nih.gov/coronavirus>.

Find NCBI SARS-CoV-2 literature, sequence, and clinical content: <https://www.ncbi.nlm.nih.gov/sars-cov-2>.

Advanced Search ?

cancer

ers

Disease/Focus (270)

Design (14)

Molecular Data Type (44)

MarkerSet (155)

Institute (21)

Consent (354)

Type (59)

try (9)

Save Results

Save Query

1/77 ▶

1 Clinical Cancer Sequencing

Accession	phs000694.v3.p2
Study Disease/Focus	Neoplasms
Study Design	Case Set
Study MarkerSet	Somatic_Mutations
Study Molecular Data Type	SNV (.MAF), WXS
Study Content	4 phenotype datasets, 10 variables, 1 molecular datasets, SRA, 15 subjects, 39 samples
NIH Institute	NHGRI
Study Consent	DS-CA-MDS --- Disease-specific (cancer, mds)
Release Date	2017-09-07
Embargo Release Date	2014-05-16
Related Terms	CA; CA - Cancer; Cancer; cell type cancer; organ system cancer; primary cancer ...

Translating whole exome sequencing (WES) for prospective clinical use may impact the care of cancer patients; however, multiple innovations are necessary for c implementation. These include: (1) rapid and robust ... alterations in 15/16 patients. Overall, this methodology may inform the widespread implementation of precision medicine. Principal Investigator: Levi...

[FileSelector](#) [RunSelector](#) [MeSH](#) [BioProject](#) [BioSample](#) [SRA](#)

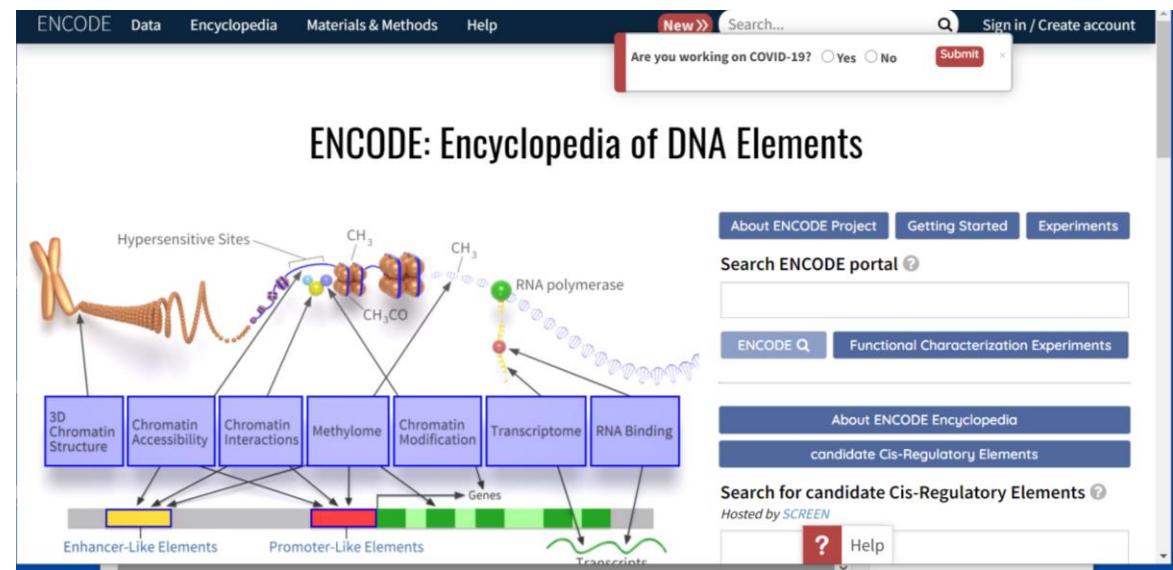
2 Breast Cancer Study

Accession	phs000807.v1.p1
Study Disease/Focus	Breast Neoplasms
Study Design	Prospective Longitudinal Cohort
Study MarkerSet	Human610-Quadv1_B
Study Molecular Data Type	SNP Genotypes (Array)
Study Content	4 phenotype datasets, 31 variables, 1 molecular datasets, 1914 subjects, 1914 samples
Ancestry (computed)	Population graph European (1612). African (7). East Asian (16). African American (169). Hispanic1 (21). Hispanic2 (58). Other Asian (16)

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>

Showing 25 of 842 results

Data Type View All (1)

Audit category: (1)

Audit category: (1)

Gene
Sin3A (<i>Drosophila melanogaster</i>) External resources: RefSeq:NM_165916.3 ↗ UniProtKB:A0A0B4K765 ↗ UniProtKB:O17142 ↗ UniProtKB:A0A0B4LF82 ↗ UniProtKB:B:077025 ↗ UniProtKB:A0A0B4LF93 ↗ UniProtKB:AIZ928 ↗ UniProtKB:O17143 ↗ UniProtKB:Q960C3 ↗ FlyBase:FBgn0022764 ↗ UniProtKB:Q5U0Y0 ↗ UniProtKB:A0A126GUN9 ↗ UniProtKB:A1Z927 ↗
Sin3a (<i>Mus musculus</i>) External resources: RefSeq:NM_001110350.1 ↗ RefSeq:NM_001110351.1 ↗ UniProtKB:Q60520 ↗ Vega:OTTMUSG00000036738 ↗ MGI:107157 ↗ ENSEMBL:ENSMUSG0000042557 ↗
SIN3B (<i>Homo sapiens</i>) External resources: MIM:607777 ↗ ENSEMBL:ENSG00000127511 ↗ UniProtKB:O75182 ↗ GeneCards:SIN3B ↗ HGNC:19354 ↗ RefSeq:NM_001297595.2 ↗ Vega:OTTHUMG00000182642 ↗ UniProtKB:MOQYC5 ↗ UniProtKB:B7Z392 ↗
SIN3A (<i>Homo sapiens</i>)

? Help

ENCODE Data Encyclopedia Materials & Methods Help New Search... Sign in / Create account

ENSEMBL:ENSG00000169375 ↗ RefSeq:NM_015477.3 ↗ GeneCards:SIN3A ↗ UniProtKB:Q96ST3 ↗ HGNC:19353 ↗ Vega:OTTHUMG00000142834 ↗ MIM:607776 ↗ RefSeq:NM_001145357.2 ↗

TF ChIP-seq of HepG2
Homo sapiens HepG2
Target: SIN3B
Lab: Michael Snyder, Stanford
Project: ENCODE

TF ChIP-seq of K562
Homo sapiens K562
Target: SIN3A
Lab: Michael Snyder, Stanford
Project: ENCODE
Experiment Series: ENCSR145QWE

</documents/31dba8f4-449d-4981-80c9-fe79d4b62f64/>
Raw data of protein groups in SIN3B_sc-13145 MS

TF ChIP-seq of HepG2
Homo sapiens HepG2
Target: SIN3A
Lab: Richard Myers, HAIB
Project: ENCODE

COSMIC (Catalog of Somatic Mutations in Cancer)

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

The screenshot shows the COSMIC website homepage. At the top, there is a navigation bar with links for Projects, Data, Tools, News, Help, About, Genome Version, a search bar, and a login link. The main content area features a banner for 'COSMIC v92, released 27-AUG-20'. Below the banner, there is a brief description of COSMIC as the world's largest resource for somatic mutations in cancer, followed by a search input field and a 'SEARCH' button. To the right, there is a 'COSMIC News' section with three items: 'COSMIC Mutational Signatures: Release V3.2 (March 2021)', 'COSMIC Actionability: Release v93 (March 2021)!', and a registration link for a COSMIC webinar. At the bottom, there is a cookie consent banner with 'Accept' and 'Cookie Preferences' buttons.

cancer.sanger.ac.uk/cosmic

COSMIC
Catalogue Of Somatic Mutations In Cancer

Projects ▾ Data ▾ Tools ▾ News ▾ Help ▾ About ▾ Genome Version ▾ Search COSMIC... **SEARCH** Login ▾

COSMIC v92, released 27-AUG-20

COSMIC, the Catalogue Of Somatic Mutations In Cancer, is the world's largest and most comprehensive resource for exploring the impact of somatic mutations in human cancer.

