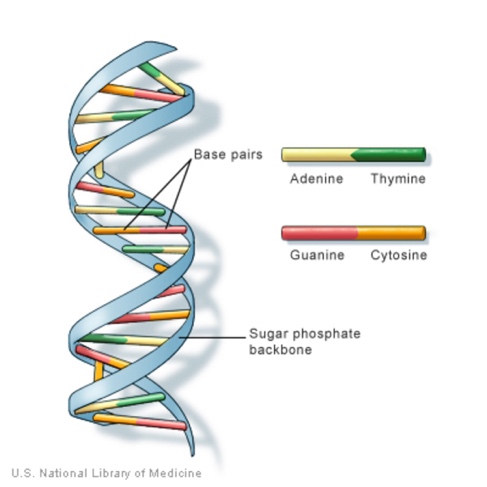
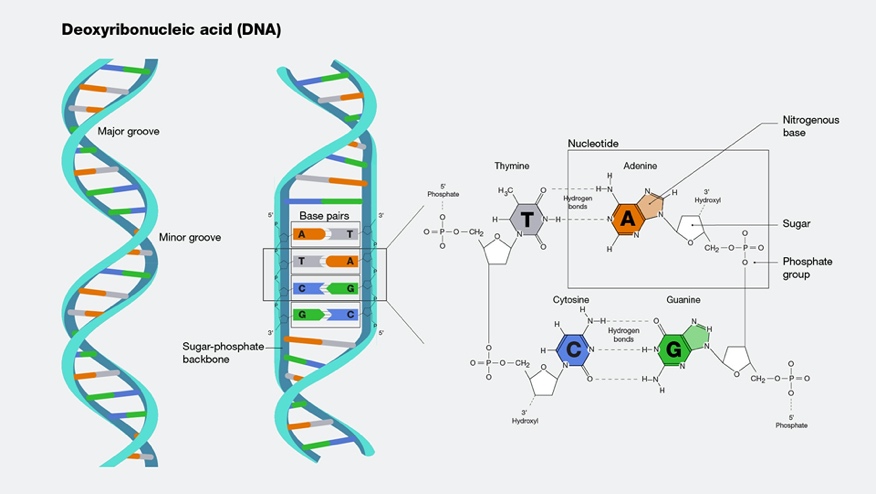
**Experimental Network (XN) project**

**Machine Learning -Find Novel Intrinsic Oncology Targets and Biology**

**Some terminologies and more project description:**

1. **DNA** is a double [helix](https://en.wikipedia.org/wiki/Helix) formed by base pairs attached to a sugar-phosphate backbone.

* Nearly every cell in a person’s body has the same DNA.
* The information in DNA is stored as a code made up of four chemical bases: adenine (A), guanine (G), cytosine (C), and thymine (T).
* Human DNA consists of about 3 billion bases, and more than 99 percent of those bases are the same in all people.
* DNA bases pair up with each other, A with T and C with G, to form units called base pairs.
* An important property of DNA is that it can replicate. Each new cell needs to have an exact copy of the DNA present in the old cell.
* Only about 1 percent of DNA is made up of protein-coding genes; the other 99 percent is noncoding. Noncoding DNA does not provide instructions for making proteins. Noncoding DNA control of gene activity, such as determining when and where genes are turned on and off.

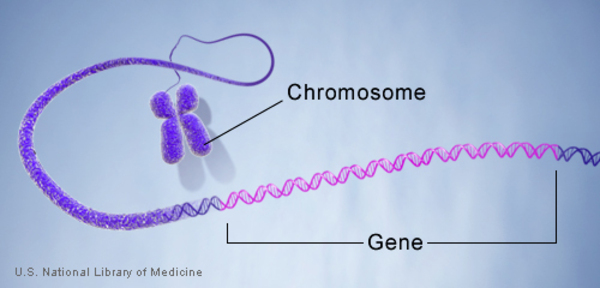
 

<https://medlineplus.gov/genetics/understanding/basics/dna/>

<https://www.genome.gov/genetics-glossary/Deoxyribonucleic-Acid>

1. **Genes** are made up of DNA.

