National Cancer Institute

National Human Genome Research Institute

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

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

September 15, 2009

Today's Presentation

THE CANCER GENOME ATLAS

A Look Back at The Cancer Genome Atlas (TCGA) Pilot Project

Significant Milestones and Lessons Learned from TCGA Pilot Project

Phase II of TCGA

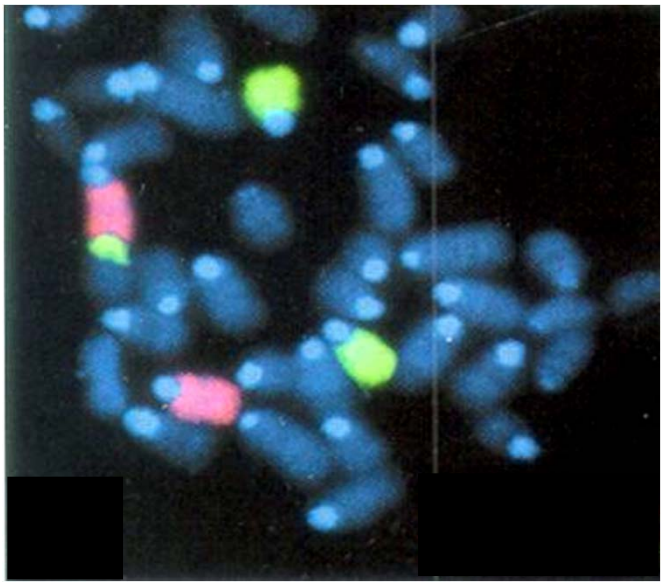
(Joined by Dr. Mark Guyer of NHGRI)

The Significance of TCGA to Cancer and Biomedical Research

TCGA Scientific Rationale

THE CANCER GENOME ATLAS

- 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

THE CANCER GENOME ATLAS

- 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

THE CANCER GENOME ATLAS

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

THE CANCER GENOME ATLAS

- 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

THE CANCER GENOME ATLAS

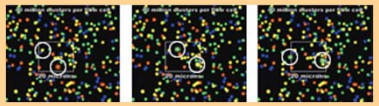
Development of New Analyses

- Tools
- Views

Data Management, Bioinformatics, and Computational Analysis

- Data Coordinating Center, DCC
- Analyses of data

Technology Development



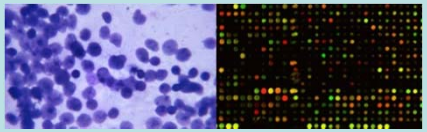
- Increased sensitivity of molecular characterization platforms
- Analysis of biomolecules from 1000 cells or less

Genome Sequencing Centers

High throughput sequencing of genes and genomic regions identified through cancer characterization

Communicate

Cancer Genome Characterization Centers



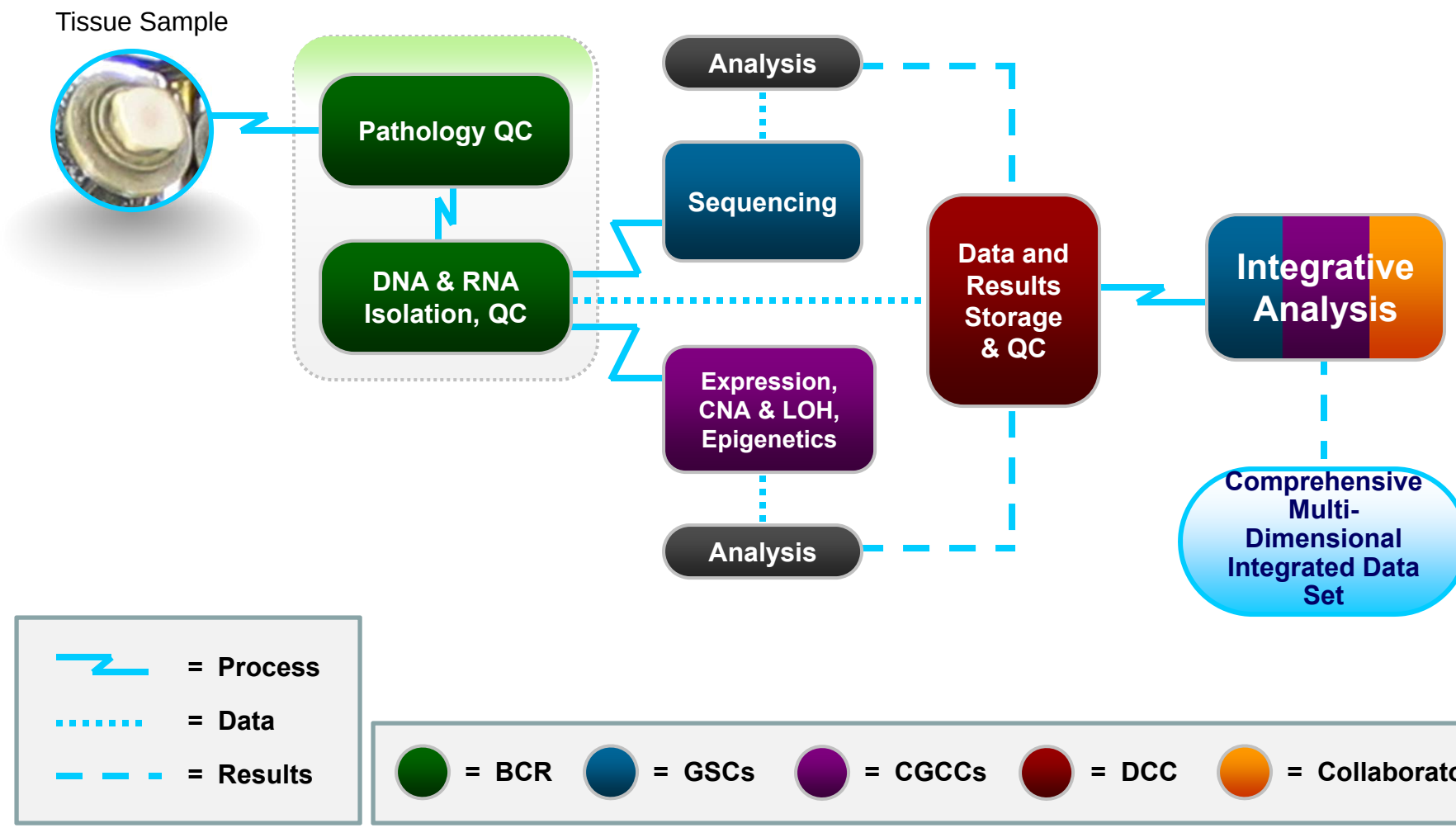
- Identification of expression alternation
- Detection of DNA fragment copy number changes and LOH
- Epigenetics

