Meeting Summary

The Cancer Genome Atlas's (TCGA) 2nd Annual Scientific Symposium: Enabling Cancer Research through TCGA

The Crystal Gateway Marriott
Arlington, VA

November 27-28, 2012

National Cancer Institute
National Human Genome Research Institute
National Institutes of Health
U.S. Department of Health and Human Services

Table of Contents

Section **Page**

Tuesday, November 27

Session I
Chair: Richard Gibbs, Ph.D.

Session II
Chair: Raju Kucherlapati, Ph.D.

Wednesday, November 28

Session III
Chair: John N. Weinstein, M.D., Ph.D.

Session IV
Chair: Charles Perou, Ph.D.

Session V
Chair: Louis M. Staudt, M.D., Ph.D.

Tuesday, November 27

Opening Remarks

Matthew Meyerson, M.D.; Dana-Farber Cancer Institute and Ilya Shmulevich, Ph.D.; Institute for Systems Biology

Dr. Meyerson began by thanking participants and attendees for their efforts to date, noting that the present represents a wonderful time for TCGA. The program is well on its way to accruing 10,000 well-characterized human cancer specimens, more than 7,000 of which have already been distributed to TCGA Genome Characterization Centers. Results from these analyses continue to be published. Dr. Meyerson stated that TCGA's true impact is symbolized by the opportunities that its data sets afford to the scientific community. He noted that this Symposium will present primary data and analysis tools as well as analyses by TCGA investigators and the broader scientific community.

Dr. Shmulevich noted that he was excited to see the TCGA project grow in terms of developing analysis tools and methods that can be used widely and applied to a large, heterogeneous, comprehensive data set. These tools enable computational biologists and cancer biologists to collaborate to gain insight into cancer biology and to inform clinical efforts. Furthermore, TCGA has grown through the number of cancers profiled, and the TCGA Pan-Cancer Working Group (WG) has been convened to develop ways to analyze multiple cancers comprehensively to provide new insights.

TCGA: Status Report and insights from Leadership

Kenna Shaw, Ph.D.; Stephen Chanock, M.D.; Louis M. Staudt, M.D., Ph.D.; National Cancer Institute

Dr. Shaw began by noting that TCGA was launched in 2006 as a pilot program to characterize three cancer types that was subsequently expanded in 2009 to cover more than twenty additional tumor types. TCGA goals include establishing an infrastructure for effective team science, developing a scalable "pipeline" that begins with obtaining the highest-quality samples, determining the feasibility of a large-scale, high-throughput approach to identify the molecular "parts list" of various cancers, evaluating statistically robust sample sets, and making the resultant data broadly available to the cancer community while protecting patient privacy. TCGA has also ventured into Rare Tumor Projects with the accrual of 50 qualified cases of chromophobe kidney cancer. TCGA analyses span a wide variety of platforms and produce diverse sets of data, including (but not limited to) clinical data, histology, gene expression/RNA sequence, chromosomal copy number, and methylation pattern analysis. The Atlas expects to accrue a total of 8,000 samples by the end of 2012, with an additional 5,000 specimens expected by the end of 2013. Samples are provided through a network of more than 150 tissue source sites worldwide and are analyzed through a unified pipeline that includes a central Biospecimen Core Resource (Nationwide Children's Hospital; Columbus, OH), six Cancer Genomic Characterization Centers, three Genome Sequencing Centers, seven Genome Data Analysis Centers, and a Data Coordinating Center (DCC).

Between the pilot and expansion phases of TCGA, each case is analyzed using Affymetrix 6.0 (SNP/copy number variation [CNV] analysis) and Infinium array (methylation analysis) platforms, RNA sequencing, and whole exome sequencing (WES). Ten percent of specimens also receive whole genome sequencing (WGS), and some are subject to proteomic analysis using reverse phase protein arrays (RPPA). At present, TCGA has shipped 7,136 cases that span more than 20 tumor types. 5,865 of these have a minimum clinical data set, and 3,893 have at least one year of follow-up data. 105,000 samples of RNA/DNA/protein have been shipped between 2006 and 2012, with an 87% return rate on data.