* Genes vary in size from a few hundred DNA bases to more than 2 million bases.
* Humans have between 20,000 and 25,000 genes, research by Human Genome Project.
* Some genes act as instructions to make molecules called proteins. However, many genes do not code for proteins.
* Every person has two copies of each gene, one inherited from each parent.
* Most genes are the same in all people, but a small number of genes (less than 1 percent of the total) are slightly different between people. These small differences contribute to each person’s unique physical features.

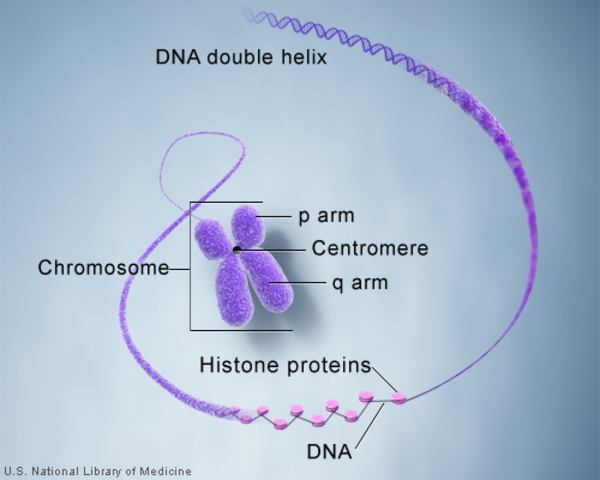


<https://medlineplus.gov/genetics/understanding/basics/gene/>

Video: <https://www.genome.gov/genetics-glossary/Gene>

1. DNA and histone proteins are packaged into structures called **chromosomes**. Each **chromosome** contains many genes.

In humans, each cell normally contains 23 pairs of chromosomes, for a total of 46. Twenty-two of these pairs, called autosomes, look the same in both males and females. The 23rd pair, the sex chromosomes, differ between males and females. Females have two copies of the X chromosome, while males have one X and one Y chromosome.



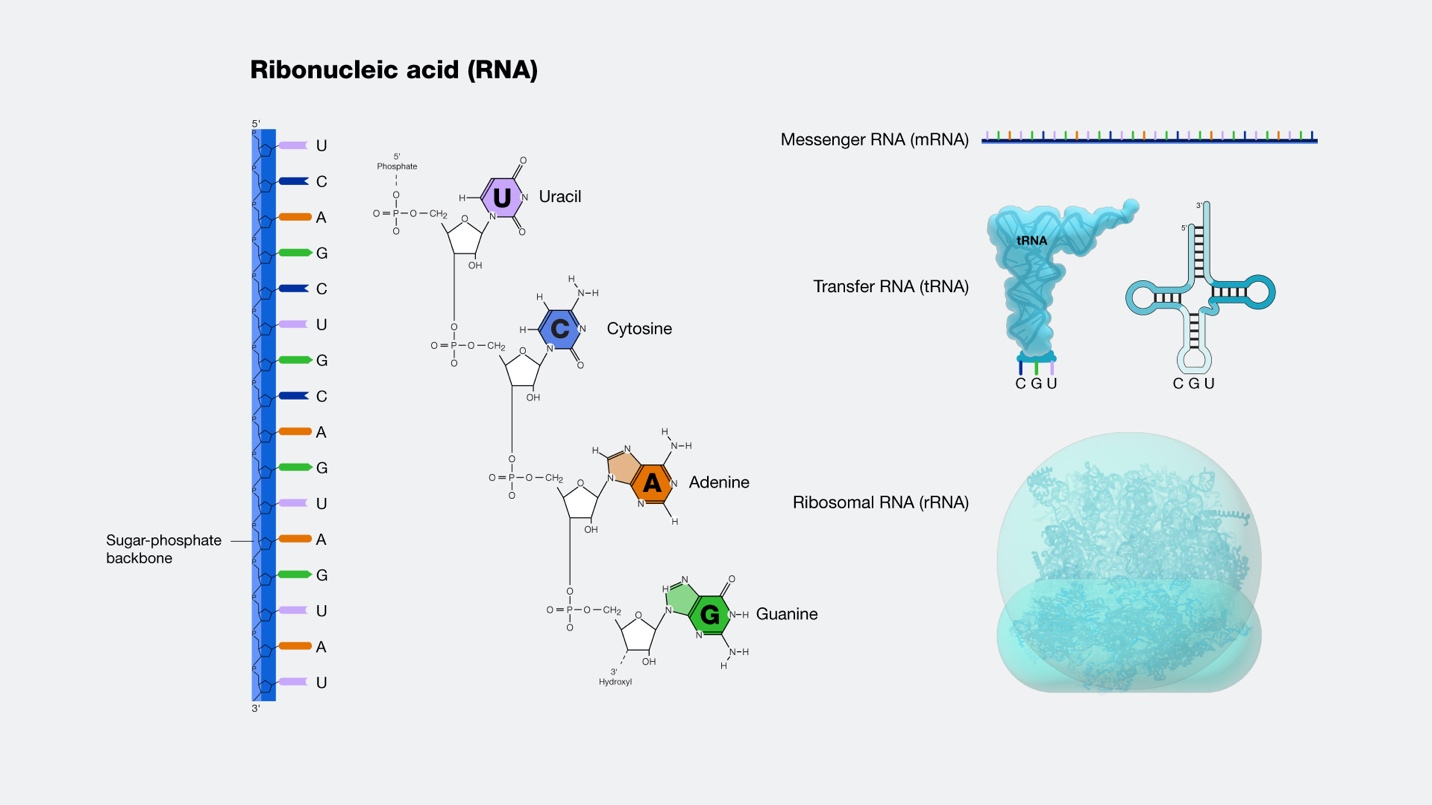
<https://medlineplus.gov/download/genetics/understanding/basics.pdf>