Human Cancer Biospecimen Core Resource

- Biospecimens-related data storage
- Histopathology confirmation performed
- Biomolecules isolated, QC'ed and distributed

TCGA Pilot Project Pipeline

THE CANCER GENOME ATLAS



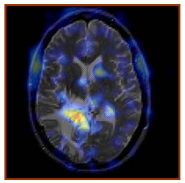
National Cancer Institute

National Human Genome Research Institute

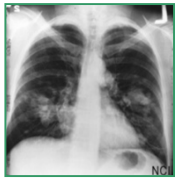
TCGA: Connecting multiple sources, experiments, and data types

Three forms of cancer

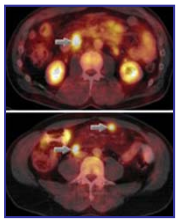
glioblastoma multiforme (brain)



squamous carcinoma (lung)



serous cystadenocarcinoma (ovarian)



Biospecimen Core Resource with more than 13 Tissue Source Sites

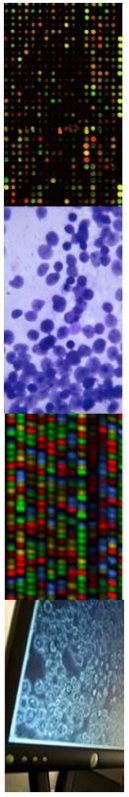
7 Cancer Genomic Characterization Centers