Comprehensive tumor characterization requires approximately five years, from initial contact by Dr. Shaw to manuscript publication. TCGA Data Portal uses a DCC coordinated by SRA International, and ongoing pipeline analysis is available through the Firehose program sponsored by the Broad Institute. TCGA sequence data have been relocated to the Cancer Genomics Hub (CGHub) at the University of California, Santa Cruz (UCSC), which currently contains two petabytes of downloadable data. Data for tumor types not represented by TCGA may be found through data repositories sponsored by the International Cancer Genome Consortium (ICGC).

Dr. Chanock noted that the present time represents an important juncture in the lifespan of TCGA. He stated that TCGA ovarian cancer analysis efforts tested germline data to assess the relationship between germline and somatic alterations, generate key preliminary findings, and assess possible clinical implications (e.g., consideration of *BRCA* status in clinical trials). As such, TCGA can bridge discovery, characterization, and clinical application. Analysis of TCGA data (Bolton KL, et.al. *JAMA* 2012;307:382-390) associated *BRCA1* and *BRCA2* mutations with survival, chemotherapy sensitivity, and gene mutator phenotype in patients with ovarian cancer. A large-multi-center follow-up study to determine the relationship between survival and *BRCA* status has shown that *BRCA* carriers are more likely than non-carriers to have a favorable response to platinum-based therapy. In summary, *BRCA* carriers show substantially improved survival compared to non-carriers, and *BRCA2* carriers show a distinct clinical course from *BRCA1* carriers. Preliminary evidence suggests that survival varies by mutation location for *BRCA1*. Dr. Chanock noted that this study can serve as a paradigm for using TCGA data to inform clinical studies using traditional therapies and those targeted to *BRCA* carriers.

He then introduced the NCI Center for Cancer Genomics (CCG), which aims to develop and apply cutting-edge genome science to improve cancer prevention, care, and detection. During the upcoming 3-6 months, the CCG will engage in active dialog about the types of studies that will follow TCGA by leveraging the lessons learned from TCGA, capitalizing on the success of TCGA structure, and continuing the partnership between NCI and NHGRI. NCI’s support of cancer genomics activities beyond 2013 will build upon the strengths of TCGA pipelines, involve hybrid projects, and plan toward clinical transition while retaining an emphasis on discovery. Major areas of interest for NCI-supported cancer genomics efforts include unraveling cancer biology (e.g., drivers versus mutations, clonality and progression, the value of epidemiology/germline data, treatment stratification, risk). The Institute will consider genome-informed trials (e.g., those in which DNA information is obtained during or after a trial closes), genome-driven trials (e.g., those in which sequencing and characterization will guide treatment), and trials that do not include requisite genomic analysis. Current TCGA goals include achieving milestones as determined per cancer site, conducting pan-cancer analyses, carrying out technical pilots for formalin-fixed, paraffin-embedded (FFPE) and other tissue types, and forging new solutions.

Dr. Staudt began by observing that directly incorporating clinical data into integrated computational analysis can inform the development of predictive and prognostic markers, identify new targets, and change clinical care. He noted that TCGA's infrastructure is suited to clinical trials, such as recent efforts to study diffuse large B-cell lymphoma (DLBCL), for which gene expression profiling has identified three clinically distinct subgroups (e.g., activated B cell-like [ABC], germinal center B cell-like [GCB], and primary mediastinal B cell lymphoma). Recent analyses have shown that 10% of ABC DLCBL cases feature chronic active B-cell receptor signaling. The B-cell receptors in ABC DLBCL are clustered and immobile. Constitutive MYD88 signaling is also observed in ABC DLBCL. Such pathway information, gained by assessing the co-occurrence or co-exclusion of mutations, has informed a multicenter clinical trial (n=70) of the *BTK* inhibitor, ibrutinib, across all types of DLBCL. Preliminary results from this trial suggest that ABC tumors respond better to ibrutinib than do GCB DLBCL tumors. Studies are now underway to determine whether analysis of recurrent genetic lesions can identify responders within ABC DLCBL. Results indicate that CD79B-mutant ABC DLBCL predicts a high rate of response to ibrutinib and that ibrutinib response does not require B cell receptor mutation. In addition, MYD88 L265P plus CD79 mutations can improve identification of patients with ABC DLBCL who will respond to ibrutinib. Furthermore, homozygous deletion of *INK4a/ARF* is recurrent in ABC DLBCL and is associated with unfavorable outcome.

