TCGA: Progress and Challenges

Lynda Chin
Belfer Institute for Applied Cancer Science
Dana-Farber Cancer Institute
Harvard Medical School
Broad Institute

September 8, 2010
Goals of cancer medicine and the promise of Cancer Genomics

- **Prevention**
  - Understanding the underlying etiology → strategy
- **Detection**
  - Identify risk alleles / genomic events for screening
- **Intervention**
  - Stratify high vs low risk patients to treat or not
  - Identify new therapeutic targets for drug discovery
  - Inform selection of the right patient for the right drug
  - Define combination / co-extinction strategies
  - Understand resistance mechanisms
Multi-dimensional Cancer Genomics

Aneuploidy; Re-arrangement; Translocation

Gene Splicing Alterations

Copy number aberrations

Methylation or histone modification

Somatic mutations

Altered expression
TCGA Pilot (2006 – 2009)

Aneuploidy; Re-arrangement; Translocation

Gene Splicing Alterations

Copy number aberrations

Methylation or histone modification

Somatic mutations

Each Sample

Altered expression
Comprehensive genomic characterization defines human glioblastoma genes and core pathways

The Cancer Genome Atlas Research Network

- A Reference GBM cancer genome
  - PIK3R1 mutation is frequent in GBM
  - NF1 is involved in sporadic GBM in human
  - TP53 is commonly mutated in primary GBM

- Unanticipated discoveries...
  - Hypothesis on a possible resistance mechanism to temozolomide (TMZ)

- Integrative analyses → Pathway knowledge
Integrated Genomic Analysis Identifies Clinically Relevant Subtypes of Glioblastoma Characterized by Abnormalities in PDGFRA, IDH1, EGFR, and NF1

Roel G.W. Verhaak1, 2, 6, 7 Katherine A. Hoadley, 2, 6, 7, 12, 15 Elizabeth Purdom, 7, 12, 15 Victoria Wang, 8, 12, 15 Yuan Qi, 4, 5 Matthew D. Wilkerson, 4, 6, 7, 12, 15 C. Ryan Miller, 4, 6, 7, 12, 15 Li Ding, 9, 12, 15 Todd Golub, 1, 10, 12, 15 Jill P. Mesirov, 1, 12, 15 Gabriele Alexe, 1 Michael Lawrence, 1, 2, Michael O’Kelly, 1, 2, Pablo Tamayo, 1, 12, 15 Barbara A. Weir, 1, 2, Stacey Gabriel, 1, Wendy Winckler, 1, 2, Supriya Gupta, 1, Lakshmi Jakkula, 11, Heidi S. Feiler, 11, J. Graeme Hodgson, 12, C. David James, 12, Jann N. Sarkaria, 13, Cameron Brennan, 14, Ari Kahn, 15, Paul T. Spellman, 11, Richard K. Wilson, 9, Terence P. Speed, 7, 16 Joe W. Gray, 11, Matthew Meyerson, 1, 2, Gad Getz, 1, Charles M. Perou, 3, 4, 8, D. Neil Hayes, 3, 4, 8, and The Cancer Genome Atlas Research Network

Proneural | Neural | Classical | Mesenchymal
---|---|---|---
TP53
IDH1
PDGFRA
EGFR
NF1
CDKN2A

Gene Expression
-2 0 2
Copy Number
-2 -1 0 1 2
Mutation
wt mut
TP53 LOH
EGFRvIII
• Occurs in Younger Patients
• Is a Subset of Proneural Expression Subtype
• Is Associated with Better Survival
• Is More Frequent in Low-Grade Gliomas
• Is Not Associated with MGMT Methylation
• Is Tightly Linked to IDH1 Mutation

Noushmehr et al. (2010) Cancer Cell, Online

Identification of a CpG Island Methylator Phenotype that Defines a Distinct Subgroup of Glioma
Clinicopathological correlation...