3 Genome Sequencing Centers

Data Coordinating Center

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

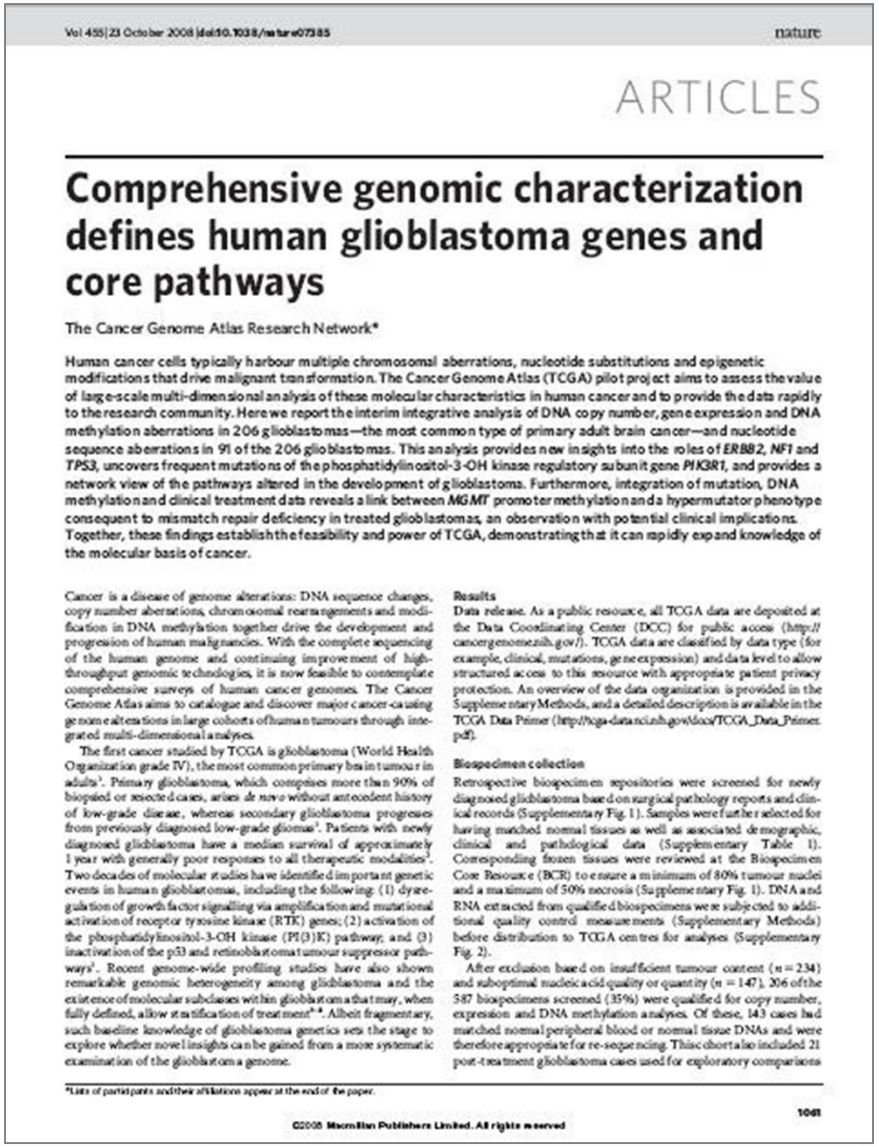


Milestones and Lessons Learned from TCGA Pilot Program

GBM Findings

THE CANCER GENOME ATLAS

- 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



Vol 455|23 October 2008|doi:10.1038/nature07385 nature

ARTICLES

Comprehensive genomic characterization defines human glioblastoma genes and core pathways

The Cancer Genome Atlas Research Network*

Human cancer cells typically harbour multiple chromosomal aberrations, nucleotide substitutions and epigenetic modifications that drive malignant transformation. The Cancer Genome Atlas (TCGA) pilot project aims to assess the value of large-scale multi-dimensional analysis of these molecular characteristics in human cancer and to provide the data rapidly to the research community. Here we report the interim integrative analysis of DNA copy number, gene expression and DNA methylation aberrations in 206 glioblastomas—the most common type of primary adult brain cancer—and nucleotide sequence aberrations in 91 of the 206 glioblastomas. This analysis provides new insights into the roles of *ERBB2*, *NF1* and *TP53*, uncovers frequent mutations of the phosphatidylinositol-3-OH kinase regulatory subunit gene *PIK3R1*, and provides a network view of the pathways altered in the development of glioblastoma. Furthermore, integration of mutation, DNA methylation and clinical treatment data reveals a link between *MGMT* promoter methylation and a hypermutator phenotype consequent to mismatch repair deficiency in treated glioblastomas, an observation with potential clinical implications. Together, these findings establish the feasibility and power of TCGA, demonstrating that it can rapidly expand knowledge of the molecular basis of cancer.

Cancer is a disease of genome alterations: DNA sequence changes, copy number aberrations, chromosomal rearrangements and modification in DNA methylation together drive the development and progression of human malignancies. With the complete sequencing of the human genome and continuing improvement of high-throughput genomic technologies, it is now feasible to contemplate comprehensive surveys of human cancer genomes. The Cancer Genome Atlas aims to catalogue and discover major cancer-causing genome alterations in large cohorts of human tumours through integrated multi-dimensional analyses.

The first cancer studied by TCGA is glioblastoma (World Health Organization grade IV), the most common primary brain tumour in adults[1]. Primary glioblastoma, which comprises more than 90% of biopsied or resected cases, arises *de novo* without antecedent history of low-grade disease, whereas secondary glioblastoma progresses from previously diagnosed low-grade glioma[2]. Patients with newly diagnosed glioblastoma have a median survival of approximately 1 year with generally poor responses to all therapeutic modalities[3]. Two decades of molecular studies have identified important genetic events in human glioblastomas, including the following: (1) dysregulation of growth factor signalling via amplification and mutational activation of receptor tyrosine kinase (RTK) genes; (2) activation of the phosphatidylinositol-3-OH kinase (PI(3)K) pathway; and (3) inactivation of the p53 and retinoblastoma tumour suppressor pathways[3]. Recent genome-wide profiling studies have also shown remarkable genomic heterogeneity among glioblastomas and the existence of molecular subclasses within glioblastoma that may, when fully defined, allow stratification of treatment[4–6]. Albeit fragmentary, such baseline knowledge of glioblastoma genetics sets the stage to explore whether novel insights can be gained from a more systematic examination of the glioblastoma genome.

Results

Data release. As a public resource, all TCGA data are deposited at the Data Coordinating Center (DCC) for public access (http://cancergenome.nih.gov/). TCGA data are classified by data type (for example, clinical, mutations, gene expression) and data level to allow structured access to this resource with appropriate patient privacy protection. An overview of the data organization is provided in the Supplementary Methods, and a detailed description is available in the TCGA Data Primer (http://tcga-data.nci.nih.gov/docs/TCGA_Data_Primer.pdf).