Dr. Staudt noted that the heterogeneity of human cancer necessitates analyzing large numbers of biopsies, suggesting that more than 10,000 specimens may be required to identify patterns of co-occurrence and exclusion among genetic lesions. As such, a pathway-centric view of genetic lesions is necessary, and it will be essential to make molecular diagnostic tests widely available and to develop a Cancer Genome Commons database to accelerate precision medicine.

One participant asked how computational scientists are involved in sample preparation for TCGA studies. Dr. Shaw replied that tissue source sites are involved at an early step in TCGA pipeline. Although each Genome Data Analysis Center carries out its own analysis, these Centers all include disease experts. Profiling for cancer subgroups uses estimates based on a signature of 100 genes. Another participant asked Dr. Staudt to elaborate on the landscape necessary to develop genomics-based diagnostics. He replied that the paradigm should include a co-diagnostic that is developed and approved simultaneously with the diagnostic agent. The bar for the quality of the diagnostic is extremely high.

Session I

Chair: Richard Gibbs, Ph.D.; Baylor College of Medicine

TCGA Clear Cell Renal Cell Carcinoma (ccRCC) Project

Chad J. Creighton, Ph.D.; Baylor College of Medicine

Dr. Creighton began by stating that genetic changes underlying ccRCC include alterations in genes controlling cellular oxygen sensing (e.g., *VHL*) and the maintenance of chromatin states (e.g., *PBRM1*). To date, 446 TCGA ccRCC tumors have been profiled to some extent in terms of clinical and pathologic features, genomic alterations, DNA methylation, and RNA and proteomic signatures; 372 of these have been fully profiled in terms of RNA and DNA analyses. Whole exome sequencing (WES) identified 39 significantly mutated genes, including *VHL*, *SETD2*, and *PBRM1*. Somatic mutations were called by three sequencing centers (the Baylor College of

Medicine, the Broad Institute, and UCSC), and efforts were focused on calls made by at least two of these centers. Ion Torrent was used as an orthogonal platform to validate significantly mutated genes. These analyses showed that ccRCC is significantly associated with frequent alterations in *VHL* and several genes involved in chromatin regulation, suggesting a major role for epigenetic reprogramming in these genes. In addition, many mutations in PI3K pathway regulators were identified. Copy number alterations (CNA) included 3p loss, although ccRCC presented fewer focal regions of copy number alteration than observed in other cancers. Several recurrent fusion RNA transcripts were also identified. Mutations in specific genes involved in chromatin regulation were associated with widespread molecular alterations, such as DNA hypomethylation associated with *SETD2* mutation. DNA hypermethylation increases with advancing tumor stage and grade, suggesting that the epigenetic state of a more aggressive cancer differs from that of a less aggressive cancer. Pathway analysis indicated that mutations involving the SWI/SNF chromatin remodeling complex show far-reaching effects on other pathways.

TCGA data have also been used to determine a gene expression signature of survival in high-grade serous ovarian cancer. A similar exercise was carried out for ccRCC that identified protein survival correlates such as AMPK and acetyl CoA carboxylase. A low level of AMPK combined with a high level of acetyl CoA carboxylase leads to a worse outcome. These results suggest that survival correlates underlie a glycolytic shift, with the PI3K pathway being highly targeted at the genetic and genomic levels. Promoter methylation of miR-21 and *GRB10* contributes to PI3K pathway deregulation. As miR21 expression increases, PTEN expression and miR21 promoter methylation decrease. An increase in PI3K activity leads to a decrease in *GRB10* expression and an increase in *GRB10* methylation. In conclusion, Dr. Creighton noted that integrative analyses highlight the importance of both the *VHL/HIF* pathway and chromatin remodeling/histone methylation pathways in ccRCC. Hallmarks of ccRCC include frequent targeting of the PI3K pathway at the genetic, genomic, and epigenetic levels, and a metabolic shift to aerobic glycolysis appears to be involved with more aggressive disease.

