# TCGA

## The Cancer Genome Atlas

Alexander Ukhatov

# The Cancer Genome Atlas (TCGA) is a…

- … project to catalogue the **genetic mutations responsible for cancer** using genome sequencing and bioinformatics. The overarching goal was to apply **high-throughput genome analysis techniques** to improve the ability to diagnose, treat, and prevent cancer through a better understanding of the genetic basis of the disease.

- TCGA was among the first to address many **ethical and logistical considerations** associated with collecting, analyzing, and providing access to data from human tissue specimens. To this end, TCGA formed an **Ethics, Law** and **Policy Group** with the goal of identifying and addressing critical ethical, legal and social questions faced by researchers and patients participating in the program. The group established informed consent guidelines for effective and **fair use of cancer genomic information**.

- The ultimate goal was to develop research policies maximizing public benefit from the data that were in accordance with these ethical and legal guidelines, ensuring:

  - Protection of human participants in the project, including their privacy

  - Secure and compliant access of TCGA data

  - Timely data release to the research community

  - Initial scientific publication by the data producers

**FOCUS ON SEQUENCING, NOT GENE EXPRESSION!!!**

- https://www.cancer.gov/ccg/research/genome-sequencing/tcga/history/timeline-milestones
- https://en.wikipedia.org/wiki/The_Cancer_Genome_Atlas

# The Cancer Genome Atlas (TCGA) is a…

## TCGA Cancers Selected for Study

The Cancer Genome Atlas (TCGA) selected the following cancers for study based on specific criteria that include:

- Poor prognosis
- Overall public health impact
- Availability of samples meeting standards for patient consent
- Availability of samples meeting standards for quality and quantity that include:
  - Primary, untreated tumor with a source of matched normal tissue or blood sample
  - Frozen, sufficiently sized, resection samples
  - Samples composed of at least 80% tumor nuclei (threshold later lowered to 60% with improved sequencing technology and computational methods)
- With support from patients, patient advocacy groups, and doctors, many rare cancers were also included

- https://www.cancer.gov/ccg/research/genome-sequencing/tcga/history/timeline-milestones
- https://en.wikipedia.org/wiki/The_Cancer_Genome_Atlas

# History

## 2005

**February**

The National Cancer Advisory Board's working group on biomedical technology recommends initiating a "Human Cancer Genome Project" with the aim of obtaining a comprehensive understanding of the genomic alterations that underlie all major cancers

**July**

The NCI and NHGRI hold a workshop, "Toward a Comprehensive Genomic Analysis of Cancer", on implementing a pilot phase of the Human Cancer Genome Project

**December**

The NIH launches The Cancer Genome Atlas with a three-year, $100 million pilot

- 2006 … **Lung, brain and ovarian cancers** will be the first mapped by TCGA
- 2008 … TCGA publishes an interim analysis on glioblastoma.
- 2009 … Genome Sequencing Centers shift from sequencing **601 genes** to hybrid capture methods that target more than **6,000 gene** and **miRNA sequences**
- 2010 … Fueled by the American Recovery and Reinvestment Act of 2009, NIH **extends** TCGA **to map 20 types of cancer**
- 2012 … TCGA publishes **marker paper on breast cancer**, revealing molecular similarities to ovarian cancer
- 2014 … TCGA publishes marker paper on **bladder cancer**
- 2014 … TCGA publishes marker papers on lung adenocarcinoma and gastric adenocarcinoma
- …

| | | | | | |
|---|---|---|---|---|---|
| mRNA Expression | mRNA sequencing | All | mRNA sequencing of tumor sampls using a poly(A) enrichment RNA preparation | *BAM*, TXT (normalized expression values per gene, isoform, exon, or splice junction) | May be labeled as RNASeqV1 and RNASeqv2 |
| | Total RNA Sequencing | Some tumor types | mRNA sequencing of tumor samples using ribosomal depletion RNA preparation | *BAM*, TXT (normalized expression values per gene, isoform, exon, or splice junction) | May be labeled as TotalRNASeqV2 |