Biospecimen collection

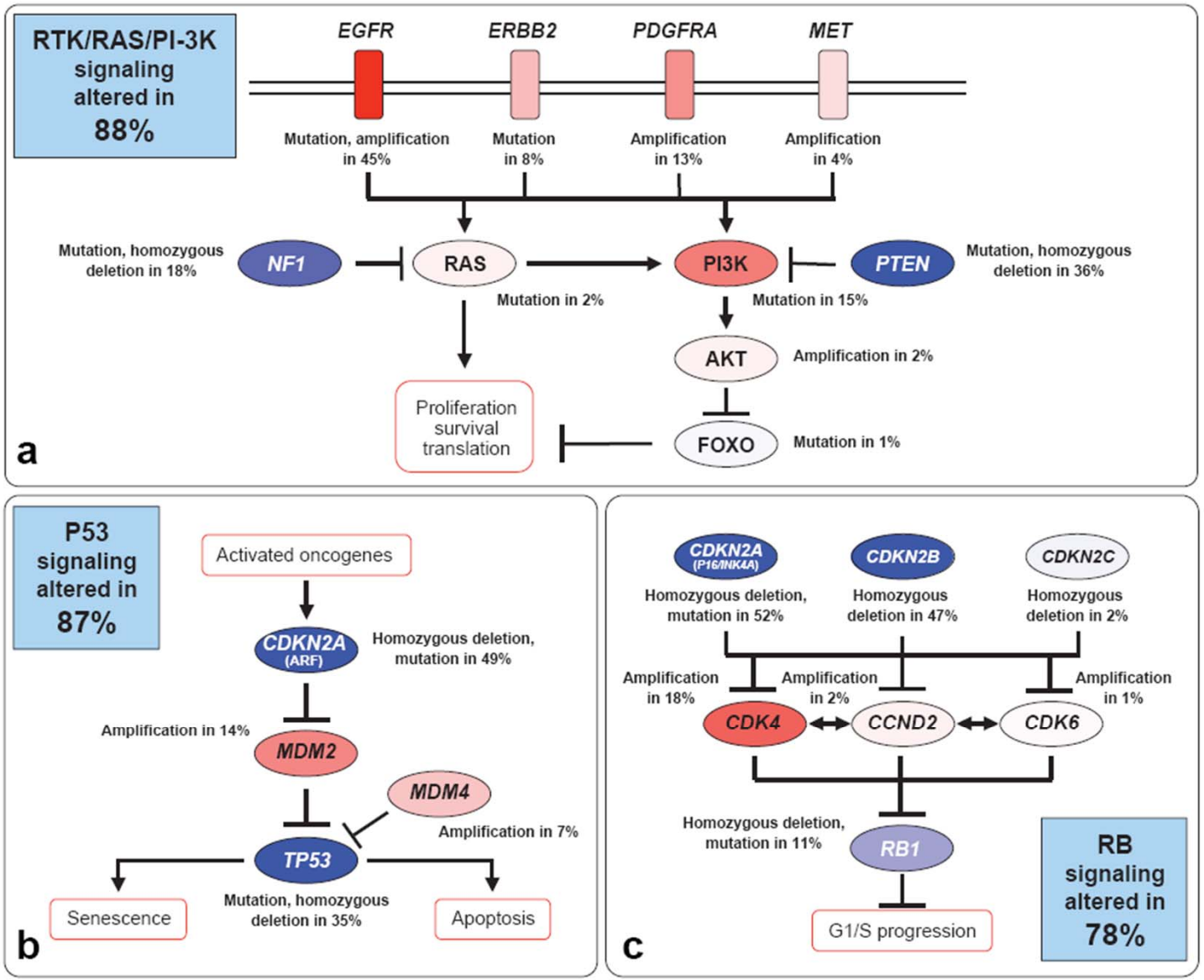
Retrospective biospecimen repositories were screened for newly diagnosed glioblastoma based on surgical pathology reports and clinical records (Supplementary Fig. 1). Samples were further selected for having matched normal tissues as well as associated demographic, clinical and pathological data (Supplementary Table 1). Corresponding frozen tissues were reviewed at the Biospecimen Core Resource (BCR) to ensure a minimum of 80% tumour nuclei and a maximum of 50% necrosis (Supplementary Fig. 1). DNA and RNA extracted from qualified biospecimens were subjected to additional quality control measurements (Supplementary Methods) before distribution to TCGA centres for analyses (Supplementary Fig. 2).

After exclusion based on insufficient tumour content ($n = 234$) and suboptimal nucleic acid quality or quantity ($n = 147$), 206 of the 587 biospecimens screened (35%) were qualified for copy number, expression and DNA methylation analyses. Of these, 143 cases had matched normal peripheral blood or normal tissue DNAs and were therefore appropriate for re-sequencing. This cohort also included 21 post-treatment glioblastoma cases used for exploratory comparisons

*Lists of participants and their affiliations appear at the end of the paper.

