The Cancer Genome Atlas (TCGA)
Faculty Presenter
William Kim, MD, University of North Carolina, Chapel Hill, NC, USA

Scholars’ Summaries

Authored by Valerie Grignol, MD, The Ohio State University, Columbus, OH, USA

The Cancer Genome Atlas (TCGA) is an NCI sponsored collaborative effort to study the molecular and genomic bases of more than 20 types of cancers. Tissues is collected for over 150 sites making a great dataset for cataloging genomic alterations but less so for clinical associations. Samples included in the database are primarily untreated patients with limited clinical data associated and short follow-up. Samples are chosen for high tumor content and purity. Tumors samples as well as normal tissue from participants are processed centrally to ensure quality. Processing includes DNA/RNA isolation, sequencing/genomic characterization, and analysis. This information is then stored in publicly available databases for that allow researchers to download and analyze the genetic profiles of specific cancer types. There are 3 levels of data, levels 1 and 2 are controlled access and level 3 is open. The level 1 allows access to FASTQ files and level 2 to BAM files. FASTQ is the raw data and BAM is the binary from for storing sequence data. BAM requires that it is opened in IGV to be usable. Level 3 are the separate data files for each sample. Using this data scientists have developed comprehensive molecular portraits of several types of cancer including breast, bladder, endometrial, colon and ovarian. They have been able to group cancers in to high and low mutational burdens and reclassify tissues based on genomic similarities as opposed to tissue of origin. In some cases this has actually split cancers from the same tissue of origin into completely different subtypes. For instance in high grade bladder cancer, similar to breast, luminal and basal subtypes have been identified. Basal bladder cancer has a poor prognosis. Examining the relationship between basal bladder and breast cancer found several genomic similarities which have treatment implications. The next step for this type of research is the development of molecular tumor boards. Patient’s tumors would be profiled and recommendations for therapy would be made based on the “Omic” profile including standard of care, novel agent off-label use and clinical trial.

Authored by Vincent Lam, MD, Greenebaum Cancer Center, University of Maryland School of Medicine, Baltimore, MD, USA

The Cancer Genome Atlas (TCGA) is an NCI-sponsored multi-institution initiative to comprehensively study the molecular and genomic basis of 20+ types of cancer. Dr. William Kim briefly explained the pipeline by which specimens are obtained, sequenced, and analyzed. In addition to exomes, other analyses available include CNA, LOH, and epigenetics. Primary data files (FASTQ and BAM) require controlled (Level 1 and 2) access. Some key characteristics to keep in mind regarding TCGA data are: it consists of primarily untreated samples with short-term follow-up, tumor purity is greater than 70% (so not optimal for studying things like tumor microenvironment), and the clinical data is often limited.

Dr. Kim mentioned that one of the most valuable applications of these data is to compare genomic characteristics across multiple cancer types. Hoadley et al (Cell 2014) is one example: integration of exome, RNA, SNP array, DNA methylation array, and RPPA data for 12 tumor types enabled description of 11 distinct molecular classifications. Bladder cancer was unique in that it had members in multiple classifications (bladder, LUAD, and squamous-like).

Bladder cancer also appears to have intrinsic subtypes similar to the basal/luminal distinction in breast cancer. This appears to have corresponding prognostic value as well (i.e. basal has poorer outcomes). Other intrinsic subtypes include claudin-low (also found in breast cancer), p53-like, and clusters I-IV, as defined in the UNC, MDACC, and TCGA cohorts, respectively. Luminal bladder cancer (thought to originate from the umbrella cells of the bladder) appears to be enriched for FGFR3 and TSC1 mutations and is more prominent in Asians. The basal subtype is associated with the female gender,

©Society for Translational Oncology 2016
higher prevalence of squamous differentiation, and RB pathway alterations. Claudin-low subtypes have increased TIC markers, higher EMT, and poorer prognosis compared to luminal.

