Introduction to (some) Biology

Bioinformatica

Febrero-Junio 2020

What is Bioinformatics?

- Term introduced by P. Hogeweg & B. Hesper in 1970
  http://en.wikipedia.org/wiki/Paulien_Hogeweg
- There are many definitions
- I will provide my own:
  - In bioinformatics we intend to develop, optimize, and parallelize algorithms, models, and production-level software for analyzing, storing, and extracting knowledge from, biological raw data.
  - Key differences to CS
    - proof-of-concept implementations are not sufficient
    - we need to produce code that can be used by biologists
    - we need to provide support for the code
    - have a look at http://groups.google.com/group/raxml
    - Most famous Bioinformaticians are known for one or more widely-used and highly cited algorithms & tools they have developed
- “Biology easily has 500 years of exciting problems to work on” – Donald Knuth

What is Bioinformatics?

Why is this exciting?

Important problems → medical applications,
Infectious diseases, genetic defects etc.
Masses of data → storage and analysis challenges
HPC → increased need for parallel codes

What are the challenges?

We can't be experts in everything → interdisciplinary collaboration
We need a culture of asking questions when we don't understand a term/concept!

Disciplines involved

Numerics
Statistics
Discrete Algorithms
Algorithm Engineering
Parallel Computing
Supercomputing
Software Engineering (in practice)

What is Biological Raw Data?

- There are many types of biological raw data
  - Images from microscopes
  - Microarray data
  - Protein structure data
  - Morphological data
  - Ecological data
  - Biogeographical data
  - …
- In this course we will mainly focus on *classic* Bioinformatics, that is, the analysis of molecular sequence data (DNA, protein data)

DNA data

- DNA data is available in public databases
- The most well-known one is GenBank
- Maintained by NCBI: National Center for Biotechnology Information, US
- Other databases for DNA data: EMBL (EU), DDBJ (Japan)

# of nucleotides/ base pairs

# of sequences

DNA data

- Genetic sequence
- Alphabet of 4 basic characters (nucleotides):
  - **A**denine
  - **C**ytosine
  - **G**uanine
  - **T**hymine
- A DNA sequence: *AACGTTTGA*
  - This sequence has 9 base pairs/nucleotides
- In RNA data: **T** is replaced by **U**racil
- A RNA sequence: *AACGUUUGA*
- We will see what RNA is later
- If we use **T** or **U** does usually not matter, computationally

DNA Sequencing

- The process of reading the nucleotide bases in a DNA molecule
- There exist various sequencing technologies
- Properties
  - Cost
  - Speed
  - Amount of data/Number of Sequences
  - Sequence length
  - Error rate

DNA Sequencing

- Sanger sequencing (*since 1977)*
  - High accuracy: 99.9%
  - Long sequences: 300-900 nucleotides
  - Expensive: $2400 per 1,000,000 nucleotides
  - Few sequences: up to ≈ 100
- Next-generation sequencing (*around 2007/2008*)
  - Lower accuracy 98-99.9%
  - Short sequences (100-400 nucleotides)
  - Inexpensive $1 - $10 per 1,000,000 nucleotides
  - Many sequences: 500 – 3,000,000,000 per sequencer run

The re

There is a lot of work to do for computer scientists to keep up with the data!

These are exciting times!

Sequencing cost versus processing cost!

DNA data

- GenBank: most-sequenced species

Some of these species are so-called *model organisms*

Model Organism

- A species that is extensively studied/sequenced to understand particular biological phenomena, with the expectation that discoveries made for the model organism will provide insight into the workings of other organisms.
- Selection criteria:
  - easy experimental manipulation
  - ease of genetic manipulation
  - easy to grow

    → short life-cycle/generation times
  - easy to extract DNA data
  - Economical importance → rice
- Often researchers reverse-engineer organisms
- Full list of model organisms: http://www.life.umd.edu/labs/mount/Models.html

