UNIVERSITY OF Cincinnati

PUBLIC REPOSITORIES FOR DEPOSITION AND SHARING OF "OMICS" DATA

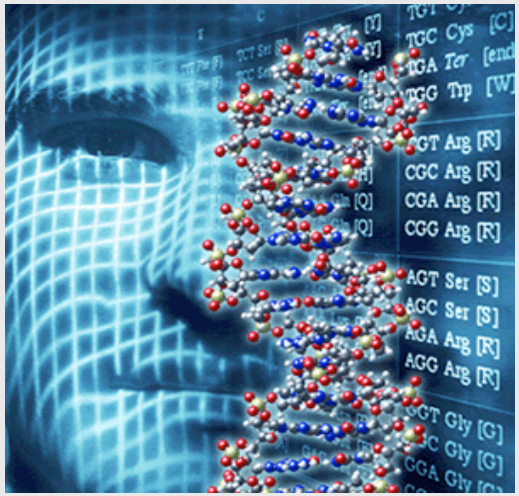
Scott M Langevin, PhD, MHA
Assistant Professor of Environmental Health
Division of Epidemiology
University of Cincinnati College of Medicine

October 16, 2015

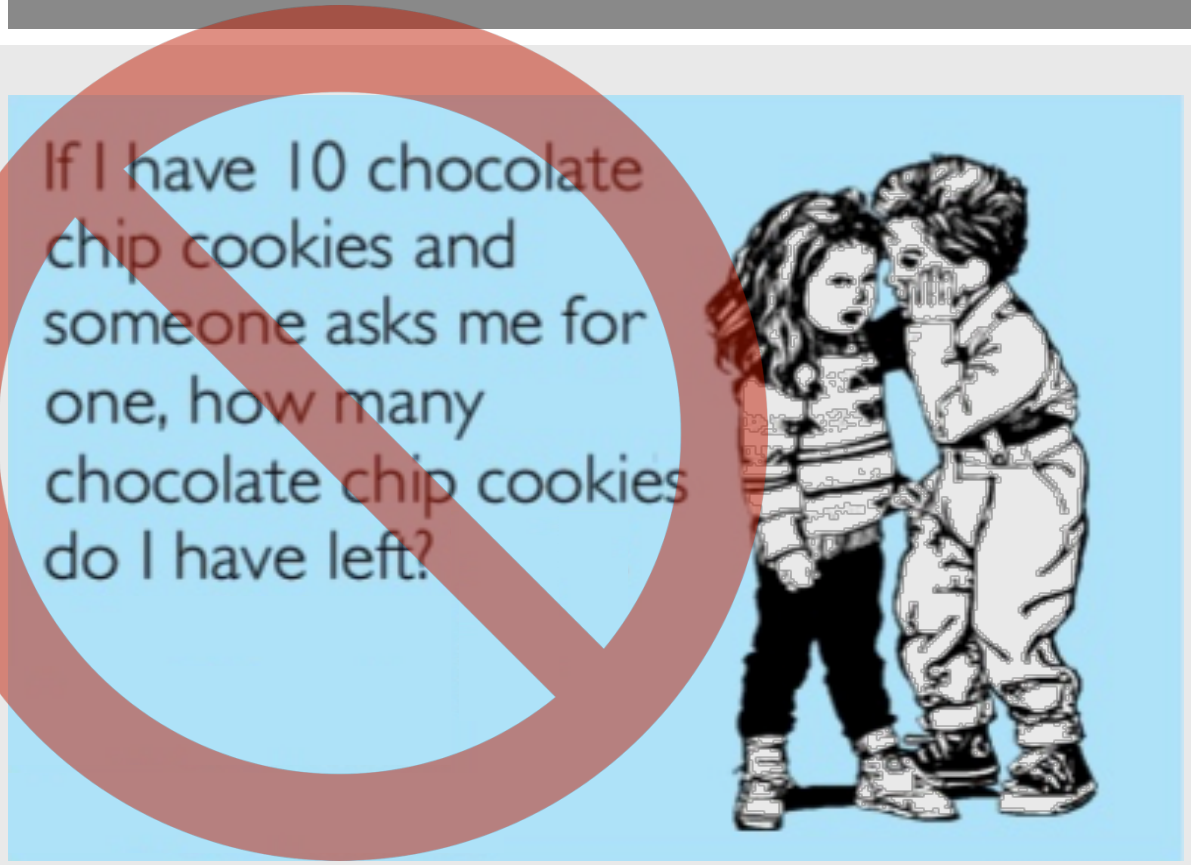
Email: langevst@uc.edu

Items To Be Discussed…

I. NIH Genomic Data Sharing Policy
II. Choosing a Repository for Data Deposition
III. How to Capitalize on Publicly-Available Data



NIH Genomic Data Sharing Policy



NO!

BAD!

New World of "Omics" and "Big Data"

- Recent explosion of genomic ("omics") technologies
- NIH has made a heavy investment in supporting these technologies
  - It is therefore understandable that they would want the fruits of these investments to become open-source

Okay - We Get It… Sharing Can Be Tough

But Sharing Omics Data is Now Compulsory!

This is the new reality, as it is now mandated for all extramural-funded projects…

So we have to deal with it!

(the truth is, many journals have required public deposition of "omics" data for some time now)

NIH GDS Policy (as of Aug 27, 2014)

http://grants.nih.gov/grants/guide/notice-files/NOT-OD-14-124.html

"The GDS Policy applies to all NIH-funded research (e.g., grants, contracts, intramural research) that generates large-scale human or non-human genomic data, **regardless of the funding level**, as well as the use of these data for subsequent research"

Effective for new applications as of: Jan 25, 2015

What Constitutes "Genomic" Data?