*Authored by Tracy L. Rose, MD, MPH, Lineberger Comprehensive Cancer Center, University of North Carolina, Chapel Hill, NC, USA*

The Cancer Genome Atlas (TCGA) is a National Cancer Institute-sponsored initiative to comprehensively study the molecular and genomic basis of over 30 types of cancer collected from over 150 sites around the world. It includes primarily untreated samples that were selected for high tumor content and purity. Tumors underwent quality evaluation and were processed for DNA and RNA isolation. DNA and RNA were sequenced and evaluated for mutations, copy number alterations, mutations, and epigenetic changes. All of this information underwent integrative analysis and all the subsequent data is publicly available in a non-identifiable format. There are 3 levels of data: Levels 1 and 2 have controlled access and level 3 is open access.

There has been extensive analysis of TCGA data within each type of cancer, and data across cancers is also emerging. For example, Hoadley et al (Cell 2014) did an integrative genomic analysis across 12 different tumor types. Using an iterative algorithm that looks for similarities in gene expression between tumors called consensus clustering, the cancers were clustered into subtypes. Some of the subtypes matched their tissue of origin. However, some tumor types clustered together such as lung squamous, head and neck cancers, and a subset of bladder cancers, suggesting similarities in these cancers on the molecular level. We now need to take our sequencing knowledge and clinically integrate the findings, as well as do more sophisticated pathway analysis, multiparametric modeling, and include information from the tissue microenvironment and host immune system.

*Authored by Debasish (Deb) Sundi, MD, The University of Texas MD Anderson Cancer Center, Houston, TX, USA*

Dr. William Kim provided a detailed and insightful summary of the structure, progress, processes, and output of The Cancer Genome Atlas (TCGA). The specific roles of several TCGA components were defined, including biospecimen core resource, cancer genome characterization center, genome sequence center, data coordinating center, and genome data analysis center. STOFF participants were also taught about the different levels of TCGA data and associated data formats (levels 1 and 2 are controlled access, and level 3 is open access). As an example of what can be learned from the TCGA, Dr. Kim, himself a leader in bladder cancer genomics, provided a state-of-the-art update on how multiplex genomics analyses have resulted in the description of several distinct bladder cancer subtypes that are associated with response to systemic therapy and oncologic outcomes.

*Authored by Adrienne (Ada) Waks, MD, Dana-Farber/Partners CancerCare, Boston, MA, USA*

The Cancer Genome Atlas (TCGA) is an NCI-sponsored, large, collaborative initiative whose objective is to comprehensively study the molecular and genomic basis of 30+ types of cancer. Data is sourced from over 150 international sites. The data can be used not just to understand the landscape within a given tumor type, but across tumor types. For example, TCGA data have been used to publish important observations about the variability in mutational load in a spectrum of different cancers.

The process of formulating TCGA data involves a biospecimen core resource (where nucleic acid is isolated, and preliminary data undergo quality control), followed by a genome sequencing center (where sequencing, expression, proteomic, and epigenetic information is generated), followed by data analysis.

The data-level descriptors produced in this process are (1) FASTQ files (the raw data directly from the sequencer, unaligned to the human genome; an application is necessary for access); (2) BAM files (sequencing data aligned to the human genome; an application is necessary for access); (3) analyzed sequencing data (in its final form; publicly available).

There are some important limitations of TCGA data to be aware of. In general, the data are derived from primary, untreated tumor samples. The samples are selected for high tumor content, and therefore are not good for studying the
tumor microenvironment i.e. immunologic characteristics. Lastly, the data are annotated with only very limited clinical data, and only short-term patient follow-up.

An important problem to note is that given the pace at which we generate all of this sequencing data, it is essentially impossible—even with highly expert, multi-member molecular tumor boards—to systematically integrate it all into patient care in a persistent way. Therefore, is there a role here for “cognitive computing”—i.e. computers that mine the molecular literature and help clinicians put their patients’ sequencing data to optimal use? This idea is not deployable at this point, but may be important in the future, as we grow increasingly sophisticated.