**Consensus path review on digital H&E images**

Daniel Brat  
Scott Vandenberg  
Roger McLendon  
David Louis  
Norm Lehman  
Mark Cohen  
Ryan Miller  
Matt Schniederjan  

**p <= 6.7 e-05**

**p <= 1.4 e-06**

---

**GBM Histological Features in Permanent Sections**

Case:  
Slide: TCGA-12-0657-01Z-00-DX1

- Yes □ No □ There is sufficient tissue on the sections to confirm GBM (necrosis & microvascular hyperplasia)
- Yes □ No □ There is sufficient tissue on the sections to collect data for this review (see criteria below)
- MIB-1 Index (surgical pathology report)

Collection method:  
- Needle Biopsy  
- Open Craniotomy  

Please select exactly one box per item below:

<table>
<thead>
<tr>
<th>Definitions:</th>
<th>Not Present</th>
<th>Present</th>
<th>Abundant</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Present</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Abundant</td>
<td></td>
</tr>
<tr>
<td>Microvascular hyperplasia</td>
<td>□</td>
<td>□*</td>
<td>□*</td>
</tr>
<tr>
<td>Complex glomeruloid</td>
<td>□</td>
<td>□*</td>
<td>□*</td>
</tr>
<tr>
<td>Circumferential endothelial hyperplasia</td>
<td>□</td>
<td>□*</td>
<td>□*</td>
</tr>
<tr>
<td>Necrosis</td>
<td>□</td>
<td>□*</td>
<td>□*</td>
</tr>
<tr>
<td>Multiple serpentine pseudopalisading pattern</td>
<td>□</td>
<td>□*</td>
<td>□*</td>
</tr>
<tr>
<td>Zonal necrosis</td>
<td>□</td>
<td>□*</td>
<td>□*</td>
</tr>
</tbody>
</table>

**Giant Cells**

<table>
<thead>
<tr>
<th></th>
<th>p53-wt</th>
<th>p53 mut/del</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>80%</td>
<td>20%</td>
</tr>
<tr>
<td>1+</td>
<td>33%</td>
<td>67%</td>
</tr>
<tr>
<td>2+</td>
<td>0%</td>
<td>100%</td>
</tr>
</tbody>
</table>

**P53 pathway intact**

<table>
<thead>
<tr>
<th>Giant Cells</th>
<th>p53 pathway intact</th>
<th>p53 pathway altered</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>75%</td>
<td>25%</td>
</tr>
<tr>
<td>1+</td>
<td>14%</td>
<td>85.7%</td>
</tr>
<tr>
<td>2+</td>
<td>0%</td>
<td>100%</td>
</tr>
</tbody>
</table>

---

**p <= 6.7 e-05**

**p <= 1.4 e-06**
Enabling resource

- Citation in 225 publications
  - Comparison with mouse models
  - Novel gene discovery and pathway analyses
  - Analysis of germline genetics
  - Novel computational algorithm development
  - In silico correlation studies
Conclusions from the pilot

- Cancer genome is highly complex and heterogeneous
  - Technologies can detect the signals above the noises
- There are new discoveries to be made
  - Detect known and discover unknown genes
  - Discovery of novel subclass, e.g. G-CIMP
- Multi-dimensional analyses enable integrative analyses
  - Pathway → Network view → translational potential
- Unbiased approach generates unanticipated hypotheses
  - Mechanism for TMZ resistance
- Reference-quality data with stringent QC as an enabling resource
  - GBM dataset has been used/referenced in > 225 publications
- The acquisition of large cohorts of high-quality clinically annotated tumor samples is critical but extremely challenging
  - Investment in biospecimen banking / infrastructure
Unique Challenges of TCGA

- Reference = Complete + Quality
  - Quality: samples $\rightarrow$ biomolecules $\rightarrow$ data $\rightarrow$ analyses
  - Complete: Multi-dimensionality; global assays
  - Complete: sufficiently powered sample size
What is the power of a discovery set of 21 samples? (Wood et al.)

- We took 100,000 subsets of 21 samples out of the 84 (non-hyper mutated GBM samples from our paper) and calculated the frequency that each of the 8 significant genes would have been detected as significant.