Start using COSMIC by searching for a gene, cancer type, mutation, etc. below.

eg *Braf*, *COLO-829*, *Carcinoma*, *V600E*, *BRCA-UK*, *Campbell* **SEARCH**

Projects

COSMIC is divided into several distinct projects, each presenting a separate dataset or view of our data:

COSMIC News

COSMIC Mutational Signatures: Release V3.2 (March 2021)
We'd like to share our exciting new developments and updates for COSMIC Mutational Signatures in our v3.2 release. [More...](#)

COSMIC Actionability: Release v93 (March 2021)!
After three years of high quality manual curation from PhD level experts, our latest product in the COSMIC suite, Mutation Actionability in Precision Oncology (Actionability), has launched. [More...](#)

Register now for the COSMIC webinar
Register now for the COSMIC webinar "Describing millions of somatic mutations at high resolution across every form of cancer underpins precision

Your choice regarding cookies on this site
We use cookies to optimise site functionality and give you the best possible experience.

Accept **Cookie Preferences**

← → ⌂ cancer.sanger.ac.uk/cosmic/search?q=WDR76

 COSMIC
Catalogue Of Somatic Mutations In Cancer

Projects ▾ Data ▾ Tools ▾ News ▾ Help ▾ About ▾ Genome Version ▾ Search COSMIC... **SEARCH** Login ▾

COSMIC search results

Your search term "**WDR76**" was an exact match for the COSMIC gene [WDR76](#).

A search of the whole COSMIC database returned results in **2** sections of the database. [More...](#)

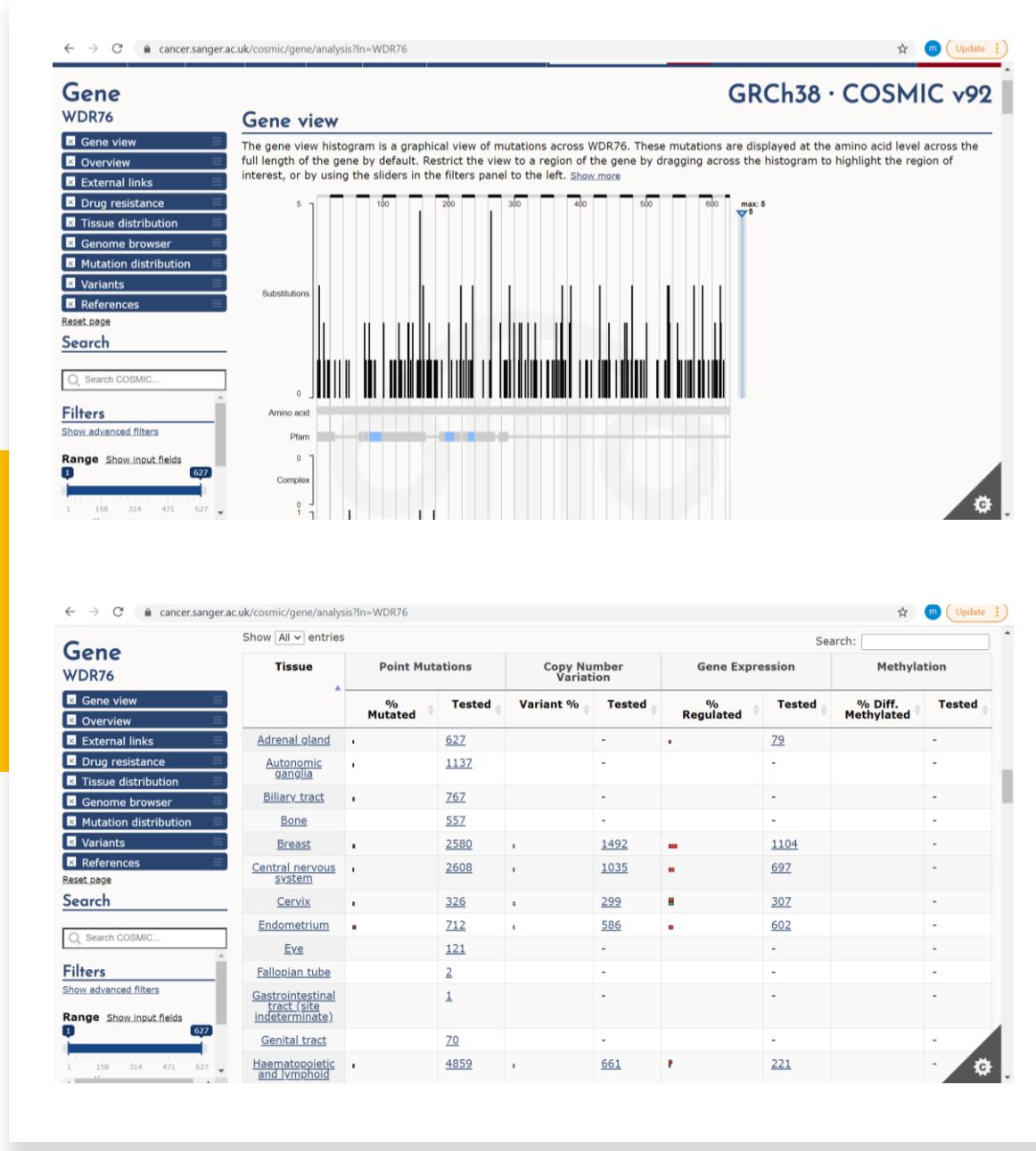
Genes (2 hits) Legacy Mutations (0) Mutations (712) SNPs (0) Cancer (0) Tumour Site (0) Samples (0) Pubmed (0) Studies (0)