- Policy applies to **human** and **non-human** data
- GDP extends beyond genes and alleles:
  - GWAS/SNP arrays
  - Genomic Sequence (including whole-exome)
  - Epigenomics
  - Transcriptomics/expression arrays
  - Metagenomics (e.g. microbiomics)

Genomic Data Sharing (GDS) Plan

- Required for any project proposing to generate large-scale genomic data
  - NIH does not explicitly define "Large-scale"…

    (…but they have provided some examples - next slide)
  - If in doubt, contact your PO
- Must include GDS plan in the *Resource Sharing* section of new grant submissions involving generation of genomic data

Examples of Large-Scale Genomic Data (Per NIH GDS Supplemental Information)

- Sequence data from >1 gene (or comparable genomic regions) in >1,000 humans
- Sequence data from >100 genes (or comparable genomic regions) in >100 humans
- Data from ≥300,000 variant sites >1,000 humans
- Sequence data from >100 isolates from infectious organisms
- Sequence data from >100 metagenomes of human or model organism microbiomes
- Sequence data from >100 metatranscriptomes of human or model organism microbiomes
- Whole-genome or -exome sequence data of >1 model organism species/strains
- Comprehensive catalog of transcripts or ncRNA from ≥1 model organism species/strains
- Catalog of >100,000 SNPs from ≥1 model organism species/strains
- Comparisons of genome-wide methylated sites across >10 cell types
- Comparisons of differentially methylated sites genome-wide at single-base resolution within a given sample (e.g., within the same subject over time or across cell types within the same subject)

GDS Plan Elements

Genomic data sharing plans must include:

1) Data source (i.e. species, tissue source, and type of genomic data)
2) The data repository where the data will be submitted
3) Data submission and release dates
4) IRB assurance (if constitutional human genomic data)
5) Justification for any data sharing restriction(s)
6) Request for an exception to submit human genomic data (only if the study will not meet NIH's institutional certification criteria)

What You Need to Share:

Genomic data (e.g. SNP, sequence, array, etc…)

AND

Relevant phenotype data

Relevant exposure data

**You should provide adequate data so that an independent investigator can reproduce your associated analyses*

Timing of the Data Release

- Most of the repositories allow you to set a date for public release of the data (often 1-year post deposition)
  - Could generate a private link to placate peer-review eds.
- **HOWEVER** – the new NIH GDS policy specifies that data should be made publicly available at the time of publication
  - This applies to both human and non-human data

De-Identification of Human Data

- Human genomic datasets should be de-identified prior to submission
  - Identifiers in the phenotype data should also be stripped in accordance with HIPAA regulations
  - De-identified IDs should be randomly generated to link genomic data with relevant phenotype or exposure data for the submission

IRB Assurance for Human Data

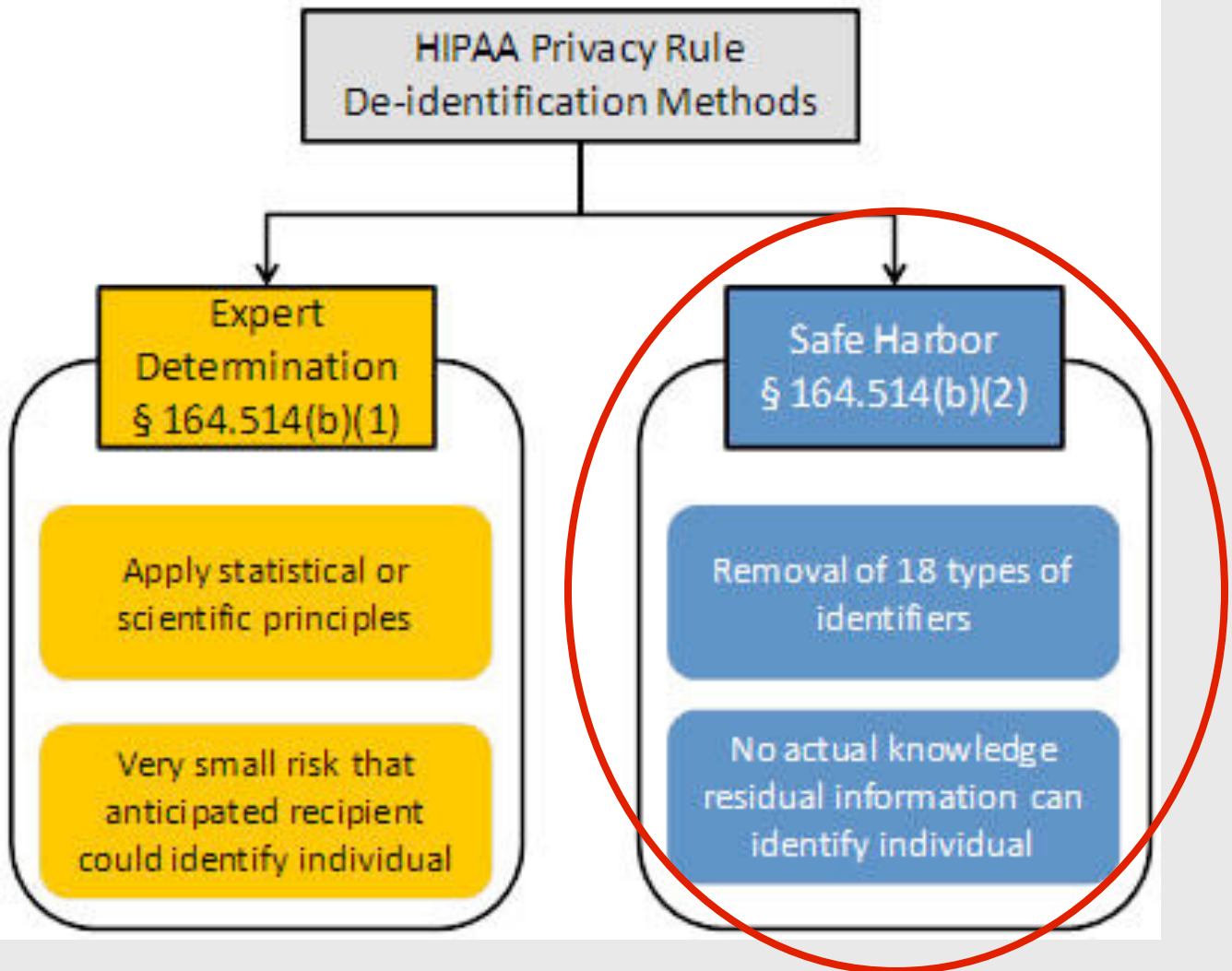
- For studies on existing cohorts, the IRB should be consulted to determine the appropriateness of GDS for secondary use
- For NIH-funded studies initiated after 01/25/2015, the expectation is that Investigators will obtain consent from participants for future use and broad sharing in accordance with the GDS Policy

