The Cancer Genome Atlas: Update for the National Cancer Advisory Board

Anna D. Barker, Ph.D.
Deputy Director, National Cancer Institute

September 15, 2009
A Look Back at The Cancer Genome Atlas (TCGA) Pilot Project

Significant Milestones and Lessons Learned from TCGA Pilot Project

Phase II of TCGA (Joined by Dr. Mark Guyer of NHGRI)

The Significance of TCGA to Cancer and Biomedical Research
Biological significance of understanding genomic changes in cancer:

- Copy number
- Expression (regulation of)
- Regulation of translation
- Mutations
- Epigenome

Cancer is a disease of genomic alterations – identification of all genomic changes would enable defining cancer subtypes – potential to transform cancer drug discovery, diagnostics and prevention.
Background for TCGA Pilot

- Cancer biology and genome sequencing technology advanced in parallel at extraordinary rates over the past several years.
- Cancer genomics developed rapidly through the efforts of individual investigators – over 300-600 genes associated with various cancers.
- Following several workshops and a specific recommendation by the National Cancer Advisory Board, TCGA was launched as a joint pilot project between the NCI and NHGRI in 2006.
- TCGA was designed as a pilot to evaluate and test several parameters (large scale genome characterization and sequencing, integration of laboratories and teams; policies ranging from data standards and access; to biospecimens and informed consent).
- The pilot explored the processes needed to perform high-throughput, large scale disease-focused genome characterization, data integration and analysis.
Goals for TCGA Pilot

Launched in 2006 as a pilot program - The Cancer Genome Atlas (TCGA) Pilot Program, a collaboration between the NCI and NHGRI the goals were to:

- Establish the needed infrastructure;
- Develop a scalable “pipeline” beginning with high quality samples;
- Determine the feasibility of a large-scale, high throughput, systematic approach to identifying all of the relevant genetic alterations in cancer;
- Systematically evaluate up to three cancers using a statistically-robust sample set (500 cancers and matched controls);
- Make the data publicly and broadly available to the cancer communities in a manner that protected patient privacy
TCGA Sample Criteria

- Primary tumor only
- Snap frozen
- ~ 200 mg
- No more than 20% necrosis; ≥ 80% tumor cells
- Normal tissue: Blood (buffy coat/white cells); adjacent normal tissue or buccal cells; or ≥ 13µg high-quality DNA
- All “Tier One” Clinical Data Elements (15 or more)
  (Goal of 500 each tumor/normal pairs for each cancer type to achieve detection of background mutations at 5% level)
TCGA Pilot Project Infrastructure

**Development of New Analyses**
- Tools
- Views

**Data Management, Bioinformatics, and Computational Analysis**
- Data Coordinating Center, DCC
- Analyses of data

**Genome Sequencing Centers**
- High throughput sequencing of genes and genomic regions identified through cancer characterization

**Technology Development**
- Increased sensitivity of molecular characterization platforms
- Analysis of biomolecules from 1000 cells or less

**Cancer Genome Characterization Centers**
- Identification of expression alternation
- Detection of DNA fragment copy number changes and LOH
- Epigenetics

**Human Cancer Biospecimen Core Resource**
- Biospecimens-related data storage
- Histopathology confirmation performed
- Biomolecules isolated, QC'ed and distributed

Communicate
TCGA Pilot Project Pipeline

- Tissue Sample
  - Pathology QC
  - DNA & RNA Isolation, QC

- Analysis
  - Sequencing
  - Expression, CNA & LOH, Epigenetics

- Data and Results Storage & QC

- Integrative Analysis
  - Comprehensive Multi-Dimensional Integrated Data Set

Symbols:
- = Process
- = Data
- = Results

Colors:
- BCR
- GSCs
- CGCCs
- DCC
- Collaborators
Three forms of cancer

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

Multiple data types

- Clinical diagnosis
- Treatment history
- Histologic diagnosis
- Pathologic status
- Tissue anatomic site
- Surgical history
- Gene expression
- Chromosomal copy number
- Loss of heterozygosity
- Methylation patterns
- miRNA expression
- DNA sequence
Milestones and Lessons Learned from TCGA Pilot Program
GBM Findings