Show 10 ▾ entries

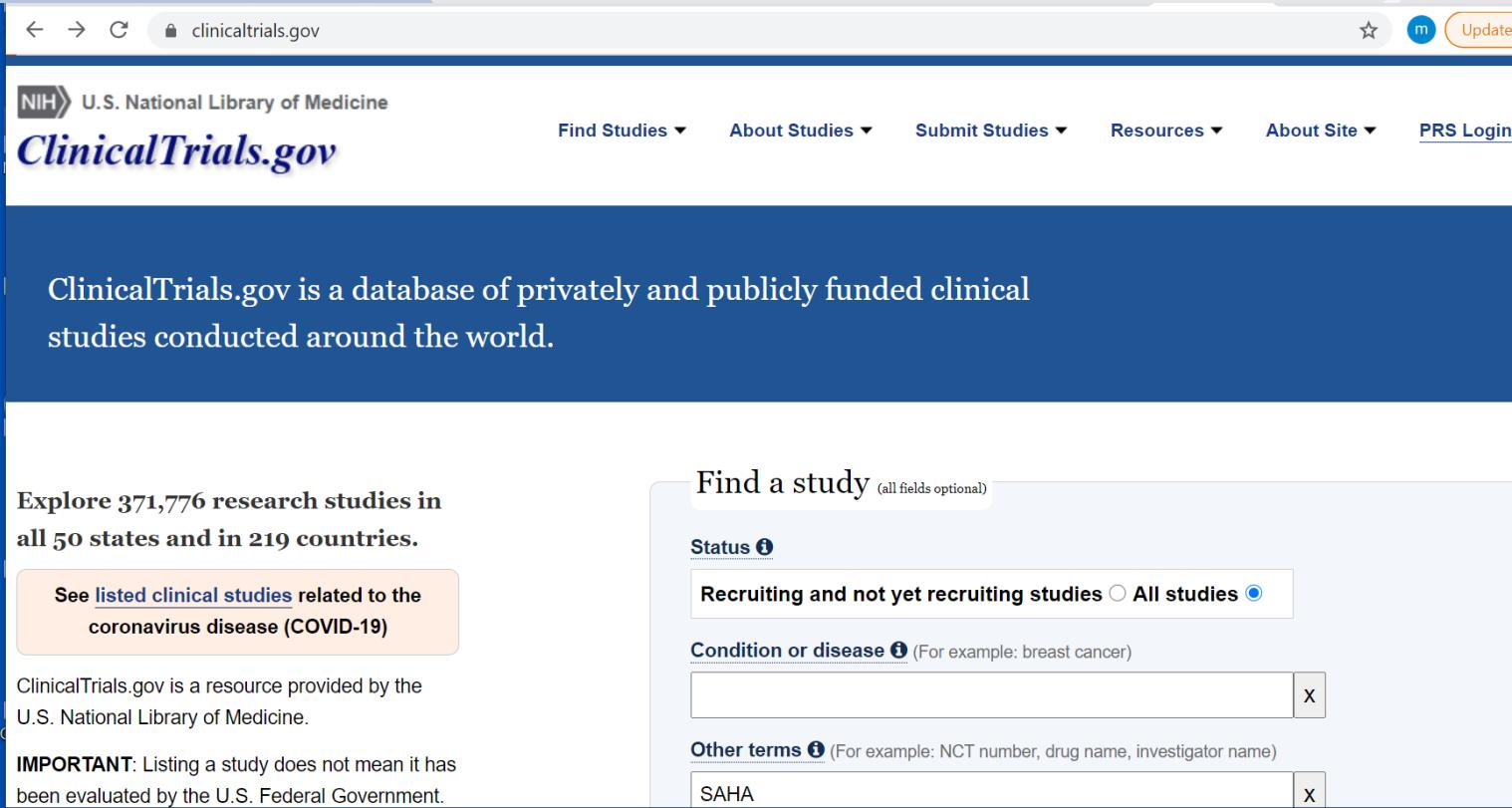
Gene	Alternate IDs	Tested samples	Simple Mutations	Fusions	Coding Mutations
WDR76	WDR76 ,ENST00000263795.10, WDR76 ...	38170	387	0	387
WDR76 ENST00000381246	WDR76 ENST00000381246,ENST00000381246.6, WDR76 ...	38170	381	0	381

Showing 1 to 2 of 2 entries First Previous **1** Next Last

Search for
WDR76



Search for WDR76



A screenshot of the ClinicalTrials.gov website. The header includes the NIH logo, U.S. National Library of Medicine, and the ClinicalTrials.gov logo. Navigation links include Find Studies, About Studies, Submit Studies, Resources, About Site, and PRS Login. A search bar at the top right has a star icon, a mobile icon, and an 'Update' button. The main content area features a dark blue banner stating: "ClinicalTrials.gov is a database of privately and publicly funded clinical studies conducted around the world." Below this, a section titled "Find a study" allows users to search by status (Recruiting and not yet recruiting studies is selected), condition or disease (e.g., breast cancer), and other terms (e.g., SAHA). A sidebar on the left promotes COVID-19 studies and provides a disclaimer about government evaluation.

clinicaltrials.gov

NIH U.S. National Library of Medicine

ClinicalTrials.gov

Find Studies ▾ About Studies ▾ Submit Studies ▾ Resources ▾ About Site ▾ PRS Login

ClinicalTrials.gov is a database of privately and publicly funded clinical studies conducted around the world.

Explore 371,776 research studies in all 50 states and in 219 countries.

See [listed clinical studies related to the coronavirus disease \(COVID-19\)](#)

ClinicalTrials.gov is a resource provided by the U.S. National Library of Medicine.

IMPORTANT: Listing a study does not mean it has been evaluated by the U.S. Federal Government.

Find a study (all fields optional)

Status i

Recruiting and not yet recruiting studies All studies

Condition or disease i (For example: breast cancer)

SAHA X

Other terms i (For example: NCT number, drug name, investigator name)

SAHA X

Clinical Trials

Resc x | Step x | Publ x | Gre x | Edit x | Get x | data x | Join x | http x | G kag x | CT Se x | +

← → ⌂ clinicaltrials.gov/ct2/results?recrs=&cond=&term=SAHA&cntry=&state=&city=&dist=

List By Topic On Map Search Details

◀ Hide Filters Download ⋮

Showing: 1-10 of 351 studies 10 studies per page

Filters

Row Saved Status Study Title Conditions Interventions

1 Terminated [Vorinostat \(SAHA\) in Uterine Sarcoma Has Results](#) • Leiomyosarcoma • Drug: Vorinostat Oral Capsule • Medical Clinic of Gynecol Graz, Au

2 Completed [Study on Efficacy and Tolerability of Vorinostat in Patients With Advanced, Metastatic Soft Tissue Sarcoma \(STS\)](#) • Soft Tissue Sarcoma • Drug: Vorinostat • Departn Oncolog Immuno Hospital Tübinge Württem

• Departn Hemost and Ster Transpla School Hannov German

Status

Recruitment :

- Not yet recruiting
- Recruiting
- Enrolling by invitation
- Active, not recruiting
- Suspended
- Terminated
- Completed
- Withdrawn
- Unknown status†

Expanded Access :

Search for
SAHA drug

List of multi-omics data repositories.