# RNAseq

**https://www.cancer.gov/ccg/research/genome-sequencing/tcga/using-tcga-data/technology**

**https://www.cancer.gov/ccg/research/structural-genomics/tcga/using-tcga-data/technology/illumina-gaiix-data-sheet**

| | | | | | |
|---|---|---|---|---|---|
| Human Genome Sequencing Center at Baylor College of Medicine | IlluminaGA_DNASeq | Illumina Genome Analyzer DNA Sequencing | Genome Analyzer IIx | N/A | N/A |
| University of California Santa Cruz | IlluminaGA_DNASeq | Illumina Genome Analyzer DNA Sequencing | Genome Analyzer IIx | N/A | N/A |
| University of North Carolina | IlluminaGA_DNASeq | Illumina Genome Analyzer DNA Sequencing | Genome Analyzer IIx | N/A | N/A |
| BC Cancer Agency | IlluminaGA_miRNASeq | Illumina Genome Analyzer miRNA Sequencing | Genome Analyzer IIx | N/A | N/A |
| Harvard Medical School | IlluminaGA_mRNA_DGE | Illumina Genome Analyzer mRNA Digital Gene Expression | Genome Analyzer IIx | N/A | N/A |
| BC Cancer Agency | IlluminaGA_RNASeq | Illumina Genome Analyzer RNA Sequencing | Genome Analyzer IIx | N/A | N/A |
| University of North Carolina | IlluminaGA_RNASeqV2 | Illumina Genome Analyzer RNA Sequencing Version 2 analysis | Genome Analyzer IIx | N/A | N/A |
| Harvard Medical School | IlluminaHiSeq_DNASeqC | Illumina HiSeq for Copy Number Variation | HiSeq 2000 | N/A | N/A |
| BC Cancer Agency | IlluminaHiSeq_miRNASeq | Illumina HiSeq 2000 miRNA Sequencing | HiSeq 2000 | N/A | N/A |

https://www.cancer.gov/ccg/research/genome-sequencing/tcga/using-tcga-data/types

## Table of Contents

| | | | | |
|---|---|---|---|---|
| Microarray | BRCA, COAD, GBM, KIRC, KIRP, LAML, LGG, LUAD, LUSC, OV, READ, UCEC | | *CEL (raw signals per probe), TXT (raw signals per probe*, normalized expression values per probe, gene, or exons) | Probe information contained in array design files for each platform |

# Microarray

https://www.cancer.gov/ccg/research/genome-sequencing/tcga/using-tcga-data/technology
https://www.cancer.gov/ccg/research/structural-genomics/tcga/using-tcga-data/technology/agilent-g3-data-sheet

TCGA Characterization Platform Specifications

| Center | TCGA Platform Code | DCC Platform Name | Instrument Support Materials | Sequence Download | TCGA ADF Download |
|---|---|---|---|---|---|
| Broad Institute of MIT and Harvard | ABI | Applied Biosystems Sequence data | 3730/3730xl DNA Analyzers | Primers | N/A |
| McDonnell Genome Institute at Washington University | ABI | Applied Biosystems Sequence data | 3730/3730xl DNA Analyzers | Primers | N/A |
| Human Genome Sequencing Center at Baylor College of Medicine | ABI | Applied Biosystems Sequence data | 3730/3730xl DNA Analyzers | Primers | N/A |
| University of North Carolina | AgilentG4502A_07_1 | Agilent 244K Custom Gene Expression G4502A-07-1 | SurePrint G3 CGH+SNP Microarray | FASTA | TCGA ADF |
| University of North Carolina | AgilentG4502A_07_2 | Agilent 244K Custom Gene Expression G4502A-07-2 | SurePrint G3 CGH+SNP Microarray | FASTA | TCGA ADF |
| University of North Carolina | AgilentG4502A_07_3 | Agilent 244K Custom Gene Expression G4502A-07-3 | SurePrint G3 CGH+SNP Microarray | FASTA | TCGA ADF |
| Memorial Sloan Kettering Cancer Center | CGH-1x1M_G4447A | Agilent SurePrint G3 Human CGH Microarray Kit 1x1M | SurePrint G3 CGH+SNP Microarray | FASTA | TCGA ADF |

Novel and powerful algorithmic analysis detects regions of copy-neutral LOH or UPD (Figure 3). With high-quality DNA samples, the SNP call rate is greater than 95% with a greater than 99% accuracy.[2] The number and quality of copy number aberrations detected on the SurePrint G3 CGH+SNP microarrays is comparable to detection using G3 CGH-only microarrays,[2] with the added benefit of simultaneous identification of copy-neutral aberrations as small as 5 Mb. The new SurePrint G3 CGH+SNP array enables efficient, high-quality discovery of chromosomal aberrations that cannot be detected in a single assay using any other method.