- September 2008, TCGA published study of glioblastoma (GBM), reported discovery of new mutations – confirmed many “maybes” (Nature)

- Data types integrated across labs and across the genome, transcriptome, epigenome – clinical data and outcomes

  - Performed in-depth, integrated characterization of the tumor genomes of 206 GBM patients
  - Identified three genes and three core biological pathways commonly altered in GBM tumors
  - Discovered possible mechanism by which GBM tumors become resistant to TMZ
RTK/RAS/PI-3K signaling altered in 88%

P53 signaling altered in 87%

RB signaling altered in 78%
Potentially Clinically-relevant Discovery in Treated GBMs

- Current standard of care for GBM is treatment with the alkylating agent temozolomide (TMZ)
  - The promoter of *O-6-methylguanine-DNA methyltransferase (MGMT)* is methylated in most treated cases
  - Most tumors which have inactivated *MGMT* are “hypermutated”, i.e. statistically increased mutations rates and many have mutations in mis-match repair (MMR) genes

- Is *MGMT* inactivation the mechanism to TMZ resistance?
  - Methylated MGMT is unable to repair alkylated guanine residues caused by TMZ
  - Inactive MMR genes can not repair the alkylating damage and move the cells into the apoptotic pathway - cells survive and multiply

- Potential for translational endpoint and impact on current GBM management

TCGA Nature 2008
Nex-Gen sequencing technology applied for ovarian cancer

Overall, the ovarian cancer genome has large numbers of rearrangements and amplifications – “noisy genomes”

Possible that P53 mutated in 100% of ovarian samples

High frequency BRAC1 and BRAC2 mutations

Number of other known oncogenes identified

Sequence data available in October – publication in process

Integrated multi-dimensional data set will set a new standard for cancer genomics
A contrast in copy number complexity

Glioblastoma

Serous ovarian cancer

Copy number abnormality statistic

Distance along the genome
Expression subtypes

Glioblastoma

Serous Ovarian
DNA Methylation Data Identifies 3 Clusters of Serous Ovarian Tumors

Overlap Between Expression and DNA Methylation Cluster Membership

<table>
<thead>
<tr>
<th>Numbers of Tumors</th>
<th>Methylation Cluster 1</th>
<th>Methylation Cluster 2</th>
<th>Methylation Cluster 3</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Expression Cluster 1</strong></td>
<td>21</td>
<td>41</td>
<td>1</td>
<td>63</td>
</tr>
<tr>
<td><strong>Expression Cluster 2</strong></td>
<td>5</td>
<td>36</td>
<td>3</td>
<td>44</td>
</tr>
<tr>
<td><strong>Expression Cluster 3</strong></td>
<td>6</td>
<td>14</td>
<td>40</td>
<td>60</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td>32</td>
<td>91</td>
<td>44</td>
<td>167</td>
</tr>
</tbody>
</table>

Consensus Clustering of 238 High-Grade Serous Ovarian Tumors with 3,226 Variant Probes

Methylation Clusters

Expression Clusters

5-Year Overall Survival
n=146 p=0.05

Slide courtesy of P. Laird/S. Baylin, Analysis Team
TCGA Pilot Program: Overall Summary

- Set up and functionalized all part of TCGA network (10 centers, over 150 scientists) – and developed pipeline from samples to data availability
- Built an unprecedented team of scientists, oncologists, pathologists, bioethicists, technologists and bioinformaticists and a working pipeline from sample to data release
- Set a high bar for sample quality and percentage of tumor nuclei – which drove data quality
- Implemented 2nd generation sequencing methods - Included intensive effort on computational methods; worked NCBI to pioneer controlled-access release of human medical sequencing large data sets