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

SEER CANCER- Demographic database

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



The screenshot shows the homepage of the SEER Incidence Data, 1975 - 2017. The top navigation bar includes links for Home, Cancer Statistics, SEER Data & Software, Registry Operations, News, and About. A search bar and user icons are also present. The main title "SEER Incidence Data, 1975 - 2017" is prominently displayed. On the left, a sidebar titled "SEER Incidence Database" lists links for Comparison of Data Products, How to Request Data Access, Frequently Asked Questions, Specialized Databases, Suggested Citations, and Contacts. A note in the center states: "Major changes were made to the SEER data release and authentication processes starting with the 1975-2017 SEER Data. Read the details on [Changes in the April 2020 SEER Data Release](#)." The main content area describes SEER's data collection from population-based cancer registries covering approximately 34.6 percent of the U.S. population. It highlights the SEER registries' collection of patient demographics, tumor site, morphology, stage at diagnosis, and treatment. The "SEER Data Products" section features a "View Comparison of Data Products" link. Below it, a paragraph explains the transition to two data products: SEER Research and SEER Research Plus, due to concerns about re-identifiability. It provides a link to the FAQ for more information.

seer.cancer.gov/data/

Home Cancer Statistics ▾ SEER Data & Software ▾ Registry Operations ▾ News About

SEER Incidence Data, 1975 - 2017

SEER Incidence Database

- Comparison of Data Products
- How to Request Data Access
- Frequently Asked Questions
- Specialized Databases
- Suggested Citations
- Contacts

i Major changes were made to the SEER data release and authentication processes starting with the 1975-2017 SEER Data. Read the details on [Changes in the April 2020 SEER Data Release](#).

SEER collects cancer incidence data from population-based cancer registries covering approximately 34.6 percent of the U.S. population. The [SEER registries](#) collect data on patient demographics, primary tumor site, tumor morphology, stage at diagnosis, and first course of treatment, and they follow up with patients for vital status.

SEER Data Products

[View Comparison of Data Products ▶](#)

There are now two data products available: SEER Research and SEER Research Plus. This was motivated because of concerns about the increasing risk of re-identifiability of individuals. The Research Plus 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](#) [Research Plus Database](#)

The screenshot shows a web browser window with the URL seer.cancer.gov/data/. The page has a dark blue header with navigation links: Home, Cancer Statistics ▾, SEER Data & Software ▾, Registry Operations ▾, News, and About. On the left, there's a sidebar with Suggested Citations and Contacts. The main content area compares two databases:

Research Database

- 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.
- 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>

The screenshot shows a web browser displaying the Census Bureau's data search results for "kansas". The URL in the address bar is data.census.gov/cedsci/all?q=kansas. The page features a dark blue header with the Census Bureau logo, a search bar containing "kansas", and a "SEARCH" button. Below the header, there are navigation links for "ALL", "TABLES" (which is underlined), "MAPS", and "PAGES". A message indicates "About 8,237 results" and a "Filter" link. On the left, a box highlights "2,913,314 Total Population in Kansas" from the "Source 2019 Population Estimates" (<https://www.census.gov/programs-surveys/popest.html>). On the right, a box titled "Kansas Profile" states that Kansas has a total area of 81,736.8 square miles, including 519.6 square miles of water, making it the 13th-largest state by area. The bottom section contains links for "Tables", "ANNUAL ESTIMATES OF THE RESIDENT POPULATION: APRIL 1, 2010 TO JULY 1, 2019 - FOR FULL ESTIMATES DETAIL, VISIT <https://www.census.gov/programs-surveys/popest.html>", "Survey/Program: Population Estimates", "Years: 2019", "Table: PEPANNRES", and "ACS DEMOGRAPHIC AND HOUSING ESTIMATES" for the American Community Survey (years 2019-2010, table DP05). A feedback link is at the bottom left, and a "Send Feedback" button is at the bottom right.

data.census.gov/cedsci/all?q=kansas

United States Census Bureau

kansas

SEARCH

ALL TABLES MAPS PAGES

About 8,237 results | Filter

EXPLORE DATA

2,913,314 Total Population in Kansas

Source 2019 Population Estimates
<https://www.census.gov/programs-surveys/popest.html>

Kansas Profile

Kansas has a total area of 81,736.8 square miles, including 519.6 square miles of water, making it the 13th-largest state by area.

Tables

ANNUAL ESTIMATES OF THE RESIDENT POPULATION: APRIL 1, 2010 TO JULY 1, 2019 - FOR FULL ESTIMATES DETAIL, VISIT <https://www.census.gov/programs-surveys/popest.html>

Survey/Program: Population Estimates
Years: 2019
Table: PEPANNRES

ACS DEMOGRAPHIC AND HOUSING ESTIMATES

Survey/Program: American Community Survey
Years: 2019,2018,2017,2016,2015,2014,2013,2012,2011,2010
Table: DP05

Send Feedback cedsci.feedback@census.gov

ON

EXPLORE DATA

Kansas Profile

Kansas has a total area of 81,736.8 square miles, including 519.6 square miles of water, making it the 13th-largest state by area.