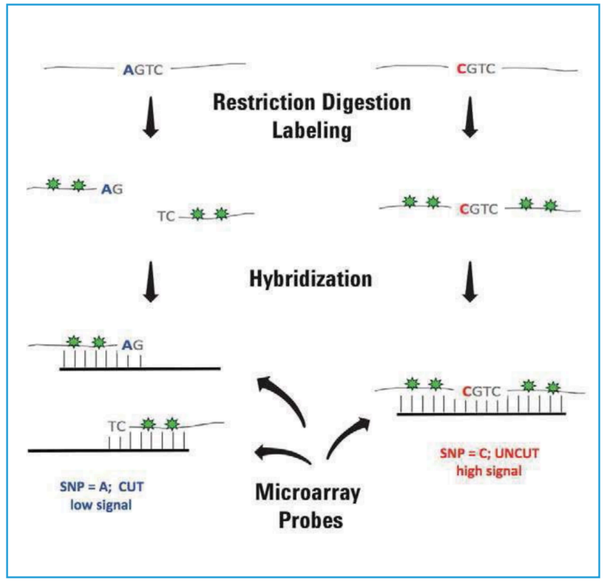
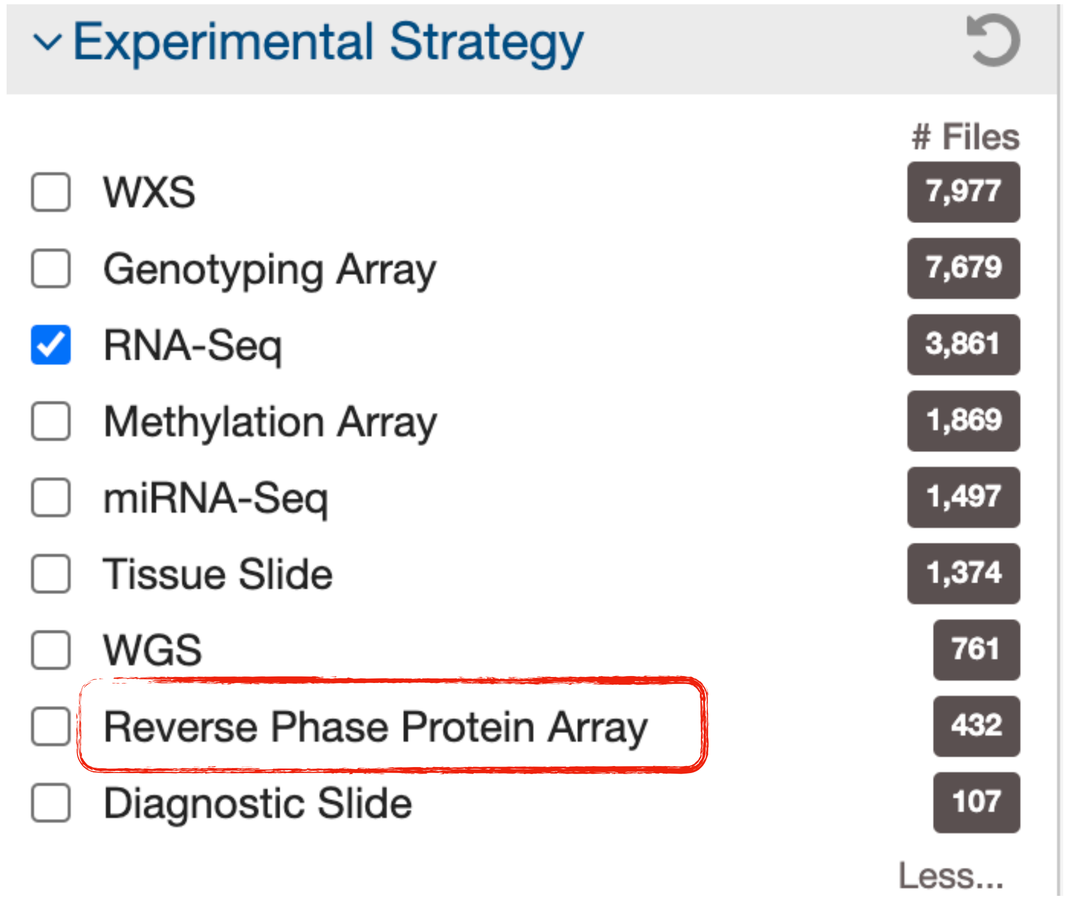



**Figure 2.** *The SurePrint G3 CGH+SNP microarrays use the same Agilent CGH workflow as the SurePrint CGH-only arrays. Restriction digestion of genomic DNA allows genotyping of SNPs located in the enzymes' recognition sites.*

https://www.cancer.gov/ccg/research/genome-sequencing/tcga/using-tcga-data/types

# Why?




Where I can get information about microarray?

https://portal.gdc.cancer.gov/repository?facetTab=files&filters=%7B%22op%22%3A%22and%22%2C%22content%22%3A%5B%7B%22content%22%3A%7B%22field%22%3A%22cases.project.project_id%22%2C%22value%22%3A%5B%22TCGA-OV%22%5D%7D%2C%22op%22%3A%22in%22%7D%2C%7B%22op%22%3A%22in%22%2C%22content%22%3A%7B%22field%22%3A%22files.experimental_strategy%22%2C%22value%22%3A%5B%22RNA-Seq%22%5D%7D%7D%5D%7D&searchTableTab=files

# Questions

- Where to find info about the place and time of the measurement? Which method was used?
- Where to get the accuracy (errors of the measurements) for each of the method?
- Why TCGA website doesn’t show information about microarray?

(https://portal.gdc.cancer.gov/repository?facetTab=files&filters=%7B%22op%22%3A%22and%22%2C%22content%22%3A%5B%7B%22content%22%3A%7B%22field%22%3A%22cases.project.project_id%22%2C%22value%22%3A%5B%22TCGA-OV%22%5D%7D%2C%22op%22%3A%22in%22%7D%2C%7B%22op%22%3A%22in%22%2C%22content%22%3A%7B%22field%22%3A%22files.data_format%22%2C%22value%22%3A%5B%22tsv%22%5D%7D%7D%2C%7B%22op%22%3A%22in%22%2C%22content%22%3A%7B%22field%22%3A%22files.experimental_strategy%22%2C%22value%22%3A%5B%22RNA-Seq%22%5D%7D%7D%5D%7D&searchTableTab=files)