Some Model Organisms

- *Escheria coli*

  gut bacterium → can cause food poisoning, grows fast, inexpensive to cultivate

- *Drosophila Melanogaster*

  fruit fly → breeds quickly

- *Arabidopsis Thaliana*

  flowering plant → small genome

Back to DNA

- What's a base pair?
- Pairing of A with T *or* C with G in double-stranded DNA

AATTGGC

TTAACCG

complement

Sloppy terminology

- The # of base pairs is frequently used as synonym for the # of nucleotides in a single-strand sequence
- This sequence has 5 nucleotides: *ACGGT*
- We can also say that it has 5 base pairs
- As in CS we use kilo, giga, etc for sequence lengths
  - kb → kilo-bases
  - Mb → Mega-bases
  - Gb → Giga-bases

Genome

- The full genetic information of an organism
  - Contains all chromosomes
  - Comprises the coding & non-coding sequence data of the organism
  - Coding sequence data → part of the genome that encodes proteins
  - Non-coding (in earlier days: junk) DNA → part of the genome that does not encode proteins but still has a function
    - The function of non-coding DNA is only partially known
    - Non-coding DNA regulates protein processes

Genome Size

- Not necessarily correlated with organism complexity
- Homo Sapiens: 3.2 Gb (Giga-bases)
- Marbled lungfish: 130 Gb (Giga-bases)
- Plants often have very large genomes → partially due to redundant information.

Why Computational Biology ?

Why <u>Computational</u> Biology: Last year's answers

- Lots of data (* lots of data)
- There are rules
- Pattern finding
- It's *all* about data
- Ability to visualize
- Simulations, temporal relationships
- Guess + verify (generate hypotheses for testing)
- Propose mechanisms / theory to explain observations
- Networks / combinations of variables
- Efficiency (reduce experimental space to cover)
- Informatics infrastructure (ability to combine datasets)
- Correlations, higher-order relationships
- Cycle from hypothesis generation to testing condensed
- Life itself is digital. Understand cellular instruction set