Institutional Certification for Human Data

The Institutional Signing Official should provide **Institutional Certification** prior to award stating:

- Controlled vs. unrestricted access
- Data submission is consistent with applicable national, tribal, and state laws and regulations as well as relevant institutional policies
- Limitations on the research use of the data (if applicable)
- Identities of research participants will not be disclosed to NIH-designated data repositories
- An IRB has reviewed the investigator's proposal for data submission and assures that the protocol for the collection of genomic and phenotypic data is consistent with the *Basic HHS Policy for Protection of Human Research Subjects* (45 CFR Part 46)

HIPAA Standards for De-Identification

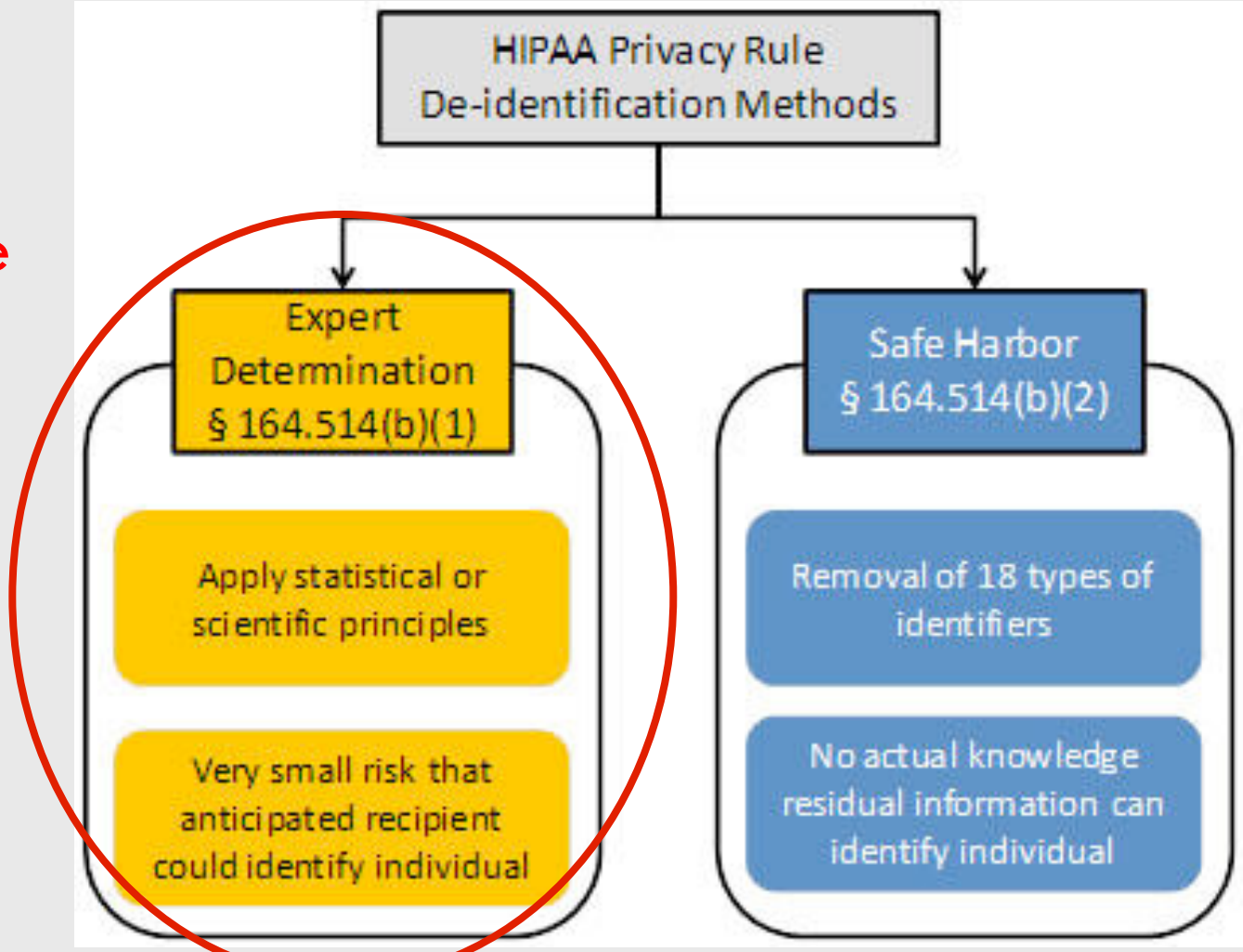


HIPAA "Safe-Harbor" Identifiers

(A) Names	
(B) All geographic subdivisions smaller than a state, including street address, city, county, precinct, ZIP code, and their equivalent geocodes, except for the initial three digits of the ZIP code if, according to the current publicly available data from the Bureau of the Census: (1) The geographic unit formed by combining all ZIP codes with the same three initial digits contains more than 20,000 people; and (2) The initial three digits of a ZIP code for all such geographic units containing 20,000 or fewer people is changed to 000	