One attendee asked if clinical data could be used with multivariate analysis to assess whether observations correlate with prognostic signatures. Dr. Creighton replied that efforts to date have focused on understanding pathways rather than searching for biomarkers. He noted that in a multivariate model, the DNA signature may perhaps be a less significant contributor, although this is difficult to determine in this set of analyses. When promoter signatures factor in tumor grade, state, and other variables, they can provide more information than clinical data alone. Methylation data have yet to be validated for these studies.

Analysis of Somatic Mutations across Many Tumor Types
Petar Stojanov; Dana-Farber Cancer Institute/Broad Institute

Dr. Stojanov began by observing that characterization efforts aim to identify the full set of genomic alterations within the cancer and germline tissue, whereas population-level interpretation seeks to identify those genomic alterations that are statistically significant in the population and the genes and pathways in which they occur. He noted that TCGA, ICGC, the Broad Institute, and others are providing a wealth of such data. The Pan-Cancer 8 (PanCan8) data set contains eight cancer types spanning 2,143 patients and more than 436,000 mutations. To identify recurrent genes within this number of mutations, the MutSig algorithm is used to calculate sample-, gene-, and context-specific background mutation rates and to evaluate base-level evolutionary conservation, positional configuration, and truncating mutations. This approach has been applied to many TCGA tumor types, including cancers of the lung, breast, and endometrium, to identify top-ranking mutations and provide insight into pathways integral to these tumor types. Combining the data from these types enables identification of genes that cannot be detected without a pan-cancer approach. In these analyses, the top-ranking genes were sorted by the maximum percent of patients mutated. Dr. Stojanov noted that combining datasets increases statistical power, and many new genes arise from the PanCan8 analysis that are not significant in any one tumor type alone but are highly significant in the combined analysis. In conclusion, he noted that combining tumor types provides significantly more power to detect putative driver genes that cannot be detected with sufficient power in each separate tumor type. However, combining datasets in a pan-cancer analysis dilutes the power to detect driver genes that are potentially important to their respective tumor types. Future steps include incorporating other information apart from conservation (e.g., polyphen2, mutation assessor, CHASM) to elucidate functional roles of mutations, performing this significance analysis on curated gene sets, extending the analysis to examine correlation and mutual exclusivity within and across tumor types and pathways, and using HotNet to assess altered gene subnetworks.

One participant asked which pathway featured the greatest number of mutated genes across tumor types. Dr. Stojanov replied that TGFβ and Wnt signaling pathways are the most commonly mutated. With regard to examining promoter regions in whole genomes, coverage in flanking regions can be used. An add-on to MutSig can assess clustering and conservation to identify genes that are less recurrent across the data set.

CRAVAT and muPIT: Web Services for High-Throughput Analysis of Sequence Variation in Cancer