GBM Pathways

THE CANCER GENOME ATLAS



TCGA: *Nature* 2008

Potentially Clinically-relevant Discovery in Treated GBMs

THE CANCER GENOME ATLAS

- 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

Ovarian Cancer Status

THE CANCER GENOME ATLAS

- 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

National Cancer Institute | National Human Genome Research Institute

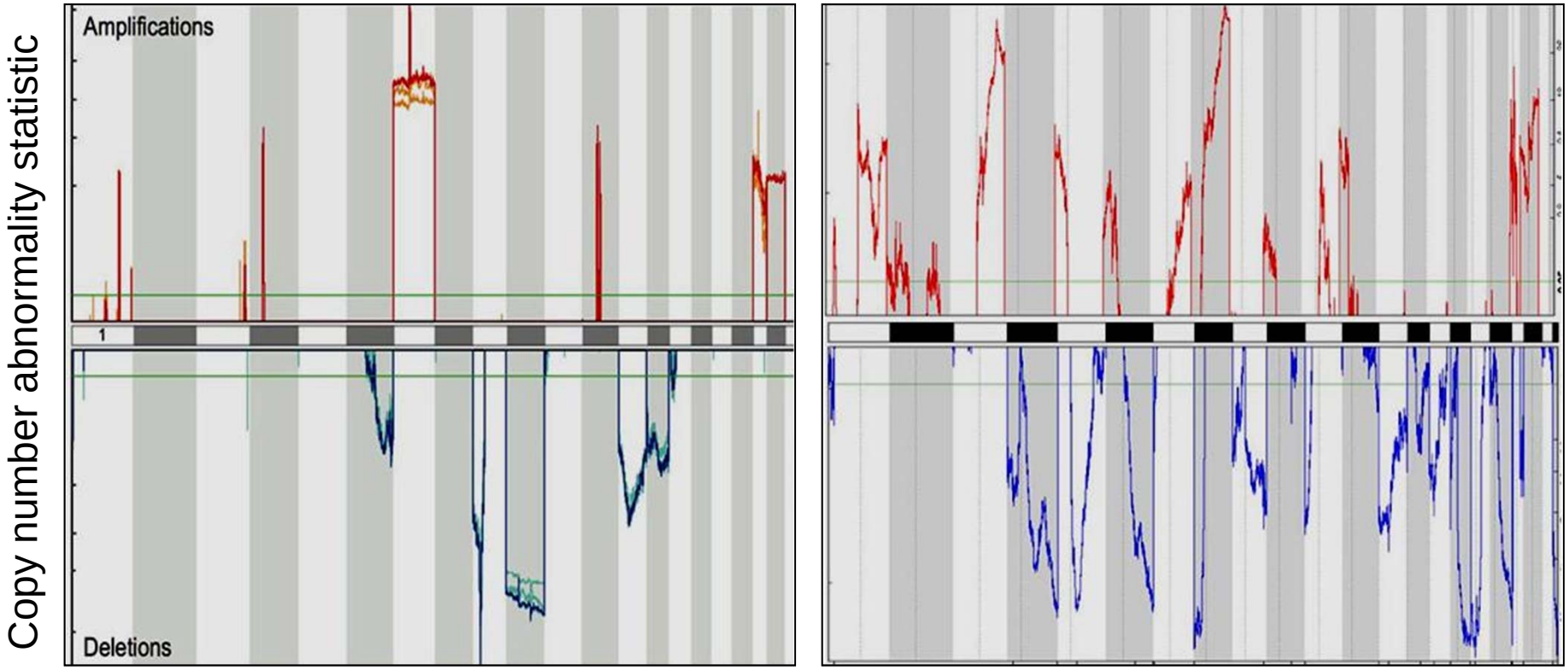
THE CANCER GENOME ATLAS

Search GO

A contrast in copy number complexity

Glioblastoma

Serous ovarian cancer



National Cancer Institute

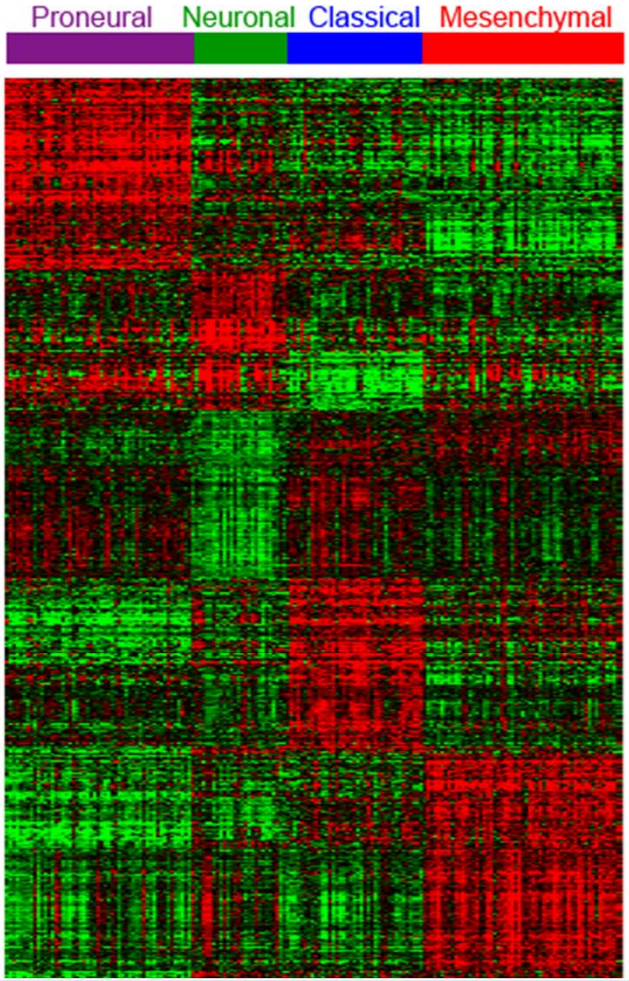
National Human Genome Research Institute