(C) All elements of dates (except year) for dates that are directly related to an individual, including birth date, admission date, discharge date, death date, and all ages over 89 and all elements of dates (including year) indicative of such age, except that such ages and elements may be aggregated into a single category of age 90 or older	
(D) Telephone numbers	(L) Vehicle identifiers and serial numbers, including license plate numbers
(E) Fax numbers	(M) Device identifiers and serial numbers
(F) Email addresses	(N) Web Universal Resource Locators (URLs)
(G) Social security numbers	(O) Internet Protocol (IP) addresses
(H) Medical record numbers	(P) Biometric identifiers, including finger and voice prints
(I) Health plan beneficiary numbers	(Q) Full-face photographs and any comparable images
(J) Account numbers	(R) Any other unique identifying number, characteristic, or code, except as permitted by paragraph (c) of this section [Paragraph (c) is presented below in the section "Re-identification"]; and
(K) Certificate/license numbers	

HIPAA Standards for De-Identification

Hmm…
This could be a problem



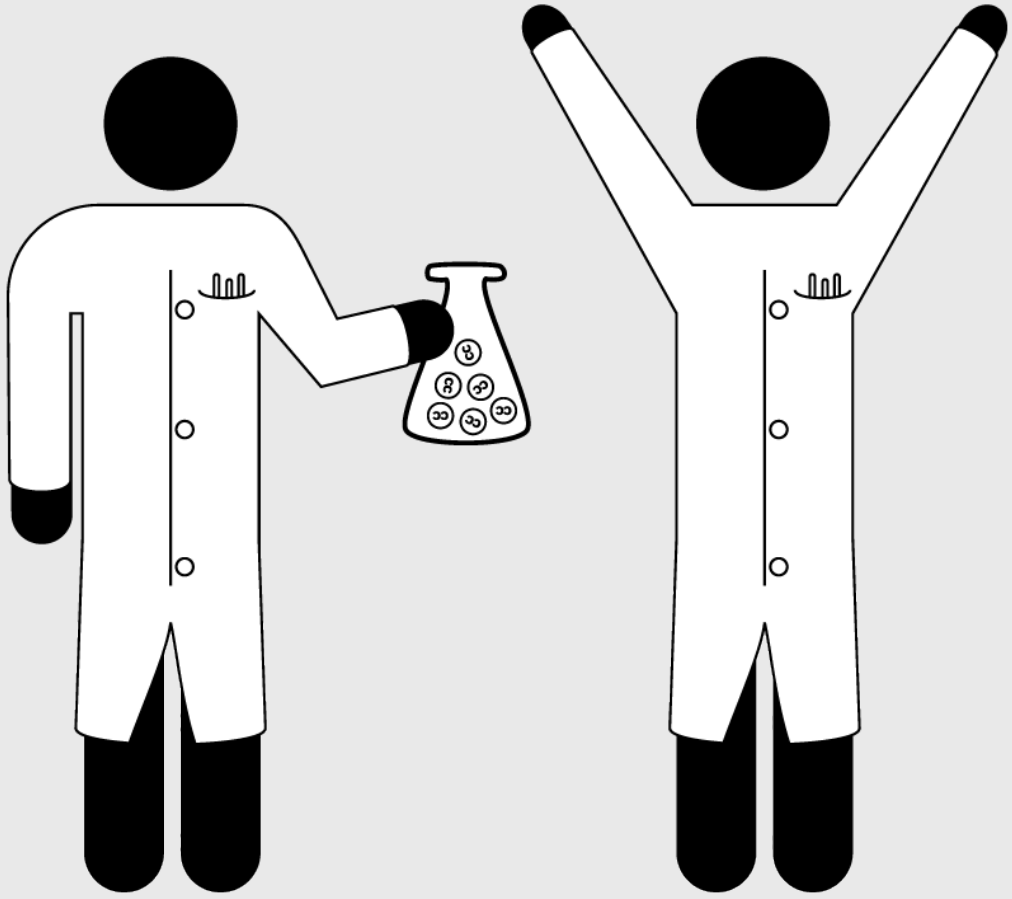
Certificate of Confidentiality

- Constitutional genomic data could potentially be used to re-identify the person
  - NIH has obtained a Certificate of Confidentiality (HG-2009-01) for dbGaP as an additional safeguard
    - Allows dbGaP to refuse to provide genotype/phenotype data to non-research entities
  - NIH further encourages investigators/institutions submitting large-scale genomic data to do the same