Rachel Karchin, Ph.D., M.S.; The Johns Hopkins University

Dr. Karchin began by noting that there is an increasing need for computational tools to analyze large-scale cancer mutation data. The goal of this work is to provide an end-to-end mutational analysis workflow that begins with a web-based application to which millions of mutations can be submitted. These mutations are then mapped to transcripts, providing the user with information about where the mutations have previously been identified, associated types of change, and predicted driver and functional mutations. Mutations can then be visualized onto three-dimensional protein structures and pathways. She noted that the majority of somatic mutations in tumor exomes are missense. The Cancer-Related Analysis of Variants Toolkit (CRAVAT), which prioritizes missense mutations, can be coupled with the Mutation Position Imaging Toolbox (muPIT), an interactive visualization tool for three-dimensional protein structure. Users can input mutations into CRAVAT in genomic or transcript coordinates, and these mutations are then mapped onto the best available transcript for a position. Transcripts are scored by coverage of coding bases and agreement with RefSeq. This tool can identify mutations in known variants, show how many times the mutation has been seen in COSMIC, and identify primary tumor tissue types that contain the mutation. Two tools for analysis, CHASM and VEST, can then be applied. CHASM is a supervised machine-learning method that attempts to discriminate drivers from passenger using a Random Forest method. Decision rules are based on a variety of sources, and 86 features are pre-computed. The training set comes from driver missense mutations from COSMIC and random passenger missense mutations generated *in silico* using a model. Tissue-specific classifiers in CRAVAT can be selected for analysis. A variant effect scoring tool (VEST), which is similar to CHASM, can be used to identify functional mutations. VEST uses a training set comprised of 50,000 functional missense mutations from the Human Gene Mutation Database that have generated clinical phenotypes. She noted that functional mutations are a subset of all somatic mutations and that drivers are a subset of functional mutations. When applied to a set of mutations, VEST and CHASM provide complementary benefits for mutation analysis by distinguishing drivers from passengers and damaging versus benign mutations. Outputs from these tools include gene annotations and gene-level scores, provided as a spreadsheet and a formatted file that can be uploaded into muPIT Interactive. muPIT Interactive users input genomic coordinates from mutations of interest and receive a table of protein structures onto which mutations can be mapped. Users can select the most interesting structures from a set of available annotations. Selection of a structure generates a details page for interactive viewing that displays mutations grouped in three-dimensional space. Future efforts will include integrating CRAVAT and muPIT into a single end-to-end service and integrating these tools with the UCSC Genome Browser and other platforms.

High-Grade Serous Ovarian Adenocarcinoma Transcriptome Sequencing

Andrew J. Mungall, Ph.D.; British Columbia Cancer Agency

Dr. Mungall began by stating that most deaths from ovarian cancer result from advanced-stage, high-grade serous ovarian carcinoma. TCGA researchers have previously published integrated genomic analyses of ovarian carcinoma (TCGA Research Network. *Nature* 2011;474;609-615), noting that the disease is characterized by a simple mutational spectrum with *TP53* present in 96% of tumors and a high frequency of somatic CN aberrations. The current study aims to use transcriptome sequencing to identify subtypes, structural variants, and alternatively-spliced transcripts in high-grade serous ovarian cancer. Data were gathered from 490 tumor samples provided by 15 tissue source sites. Efforts include sequencing of 420 RNA Seq libraries, for which 300 expression datasets passed quality control (QC) and have been submitted to the DCC. Moreover, 485 samples that passed QC have miRNA-Seq. Data for these samples have been submitted to CGHub and the DCC. Analyses included unsupervised non-negative matrix factorization (NMF) consensus clustering, miRNA anti-correlations with mRNA isoform expression, and identifications of gene fusions using Trans-ABySS and the UC-Fusion-Finder. Sequence-based mRNA expression profiling suggests two additional groupings not observed in microarray-based profiling, and miRNA expression profiling identifies at least three clusters not identified in microarray-based profiling. These expression signatures enable investigators to examine the interplay between microRNA and messenger RNA, noting that some interplay exists between miR-29a and *DNMT3A* transcript isoforms. Only *DNMT3A* mRNA isoforms harboring the miR-29a binding site have negatively-correlated expression profiles with miR-29a. Orthogonal analysis methods were used to investigate gene fusions because no tumor total RNA was available for verification. More than 1,500 gene fusions were called by Trans-ABySS and the UC Fusion Finder, with 64 recurrent gene fusions identified. The *MDS1* and *EVI1* complex locus (*MECOM*), which is a target of therapeutics such as aurintricarboxylic acid and arsenic trioxide, was focally amplified in more than 20% of ovarian cancer tumors. In this study, *MECOM* in-frame fusions with several different gene partners were identified in at least 3% of ovarian cancer cases; *MECOM* and *LRRC31* have a recurrent, in-frame fusion. Dr. Mungall noted that many cancer-related pathways are significantly enriched with fusion genes. In summary, he stated that efforts to date have generated mRNA-seq and miRNA-seq for 420 and