Related Searches

Kansas Business and Economy

Kansas Education

Kansas Employment

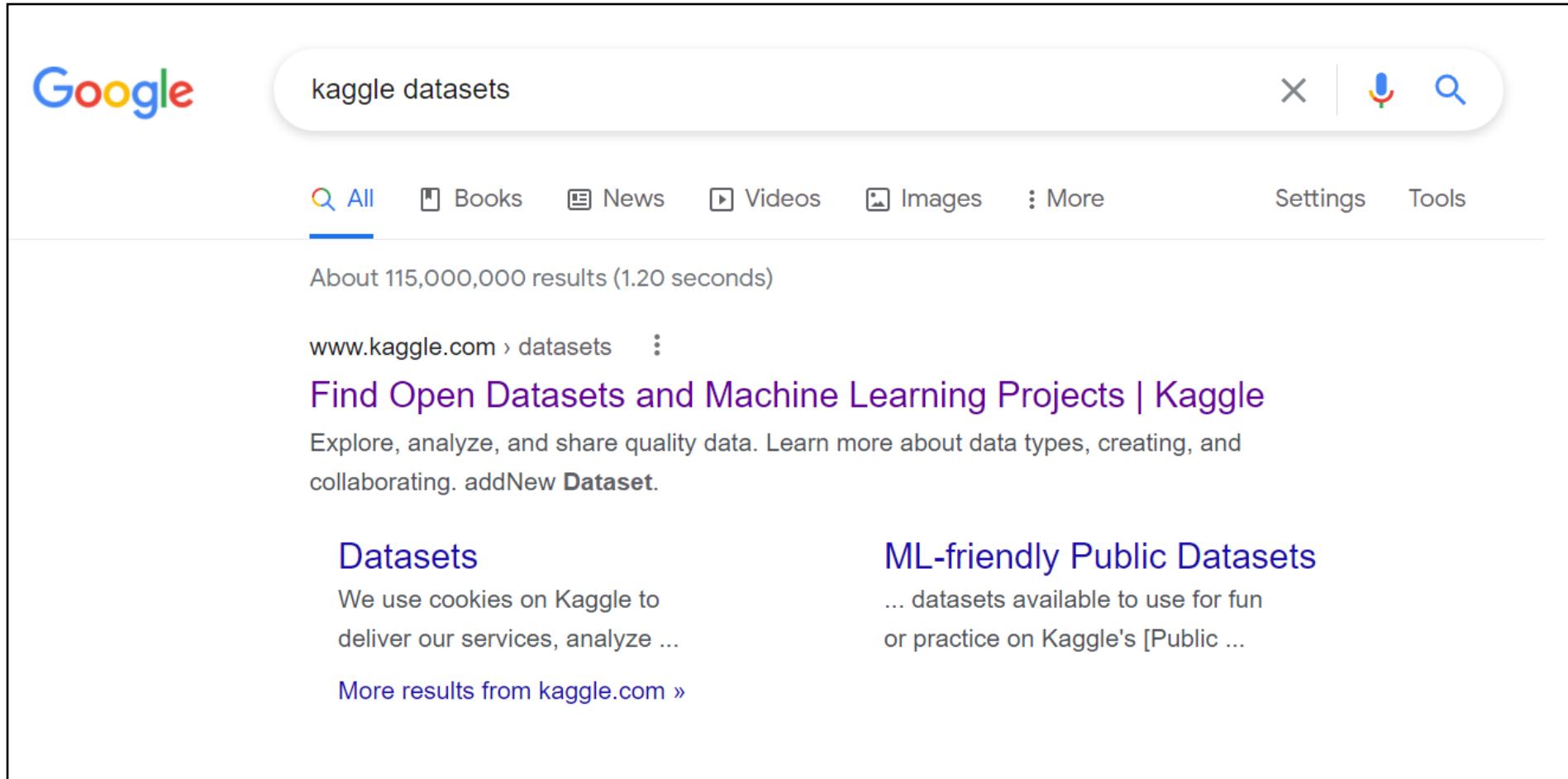
Kansas Families and Living Arrangements

Kansas Government

Kansas Health

Kansas Housing

Kaggle databases for Machine Learning



A screenshot of a Google search results page. The search bar at the top contains the query "kaggle datasets". Below the search bar, the "All" filter is selected, along with other options like Books, News, Videos, Images, More, Settings, and Tools. A message indicates there are about 115,000,000 results found in 1.20 seconds. The top result is a link to the Kaggle website: "Find Open Datasets and Machine Learning Projects | Kaggle". The description for this result mentions exploring, analyzing, and sharing quality data, learning about data types, creating, and collaborating, with a link to "addNew Dataset". Below this, two sections are visible: "Datasets" (describing cookie usage for service delivery) and "ML-friendly Public Datasets" (mentioning datasets available for fun or practice). A link "More results from kaggle.com »" is also present.

Google

kaggle datasets

All Books News Videos Images More Settings Tools

About 115,000,000 results (1.20 seconds)

www.kaggle.com › datasets

[Find Open Datasets and Machine Learning Projects | Kaggle](#)

Explore, analyze, and share quality data. Learn more about data types, creating, and collaborating. [addNew Dataset](#).

Datasets

We use cookies on Kaggle to deliver our services, analyze ...

[More results from kaggle.com »](#)

ML-friendly Public Datasets

... datasets available to use for fun or practice on Kaggle's [Public ...]

Datasets

Datasets X

76,684 Datasets

Hottest ▾

Grid List

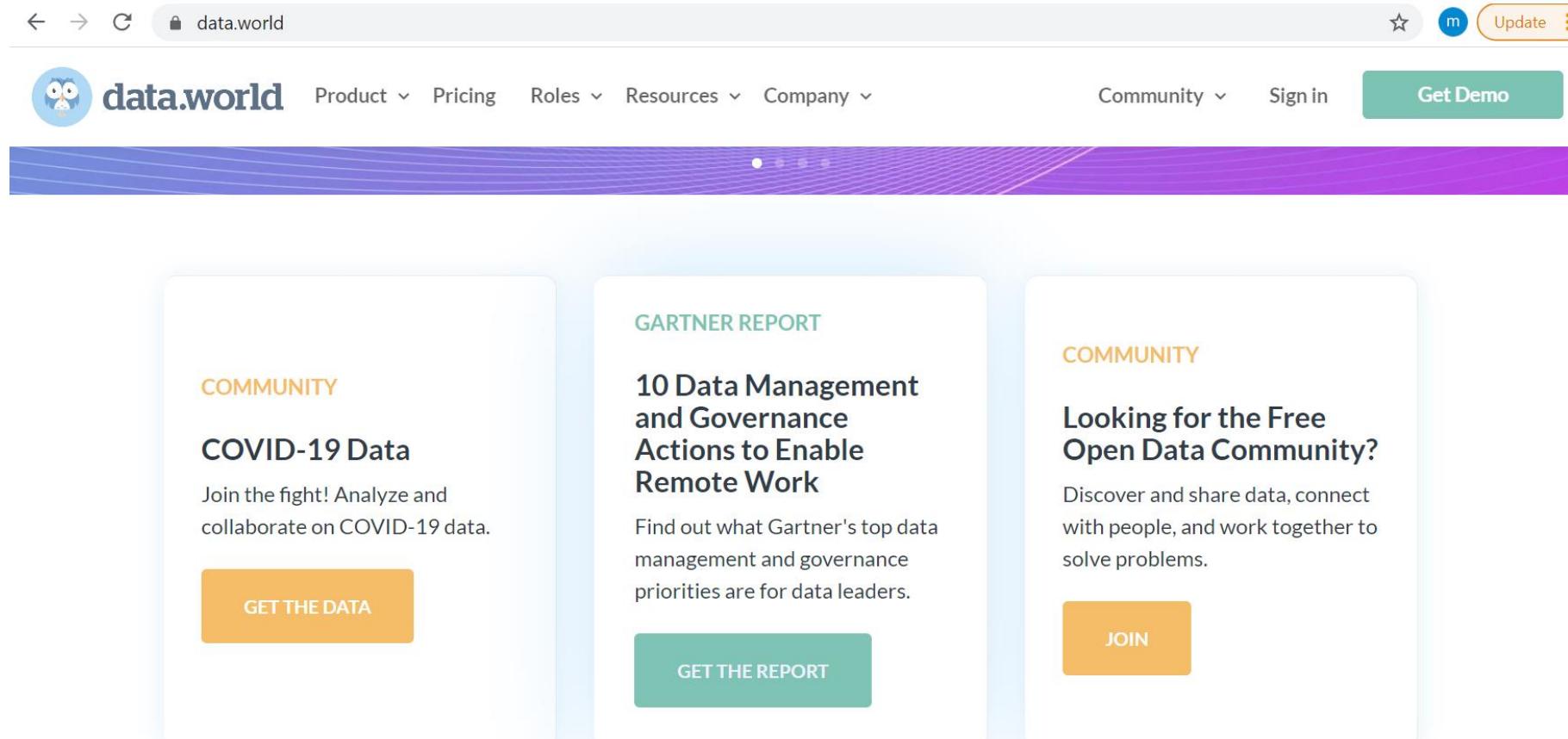
 **Reddit Vaccine Myths**
Gabriel Preda · Updated 6 hours ago
Usability **10.0** · 1 File (CSV) · 217 KB · 1 Task

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

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

Data.world

- <https://data.world/>



The screenshot shows the Data.world homepage. At the top, there's a navigation bar with links for Product, Pricing, Roles, Resources, Company, Community, Sign in, and a prominent green "Get Demo" button. Below the navigation is a banner featuring a blue and purple gradient background with abstract white lines. The main content area is divided into three sections:

- COMMUNITY**
COVID-19 Data

Join the fight! Analyze and collaborate on COVID-19 data.

[GET THE DATA](#)
- GARTNER REPORT**
10 Data Management and Governance Actions to Enable Remote Work

Find out what Gartner's top data management and governance priorities are for data leaders.

[GET THE REPORT](#)
- COMMUNITY**
Looking for the Free Open Data Community?