THE CANCER GENOME ATLAS

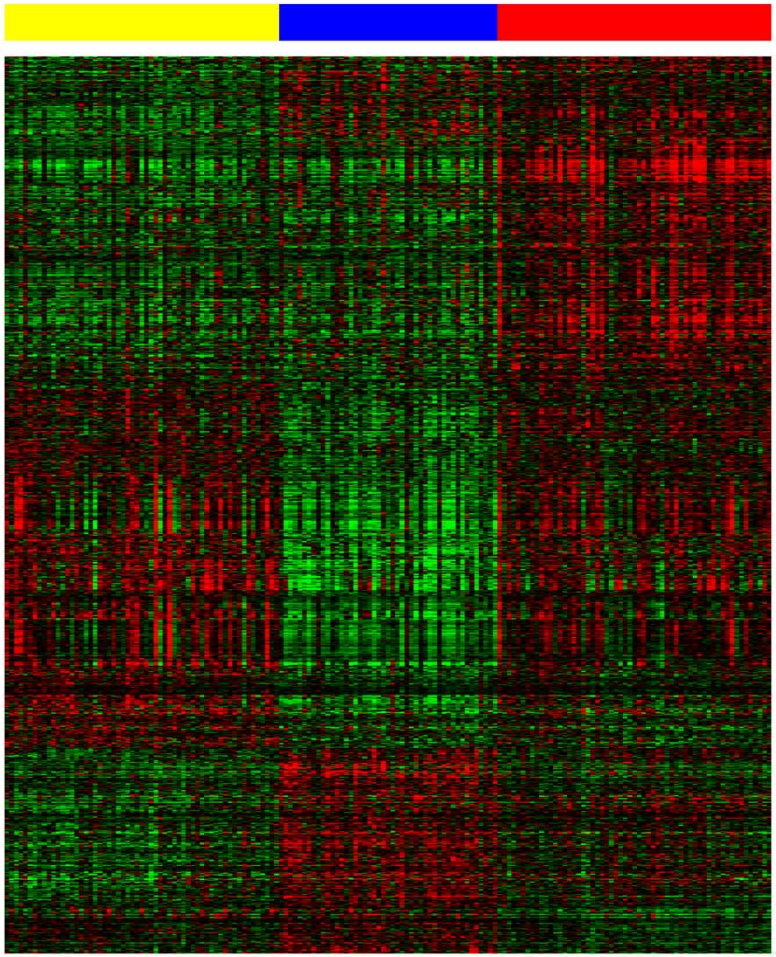
Search GO

Expression subtypes

Glioblastoma



Serous Ovarian

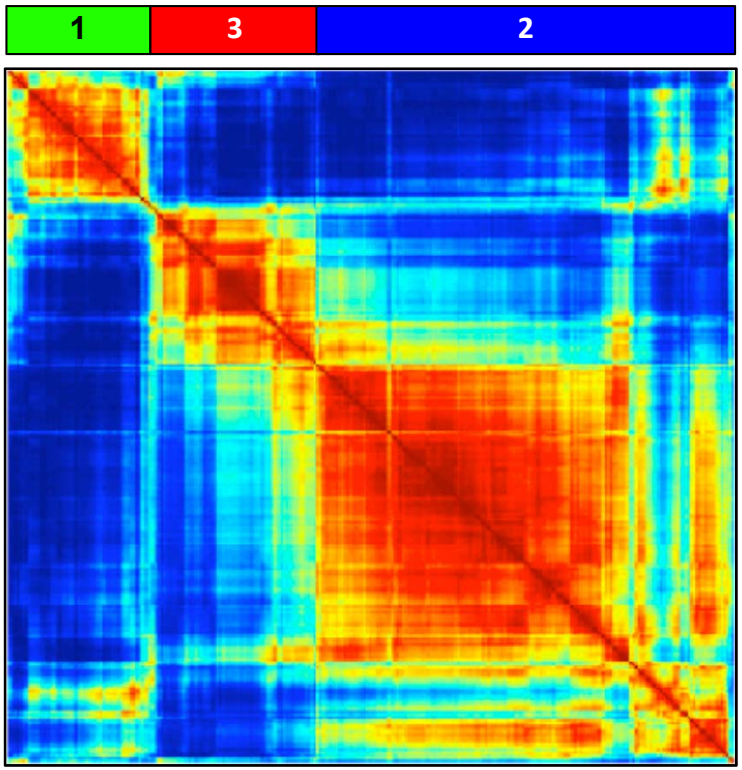


Epigenetic and Expression Profiles: Identify Clusters of High-Grade Serous Ovarian Tumors- With Differences in Five-Year Survival Rates

