TCGA: A Community Resource Looking for a Broader Community

Kenna Shaw, Ph.D.
Director
The Cancer Genome Atlas
Launched in 2006 as a pilot and expanded in 2009, the goals of TCGA are to:

• Establish infrastructure for effective team science

• Develop a scalable “pipeline” beginning with highest quality samples

• Determine the feasibility of a large-scale, high throughput approach to identifying the molecular ‘parts-list’

• Evaluate using statistically-robust sample sets

• Make the data publicly and broadly available to the cancer community while protecting patient privacy
TCGA: “No Platform Left Behind”

25* forms of cancer

glioblastoma multiforme (brain)
squamous carcinoma (lung)
serous cystadenocarcinoma (ovarian)

Multiple data types
- Clinical diagnosis
- Treatment history
- Histologic diagnosis
- Pathologic report/images
- Tissue anatomic site
- Surgical history
- Gene expression/RNA sequence
- Chromosomal copy number
- Loss of heterozygosity
- Methylation patterns
- miRNA expression
- DNA sequence
- RPPA (protein)
- Subset for Mass Spec

Biospecimen Core Resource with more than 150 Tissue Source Sites
6 Cancer Genomic Characterization Centers
3 Genome Sequencing Centers
7 Genome Data Analysis Centers

Data Coordinating Center
Tumor Project Progress

- Manuscript submitted or published
- Analysis underway
- Sample acquisition phase
- Rare tumor project
- Only accepting AA cases
Rare Tumor Project (Initiated March 2012)

- Adrenocortical Carcinoma
- Adult ALL (B-cell and T-Cell)
- Anaplastic Thyroid
- Cholangiocarcinoma
- Chromophobe kidney
- High Risk MDS (del 5q- cases)
- Mesothelioma
- Paraganglioma/Pheochromocytoma
- Testicular Germ Cell
- Thymoma
- Uterine Carcinosarcoma
- Sarcomas
- Others??
## TCGA: Platforms- Then and Now

<table>
<thead>
<tr>
<th>Platform</th>
<th>Pilot</th>
<th>Expansion</th>
</tr>
</thead>
<tbody>
<tr>
<td>SNP/CNV</td>
<td>Affy SNP 6.0</td>
<td>Affy SNP 6.0</td>
</tr>
<tr>
<td></td>
<td>Agilent CGH Array</td>
<td>Low Pass Sequencing*</td>
</tr>
<tr>
<td></td>
<td>Illumina 1M Duo</td>
<td></td>
</tr>
<tr>
<td>Methylation</td>
<td>Infinium Array</td>
<td>Infinium Array</td>
</tr>
<tr>
<td>mRNA</td>
<td>Agilent 244K Array</td>
<td>RNAseq</td>
</tr>
<tr>
<td></td>
<td>Affy Human Exon Array</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Affy U133 Array</td>
<td></td>
</tr>
<tr>
<td>miRNA</td>
<td>Agilent 8 x 15K Array</td>
<td>RNAseq</td>
</tr>
<tr>
<td>Mutation</td>
<td>600-1000 genes</td>
<td>DNAseq:</td>
</tr>
<tr>
<td></td>
<td></td>
<td>100% whole exomes</td>
</tr>
<tr>
<td></td>
<td></td>
<td>10% whole genomes</td>
</tr>
<tr>
<td>Proteomics</td>
<td>None</td>
<td>Reverse Phase Protein Arrays</td>
</tr>
<tr>
<td>Clinical Data</td>
<td>Minimum Enrollment &amp; Follow-Up*</td>
<td>Minimum Enrollment &amp; Follow-Up*</td>
</tr>
<tr>
<td></td>
<td>H&amp;E from Frozen Section Images</td>
<td>Pathology Reports</td>
</tr>
<tr>
<td></td>
<td>Treatment Data</td>
<td>H&amp;E from Frozen &amp; Diagnostic Images</td>
</tr>
</tbody>
</table>

* Not a core platform; Not all samples will have data file for this platform

TCGA: Data Availability

- 7,136 cases across 20+ tumor types
- 5,865 with minimum clinical data set
- 3,893 with at least 1 year follow-up; ~50% with treatment data
- 105,000 samples of RNA/DNA/protein shipped between 2006 & 6/2012
- All but 13,000 samples have data returned:
  - ~87% of all samples have Level 1 data publicly available
  - TCGA Program Office to complete accounting with Batelle/QMS by end of 2012
TCGA: The Pipeline for Comprehensive Characterization