485 samples, respectively, from TCGA high-grade serous ovarian adenocarcinoma cohort. Unsupervised clustering of mRNA/miRNA expression profiles has identified additional sample groups. An exploration of putative miRNA and mRNA interactions has identified significant expression anti-correlations, including miR-29a, with specific isoforms of *DNMT3A*. In contrast to other cancers such as AML, duplication is the primary rearrangement leading to gene fusions, with *MECOM* fusions representing the most commonly recurrent in-frame events. Future work will include recurrent partial and internal tandem duplication rearrangements such as *MECOM*, differential expression and discriminatory gene analysis for unsupervised clusters and for gene rearrangements, and additional integrative analysis with TCGA collaborators. One participant commented that *TP53* mutation is a critical aspect of ovarian carcinoma and that genomic rearrangement is an integral part of this disease, which differs from other cancers.

Clinical and Pathologic Associations of Chromatin-Modifying Tumor Suppressors in Clear Cell Renal Cell Carcinoma (ccRCC)

Ari Hakimi, M.D.; Memorial Sloan-Kettering Cancer Center

Dr. Hakimi began by noting that renal cell carcinoma is the sixth leading cause of cancer death in the US and includes several malignant subtypes, including ccRCC, papillary, and chromophobe cancers. Nearly one-third of patients with RCC present with metastatic disease, and RCC is resistant to chemotherapy and radiation. Four prevalent mutations have been recently identified (*VHL*, *PBRM1*, *SETD2*, *BAP1*), all of which are located on the 3p21 locus. These genes are likely tumor suppressors, given that 3p is lost in approximately 90% of RCC tumors. *PBRM1* is frequently mutated in ccRCC. These genes are all in the histone modifying/chromatin remodeling pathway. However, *VHL* alteration alone does not have any bearing on prognosis, as *VHL* is lost in 80-90% of RCC cases (Gossage L, Eisen T. *Nat Rev Clin Oncol* 2010;7:277-288). This study sequenced 188 non-TCGA ccRCC cases and validated key findings using TCGA data (n=424). Results showed that *BAP1* is associated with nearly all tumor prognostic factors, and *SETD2* and *BAP1* are associated with poor survival and disease recurrence. Relatively few additional events appear as the tumor progresses. In conclusion, Dr. Hakimi noted that these efforts confirm the frequency of these novel mutations in ccRCC and suggest that they are associated with advanced stage, grade, tumor invasiveness, and high pathologic stage for smaller tumors. Tumors of less than four centimeters were more likely to acquire invasive characteristics if they contained these mutations. In particular, *BAP1* and *SETD2* were associated with worse cancer-specific survival. However, additional studies with longer follow-up will be necessary to assess the clinical impact of these and other mutations and to develop targeted therapies.

Individual Patient Cancer Profiles in the cBIO Cancer Genomics Portal

Jianjiong Gao, Ph.D.; Memorial Sloan-Kettering Cancer Center

Dr. Gao noted that the cBIO Cancer Genomics Portal aims to lower barriers between cancer genomics data and analysis, make genomics data available to a wide audience, facilitate explorative data analysis through facile viewing, and identify genes altered in tumorigenesis. Because tumor samples have hundreds or thousands of mutations, many of which are not disease-relevant, it is important to automate the process of identifying relevant mutations. The cBIO Cancer Genomics Portal is an interactive system to facilitate data mining that provides a patient-centric view that integrates clinical and genomic data (e.g., mutation profiles, gene annotation) with functional annotation. Data that can be visualized include a scatter plot of mutations versus copy number alteration, mutations and copy number alterations of specific interest to a case, a cancer study view that presents cohort data, and lists of significantly mutated genes.