<table>
<thead>
<tr>
<th>Gene</th>
<th>Detected in fraction</th>
</tr>
</thead>
<tbody>
<tr>
<td>TP53</td>
<td>97.5%</td>
</tr>
<tr>
<td>PTEN</td>
<td>95.6%</td>
</tr>
<tr>
<td>EGFR</td>
<td>70.9%</td>
</tr>
<tr>
<td>NF1</td>
<td>69.5%</td>
</tr>
<tr>
<td>RB1</td>
<td>48.9%</td>
</tr>
<tr>
<td>ERBB2</td>
<td>51.4%</td>
</tr>
<tr>
<td>PIK3R1</td>
<td>33.3%</td>
</tr>
<tr>
<td>PIK3CA</td>
<td>21.8%</td>
</tr>
</tbody>
</table>

- Stage 1 = 200 Discovery set
  - 20 whole genomes + 180 whole exome
- Stage 2 = 300 Extension validation set
  - targeted sequencing of ~3000-6000 most significant genes

>80% power to detect 3% frequency event
• Reference = Complete + Quality
  • Quality: samples → biomolecules → data → analyses
  • Complete: Multi-dimensionality; global assays
  • Complete: sufficiently powered sample size

→ Transformative Technology Revolution
Massively Parallel Sequencing
Massively Parallel Sequencing

- Point mutation
- Indel
- Homozygous deletion
- Hemizygous deletion
- Gain
- Translocation
- Breakpoint
- Pathogen
- Copy-number alterations
Example of a cancer genome

GLIOBLASTOMA

Coverage(T/N) | Callable | Purity | Ploidy
--- | --- | --- | ---
30x / 30x | 85% | 65% | 5.5

Point Mutations

Rate/ Mb Total Coding
--- --- ---
1.21 3164 27

<table>
<thead>
<tr>
<th>Gene</th>
<th>Type</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>TP53</td>
<td>DNP Splice_site</td>
<td>Tumor suppressor</td>
</tr>
<tr>
<td>PTPRB</td>
<td>Missense</td>
<td>Tumor suppressor family member</td>
</tr>
<tr>
<td>PTEN</td>
<td>Indel</td>
<td>Tumor suppressor</td>
</tr>
<tr>
<td>TNC</td>
<td>Missense</td>
<td>Glioma associated extracellular matrix antigen. Involved in migration of neurons and axons during development</td>
</tr>
</tbody>
</table>

Chr. Aberrations

CNA Breaks TX-Inter TX-Intra
--- 6 84

HIGHLIGHTS

Major rearrangements in chr1 including CDKN2C and FAF1
Scale of Growth is unprecedented

Examples of technical challenges:

- IT infrastructure
- Optimization of library generation
- Input requirement
- Alignment to genome
- Variance calling algorithms
Validation Challenges

- Currently every variant must be ‘validated’
  - For a whole genome, this is thousands of variants and the cost can dwarf discovery cost,
  - Focus on coding regions – still hundreds per tumor type
  - Need to improve error models and practicality of mass-validation

- OVCa MS: all 20,398 somatic variants are being (already) validated in a 2\textsuperscript{nd} assay (by genotyping or repeat sequencing) in all samples
TCGA IS GENERATING NEW KNOWLEDGE

In the midst of a technology revolution
**Significantly mutated genes in serous ovarian cancer (n=316)**

<table>
<thead>
<tr>
<th>Gene</th>
<th># of Mutations</th>
</tr>
</thead>
<tbody>
<tr>
<td>TP53</td>
<td>277</td>
</tr>
<tr>
<td>FAT3</td>
<td>19</td>
</tr>
<tr>
<td>CSMD3</td>
<td>18</td>
</tr>
<tr>
<td>NF1</td>
<td>14</td>
</tr>
<tr>
<td>BRCA1</td>
<td>10</td>
</tr>
<tr>
<td>RB1</td>
<td>9</td>
</tr>
<tr>
<td>CDK12</td>
<td>9</td>
</tr>
<tr>
<td>BRCA2</td>
<td>9</td>
</tr>
<tr>
<td>RB1CC1</td>
<td>7</td>
</tr>
<tr>
<td>GABRA6</td>
<td>6</td>
</tr>
<tr>
<td>TACC3</td>
<td>5</td>
</tr>
</tbody>
</table>

- **TP53** was mutated in 96.5%
- **BRCA1/2** were mutated in 21% of tumors due to germline (9%/6%) or somatic (3%) mutations.
- Other significantly mutated genes in serous OvCa were present in only 1-6% of tumors.