Discover and share data, connect with people, and work together to solve problems.

[JOIN](#)

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 Communication
 Labour	 Language and Linguistics	 Law	 Military	 Nature and Environment	 Persons	 Processes	 Science and Technology
 Society and Culture	 Transport						

Open Government Portal

Found 9969 records

 Clear all choices

green house emission facilities

Search

Order By

Best match ▾

 Download Search Results

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

Learn how the World Bank Group is helping countries with COVID-19 (coronavirus). [Find Out ➔](#)



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 absolute paths
- **Relative** – `setwd("./data")`, `setwd("../")`
- **Absolute** – `setwd("users/Mihaela/data/")`
- Important difference in Windows `setwd("C:\\Users\\Mihaela\\Downloads")`

Checking for and creating directories

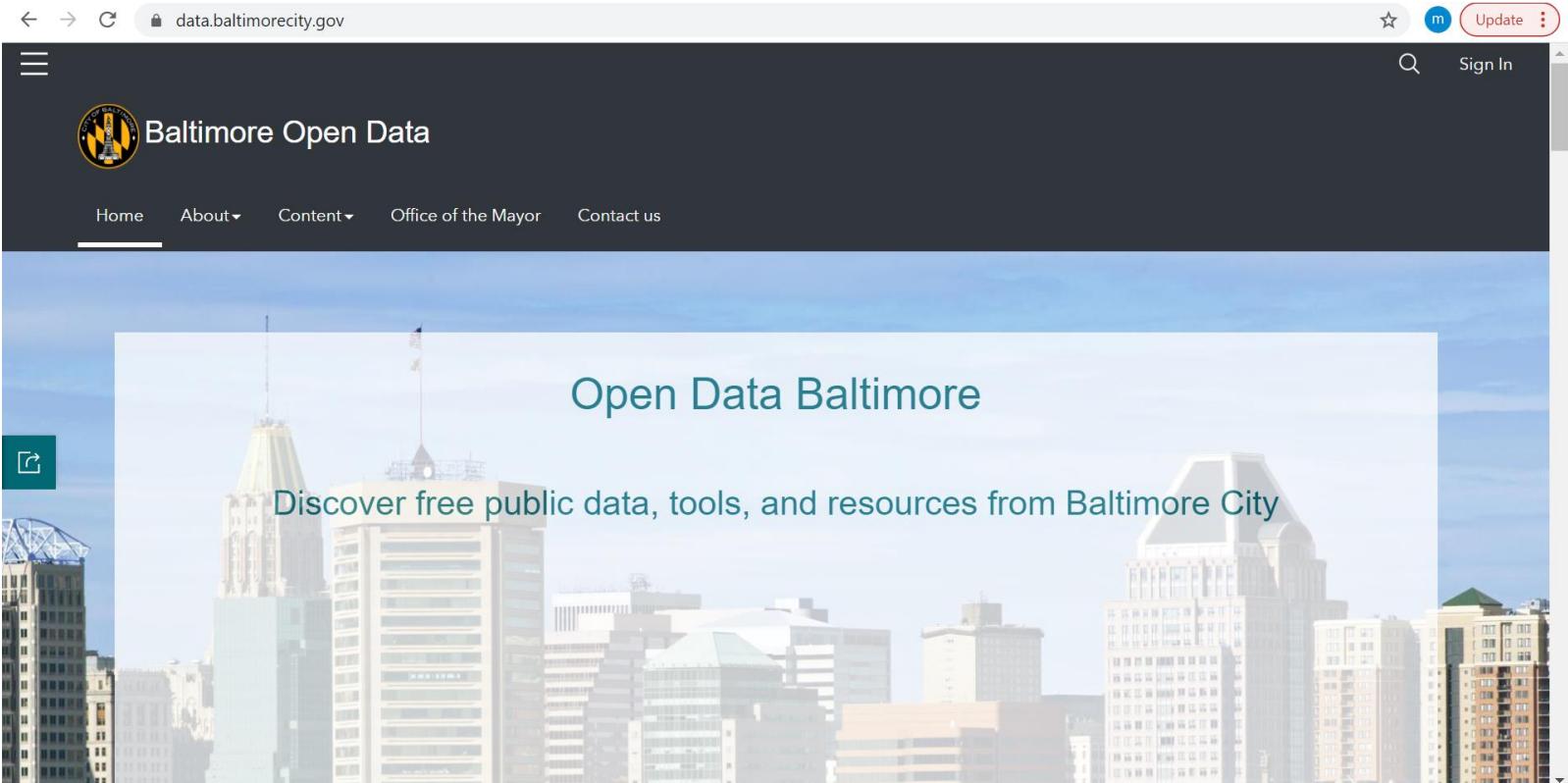
- `file.exists("directoryName")` will check to see if the directory exists
- `dir.create("directoryName")` will create a directory if it doesn't exist
- Here is an example checking for a "data" directory and creating if it doesn't exist

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

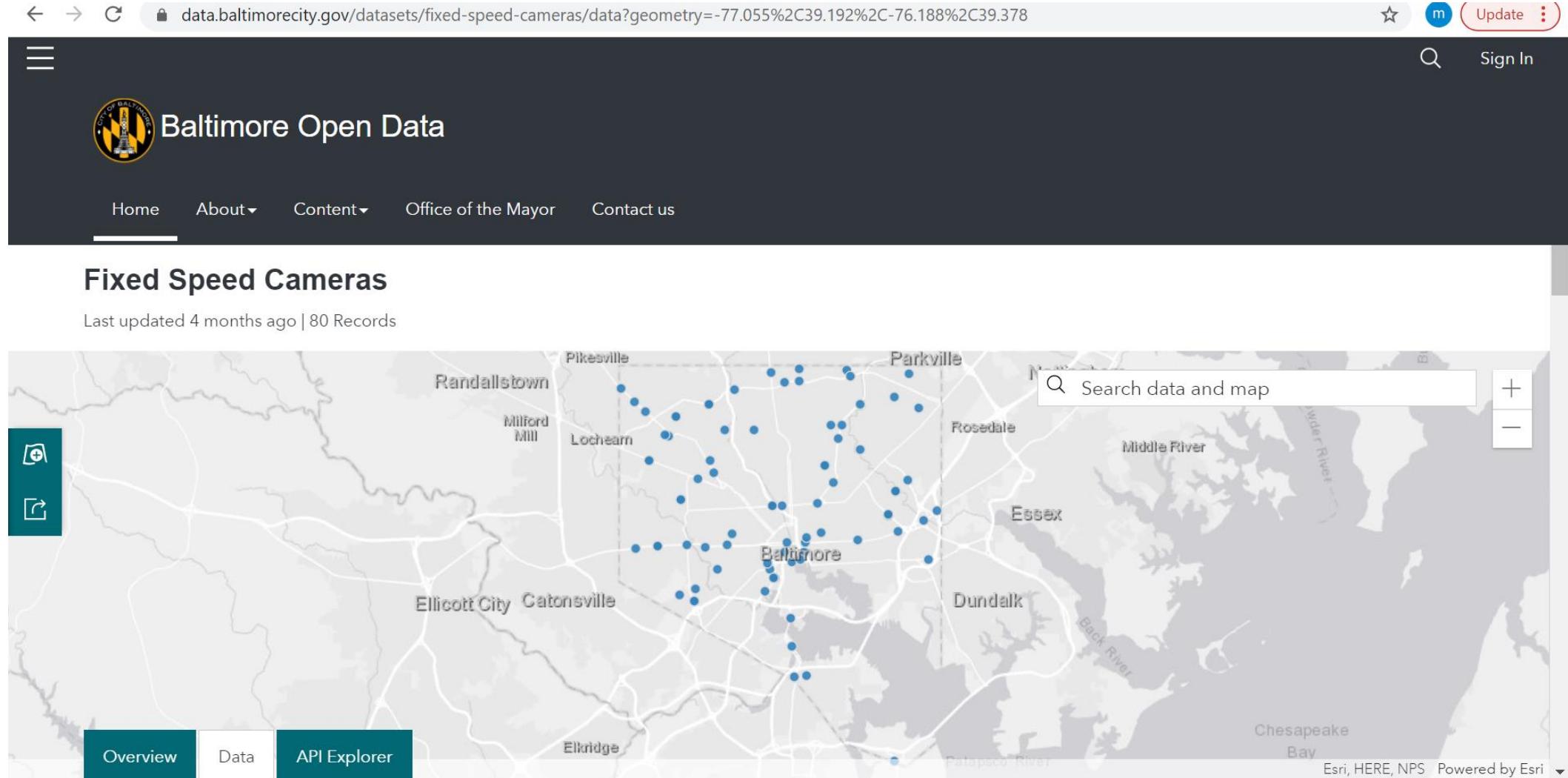
Getting data from the internet with download.file()