Synopsis of GDS Requirements

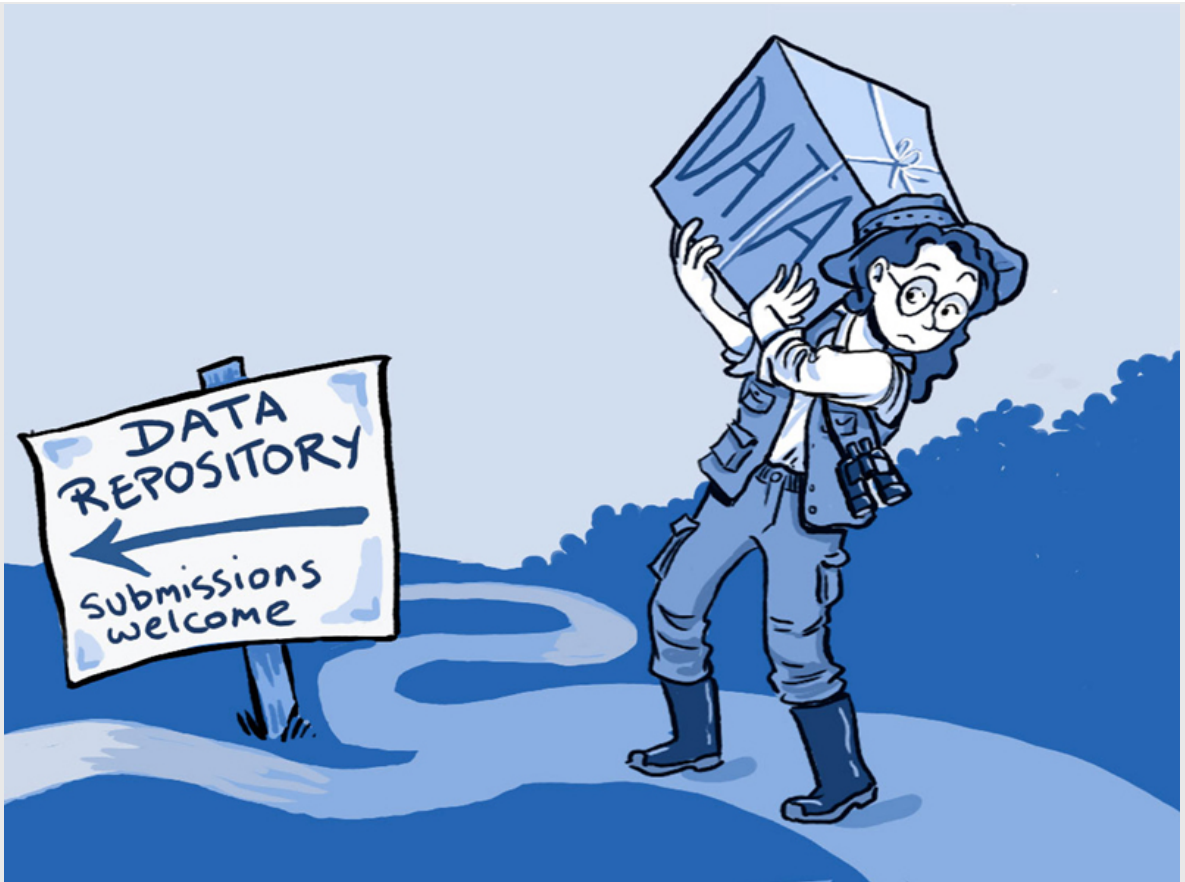
- GDS plan is now required for all large-scale proposals involving generation of genomic data
  - Includes both human and non-human sources
  - "Genomic" = GWAS, sequencing, epigenomic, transcriptomic, & metagenomic
- PI is responsible for the GDS plan
  - Check with PO ahead of submission to discuss expectations for the project GDS plan and timeline
- Data should be made available no later than the date of publication
- For work involving human subjects: Inquire with IRB to determine appropriateness of public deposition of data from existing cohorts and to craft adequate consent forms moving forward

Remember:



sharing is caring

Choosing a Repository



Repository Selection

- Non-human data can be made available through *any* commonly-used data repositories, regardless of whether or not it is NIH-funded
- Human data must be deposited in a NIH-designated repository

NIH-Designated Repositories

- Database of Genotypes and Phenotypes (dbGaP): NIH database at NCBI originally designed to archive and distribute coded genotype, phenotype, exposure, and pedigree data from genome-wide association studies. dbGaP now accepts additional types of data such as copy number variants and large-scale sequencing
- European Nucleotide Archive (ENA): database at the European Molecular Biology Laboratory -European Bioinformatics Institute (EMBL-EBI) that collects, maintains, and presents comprehensive sequencing information--including raw sequencing data, sequence assembly information, and functional annotation--as part of the permanent public scientific record

NIH-Designated Repositories

- Gene Expression Omnibus (GEO): NIH data repository that archives and distributes microarray, next-generation sequencing, and other forms of high-throughput functional genomic data
- Array Express: NIH-funded database at the EMBL-EBI that collects and disseminates microarray gene-expression data
- Sequence Read Archive (SRA): NIH's primary archive of high-throughput sequencing data at the National Center for Biotechnology Information (NCBI). SRA stores raw sequencing data as well as alignment information in the form of read placements on a reference sequence.