➤ **OVCa** is a disease of genomic instability driven by p53 mutation and defects in HR.
Patterns of somatic genomic alterations

- 68 amplified putative oncogenes in OVCa that are targets or putative targets of drugs or inhibitors in development
Mutated BRCA1/2 have defective HR and are sensitive to PARP inhibitors

HR defects occur in approximately half of serous OvCa
- Core HR genes that are genomically altered
- Mutation vs genomic amplification/deletion vs methylation
Fusion transcripts by RNA-seq in AML

- Identify by assembly and read pairs
  - AML1-ETO 5% of samples
  - PML-RARα 9% of samples
  - BCR-ABL 2% of samples
  - CBFB-MYH11 7% of samples
  - MLL fusions 5 to date
  - Other known 2 to date (CALM/AF10, MYST3/CREBBP)
  - Novel fusions 2 to date

Read assembly evidence for MYST3/CREBBP fusion
Translocation in CRC by sequencing

Process:

~50bp Paired Ends Reads, Illumina HiSeq or GAII (4X Seq Coverage) → BWA alignment (.bam file) → BreakDancer (structural variants)

Results:

Sequenced 10 pairs of Colorectal Cancer Pairs. We have analyzed 8 pairs so far. In red, these translocations are observed in multiple samples.

<table>
<thead>
<tr>
<th>Translocations Detected by BreakDancer</th>
<th>Gene Name(s)</th>
</tr>
</thead>
<tbody>
<tr>
<td>HFM1-SLC05A1</td>
<td>DNA helicase-SLC05A1</td>
</tr>
<tr>
<td>HFM1-PPEF2</td>
<td>DNA helicase-Protein Phosphatase</td>
</tr>
<tr>
<td>HFM1-SEC14L1</td>
<td>DNA helicase-Sec 14 like</td>
</tr>
<tr>
<td>THEM2-SYN3</td>
<td>Thioesterase-Synapsin</td>
</tr>
<tr>
<td>STMN3-KIAA1667</td>
<td>Stathmin like-Herman Pudalski gene</td>
</tr>
<tr>
<td>MTMR2-LRCH1</td>
<td>Myotubularin-Larch</td>
</tr>
<tr>
<td>PRDM9-PRDM7</td>
<td>histone methyltransferase-histone methyltransferase</td>
</tr>
<tr>
<td>THSD7B-C12orf32</td>
<td>Thrombospordin-Orf</td>
</tr>
<tr>
<td>NR5A2-KLHL29</td>
<td>Nuclear Receptor-Kelch</td>
</tr>
<tr>
<td>WDR70-NXPH1</td>
<td>WD Repeat-Neuroexophilin</td>
</tr>
<tr>
<td>C8orf37-CRTC1</td>
<td>Orf-CREB related trx factor</td>
</tr>
<tr>
<td>SEMA5B-SPATS2</td>
<td>Semaphorin-Spermatogenesis associated</td>
</tr>
</tbody>
</table>
Challenges ahead

- **Reference = Complete + Quality**
  - Quality: samples → biomolecules → data → analyses
  - Complete: Multi-dimensionality; global assays
  - Complete: sufficiently powered sample size

- **Analysis and Enablement**
  - Rapid data release
  - Analyses and Publication
  - Knowledge dissemination (Results, Tools)
Data analysis and dissemination

- **Technical challenge:**
  - Cancer genomic data are noisy and complex, particularly challenging amidst rapid evolution in technological platforms
  - Better computational tools to make sense of the data
- **Biological challenge:**
  - Cancer is biologically complex
  - Cancer gene functions are context specific