<https://medlineplus.gov/genetics/>

**Genetic Code Video:** <https://youtu.be/98s4nnD-leQ>

1. **RNA** is a nucleic acid present in all living cells

* structural similarities to DNA, but often single-stranded
* RNA molecule has a backbone made of alternating phosphate groups and the sugar ribose, rather than the deoxyribose found in DNA.
* Attached to each sugar is one of four bases: adenine (A), uracil (U), cytosine (C) or guanine (G)
* Different types of RNA exist in cells: messenger RNA (mRNA), ribosomal RNA (rRNA) and transfer RNA (tRNA).



<https://www.genome.gov/genetics-glossary/RNA-Ribonucleic-Acid>

More topics:

From DNA to protein: <https://www.youtube.com/watch?v=gG7uCskUOrA>

1. **Cells** are the basic building blocks of all living things. In the nucleus of each cell, the DNA molecule is packaged into chromosomes.

Diagram, radar chart

Description automatically generated

<https://prevention.cancer.gov/news-and-events/blog/taking-advantage>

**XN Project Data**: **genomic data sets**. (public)

**Interests/Goal of Merck Company**

Broadly the main questions are, which algorithms are suitable for mixed data distribution types for classification and scalable for effective computing. More concretely,

1. Identify cancer vulnerabilities, differential cancer gene dependencies.
2. Classify predictive models that share common genetic determinants using *CRISPR KO gene dependencies, mutational and copy number profiles and RNA expression data* obtained from over 1000 cancer cell lines.

This task requires application of methods that can integrate diverse data types and machine learning algorithms to classify the **lineages** that are dependent on essential genes for tumorigenesis.

**Terminologies in the description:**

* **CRISPR (clustered regularly interspaced short palindromic repeats**) is a technology that research scientists use to selectively modify the DNA of living organisms.
* **KO (knockout):** refers to the use of genetic engineering to inactivate or remove one or more specific genes from an organism.

CRISPR Knockout videos: <https://www.youtube.com/watch?v=1BSiFfyObTY>

* Each **mutational** **process** may involve components of DNA damage or modification, DNA repair and DNA replication (which may be normal or abnormal), and generates a characteristic mutational signature that potentially includes base substitutions, small insertions and deletions (indels), genome rearrangements and chromosome copy-number changes.
* **copy number profiles:** A copy-number profile (CNP) is a vector u=(u1,…,un) of non-negative integers representing the copy-number of each position in a clone.

**Project Deliverable/Tasks:**

1. Identify a machine learning algorithm that is best suitable for data and classifies lineages into significant clusters.
2. Test if the classification schemes derived by associating gene dependencies with other data sets such as RNA expression, pathway signatures, DNA mutation and copy number data are unique.
3. Determine whether the machine learning algorithm can be extended to identification of predictors of general viability loss rather than specific priors lineage-specific genes only.