Tissue Sample

- Pathology QC
- DNA & RNA Isolation, QC
- Sequencing
  - Expression, CNA & LOH, Epigenetics
  - Data Storage at DCC & CGHub

GDAC

Integrative Analysis

Comprehensive Characterization of a Cancer Genome

- SNP 6.0 ~45d
- Methylation ~60d
- miRNAseq ~105d
- mRNAseq ~120d
- DNAseq Exome ~180d

3 months – 2 years

~90d

~12-24 months
TCGA: The Pipeline for Comprehensive Characterization

Tissue Sample

- Pathology & Clinical data QC
- DNA & RNA Isolation, QC

Sequencing

Data Storage at DCC & CGHub

Expression, CNA & LOH, Epigenetics

Integrative Analysis

GDAC

Comprehensive Characterization of a Cancer Genome
Sample Criteria Limit ‘Askable’ Questions

- **Primary, adult** tumors (except for melanoma and triplets)
- **Malignant** (no *in situ* cases)
- Snap frozen, <60min from clamp to LN2
- ~ 50 mg (*biopsies starting to be feasible*)
- Pathology review of tissue sent to TCGA
- No more than 20% necrosis; ≥ 60%* tumor cells
- **No prior treatment**
- *Matched source of germline*: Blood (buffy coat/white cells)/saliva or skin for liquid tumors
- Clinical annotation; but not pre-analytic variables
- IRB approval for use in TCGA; proactive consenting for genomic studies
- MTA w/out retention of IP
End Goal: Making an Exhaustible Resource Inexhaustible
TCGA Data Portal

TCGA Data Portal Overview

We provide 3 ways to download data. The Cancer Genome Atlas (TCGA) Data Portal provides a platform for researchers to search, download, and analyze data generated by TCGA. It contains clinical information, genetic characterization data, and high-throughput sequencing analysis of the tumor genomes.

Available Cancer Types

<table>
<thead>
<tr>
<th>Available Cancer Types</th>
<th>Patients with Samples</th>
<th>Downloadable Tumor Samples</th>
<th>Date Last Updated (mm/dd/y)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Acute Myeloid Leukemia (AML)</td>
<td>209</td>
<td>209</td>
<td>03/01/12</td>
</tr>
<tr>
<td>Bladder Urothelial Carcinoma (BUC)</td>
<td>89</td>
<td>76</td>
<td>03/01/12</td>
</tr>
<tr>
<td>Brain Lower Grade Glioma (BLG)</td>
<td>144</td>
<td>80</td>
<td>03/01/12</td>
</tr>
<tr>
<td>Breast Invasive carcinoma (BRCA)</td>
<td>864</td>
<td>917</td>
<td>03/01/12</td>
</tr>
<tr>
<td>Central Nervous System Tumors (CNS)</td>
<td>99</td>
<td>37</td>
<td>03/01/12</td>
</tr>
<tr>
<td>Colorectal adenocarcinoma (CRA)</td>
<td>423</td>
<td>422</td>
<td>03/01/12</td>
</tr>
<tr>
<td>Glioblastoma multiforme (GBM)</td>
<td>599</td>
<td>592</td>
<td>03/01/12</td>
</tr>
<tr>
<td>Head and Neck squamous cell carcinoma (HNSC)</td>
<td>204</td>
<td>202</td>
<td>03/01/12</td>
</tr>
<tr>
<td>Kidney renal clear cell carcinoma (KRCC)</td>
<td>502</td>
<td>501</td>
<td>03/01/12</td>
</tr>
<tr>
<td>Kidney renal papillary cell carcinoma (KRPP)</td>
<td>103</td>
<td>97</td>
<td>03/01/12</td>
</tr>
<tr>
<td>Liver hepatocellular carcinoma (LHCC)</td>
<td>84</td>
<td>55</td>
<td>03/01/12</td>
</tr>
<tr>
<td>Lung adenocarcinoma (LUAD)</td>
<td>374</td>
<td>351</td>
<td>03/01/12</td>
</tr>
</tbody>
</table>

3/01/12 - Server Migration Complete

We are pleased to announce that the migration to new hardware is complete and all DCC software is now running in the new environment. Because this is a new hardware installation, there is the possibility for unexpected issues to arise. If you have any problems logging into or using the systems once service resumes, please email bioinformatics@nci.nih.gov.