THE CANCER GENOME ATLAS

Slide courtesy of P. Laird/S. Baylin, Analysis Team

DNA Methylation Data Identifies 3 Clusters of Serous Ovarian Tumors

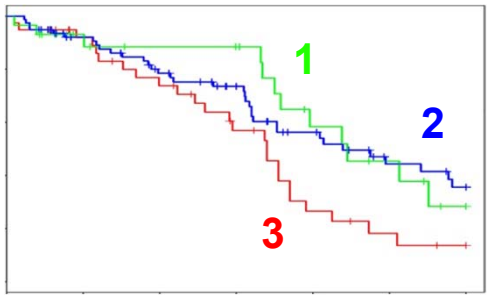


Consensus Clustering of 238 High-Grade Serous Ovarian Tumors with 3,226 Variant Probes

Overlap Between Expression and DNA Methylation Cluster Membership

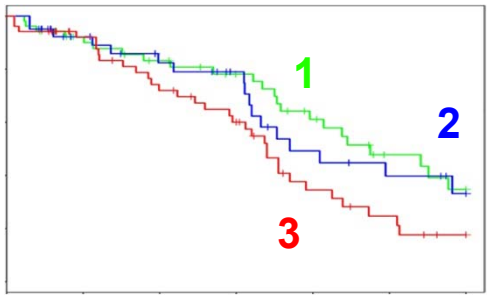
Numbers of Tumors	Methylation Cluster 1	Methylation Cluster 2	Methylation Cluster 3	Total
Expression Cluster 1	21	41	1	63
Expression Cluster 2	5	36	3	44
Expression Cluster 3	6	14	40	60
Total	32	91	44	167

Methylation Clusters



5-Year Overall Survival
n=146 p=0.05

Expression Clusters



5-Year Overall Survival
n=146 p=0.07

Ovarian Cancer: The Analysis Team

THE CANCER GENOME ATLAS

Methylation
Peter Laird

Dan Weisenberger
Mike Lawrence
Dave Larson
XiaoQi Shi
Houtan Noushmehr
Pierre Neuvial

Pathways
Chris Sander

Niki Schultz
Rachel Karchin
Mike Lawrence
Li Ding
Yonghong Xiao
Ethan Cerami
Larry Donehower
Lincoln Stein
Wendy Winckler
Mike Wendl
Svetlana Tyekucheva
Chad Creighton
David Wheeler
Janet Rader
Barry Taylor

Copy Number
Gaddy Getz

Adam Olshen
Barry Taylor
Chad Creighton
Devin Absher
Henrik Bengtsson
Jun Li
Nick Gauthier
Peter Park
Ronglai Shen
Scott Morris
Xiaoqi Shi
Carolyn Compton
David Wheeler
Hailei Zhang
John Zhang
Ken Chen
Nick Socci
Qunyan Zhang
Scott Carter
Wendy Winckler

Coordination
Paul Spellman

Julia Zhang/NCI Staff

Expression
Roel Verhaak

Katie Hoadley
Dan Weisenberger
Nick Socci
Hailei Zhang
Chad Creighton
Ronglai Shen
Elizabeth Purdom
Neil Hayes
Nick Gauthier
Xiaoqi Shi
Pierre Neuvial
Qunyan Zhang

Mutation Detection and Significance
Li Ding

Gaddy Getz
Kristian Cibuluskis
Larry Donehower
Rachel Karchin
Gavin Sherlock
Jinghui Zhang
Dave Larson
Carrie Sougnez
David Wheeler
Mike Wendl
Hannah Carter
Boris Reva
Anil Sood
Dan Koboldt

Whole Genome Analysis
Elaine Mardis

Jinghui Zhang
Barry Taylor
Kristian Cibuluskis
Carrie Sougnez
Li Ding
Sachet Shukla
Ben Raphael
Gaddy Getz
David Wheeler
Houton Noushmehr

miRNAs
Neil Hayes

Dave Wheeler
Todd Wylie
Robert Sheridan
Doug Levine
Preethi Gunaratnee
Laura Heiser
Shaowu Ming
Anil Sood
Dan Koboldt