**Resources DepMap Portal: (Cancer Dependency Map)**

* **Website**: <https://depmap.org/portal/>
* **Data Download:** [**https://depmap.org/portal/download/**](https://depmap.org/portal/download/)
* **Data Explorer:** [**https://depmap.org/portal/interactive/**](https://depmap.org/portal/interactive/)
* **Related publications:** [**https://depmap.org/portal/publications/**](https://depmap.org/portal/publications/)

**Regarding data:**

* Each row of the data is a **Cancer** **cell** line.
* CRISPR knockout each **gene** and collect the gene effect.
* For example, **BRAF(673)-** a human gene that encodes a protein called B-Raf. Gene ID: 673. This is a column in the CRISPR\_gene\_effect data. See for details about **BRAF**: <https://depmap.org/portal/gene/BRAF?tab=overview>
* Another example, KRAS

<https://depmap.org/portal/gene/KRAS?tab=characterization>

**Data type:**

*CRISPR KO* ***gene dependencie*** *data:*

***Mutational*** *data:*

***Copy number profiles*** *data: quantify? Log to -2, -1, 0(normal), 1, 2*

***RNA*** *expression data: count?*

**Mu (binary data 0,1)**

**lineages**

## Data set 1: CRISPR\_gene\_effect.csv

**Gene Effect scores** derived from **CRISPR knockout** screens published by Broad’s Achilles and Sanger’s SCORE projects.

Negative scores imply cell growth inhibition and/or death following gene knockout. Scores are normalized such that nonessential genes have a median score of 0 and independently identified common essentials have a median score of -1.

Gene Effect scores were inferenced by Chronos ( <https://genomebiology.biomedcentral.com/articles/10.1186/s13059-021-02540-7> )

Integration of the Broad and Sanger datasets was performed as described in <https://doi.org/10.1038/s41467-021-21898-7>, except that quantile normalization was not performed.

**Genes:17386**

**Cell Lines:1086**

**Primary Diseases:31**

**Lineages:28**

Source: **Broad Institute**

## Data set 2: CRISPR\_gene\_dependency.csv

Gene Dependency Probabilities represent the likelihood that knocking out the gene has a cell growth inhibition or death effect. These probabilities are derived from the scores in CRISPR\_gene\_effect.csv as described in <https://doi.org/10.1101/720243>.

Data Set 3: CCLE\_expression.csv

Gene expression TPM values of the protein coding genes for DepMap cell lines. Values are inferred from RNA-seq data using the RSEM tool and are reported after log2 transformation, using a pseudo-count of 1; log2(TPM+1).

Additional RNA-seq-based expression measurements are available for download as part of the full DepMap Data Release

More information on the DepMap Omics processing pipeline is available at <https://github.com/broadinstitute/depmap_omics>.

**Genes:19221**

**Cell Lines:1406**

**Primary Diseases:33**

**Lineages:30**

Source: **Broad Institute**

**CCLE:(Cancer Cell Line Encyclopedia)** [**https://depmap.org/portal/ccle/**](https://depmap.org/portal/ccle/)

## Data Set 4: CCLE\_mutations.csv

MAF file containing information on all the somatic point mutations and indels called in the DepMap cell lines. The calls are an ensemble of calls from MuTect1, MuTect2, and Strelka. A description of the various columns is in the DepMap Release README file.

Additional processed mutation datasets containing binary mutation calls are available for download as part of the full DepMap Data Release.

More information on the DepMap Omics processing pipeline is available at <https://github.com/broadinstitute/depmap_omics>.

Columns:

For all columns with AC, the allelic ratio is presented as [ALTERNATE:REFERENCE].