TATTGAATTTTCAAAAATTCTTACTTTTTTTTTGGATGGACGCAAAGAAGTTTAATAATCATATTACATGGCATTACCACCATATA
ATCCATATCTAATCTTACTTATATGTTGTGGAAATGTAAAGAGCCCCATTATCTTAGCCTAAAAAAACCTTCTCTTTGGAACTTTC
AATACGCTTAACTGCTCATTGCTATATTGAAGTACGGATTAGAAGCCGCCGAGCGGGCGACAGCCCTCCGACGGAAGACTCTCCTC
GCGTCCTCGTCTTCACCGGTCGCGTTCCTGAAACGCAGATGTGCCTCGCGCCGCACTGCTCCGAACAATAAAGATTCTACAATACT
TTTTATGGTTATGAAGAGGAAAAATTGGCAGTAACCTGGCCCCACAAACCTTCAAATTAACGAATCAAATTAACAACCATAGGATG
ATGCGATTAGTTTTTTAGCCTTATTTCTGGGGTAATTAATCAGCGAAGCGATGATTTTTGATCTATTAACAGATATATAAATGGAA
CTGCATAACCACTTTAACTAATACTTTCAACATTTTCAGTTTGTATTACTTCTTATTCAAATGTCATAAAAGTATCAACAAAAAAT
TAATATACCTCTATACTTTAACGTCAAGGAGAAAAAACTATAATGACTAAATCTCATTCAGAAGAAGTGATTGTACCTGAGTTCAA
TAGCGCAAAGGAATTACCAAGACCATTGGCCGAAAAGTGCCCGAGCATAATTAAGAAATTTATAAGCGCTTATGATGCTAAACCGC
TTGTTGCTAGATCGCCTGGTAGAGTCAATCTAATTGGTGAACATATTGATTATTGTGACTTCTCGGTTTTACCTTTAGCTATTGAT
GATATGCTTTGCGCCGTCAAAGTTTTGAACGAGAAAAATCCATCCATTACCTTAATAAATGCTGATCCCAAATTTGCTCAAAGGAA
CGATTTGCCGTTGGACGGTTCTTATGTCACAATTGATCCTTCTGTGTCGGACTGGTCTAATTACTTTAAATGTGGTCTCCATGTTC
ACTCTTTTCTAAAGAAACTTGCACCGGAAAGGTTTGCCAGTGCTCCTCTGGCCGGGCTGCAAGTCTTCTGTGAGGGTGATGTACCA
GGCAGTGGATTGTCTTCTTCGGCCGCATTCATTTGTGCCGTTGCTTTAGCTGTTGTTAAAGCGAATATGGGCCCTGGTTATCATAT
CAAGCAAAATTTAATGCGTATTACGGTCGTTGCAGAACATTATGTTGGTGTTAACAATGGCGGTATGGATCAGGCTGCCTCTGTTT
GTGAGGAAGATCATGCTCTATACGTTGAGTTCAAACCGCAGTTGAAGGCTACTCCGTTTAAATTTCCGCAATTAAAAAACCATGAA
AGCTTTGTTATTGCGAACACCCTTGTTGTATCTAACAAGTTTGAAACCGCCCCAACCAACTATAATTTAAGAGTGGTAGAAGTCAC
AGCTGCAAATGTTTTAGCTGCCACGTACGGTGTTGTTTTACTTTCTGGAAAAGAAGGATCGAGCACGAATAAAGGTAATCTAAGAG
TCATGAACGTTTATTATGCCAGATATCACAACATTTCCACACCCTGGAACGGCGATATTGAATCCGGCATCGAACGGTTAACAAAG
CTAGTACTAGTTGAAGAGTCTCTCGCCAATAAGAAACAGGGCTTTAGTGTTGACGATGTCGCACAATCCTTGAATTGTTCTCGCGA
ATTCACAAGAGACTACTTAACAACATCTCCAGTGAGATTTCAAGTCTTAAAGCTATATCAGAGGGCTAAGCATGTGTATTCTGAAT
TAAGAGTCTTGAAGGCTGTGAAATTAATGACTACAGCGAGCTTTACTGCCGACGAAGACTTTTTCAAGCAATTTGGTGCCTTGATG
GAGTCTCAAGCTTCTTGCGATAAACTTTACGAATGTTCTTGTCCAGAGATTGACAAAATTTGTTCCATTGCTTTGTCAAATGGATC
TGGTTCCCGTTTGACCGGAGCTGGCTGGGGTGGTTGTACTGTTCACTTGGTTCCAGGGGGCCCAAATGGCAACATAGAAAAGGTAA
AAGCCCTTGCCAATGAGTTCTACAAGGTCAAGTACCCTAAGATCACTGATGCTGAGCTAGAAAATGCTATCATCGTCTCTAAACCA
TTGGGCAGCTGTCTATATGAATTATAAGTATACTTCTTTTTTTTACTTTGTTCAGAACAACTTCTCATTTTTTTCTACTCATAACT
GCATCACAAAATACGCAATAATAACGAGTAGTAACACTTTTATAGTTCATACATGCTTCAACTACTTAATAAATGATTGTATGATA
TTTTCAATGTAAGAGATTTCGATTATCCACAAACTTTAAAACACAGGGACAAAATTCTTGATATGCTTTCAACCGCTGCGTTTTGC
CCTATTCTTGACATGATATGACTACCATTTTGTTATTGTACGTGGGGCAGTTGACGTCTTATCATATGTCAAAGTCATTTGCGAAG
TTGGCAAGTTGCCAACTGACGAGATGCAGTAAAAAGAGATTGCCGTCTTGAAACTTTTTGTCCTTTTTTTTTTCCGGGGACTCTAC
AACCCTTTGTCCTACTGATTAATTTTGTACTGAATTTGGACAATTCAGATTTTAGTAGACAAGCGCGAGGAGGAAAAGAAATGACA
AAATTCCGATGGACAAGAAGATAGGAAAAAAAAAAAGCTTTCACCGATTTCCTAGACCGGAAAAAAGTCGTATGACATCAGAATGA
ATTTTCAAGTTAGACAAGGACAAAATCAGGACAAATTGTAAAGATATAATAAACTATTTGATTCAGCGCCAATTTGCCCTTTTCCA
TCCATTAAATCTCTGTTCTCTCTTACTTATATGATGATTAGGTATCATCTGTATAAAACTCCTTTCTTAATTTCACTCTAAAGCAT
CCATAGAGAAGATCTTTCGGTTCGAAGACATTCCTACGCATAATAAGAATAGGAGGGAATAATGCCAGACAATCTATCATTACATT
GCGGCTCTTCAAAAAGATTGAACTCTCGCCAACTTATGGAATCTTCCAATGAGACCTTTGCGCCAAATAATGTGGATTTGGAAAAA
TATAAGTCATCTCAGAGTAATATAACTACCGAAGTTTATGAGGCATCGAGCTTTGAAGAAAAAGTAAGCTCAGAAAAACCTCAATA
CTCATTCTGGAAGAAAATCTATTATGAATATGTGGTCGTTGACAAATCAATCTTGGGTGTTTCTATTCTGGATTCATTTATGTACA
AGGACTTGAAGCCCGTCGAAAAAGAAAGGCGGGTTTGGTCCTGGTACAATTATTGTTACTTCTGGCTTGCTGAATGTTTCAATATC
ACTTGGCAAATTGCAGCTACAGGTCTACAACTGGGTCTAAATTGGTGGCAGTGTTGGATAACAATTTGGATTGGGTACGGTTTCGT