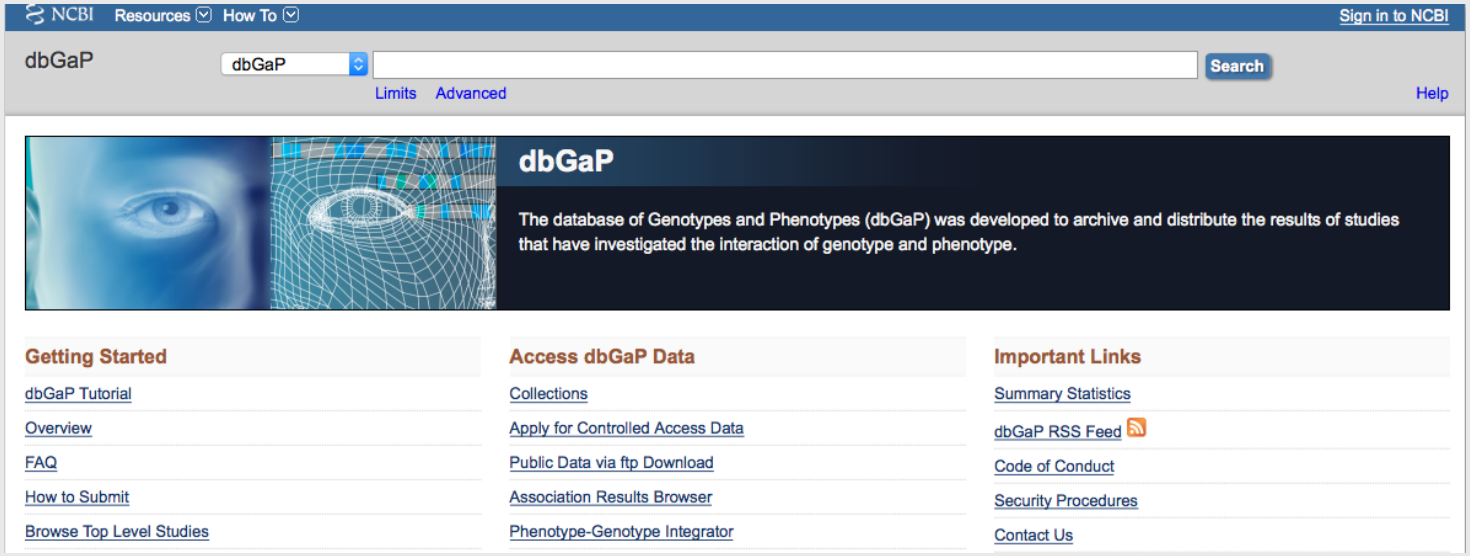
NIH-Designated Repositories

For a more extensive list of NIH-designated repositories, please refer to:

https://gds.nih.gov/02dr2.html

dbGaP Registration

- All GDS plans must be *registered* with dbGaP regardless of the repository in which they will ultimately be *deposited*
  - This should be done by the time the data cleaning and QC process begins (i.e. preprocessing)



Plan Ahead!

- Know the required format for the data that you plan to deposit
  - For example, for methylation arrays, GEO requires:
    - Raw signal intensity, detection p-value and average beta for each sample
    - Beta values post-processing (i.e. after any normalization, filtering, batch-adjustment, etc.)
  - Easier to procure at the time of preprocessing
- Makes life MUCH simpler when it comes time to submit the data for deposition!!!

Putting Publicly Available Data to Work

What Resources Are Out There?

- Remember: any of the aforementioned Repositories can also be searched for downloadable data

GENOMIC DATA SHARING

- Excellent resources for validation or generation of preliminary data
- Vast amounts of integrated genomic data is also available through TCGA

The Cancer Genome Atlas (TCGA)

- Joint effort between NCI/NHGRI
- Publically available genomic, epigenomic, RNA/ miRNA expression, proteomic, and clinical data
- Some "normal" tissues but mostly matched
- Can download data via TCGA Data Portal:

  https://tcga-data.nci.nih.gov/tcga/

Home

TCGA Data Portal Overview

The Cancer Genome Atlas (TCGA) Data Portal provides a platform for researchers to search, download, and analyze data sets generated by TCGA. It contains clinical information, genomic characterization data, and high level sequence analysis of the tumor genomes.

Please note some data on the TCGA Data Portal are in controlled-access. Please visit the Access Tiers page for more information.

The TCGA Data Portal does not host lower levels of sequence data. NCI's Cancer Genomics Hub (CGHub) is the new secure repository for storing, cataloging, and accessing BAM files and metadata for sequencing data.

Download Data

Choose from four ways to download data

Available Cancer Types	# Cases Shipped by BCR*	# Cases with Data*	Date Last Updated (mm/dd/yy)
Acute Myeloid Leukemia [LAML]	200	200	10/08/15
Adrenocortical carcinoma [ACC]	80	80	10/08/15
Bladder Urothelial Carcinoma [BLCA]	412	412	10/08/15
Brain Lower Grade Glioma [LGG]	516	516	10/08/15
Breast invasive carcinoma [BRCA]	1100	1098	10/08/15
Cervical squamous cell carcinoma and endocervical adenocarcinoma [CESC]	308	308	10/08/15
Cholangiocarcinoma [CHOL]	36	36	10/08/15
Colon adenocarcinoma [COAD]	461	461	10/08/15
Esophageal carcinoma [ESCA]	185	185	10/08/15
FFPE Pilot Phase II [FPPP]	38	38	10/08/15
Glioblastoma multiforme [GBM]	529	528	10/08/15
Head and Neck squamous cell carcinoma [HNSC]	528	528	10/08/15
Kidney Chromophobe [KICH]	66	66	10/08/15
Kidney renal clear cell carcinoma [KIRC]	536	536	10/08/15
Kidney renal papillary cell carcinoma [KIRP]	291	291	10/08/15
Liver hepatocellular carcinoma [LIHC]	377	377	10/08/15
Lung adenocarcinoma [LUAD]	521	521	10/08/15
Lung squamous cell carcinoma [LUSC]	510	504	10/08/15
Lymphoid Neoplasm Diffuse Large B-cell Lymphoma [DLBC]	48	48	10/08/15
Mesothelioma [MESO]	87	87	10/12/15
Ovarian serous cystadenocarcinoma [OV]	586	586	10/08/15
Pancreatic adenocarcinoma [PAAD]	185	185	10/08/15
Pheochromocytoma and Paraganglioma [PCPG]	179	179	10/08/15
Prostate adenocarcinoma [PRAD]	498	498	10/08/15
Rectum adenocarcinoma [READ]	172	171	10/08/15
Sarcoma [SARC]	261	261	10/08/15
Skin Cutaneous Melanoma [SKCM]	470	470	10/08/15
Stomach adenocarcinoma [STAD]	445	443	10/08/15
Testicular Germ Cell Tumors [TGCT]	150	150	10/08/15
Thymoma [THYM]	124	124	10/08/15
Thyroid carcinoma [THCA]	507	507	10/08/15
Uterine Carcinosarcoma [UCS]	57	57	10/08/15
Uterine Corpus Endometrial Carcinoma [UCEC]	548	548	10/08/15
Uveal Melanoma [UVM]	80	80	10/08/15

* Excludes non-canonical cases

Announcements

07/27/2015 - Software release

On July 28th, 2015 the DCC will have a software release that will start at 8AM EST (GMT -5) and last for approximately one hour. During this time the TCGA Data Portal will be unavailable. This release will address a bug that was preventing compressed VCF files (.vcf.gz) from being displayed in the Data Matrix and File Search.