* CGA\_WES\_AC: the allelic ratio for this variant in all our WES/WGS(exon only) using a cell line adapted version of the 2019 CGA pipeline that includes germline filtering.
* SangerWES\_AC: in Sanger WES (called by sanger) (legacy)
* SangerRecalibWES\_AC: in Sanger WES after realignment at Broad (legacy)
* RNAseq\_AC: in Broad RNAseq data from the CCLE2 project (legacy)
* HC\_AC: in Broad Hybrid capture data from the CCLE2 project (legacy)
* RD\_AC: in Broad Raindance data from the CCLE2 project (legacy)
* legacy\_wgs\_exon\_only: in Broad WGS data from the CCLE2 project (legacy)

Additional columns:

* isTCGAhotspot: is this mutation commonly found in TCGA
* TCGAhsCnt: number of times this mutation is observed in TCGA
* isCOSMIChotspot: is this mutation commonly found in COSMIC
* COSMIChsCnt: number of samples in COSMIC with this mutation
* ExAC\_AF: the allelic frequency in the Exome Aggregation Consortium (ExAC)

Descriptions of the remaining columns in the MAF can be found here: <https://docs.gdc.cancer.gov/Data/File_Formats/MAF_Format/>

**Genes:18784**

**Cell Lines:1771**

**Primary Diseases:33**

**Lineages:30**

Source: **Broad Institute**

## Data Set 5 : CCLE\_gene\_cn.csv

Gene-level **copy number data** that is **log2** transformed with a pseudo-count of 1; log2(CN ratio + 1) . Inferred from WGS, WES or SNP array depending on the availability of the data type. Values are calculated by mapping genes onto the segment level calls and computing a weighted average along the genomic coordinate.

Additional copy number datasets are available for download as part of the full DepMap Data Release.

More information on the DepMap Omics processing pipeline is available at <https://github.com/broadinstitute/depmap_omics>.

**Genes:25368**

**Cell Lines:1766**

**Primary Diseases:33**

**Lineages:30**

Source: **Broad Institute**

## Data Set 6: CCLE\_wes\_gene\_cn.csv

Gene-level **copy number data** that is log2 transformed with a pseudo-count of 1. Inferred from only WES data by mapping genes onto the segment level calls.

Additional copy number datasets are available for download as part of the full DepMap Data Release.

More information on the DepMap Omics processing pipeline is available at <https://github.com/broadinstitute/depmap_omics>.

**Genes:59267**

**Cell Lines:1635**

**Primary Diseases:33**

**Lineages:30**

Source: **Broad Institute**

## Label file: sample\_info.csv

Metadata for all of DepMap’s cancer models/cell lines. A full description of each column is available in the DepMap Release README file.

* Columns:
* DepMap\_ID: Static primary key assigned by DepMap to each cell line
* cell\_line\_name: Original cell line name, including punctuation
* stripped\_cell\_line\_name: Cell line name with alphanumeric characters only
* CCLE\_Name: Previous naming system that used the stripped cell line name followed by the lineage; no longer assigned to new cell lines
* alias: Additional cell line identifiers (not a comprehensive list)
* COSMICID: Cell line ID used in Cosmic cancer database
* sex: Sex of tissue donor if known
* source: Source of cell line vial used by DepMap
* RRID: Cellosaurus research resource identifier
* WTSI\_Master\_Cell\_ID: ID of corresponding record in Sanger Drug dataset
* sample\_collection\_site: Tissue collection site
* primary\_or\_metastasis: Indicates whether tissue sample is from primary or metastatic site
* primary\_disease: General cancer lineage category
* Subtype: Subtype of disease; specific disease name
* age: If known, age of tissue donor at time of sample collection
* Sanger\_Model\_ID: Sanger Institute Cell Model Passport ID
* depmap\_public\_comments: Further information about the cell line
* lineage, lineage\_subtype, lineage\_sub\_subtype, lineage\_molecular\_subtype: Cancer type classifications in a standardized form
* default\_growth\_pattern: Typical growth pattern of the cell line
* model\_manipulation: Cell line modifications including drug resistance and gene knockout
* model\_manipulation\_details: Additional information about the model manipulation
* patient\_id: Identifier indicating which cell lines come from the same patient
* parent\_depmap\_id: If known, DepMap ID of parental cell line
* Cellosaurus\_NCIt\_disease: From Cellosaurus, NCI thesaurus disease term
* Cellosaurus\_NCIt\_id: From Cellosaurus, NCI thesaurus code
* Cellosaurus\_issues: From Cellosaurus, documented issues with cell line'

**Cell Lines:1840**

**Primary Diseases:33**

**Lineages:30**

Source: **Broad Institute**