- 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"
```

```
> 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")
   % Total    % Received % Xferd  Average Speed   Time     Time      Current
          Dload  Upload Total   Spent    Left Speed
100 70668  100 70668     0      0  70668      0  0:00:01  0:00:01  --:--:-- 44361
>
```

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.

Reproducible Research

with R & RStudio

Christopher
Gandrud



[Overview](#) [Who is it for?](#) [Table of Contents](#) [Purchase](#) [Extra Materials](#) [Author](#) [Updates/Errata](#)

Overview of the Book

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 Chan⁴, Robyn Sayer⁴, Janiel Cragun⁴, Hope Cottrell⁴, Michael J Kelley², Rebecca Petersen⁵, David Harpole⁵, Jeffrey Marks⁵, Andrew Berchuck^{1, 6}, Geoffrey S Ginsburg^{1, 2}, Phillip Febbo^{1, 2, 3}, Johnathan Lancaster⁴ & 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.

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

Keith A. Baggerly and Kevin R. Coombes

Full-text: Open access

Enhanced PDF (1017 KB)

Abstract

Article info and citation

First page

References

Supplemental materials

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

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

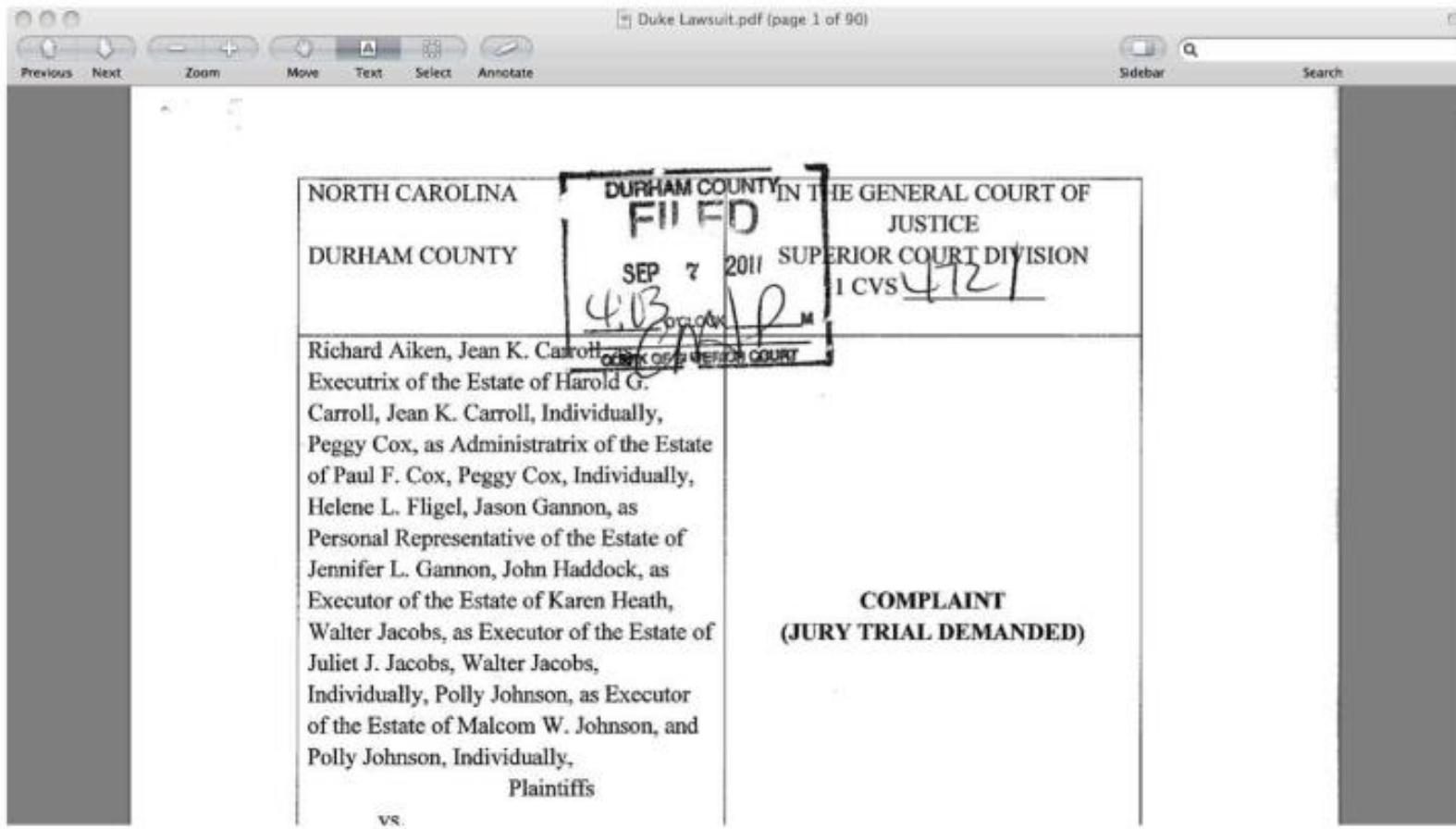
PO Box 9905 Washington DC 20016 Telephone 202-362-1809

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

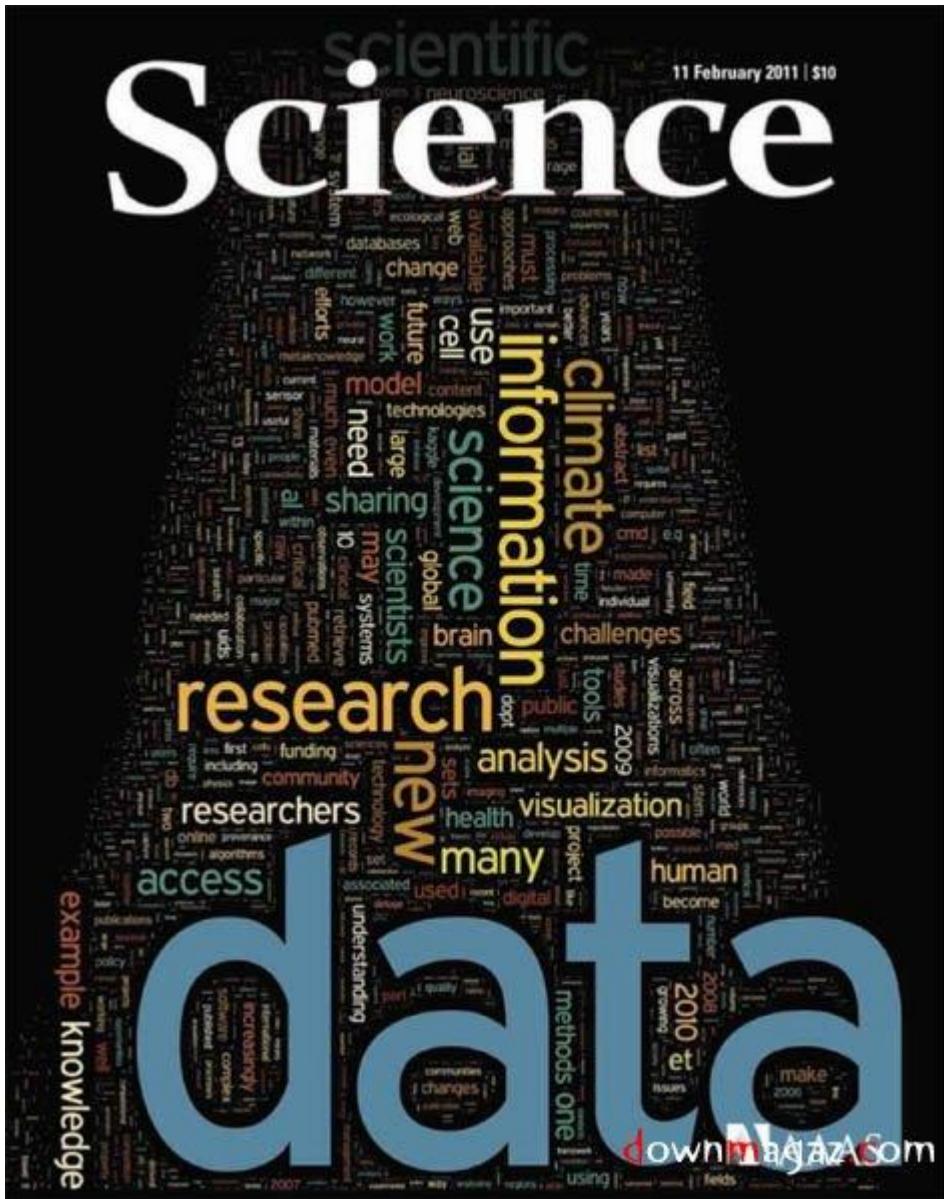
By Paul Goldberg

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

Science AAAS.org | FEEDBACK | HELP | LIBRARIANS All Science Journals Enter Search Term STOWERS INSTITUTE ALERTS ACCESS R

AAAS NEWS SCIENCE JOURNALS CAREERS MULTIMEDIA COLLECTIONS

Science The World's Leading Journal of Original Scientific Research, Global News, and Commentary.

Science Home Current Issue Previous Issues Science Express Science Products My Science About the Journal