Genes
Encode proteins
Regulatory motifs
Control gene expression

Extracting signal from noise

The components of genomes and gene regulation

Goal: A systems-level understanding of genomes and gene regulation:

- The genome: Map reads, align genes/genomes, assembly strategies
- The genes: Protein-coding exons, introns, non-coding RNA, RNA folding
- The control regions: Promoters, enhancers, insulators, chromatin states
- The actual words: Regulatory motifs, high-resolution accessibility maps
- The regulators: Transcription factors, chromatin modifiers, nucleosomes
- The dynamics: Changing maps between cell types, across development
- The networks: regulator→enhancer→target, ChIP-seq, correlated activity
- The grammars: TF/motif/mark combinations, predictive models
- Human variation: Human diversity, population genomics, linkage maps
- Evolution: Phylogenetics, phylogenomics, coalescent, human ancestry
- GWAS/QTLs: Genome variation ⇔ organismal/molecular phenotypes
- Disease: Personal (epi)genomics, pharmacogenomics, synthetic biology

Challenges in Computational Biology

4 Genome Assembly

5 Regulatory motif discovery

1 Gene Finding

DNA

2 Sequence alignment

6 Comparative Genomics

3 Database lookup

7 Evolutionary Theory

TCATGCTAT
TCGTGATAA
TGAGGATAT
TTATCATAT
TTATGATTT

8 Gene expression analysis

RNA transcript

9 Cluster discovery

10 Gibbs sampling

11 Protein network analysis

12 Metabolic modelling

13 Emerging network properties

“Central dogma” of Molecular Biology

DNA: The double helix

- The most noble molecule of our time

Image by MIT OpenCourseWare.

DNA: the molecule of heredity

- Self-complementarity sets molecular basis of heredity
  - Knowing one strand, creates a template for the other
  - "It has not escaped our notice that the specific pairing we have postulated immediately suggests a possible copying mechanism for the genetic material." Watson & Crick, 1953

Image by MIT OpenCourseWare.

DNA: chemical details

- Bases hidden on the inside
- Phosphate backbone

outside

- Weak hydrogen bonds hold the two strands together
- This allows low-energy opening and re-closing of two strands

- Anti-parallel strands
- Extension 5' →3' tri-phosphate coming from newly added nucleotide

The only parings are:

- A with T
- C with G

DNA: the four bases

The Nucleotides of DNA

Adenine	Guanosine	Thymine	Cytosine
Purine	Purine		
		Pyrimidine	Pyrimidine
Weak		Weak	
	Strong		Strong
Amino			Amino
	Keto	Keto	

Alignment: all species/genes share common ancestry

© Various sources. All rights reserved. This content is excluded from our Creative Commons license. For more information, see http://ocw.mit.edu/help/faq-fair-use/.