TCGA Pilot Program: Overall Summary

THE CANCER GENOME ATLAS

- **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

THE CANCER GENOME ATLAS

- 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

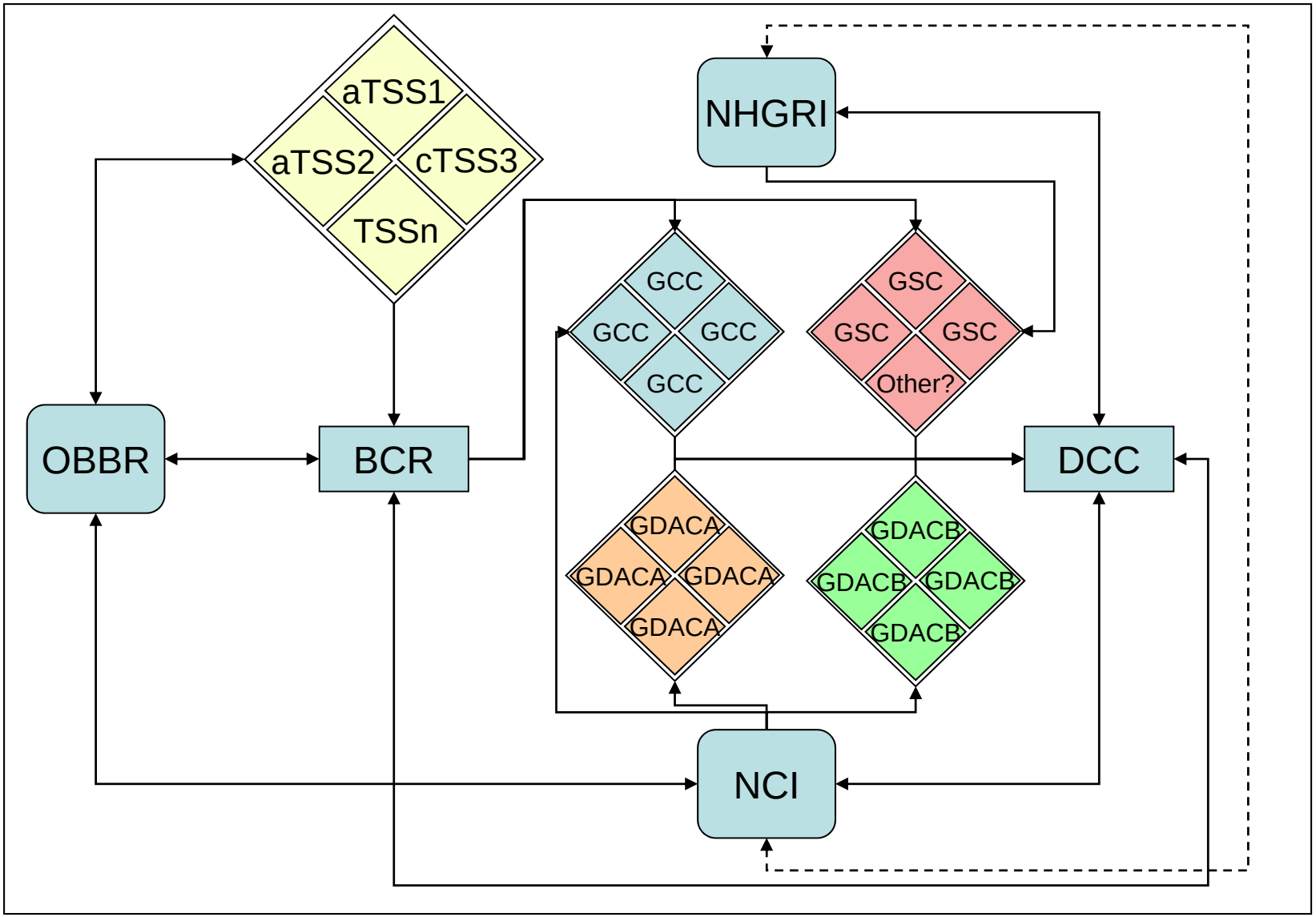
TCGA Phase II: Goals

THE CANCER GENOME ATLAS

- 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

TCGA Phase II: Approach

THE CANCER GENOME ATLAS



TCGA Phase II: Tissue Accrual Plan

THE CANCER GENOME ATLAS

- 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

THE CANCER GENOME ATLAS

- 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

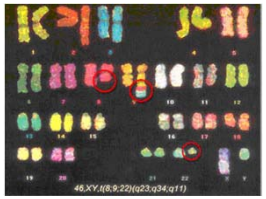
THE CANCER GENOME ATLAS

- 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

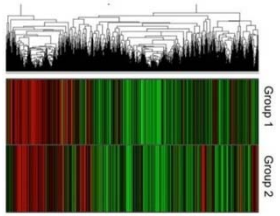
THE CANCER GENOME ATLAS

X



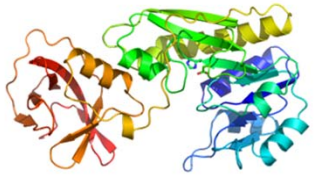
Genetics/Genomics

X



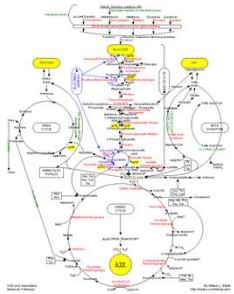
Gene Expression

x

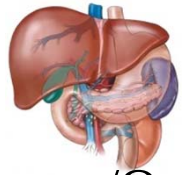


Proteomics

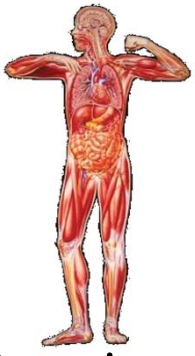
Cancer Biology



Metabolomics



Tissues/Organs



Organisms

Environment

Some Questions for Discussion

THE CANCER GENOME ATLAS

- 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?