3/02/12 - Wiki User Guide

The TCGA Members Wiki User Guide has been posted to the wiki. This guide contains information on gaining access to the wiki, how to navigate, how to upload files, the structure of the wiki, and rules for posting files and metadata. The user guide may be accessed at https://wiki.nci.nih.gov/TCGA/wiki/index.php?title=TCGA

More TCGA Information

For more information about The Cancer Genome Atlas project, please visit the TCGA website (www.cancer.gov/tcga).

Data Matrix Datasets

- Cancer Cell Lines
- Cancer Genes
- Cancer Mutations
- Cancer Proteins
- Cancer Signatures

TCGA Website

https://tcga-data.nci.nih.gov/tcga/
Ongoing Pipeline Analysis Available

http://gdac.broadinstitute.org
Where to find TCGA Sequence Data

• Moved from Short Read Archive (SRA at NCBI) to UCSC
• Open for downloads as of January 2012

- 2 petabytes now, 5Pb \( (5 \times 10^{15}) \) total by 2014
- General Parallel File System, Dual RAID 6 subsystems, Redundant I/O paths
- Currently holds 10,000 files; expected to double in next 3 months
- Co-location opportunities in same data center for groups who want to compute on the data
- User support: support@cghub.ucsc.edu

https://cghub.ucsc.edu
March 2011: International Cancer Genome Consortium Projects

**ICGC**

Chair of Secretariat: Thomas Hudson, MD
Ontario Institute of Cancer Research

**Canada**
- Pancreatic cancer (Ductal adenocarcinoma)
- Prostate cancer (Adenocarcinoma)

**United States**
- Bladder cancer
- Blood cancer (Acute myeloid leukemia)
- Brain cancer (Glioblastoma multiforme/low grade glioma)
- Breast cancer (Ductal & lobular)
- Cervical cancer (Squamous)
- Colon cancer (Adenocarcinoma)
- Endometrial cancer (Endometrial carcinoma)
- Gastric cancer (Adenocarcinoma)
- Head and neck cancer (Squamous cell carcinoma/Thyroid carcinoma)
- Renal cancer (Renal cell carcinoma / Renal papillary carcinoma)
- Liver cancer (Hepatocellular carcinoma)
- Lung cancer (Adenocarcinoma / Squamous cell carcinoma)
- Ovarian cancer (Serous cystadenocarcinoma)
- Prostate cancer (Adenocarcinoma)
- Rectal cancer (Adenocarcinoma)
- Skin cancer (Cutaneous melanoma)

**EU/United Kingdom**
- Breast cancer (ER positive, HER2 negative)

**EU/France**
- Renal cancer (Renal cell carcinoma) (Focus on but not limited to clear cell subtype)

**France**
- Breast cancer (Subtype defined by an amplification of the HER2 gene)
- Liver cancer (Hepatocellular carcinoma) (Secondary to alcohol and adipsis)
- Prostate cancer (Adenocarcinoma)

**Spain**
- Chronic lymphocytic leukemia (CLL with mutated and unmutated IgVH)

**Germany**
- Malignant lymphoma (Germinal center B-cell derived lymphomas)
- Pediatric brain tumors (Medulloblastoma and Pediatric pilocytic astrocytoma)
- Prostate cancer (Early onset)

**China**
- Gastric cancer (Intestinal- and diffuse-type)

**Japan**
- Liver cancer (Hepatocellular carcinoma) (Virus-associated)

**Australia**
- Ovarian cancer (Serous cystadenocarcinoma)
- Pancreatic cancer (Ductal adenocarcinoma)
- Prostate cancer
Acknowledgements

Center for Cancer Genomics
Stephen Chanock & Lou Staudt

NCI Program Office
Kenna Shaw
Liming Yang
Roy Tarnuzzer
Zhining Wang
Emma Spaulding
Margi Sheth
John Demchok
Julia Zhang
Martin Ferguson
Greg Eley

NHGRI
Brad Ozenberger
Heidi Sofia
Lindsay Lund
Mark Guyer

The Cancer Genome Atlas