One participant asked about the status of the portal's clinical interface for physicians to use. Dr. Gao replied that the portal is currently designed for researchers, although clinical elements are being added. Cohort information from GISTIC is being integrated for deriving case-specific CNA. Data in the portal are obtained from Firehose with some preprocessing and annotation to the MAF files. Dr. Gao stated that less than one hour is required to preprocess 100 samples.

Session II

Chair: Raju Kucherlapati, Ph.D.; Harvard Medical School

Comprehensive Genomic Characterization of Squamous Cell Carcinoma of the Head and Neck

David Neil Hayes, M.D.; The University of North Carolina at Chapel Hill

Dr. Hayes began by stating that head and neck squamous cell carcinoma (HNSCC) is the fifth most common cancer worldwide (the most common in central Asia) and the sixth most common in the US. Risk factors include smoking and infection with the human papilloma virus (HPV) 16, which makes E6 and E7 oncoproteins. HPV-infected cells express high levels of p16, the most prevalent clinical diagnostic for infection. Dr. Hayes then discussed the characterization of 279 TCGA HNSCC specimens from patients who predominantly have advanced-stage disease. These samples have been characterized fully using DNA, RNA, and miRNA sequencing along with DNA copy number profiling, mRNA quantification, and methylation analysis. Of these, 212 cases have undergone RPPA analysis. Although more than 200 additional samples will ultimately be collected and analyzed, the cohort of 279 specimens represents the largest genomic dataset ever assembled for each of the individual cancer sites (e.g., oral cavity, larynx, hypopharynx, oropharynx) by a factor of at least two. Limitations include the fact that this collection represents a surgical cohort and thus contains relatively few HPV-positive samples, although HPV status can be defined using seven distinct criteria. He noted that non-squamous cell lung carcinomas (NSCLCs) are among the most genomically deranged of all cancers, featuring a high mutation rate and many significantly mutated genes (e.g., *TP53*, *CASP8*, *HLA*). Significantly mutated genes in HPV-negative HNSCC highly overlap with those observed in lung squamous cell carcinoma; the two diseases appear similar in terms of mutations, CN, expression patterns, and pathways. The patterns of somatic copy number aberration in HNSCC are similar to those of lung squamous cell carcinoma, and there is some similarity in recurring focal amplifications between lung cancer and HNSCC. Dr. Hayes noted that HPV-positive HNSCC tumors feature a striking lack of oncogenes other than *PIK3CA*. While the CN landscape for HNSCC is rich, confident attribution of a gene is difficult. In these analyses, RNA Seq data were used to validate mutations and to look at deeper coverage of structural variants. Early analysis does not provide convincing evidence for recurrent in-frame gene fusions, although structural gene rearrangements are common. Functional events appear more likely than tumor suppressor genes to be inactivating events.

Dr. Hayes stated that expression analysis should identify patterns that are statistically significant, reproducible/valid, and of genomic/clinical relevance. Expression profiling in HNSCC reveals four subtypes that reflect structural rearrangements (e.g., basal, mesenchymal, atypical, and classical). Assessment of marker genes in these subtypes suggests roles for cell death and apoptosis pathways. Unbiased sequencing has been used to detect viral RNA, finding that 20% of patients express HPV16 viral RNA, although 11% have evidence of the virus itself.

Genomic Characterization of Cancer-Adjacent Tissue: Evidence of Field Effects and Expression Subtypes