Slide credit: Serafim Batzoglou

Tree of Life

© Neal Olander. All rights reserved. This content is excluded from our Creative Commons license. For more information, see http://ocw.mit.edu/help/faq-fair-use/.

© Neal Olander. All rights reserved. This content is excluded from our Creative Commons license. For more information, see http://ocw.mit.edu/help/faq-fair-use/.

“Central dogma” of Molecular Biology

Epigenomics

Chromosomes inside the cell

Figures by MIT OpenCourseWare.

DNA packaging

- Why packaging
  - DNA is very long
  - Cell is very small
- Compression
  - Chromosome is 50,000 times shorter than extended DNA
- Using the DNA
  - Before a piece of DNA is used for anything, this compact structure must open locally
- Now emerging:
  - Role of accessibility
  - State in chromatin itself
  - Role of 3D interactions

Image removed due to copyright restrictions.
Please see: Figure 8-10 from Alberts, Bruce, and Martin Raff. Essential Cell Biology. New York, NY: Garland Publishing Inc., 1997. ISBN: 0815320450.

Diverse epigenetic modifications

Courtesy of the National Institutes of Health; in the public domain.

Image source: http://nihroadmap.nih.gov/epigenomics/

“Central dogma” of Molecular Biology

Genes control the making of cell parts

- The gene is a fundamental unit of inheritance
  - Each DNA molecule ⇔ 10,000+ genes
  - 1 gene ⇔ 1 functional element (one "part" of cell machinery)
  - Every time a "part" is made, the corresponding gene is:
    - Copied into mRNA, transported, used as blueprint to make protein
- RNA is a temporary copy
  - The medium for transporting genetic information from the DNA information repository to the protein-making machinery is an RNA molecule
  - The more parts are needed, the more copies are made
  - Each mRNA only lasts a limited time before degradation

mRNA: The messenger

- Information changes medium
  - single strand vs. double strand
  - ribose vs. deoxyribose sugar

```
A  T  T  A  C  G  G  T  A  C  C  G  T
||  ||  ||  ||  ||| ||| |||  ||  ||  ||| ||| |||  ||
U  A  A  U  G  C  C  A  U  G  G  C  A
```

  - Compatible base-pairing in hybrid

Image by MIT OpenCourseWare.

From pre-mRNA to mRNA: Splicing

- In Eukaryotes, not every part of a gene is coding
  - Functional exons interrupted by non-translated introns
  - During pre-mRNA maturation, introns are spliced out
  - In humans, primary transcript can be 10^6 bp long

Image removed due to copyright restrictions.
Please see: Figure 7-16 from Alberts, Bruce, and Martin Raff. Essential Cell Biology.
New York, NY: Garland Publishing Inc., 1997. ISBN: 0815320450.

- Alternative splicing can yield different exon subsets for the same gene, and hence different protein products

RNA can be functional

- Single Strand allows complex structure
  - Self-complementary regions form helical stems
  - Three-dimensional structure allows functionality of RNA
- Four types of RNA
  - mRNA: messenger of genetic information
  - tRNA: codon-to-amino acid specificity
  - rRNA: core of the ribosome
  - snRNA: splicing reactions
- To be continued…
  - We'll learn more in a dedicated lecture on RNA world
  - Once upon a time, before DNA and protein, RNA did all

RNA structure: 2ndary and 3rdary

Courtesy of SStructView

“Central dogma” of Molecular Biology

Proteins carry out the cell's chemistry

Image by MIT OpenCourseWare.

- More complex polymer
  - Nucleic Acids have 4 building blocks
  - Proteins have 20. Greater versatility
  - Each amino acid has specific properties

Sequence → Structure → Function

  - The amino acid sequence determines the three-dimensional fold of protein
  - The protein's function largely depends on the features of the 3D structure

- Proteins play diverse roles
  - Catalysis, binding, cell structure, signaling, transport, metabolism

Protein structure

Image by MIT OpenCourseWare.