If you have any questions or concerns, contact tcga-dcc-binf-l@list.nih.gov.

07/09/2015 - Software release

The DCC has successfully completed the software release scheduled for today. Details about this release can be found on the TCGA Wiki: https://wiki.nci.nih.gov/x/coCyEQ.

Questions or concerns about this release can be directed to tcga-dcc-binf-l@list.nih.gov.

See all announcements

More TCGA Information

More information about The Cancer Genome Atlas program can be found by following the links below:

TCGA website

TCGA Publications

Publications using TCGA Data

TCGA publication guidelines

Head and Neck squamous cell carcinoma: Case Counts

Target number of Head and Neck squamous cell carcinoma samples:
500 (number subject to change)

Head and Neck squamous cell carcinoma [HNSC]	Total	Exome[1]	SNP	Methylation	mRNA	miRNA	Clinical
Cases	528	510	526	528	520	523	528
Organ-Specific Controls[2]	N/A	N/A	N/A	N/A	N/A	N/A	N/A

[1] Raw exome data are available at CGHub. Variant calling data are available via the links under Exome above.

[2] Organ-Specific Controls are derived from donor material taken from individuals not matched to the tumors in this study. Specifically, these tissues would be from individuals that did not have cancer but were able to donate tissue for other reasons (e.g. rapid autopsy programs, organ procurement programs, etc). N/A means that organ-specific tissue control data have not yet been collected for this tumor type by The Cancer Genome Atlas.

The TCGA Data Portal does not host lower levels of sequence data. NCI's Cancer Genomics Hub (CGHub) is the new secure repository for storing, cataloging, and accessing BAM files and metadata for sequencing data.

Announcements

07/27/2015 - Software release

On July 28th, 2015 the DCC will have a software release that will start at 8AM EST (GMT -5) and last for approximately one hour. During this time the TCGA Data Portal will be unavailable. This release will address a bug that was preventing compressed VCF files (.vcf.gz) from being displayed in the Data Matrix and File Search.

If you have any questions or concerns, contact tcga-dcc-binf-l@list.nih.gov.

07/09/2015 - Software release

The DCC has successfully completed the software release scheduled for today. Details about this release can be found on the TCGA Wiki: https://wiki.nci.nih.gov/x/coCyEQ.

Questions or concerns about this release can be directed to tcga-dcc-binf-l@list.nih.gov.

See all announcements

More TCGA Information

More information about The Cancer Genome Atlas program can be found by following the links below:

- TCGA website
- TCGA Publications
- Publications using TCGA Data
- TCGA publication guidelines

Data Matrix Datasets

The Data Matrix only provides the latest revision of each archive; older revisions are available through bulk download or HTTP access. Also, it does not allow for querying across multiple disease studies.

