A description of The Cancer Genome Atlas (TCGA) selection process for the glioblastoma multiforme target gene list

An interim target selection committee was established in the late summer of 2006 in order to produce an initial sequencing target list for the glioblastoma multiforme (GBM), the first tumor of TCGA pilot project. The committee included cancer and genome experts (see membership below) who were asked to propose a prioritized list of targets (genes and chromosome regions) that were:

- GBM-specific
- cancer-specific
- genes that are members of the same gene families as above
- genes that are members of the same pathways as above

It was envisioned that up to 1000 targets would be selected for this initial list. Another list of targets, up to 2000 will be selected on the basis of research community results, i.e. data from the Cancer Genome Characterization Centers. In the future, sequencing costs may decrease so that it will be economically and technically feasible, as well as practical, to sequence the entire genome and thus a partial gene list to guide sequencing efforts will not be necessary.

The initial list presented by the Selection Committee was a compilation of several curated gene lists, as well as genes added from several sources based upon published and unpublished data. The following provides details about the curated gene lists included by the Target Selection Committee.

1. **NCI Gene List:** The Cancer Gene Data Curation Pilot at the NCICB (http://ncicb.nci.nih.gov/NCICB/projects/cgdcp) is an attempt to create a database of associations between genes and diseases and genes and drug compounds derived from the biomedical literature. The project involves a mixture of automatic text mining, semi-automatic verification, and manual validation/scoring of results. Phase I curation led to the identification of 1000 genes by the NCICB in March 2005. Phase II began September 2005 and resulted in the curation of 1500 genes. Genes from both Phase I and II were considered in draft TCGA gene lists.

In Phase I, 8.8 million MEDLINE abstracts containing ~58 million sentences, from 1975 to June 2004, were screened for occurrence of gene-cancer and gene-compound relations. In total about 17,000 gene names co-occurred with at least one cancer term, which was reduced to approximately 8,000 individual genes after accounting for aliases. To identify genes that are truly associated with cancer, MEDLINE sentences for each gene were manually screened. In the first pass, a gene was scored as a true cancer-related gene when at least one sentence contained a true association, i.e. it mentioned the gene together with a well validated cancer term. Of the initial 7,867 genes identified by the automated process, 4,685 were scored as "true" cancer associated genes, 1,083 as "suspect" and 2,099 as "false". One thousand of the genes scored as being true cancer associated genes were selected randomly for a detailed manual annotation in Phase I.

The 1000 selected genes underwent a careful and thorough manual annotation process. For each of these genes, annotators read all the cancer-term and compound-term associated statements (sentences and/or abstracts) and added evidence and roles to each relevant statement. Evidence codes were assigned to qualify the assertion made in the statement in respect to the association of the cancer or compound term to gene name.

The Phase II list consists of 1500 genes from the initial set of 4,685 that have been fully annotated. The selection criteria for this group emphasized recently discovered gene to disease associations, with the remainder of the selected targets having a moderate number of citations (between 10 and 100). Overall, this set of genes had a mean number of sentences with significant co-occurrence of gene and disease or cancer terms of 200. The cut-off date for MEDLINE abstracts included in Phase II processing was October 20, 2005.

2500 manually-curated cancer genes from Phases I and II of the NCI Cancer Gene data curation project were further processed. Any gene that was annotated 'false positive', had a sentence status flag of 'no_fact', 'unclear' or that had a negative indicator of 'yes' were removed. The final number is **2313** genes which were labeled as the NCI list.

2. Sanger Institute's Gene Census: The Cancer Gene Census is an ongoing effort at the Sanger Institute as part of their Cancer Genome Project to catalogue, by manual curation of the literature, those genes for which mutations have been causally implicated in cancer. The list includes mostly genes that are mutated in cancer and does not necessarily include genes that have been reported to have altered expression or methylation patterns in cancer tissue. The original census and analysis was published in Nature Reviews (*Nature Reviews Cancer* 4, 177-183 (2004)). The complete gene list (http://www.sanger.ac.uk/genetics/CGP/Census/Table_1_full_2006-11-02.xls), containing **355** unique gene symbols, was downloaded from the Sanger Institute web site and submitted for inclusion in TCGA target selection process.

3 & 4. Cancer Sequencing Project (CSP) gene lists: NHGRI has funded the Tumor Sequencing Project (TSP) consortium and the J. Craig Venter Institute (JCVI) (http://www.genome.gov/cancersequencing/#1) as demonstration projects to pilot approaches for detecting large scale genomic changes in tumors using high-throughput DNA sequencing. The TSP consortium chose **949** genes for sequencing in lung adenocarcinoma samples. The JCVI project selected **37** receptor kinase genes to be sequenced in glioblastoma samples. The TSP and JCVI gene lists were also included for TCGA target selection process.