Outcomes to date:
- Signal can be differentiated from “noise”
- New cancer genes have been discovered – beyond the “streetlamps”
- Tumor subtypes can be differentiated based on comprehensive knowledge of genomic alterations
- The integrated teams can be built – and it will take teams to analyze multi-dimensional data
- Clinically relevant data has/will come from this comprehensive approach
- High-throughput large-scale comprehensive characterization is possible and a prerequisite to defining the range and biologic effects of genomic alterations (and their expression) in cancer
- Single targets – unlikely – pathway biology in cancer is likely our best hope – argues strongly for rational combinations and/or new generations of interventions
Phase II TCGA
TCGA Phase II: Overview

- ARRA funding will be employed for 2 years to collect tissues for years 1-5 of TCGA – and scale up the Biospecimen Core Resource
- During two years of ARRA funding – plan to complete comprehensive genome characterization of 10 tumor types (at 200 cases/tumor type as a discovery set and more depending on tumor type); 200 exomes; 20 whole genomes/tumor
- GCCs will perform expression, CN, SNP analysis, Methylation and miRNA characterization
- Genome Sequencing Centers will use Nex-Gen sequencing technologies – exomes and whole genomes (cost dependent)
- Genome Data Analysis Centers will integrate data from GCCs – GDAC-Bs will further integrate data, create new models and tools to refine and further add value to data for communities
TCGA Phase II: Goals

- Project will scale – production level pipeline for 20 tumors
- Increased emphasis on an analysis pipeline
- Integration of next generation genome characterization/sequencing technologies

Specific Phase II goals:

- Standards and SOPs for biospecimen acquisition - high quality of all aspects of samples, clinical information and data
- Mix of common and rare tumors – emphasis on highly lethal tumors – focus on subtypes as appropriate
- Complete genome characterization each cancer case
- Two levels of data integration and analysis – advanced approaches and tools for visualization and management of data
- Quality management system
NCI’s ARRA investment is focused on the front end of TCGA pipeline – tissue accrual and biomolecule preparation.

Samples will be procured through competitive RFPs for retrospective samples and prospective networks.

TCGA Phase II requires approximately 20,000 cases from 20 different tumor types.

Final goals for accrual assumes a 50% failure rate in production.

Accrual through prospective networks will be based on prevalence of disease.

BCR expansion – addition of second core resource.
NHGRI - Next Generation Genome Sequencing for TCGA

(Dr. Mark Guyer, NHGRI)
Impact of TCGA

http://cancergenome.nih.gov
Lessons Learned to Date from TCGA Pilot Project

- This is really hard – but with dedication to quality at all levels – it is one of our best bets to generate the knowledge we need in the biological space
- Quality of tissue impacts directly on the quality of molecular characterization data generated
- ~500 cases per cancer studied provides enough power to detect changes at the 3-5% level
- Retrospective cancer cases which have high quality samples and clinical annotation, including treatment and outcome are difficult to find and procure – so prospective collections and characterization are a better bet to maximize investment and produce dependable data
- Large scale data generation requires an analytical pipeline to ensure close to a “real-time” interpretation of the results
- If the data are good enough – and the problem is really hard – the analysis teams emerge
TCGA: Driving a New Model for Drug/Diagnostics Development

- TCGA is developing the required high quality multi-dimension data
- Cancer genomes are digital – knowable); not known - how much we have to know (We need the “parts list”)
- Discovering genes one at a time…no longer makes sense
- Support making it all public – the IP will come from the analysis – and integrating the genome characterization with clinical data and outcomes
- We need translational infrastructure turned to the analysis and translation of the data – private sector should significantly engage
- Need virtual translational genomics “centers” – could be next generation, mutually beneficial public-private partnership
TCGA: Filling in the Biologic Knowledge Space

- Genetics/Genomics
- Gene Expression
- Proteomics
- Metabolomics
- Tissues/Organs
- Environment
- Organisms
- Cancer Biology
Some Questions for Discussion

- Is it a better strategy to completely characterize a cancer (e.g., the GBM cancer genome) – or move on to discovery sets in a number of cancers?

- How can NCI engage its translational infrastructures to do in depth analysis of TCGA data? Let it happen – or be more proactive?

- Can tumor subtype information be translated more directly and quickly – so that patients can benefit as soon as possible?