Helix-turn-helix

Common motif for DNA-binding proteins that often play a regulatory role as mRNA level transcription factors

Image by MIT OpenCourseWare.

Beta-barrel

Some antiparallel b-sheet domains are better described as b-barrels rather than b-sandwiches, for example streptavadin and porin. Note that some structures are

intermediate between the extreme barrel and sandwich arrangements.

Image by MIT OpenCourseWare.

Alpha-beta horseshoe

this placental ribonuclease inhibitor is a cytosolic protein that binds extremely strongly to any ribonuclease that may leak into the cytosol. 17-stranded parallel b sheet curved into an open horseshoe shape, with 16 a-helices packed against the outer surface. It doesn't form a barrel although it looks as though it should. The strands are only very slightly slanted, being nearly parallel to the central `axis'.

Protein building blocks

- Amino Acids

From RNA to protein: Translation

- Ribosome

Image by MIT OpenCourseWare.

- tRNA

The Genetic Code

SECOND POSITION

FIRST POSITION	U	C	A	G	THIRD POSITION
U	phenyl-alanine	serine	tyrosine	cysteine	U
					C
	leucine		stop	stop	A
			stop	tryptophan	G
C	leucine	proline	histidine	arginine	U
					C
			glutamine		A
					G
A	isoleucine	threonine	asparagine	serine	U
					C
			lysine	arginine	A
	* methionine				G
G	valine	alanine	aspartic acid	glycine	U
					C
			glutamic acid		A
					G

* and start

➔ Use evolutionary and compositional properties to computationally discover protein-coding genes

Summary: The Central Dogma

DNA makes RNA makes Protein

Inheritance

Messages

Reactions

Image by MIT OpenCourseWare.

Cellular dynamics and regulation

How cells move through this Central Dogma

Animal/Human gene regulation: One genome ⇔ Many cell types

```
ACCAGTTACGACGGTCA
GGGTACTGATACCCCAA
ACCGTTGACCGCATTTA
CAGACGGGGTTTGGGTT
TTGCCCCACACAGGTAC
GTTAGCTACTGGTTTAG
CAATTTACCGTTACAAC
GTTTACAGGGTTACGGT
TGGGATTTGAAAAAAAG
TTTGAGTTGGTTTTTTC
ACGGTAGAACGTACCGT
      TACCAGTA
```

Image in the public domain.

Images of a heart, red blood cell, and a brain removed due to copyright restrictions.

Image Source wikipedia

Diverse roles for regulatory non-coding RNAs

- **Small RNA pathways (18-21 nt)**
  - microRNAs:
    - Repress genes by targeting their 3' UTRs by complementarity
    - Double-stranded RNA is then recognized and degraded
    - Recently found to also target promoter regions in rare cases
  - piwiRNAs
    - Target and repress transposable elements in germline
  - snoRNAs
  - 21U-RNAs
- **Long non-coding RNAs (1000s nt, many exons)**
  - Scaffolds for protein/TF binding
  - Scaffolds for 3D structure of RNA

Regulation of Gene Expression

Examples:

- Upstream of genes are *promoter* regions
- Contain promoter sequences or *motifs*
- *Transcription factors* (TFs) bind to motifs
- TFs recruit *RNA polymerase*
- Gene transcription

Over-express a single microRNA leads to new wing

Note: C,D,E same magnification

- Discovery of sense/anti-sense miRNAs
- Regulatory switch selects between two developmental programs
- By over-expressing one strand (miRNAas) the balance is tilted
- Wing program launched vs. haltere

Stark *et al*, Genes&Development 2007

© source unknown. All rights reserved. This content is excluded from our Creative Commons license. For more information, see http://ocw.mit.edu/help/faq-fair-use/.

The role of genetic alterations

Brief intro to human genetics

- **Human genome**: 3.2B letters, 2 copies, 23 chromosomes, 20k genes, ~3M common SNPs, ~500k haplotype blocks

Figure in the public domain. Created by Darryl Leja and Teri Manolio, NHGRI; Tony Burdett, Dani Welter, and Helen Parkinson, EBI.