In addition to these lists, other target lists were submitted for consideration from investigators on the Target Selection Committee, as well as within TCGA Research Network. These lists were not unique, but had overlapping genes. Those lists include:

1. 62 targets submitted by Howard Fine based upon GBM single nucleotide polymorphism (SNP) and expression data,
2. 46 targets submitted by Ken Aldape based upon expression and clinical data correlation,
3. 2 targets submitted by Lynda Chin and Ron DePinho based upon functional and clinical GBM data,
4. 51 targets submitted by Lynda Chin and Cameron Brennan based upon GBM-specific copy number and expression data,
5. 189 genes included from Victor Vesculescu's published list of CANgenes which contain somatic mutations in breast and colon cancers,
6. 66 genes included from the Sanger Institute's cancer kinome sequencing project,
7. 58 genes included from Ron DePinho and Lynda Chin based upon GBM-specific comparative genomic hybridization data,
8. 20 genes from Ron DePinho's laboratory based upon receptor tyrosine kinases that are overexpressed in GBM based upon focal minimal common regions (MCR) data in GBM,
9. 345 genes submitted by Lynda Chin and Cameron Brennan based on mapping to focal amplification/deletion in GBM;
10. 202 genes included from Matthew Meyerson's laboratory based upon GBM SNP data,
11. 62 genes included from National Institute of Neurological Disorders and Stroke's PubMed text mining,
12. 269 genes included from Bill Hahn's laboratory based upon GBM-cell line RNA Interference screen data,
13. 284 genes included from Ken Aldape's laboratory based upon expression data,
14. 279 genes proposed by Charles Sawyers at Memorial Sloan Kettering Cancer Center based upon cancer pathways,
15. 46 genes proposed by John Minna based upon cancer-associated pathways,
16. 324 genes submitted by David Housman at the Massachusetts Institute of Technology,
17. 223 genes submitted by Michael Hemann and
18. 200 genes submitted by Stan Gertler.

Prioritization and Selection Process

In total, 2,124 unique target candidates from the above gene nomination lists were collated for assessment. Prioritization was facilitated primarily by Dr. Lynda Chin and her bioinformatics team (Drs. Yonghong Xiao, Bin Feng and Vladimir Morozov). The goal was to prioritize these candidates into three categories:

A-List = candidates already with sufficient evidence to warrant re-sequencing immediately;

B-List = candidates with strong enough evidence to justify re-sequencing without additional data;

C-List = candidates that will require additional supporting data prior to re-sequencing.

Before prioritization, each of the target candidate lists was assigned a "strong", "good" or "okay" score based upon the strength of data used to generate the list. The criteria used for were the following: (1) "strong" score required TWO lines of evidences, one of which was genetic; for example, genomic alteration or mutation data plus expression data or clinical outcome correlation data; (2) "good" score required genetic evidence; for example, mapping to regions of amplification or deletion in GBM, or known somatic mutations in other tumor types; and (3) "okay" score included candidates based predominantly on correlative data, such as expression profiling gene lists, PubMed text mining and/or pathway components. Each of the 2,124 unique gene candidates was assigned "strong", "good" or "okay" score based on its source (list). Genes nominated by multiple lists were associated with multiple "strong", "good" or "okay" scores.

Next, these 2,124 candidates were separated into A, B or C Lists in the following manner.

The "**A-List**" included all of the genes that had: (1) at least one "strong" score (n=161); (2) two or more "good" scores (n=84); (3) one "good" score AND at least two "okay" scores (n=39); and (4) B-List genes promoted to A-List based on association with top gene ontology (GO) terms enriched in the A-List (n=18). In total, the A-List contains **302** genes.

The "**B-List**" included genes that had: (1) at least one "good" plus one or more "okay" score (n=91); (2) at least three "okay" scores (n=151); and (3) C-List genes promoted to B-List based on association with top GO terms enriched in the B-List (n=75). In total, the B-List contains **299** genes (accounting for the 18 genes promoted to A-List).

The remaining genes were considered C-List.

The prioritized A-, B- and C-lists were sent to TCGA Steering Committee for input on January 17, 2007. On Thursday, January 18, 2007, the Steering Committee discussed the generation of the Gene Target List and the Selection Process. It was agreed upon that the initial sequence targets for TCGA would consist of genes on the combined "A" and "B" lists, a sum of **601** genes. In addition, it was recommended that all known microRNAs (approximating about 40-50 genes) be included for re-sequencing.

TCGA Target Selection Committee

Co-Chairs:

Thea Tlsty, Ph.D.
Professor, Department of Pathology
Comprehensive Cancer Center
University of California-San Francisco

Richard K. Wilson, Ph.D.
Director, Genome Sequencing Center
Professor, Genetics and Microbiology
Washington University School of Medicine

Howard A. Fine, M.D.
Branch Chief, Neuro Oncology Branch
Center for Cancer Research
National Cancer Institute
National Institute of Neurological Disorders and Stroke

Todd R. Golub, M.D.
Associate Professor of Pediatrics
Department of Pediatric Oncology
Harvard Medical School
Dana-Farber Cancer Institute
Cancer Program Director, Broad Institute

Mary C. Hendrix, Ph.D.
President and Scientific Director
Medical Research Institute Council Professor
Children's Memorial Research Center
Northwestern University Feinberg School of Medicine

David Housman, Ph.D.
Ludwig Professor of Biology
Center for Cancer Research
Massachusetts Institute of Technology

Gordon Mills, M.D., Ph.D.
Chair, Department of Molecular Therapeutics
Professor of Medicine
MD Anderson Cancer Center
University of Texas

John Minna, M.D.
Sarah M. And Charles E. Seay Distinguished Chair in Cancer Research
Max L. Thomas Distinguished Chair in Molecular Pulmonary Oncology
Professor & Director, Hamon Center for Therapeutic Oncology Research
The Harold C. Simmons Comprehensive Cancer Center
University of Texas Southwestern Medical Center

Charles Sawyers, M.D.
Howard Hughes Medical Institute Investigator
Marie Josée and Henry R. Kravis Chair, Human Oncology Pathogenesis Program
Memorial Sloan-Kettering Cancer Center

Victor Velculescu M.D., Ph.D.
Associate Professor, Oncology
Sidney Kimmel Comprehensive Cancer Center
Johns Hopkins University