---

**Genome Data Analysis Centers**
- Broad Institute, Cambridge, Mass.
- Institute for Systems Biology, Seattle, Wash.
- University of Texas/M.D. Anderson Cancer Center, Houston, Texas
- Lawrence Berkeley National Laboratory, Berkeley, Calif.
- Memorial Sloan-Kettering Cancer Center, New York, N.Y.
- University of California at Santa Cruz, Calif.
- University of North Carolina, Chapel Hill

- Develop new computational tools for integrative cancer genome analyses
- *Generate TCGA data analysis results in an “accessible” format for the cancer biology community*
- Disseminate results rapidly
# Analysis is a bottleneck

<table>
<thead>
<tr>
<th>Tumor Type</th>
<th>GCC assays</th>
<th>Whole Exomes</th>
<th>Whole Genomes</th>
</tr>
</thead>
<tbody>
<tr>
<td>GBM</td>
<td>380</td>
<td>109</td>
<td>8</td>
</tr>
<tr>
<td></td>
<td></td>
<td>76 in progress</td>
<td>12 in progress</td>
</tr>
<tr>
<td>Ovarian</td>
<td>560</td>
<td>434</td>
<td>10</td>
</tr>
<tr>
<td></td>
<td></td>
<td>86 in progress</td>
<td>17 in progress</td>
</tr>
<tr>
<td>AML</td>
<td>162</td>
<td>15</td>
<td>26</td>
</tr>
<tr>
<td></td>
<td>39 in progress</td>
<td>135 in progress</td>
<td>29 in progress</td>
</tr>
<tr>
<td>Colon</td>
<td>103</td>
<td>52</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>41 in progress</td>
<td>51 in progress</td>
<td></td>
</tr>
<tr>
<td>Rectal</td>
<td>50</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>17 in progress</td>
<td>67 in progress</td>
<td></td>
</tr>
<tr>
<td>Breast ductal</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>233 in progress</td>
<td>186 in progress</td>
<td></td>
</tr>
<tr>
<td>Lung adeno</td>
<td>21</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>74 in progress</td>
<td>95 in progress</td>
<td></td>
</tr>
<tr>
<td>Lung scc</td>
<td>69</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>45 in progress</td>
<td>114 in progress</td>
<td></td>
</tr>
<tr>
<td>Endometrial</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>70 in progress</td>
<td>70 in progress</td>
<td></td>
</tr>
<tr>
<td>Renal</td>
<td>32</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>32 in progress</td>
<td>32 in progress</td>
<td></td>
</tr>
<tr>
<td>Gastric</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>82 in progress</td>
<td>82 in progress</td>
<td></td>
</tr>
</tbody>
</table>
Example workflow of an analysis run

Input

Automated Pre-defined Integrative Analyses

Output

Mutation, copy number analysis; subclassification; pathway...
Streamlined Tumor Project Model

- SC Chair
- Co-Chair
- DWG
- Disease Experts

Pre-defined Analysis

Customized analysis & MS writing

Timeline

- Project Starts with Tissue Accrual
- 100 Cases Shipped from BCR
- 200 Cases Data to DCC
- Discovery Phase Ends with Publication
In the face of the evolving technologies...

• TCGA is generating new knowledge, enabling and impacting diverse research endeavors

• ‘Genome Paradigm’ brought to cancer
  • Completeness
  • Standardization
  • Open data release

• ‘Field Enhancement’ is evident
  • Methods improving
  • Costs driven down
  • Community engagement increasing
  • Log-changes being accepted and expected
Acknowledgement

Genome Characterization Centers
SNP – Broad Institute
Genome Copy Number - Harvard
mRNA - Univ. North Carolina,
miRNA - Univ. British Columbia
Methylation - Univ. Southern Cali.
Adv. Genomics - Harvard, Baylor

Project Team
NCI and NHGRI

Genome Sequencing Centers
Broad Institute
Washington University
Baylor College of Medicine

Data Coordinating Center

Genome Data Analysis Centers
Broad Institute, Institute for Systems Biology, MD
Anderson Cancer Center, Lawrence Berkeley Nat’l Lab., Memorial Sloan Kettering Cancer Center, Univ. California, Santa Cruz, Univ. North Carolina

Public Data Portal
http://cancergenome.nih.gov/dataportal/