Genomewide association in schizophrenia with 40,000 cases

More than 100 distinct regions of the genome associated to schizophrenia!!!

Courtesy of Macmillan Publishers Limited. Used with permission.
Source: Ripke, Stephan et al. "Biological insights from 108 schizophrenia-associated genetic loci." Nature 511, no. 7510 (2014): 421.

Characterizing sub-threshold variants in heart arrhythmia

© source unknown. All rights reserved. This content is excluded from our Creative Commons license. For more information, see http://ocw.mit.edu/help/faq-fair-use/.

Courtesy of Macmillan Publishers Limited. Used with permission.
Source: Arking, Dan E. et al. "Genetic association study of QT interval highlights role for calcium signaling pathways in myocardial repolarization." Nature Genetics 46, no. 8 (2014): 826-836.

Trait: QRS/QT interval

Focus on sub-threshold variants (e.g. rs1743292 $P=10^{-4.2}$)

(1) Large cohorts, (2) many known hits
(3) well-characterized tissue drivers

TCGA Data:
Access, types and usability

Arunima Srivastava
srivatava.1@osu.edu
Hao Ding
ding.96@osu.edu

Overview

The Cancer Genome Atlas (TCGA) 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.

By 11/18/2015

- 34 cancer types
- 11079 sample

TCGA Barcode

Label	Identifier for	Value	Value description	Possible values
Project	Project name	TCGA	TCGA project	TCGA
TSS	Tissue source site	02	GBM (brain tumor) sample from MD Anderson	See Code Tables Report
Participant	Study participant	0001	The first participant from MD Anderson for GBM study	Any alpha-numeric value
Sample	Sample type	01	A solid tumor	Tumor types range from 01 - 09, normal types from 10 - 19 and control samples from 20 - 29. See Code Tables Report for a complete list of sample codes
Vial	Order of sample in a sequence of samples	C	The third vial	A to Z
Portion	Order of portion in a sequence of 100 - 120 mg sample portions	01	The first portion of the sample	01-99
Analyte	Molecular type of analyte for analysis	D	The analyte is a DNA sample	See Code Tables Report
Plate	Order of plate in a sequence of 96-well plates	0182	The 182nd plate	4-digit alphanumeric value
Center	Sequencing or characterization center that will receive the aliquot for analysis	01	The Broad Institute GCC	See Code Tables Report

https://wiki.nci.nih.gov/display/TCGA/TCGA+barcode

TCGA Sample Code

Code	Definition	Short Letter Code
01	Primary solid Tumor	TP
02	Recurrent Solid Tumor	TR
03	Primary Blood Derived Cancer - Peripheral Blood	TB
04	Recurrent Blood Derived Cancer - Bone Marrow	TRBM
05	Additional - New Primary	TAP
06	Metastatic	TM
07	Additional Metastatic	TAM
08	Human Tumor Original Cells	THOC
09	Primary Blood Derived Cancer - Bone Marrow	TBM
10	Blood Derived Normal	NB
11	Solid Tissue Normal	NT
12	Buccal Cell Normal	NBC
13	EBV Immortalized Normal	NEBV
14	Bone Marrow Normal	NBM
20	Control Analyte	CELLC
40	Recurrent Blood Derived Cancer - Peripheral Blood	TRB
50	Cell Lines	CELL
60	Primary Xenograft Tissue	XP
61	Cell Line Derived Xenograft Tissue	XCL

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	10/19/15
Bladder Urothelial Carcinoma [BLCA]	412	412	10/21/15
Brain Lower Grade Glioma [LGG]	516	516	11/18/15
Breast invasive carcinoma [BRCA]	1100	1098	11/18/15
Cervical squamous cell carcinoma and endocervical adenocarcinoma [CESC]	308	308	11/05/15
Cholangiocarcinoma [CHOL]	36	36	10/08/15
Colon adenocarcinoma [COAD]	461	461	11/04/15
Esophageal carcinoma [ESCA]	185	185	10/19/15
FFPE Pilot Phase II [FPPP]	38	38	10/08/15
Glioblastoma multiforme [GBM]	529	528	11/03/15