Home > Science Magazine > 2 December 2011 > Peng, 334 (6060): 1226-1227

Article Views < Prev | Table of Contents | Next >
Abstract Leave a comment (2)

Full Text PERSPECTIVE
Full Text (PDF)
Figures Only

Article Tools Roger D. Peng
Author Affiliations
Leave a comment (2)
Save to My Folders
Download Citation
Alert Me When Article is Cited
Post to CiteULike
Article Usage Statistics
Email This Page

To whom correspondence should be addressed. E-mail: rpeng@hsph.edu

ABSTRACT

Computational science has led to exciting new developments, but the nature of the work has exposed limitations in our ability to evaluate published findings. Reproducibility has the potential to serve as a minimum standard for judging scientific claims when full independent replication of a study is not possible.

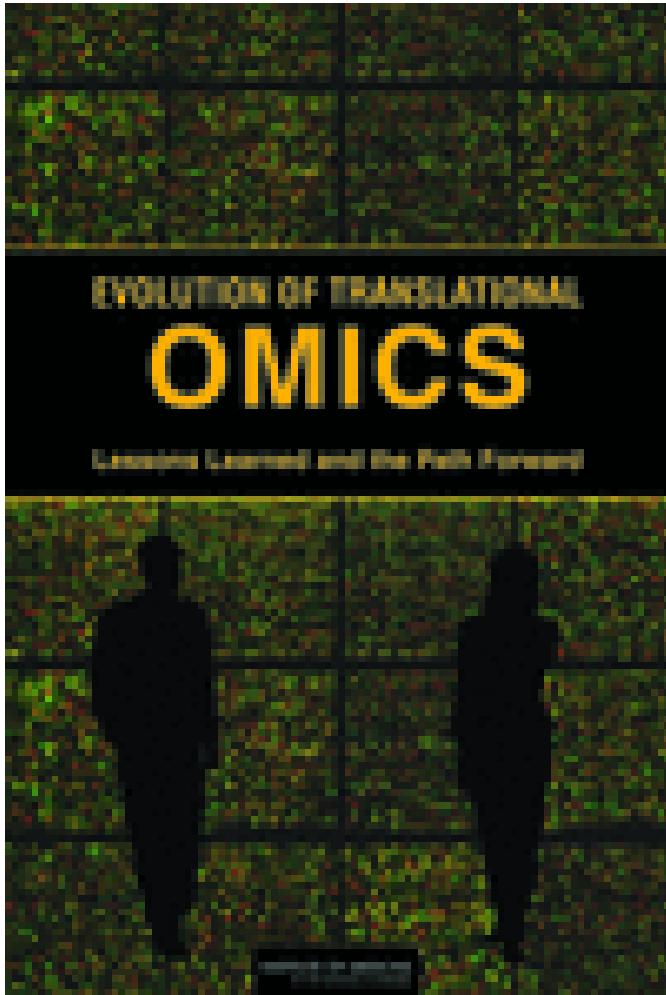
The journal *Biostatistics*, has implemented a policy for encouraging authors of accepted papers to make their work reproducible by others. Authors can submit their code or data to the journal for posting as supporting online material and can additionally request a “reproducibility review,” in which 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

Recommended to be done to improve reproducibility

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 | Press Room | Log In

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

ABOUT THE IOM REPORTS ACTIVITIES MEETINGS Explore by Topic

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 x PA1_template.Rmd x Knit HTML

```
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(suppressMessages(require(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$set(tidy=FALSE, fig.path='figures/')
17
18
19 unzip("U:/mis/HOPKINS/redata_data_activity.zip")
20 activity <- read.csv(file="U:/mis/HOPKINS/activity.csv",stringsAsFactors =
21 FALSE)
21 activity$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)
```