HNSC Data Matrix Options: Reset Matrix | Edit Filter | Remove Filter Help

Preservation: Frozen

Color Cells By: Availability Scroll Size: Standard

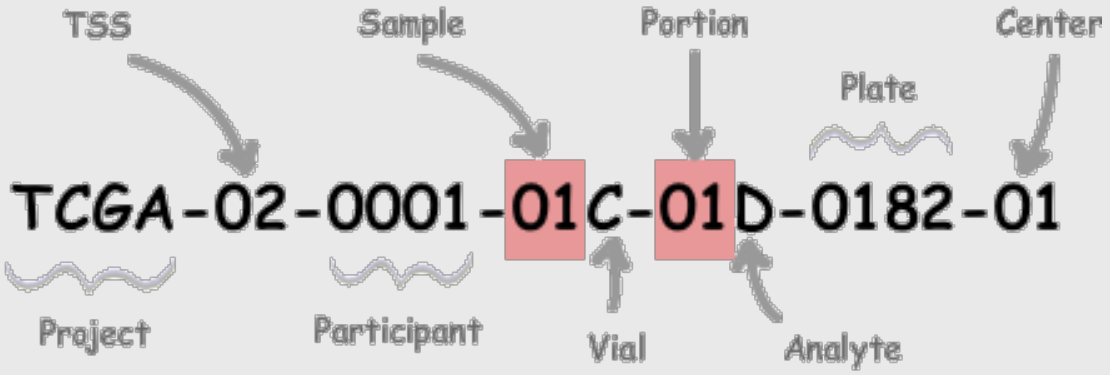
Build Archive

Legend:
- A Available
- P Pending
- N Not Available
- Not Applicable

*Protected data

		Clinical		Methyl	CNV (SNP Array)	Somatic Mutations				RNASeq	miRNASeq		Exp-Protein			RNASeqV2	CNV (Low Pass DNASeq)	Protected Mutations			
		XML	Biotab	JHU-USC HumanMethylation450	BI Genome_Wide_SNP_6	BI Mutation Calling	BI Automated Mutation Calling	BI Curated Mutation Calling	BCGSC Automated Mutation Calling	UNC IlluminaHiSeq_RNASeq	BCGSC IlluminaGA_miRNASeq	BCGSC IlluminaHiSeq_miRNASeq	MDA MDA_RPPA_Core			UNC IlluminaHiSeq_RNASeqV2	HMS IlluminaHiSeq_DNASeqC	BI Mutation Calling	BI Automated Mutation Calling	UCSC Mutation Calling	BCGSC Automated Mutation Calling
Batch/Sample	Level			1	1*	2	2	2	2	3	3	3	1	2	3	3	2*	2*	2*	2*	2*
Batch 54	TCGA-BA-4074-01	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A
	TCGA-BA-4075-01	A	A	A	A	A	A	A	A		A	A				A	A	A	A	A	A
	TCGA-BA-4076-01	A	A	A	A	A	A	A	A	A	A	A	A			A	A	A	A	A	A
	TCGA-BA-4077-01	A	A	A	A	A	A	A	A	A	A	A	A			A	A	A	A	A	A
	TCGA-BA-4078-01	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A
	TCGA-BA-5151-01	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A
	TCGA-BA-5153-01	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A
	TCGA-BB-4223-01	A	A	A	A	A	A	A	A	A	A	A				A	A	A	A	A	A
	TCGA-BB-4224-01	A	A	A	A	A	A	A	A	A	A	A				A	A	A	A	A	A
	TCGA-BB-4225-01	A	A	A	A	A	A	A	A	A	A	A				A	A	A	A	A	A
	TCGA-BB-4228-01	A	A	A	A	A	A	A	A	A	A	A				A		A	A	A	A
	TCGA-CN-4722-01	A	A	A	A	A				A	A		A	A	A	A	A				
	TCGA-CN-4723-01	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A		A	A	A	A
	TCGA-CN-4725-01	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A
	TCGA-CN-4726-01	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A		A	A	A	A
	TCGA-CN-4727-01	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A
	TCGA-CN-4728-01	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A		A	A	A	A
	TCGA-CN-4729-01	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A
	TCGA-CN-4730-01	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A
	TCGA-CN-4731-01	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A
	TCGA-CN-4734-01	A	A	A	A	A		A	A		A	A		A	A	A	A	A	A		A
	TCGA-CN-4735-01	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A
	TCGA-CN-4736-01	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A

Deciphering the TCGA Barcode

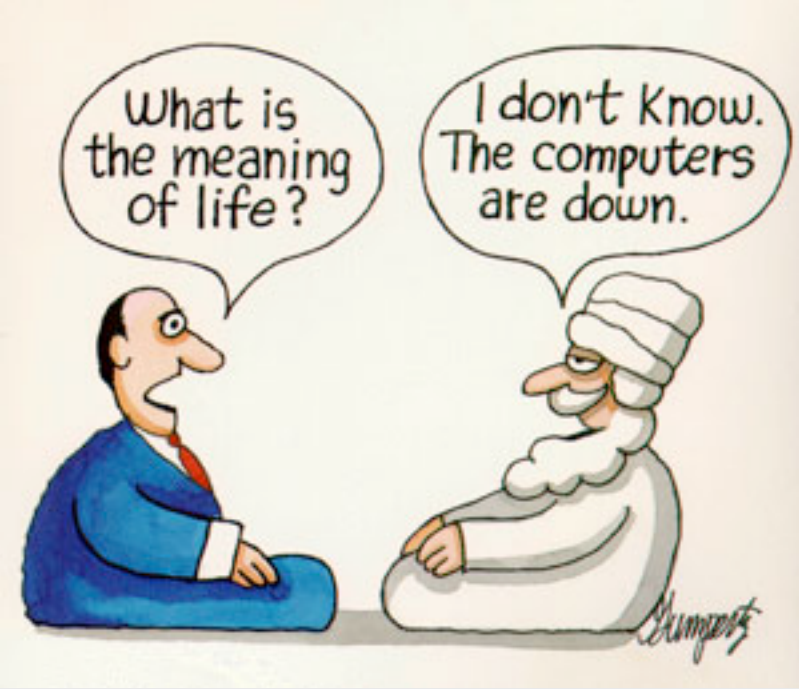


Label	Identifier for	Value	Value description
Participant	Study participant	0001	The first participant from MD Anderson for GBM study
Sample	Sample type	01	A solid tumor
Portion	Order of portion in a sequence of 100 - 120 mg sample portions	01	The first portion of the sample
Center	Sequencing or characterization center that will receive the aliquot for analysis	01	The Broad Institute GCC
Plate	Order of plate in a sequence of 96-well plates	0182	The 182nd plate
TSS	Tissue source site	02	GBM (brain tumor) sample from MD Anderson
Vial	Order of sample in a sequence of samples	C	The third vial
Analyte	Molecular type of analyte for analysis	D	The analyte is a DNA sample
Project	Project name	TCGA	TCGA project

TCGA Sample Code List

Code	Definition
01	Primary solid Tumor
02	Recurrent Solid Tumor
03	Primary Blood Derived Cancer - Peripheral Blood
04	Recurrent Blood Derived Cancer - Bone Marrow
05	Additional - New Primary
06	Metastatic
07	Additional Metastatic
08	Human Tumor Original Cells
09	Primary Blood Derived Cancer - Bone Marrow
10	Blood Derived Normal
11	Solid Tissue Normal
12	Buccal Cell Normal
13	EBV Immortalized Normal
14	Bone Marrow Normal
20	Control Analyte
40	Recurrent Blood Derived Cancer - Peripheral Blood
50	Cell Lines
60	Primary Xenograft Tissue
61	Cell Line Derived Xenograft Tissue

Thank You!


What is the meaning of life?
I don't know. The computers are down.