Home > Cancer Details

Acute Myeloid Leukemia: Case Counts

Target number of Acute Myeloid Leukemia samples:
500 (number subject to change)

Acute Myeloid Leukemia [LAML]	Total	Exome[1]	SNP	Methylation	mRNA	miRNA	Clinical
Cases	200	150	200	194	179	188	200
Organ-Specific Controls[2]	N/A	N/A	N/A	N/A	N/A	N/A	N/A

Data Matrix Datasets

The Data Matrix only provides the latest revision of each archive; older revisions are available through bulk download or HTTP access. Als
querying across multiple disease studies.

LAML Data Matrix Options: Reset Matrix Set Filter

Color Cells By: Tumor/Normal Scroll Size: Standard

Build Archive

Legend:
- TN Tumor, Matched Normal
- T Tumor, no Matched Normal
- NT Normal, Matched Tumor
- N Organ-Specific Control
- C Cell Line Control
- No Available Dataset

*Protected data

		Clinical		Exp-Gene			Methyl						CNV (SNP Array)						Somatic Mutations	RNASeq		miRNASeq	RNASeqV2
		XML	Biotab	WUSM HG-U133_Plus_2			JHU-USC HumanMethylation27			JHU-USC HumanMethylation450			BI Genome_Wide_SNP_6			WUSM Genome_Wide_SNP_6			WUSM Mutation Calling	BCGSC IlluminaGA_RNASeq		BCGSC IlluminaGA_miRNASeq	UNC IlluminaHiSeq_RNASeqV2
Batch/Sample	*Level*			1	2	3	1	2	3	1	2	3	1*	2*	3	1*	2*	3	2	2*	3	3	3
Batch 25	TCGA-AB-2802-03	TN	TN	TN	TN	TN	TN	TN	TN	TN	TN	TN	TN	TN	TN	TN	TN	TN	TN			TN	
	TCGA-AB-2803-03	TN	TN	TN	TN	TN	TN	TN	TN	TN	TN	TN	TN	TN	TN	TN	TN	TN	TN	TN	TN	TN	TN
	TCGA-AB-2804-03	TN	TN	TN	TN	TN	TN	TN	TN	TN	TN	TN	TN	TN	TN	TN	TN	TN	TN				
	TCGA-AB-2805-03	TN	TN	TN	TN	TN	TN	TN	TN	TN	TN	TN	TN	TN	TN	TN	TN	TN	TN	TN	TN	TN	TN
	TCGA-AB-2806-03	TN	TN	TN	TN	TN	TN	TN	TN	TN	TN	TN	TN	TN	TN	TN	TN	TN	TN	TN	TN	TN	TN
	TCGA-AB-2807-03	TN	TN	TN	TN	TN	TN	TN	TN	TN	TN	TN	TN	TN	TN	TN	TN	TN	TN	TN	TN	TN	TN
	TCGA-AB-2808-03	TN	TN	TN	TN	TN	TN	TN	TN	TN	TN	TN	TN	TN	TN	TN	TN	TN	TN	TN	TN	TN	TN
	TCGA-AB-2809-03	TN	TN	TN	TN	TN	TN	TN	TN	TN	TN	TN	TN	TN	TN	TN	TN	TN	TN				

Transcriptomics

- Microarray
- RNASeq
- Depending on platform
  - Gene Level expression
  - Exon Level quantitation
  - Splice Junction Counts

Usability

- Comparison of Normal vs Tumor
  - What are driving factors of cancer?
  - What level of data show most differentiation?
  - Can I make disease/subtype predictions using the data?
- Combining Two levels of data
  - Do I see stronger evidence of cancer drivers when I combine data?
  - Is one level of data enough to make predictions?
  - Which type of data is more effective for predictive analysis?

Download Tissue Slides Images

- Can only be download through **Bulk Download** and or **File Search**
- file size: **1-2GB** per slide
- Viewing image online: [Lawrence Berkeley BioSig]