Melissa Troester, Ph.D., M.P.H.; The University of North Carolina at Chapel Hill

Dr. Troester began by stating that breast cancer recurrence rates increase with breast-conserving therapy (Veronesi U, et.al. *N Engl J Med* 2002;347:1227-1232) and that local recurrence commonly occurs in the lumpectomy bed. Local recurrence rates are higher among basal-like breast cancers. Field cancerization explains the development of multiple primaries and local recurrences, and epithelial changes may create a normal-looking area that, if not removed, can create a second primary tumor. Also, cancer-adjacent tissue responds to the tumor through a host of mechanisms, including stress and immune responses, wound response, angiogenesis, and chemotaxis (Troester MA, et.al. *Clin Cancer Res* 2009;15:7020-7028). Current efforts use WES, CNA, methylation, microRNA seq and RNA seq to assess field effects. Field effects occur in all types of breast cancer tumors; 7% of samples have some evidence of a field effect (or tumor contamination) as assessed by CNA. Variant allele frequencies were compared between tumor and adjacent tissues using WES, indicating that 25% of tumors have strong evidence for a field effect or contamination. Methylation analysis shows that 7-10% of samples exhibit a field effect or contamination by tumor cells. Distinguishing between contamination and field effects may require histopathologic assessment. Dr. Troester noted that field effects will not be identified in cases for which there is good evidence that no tumor is present in the normal tissue. Methylation marks correlate with tissue stromal or epithelial content, a finding that could help to interpret methylation profiles. She noted that there are two subtypes of cancer-adjacent tissue--active and inactive. Active microenvironment occurs in all tumor subtypes and predicts survival. RNA expression analysis indicates two main clusters by microRNA seq and RNA seq. In these analyses, RNA and miRNA results were concordant, and tumor characteristics were not strongly associated with the main clusters. Samples with “probable contamination” were not readily detected. In conclusion, Dr. Troester noted that DNA analysis can show field effects or tumor contamination, and RNA identifies breast cancer expression subtypes. Future efforts will aim to distinguish field effects from contaminating tumor cells.

One attendee asked if these analyses incorporated hyperplasia and other changes in benign breast tissue, to which Dr. Troester replied that a pathologist’s score was provided for benign conditions. In these studies, younger age was associated with a more active phenotype.

Genome-Wide Analysis of Expression Quantitative Trait Loci in Breast Cancer

Nicholas W. Knoblauch; Harvard Medical School

Mr. Knoblauch began by noting that genome-wide association studies (GWAS) of breast cancer have identified a number of single nucleotide polymorphisms (SNPs) associated with increased risk. As such, gene expression can be considered as a phenotype. Expression quantitative trait loci (eQTL) analysis has been carried out on 382 TCGA basal breast cancer cases using germline SNP data generated by the Affymetrix 6.0 array platform to assess the relationship between

SNPs and quantitative traits. Imputation was used to estimate genotype for un-genotyped markers using a genotyped reference panel and a linear model with parameters for intercept, genotype, and covariant (e.g., ER status). Of the approximately eight million SNPs identified, 140,000 are eQTL, some of which were ER-dimorphic. Of the one trillion SNP-transcript interactions, 375,000 eQTL were identified. One attendee asked where the dimorphic eQTLs were highly expressed, to which Dr. Knoblauch replied that these appear to be lower in the minor allele. Another attendee noted that germline risk alleles are not associated with eQTLs, suggesting an alternate hypothesis for their role in cancer. One possible explanation for this observation is that SNPs could lead to cancer yet not promote any changes in expression.

Pathway and Network Analysis of Somatic Mutations across Cancer Types in TCGA
Ben Raphael, Ph.D.; Brown University

Dr. Raphael began by noting that identifying significantly mutated genes poses a challenge. Enriched gene lists often contain few genes, and many genes are mutated at modest frequencies. Because genes act in pathways, they can be interrogated at several points along the continuum from individual genes to networks. To this end, the Raphael laboratory has developed two algorithms, HotNet and Dendrix. HotNet aims to identify connected subnetworks of a gene interaction network that are mutated in a significant number of patients. Although individual genes can be individually significant, they can also be highly connected through a high-degree node. HotNet is an algorithm based on the concept that gene networks are “heated” entities in which the “heat” diffuses across neighboring genes along the edges of the network. As such, it incorporates the mutation frequency of a given gene and the topology of the interactions between genes. HotNet has been applied to identify subnetworks mutated across all cancer types as well as those mutated in a subset of cancer types. Given the noise within the interaction network, some filtering is required (e.g., selecting SNVs of frequency above 80%, using GISTIC maximum peak predictions). HotNet has been applied to pan-cancer analysis of 1,984 samples across 765 genes. It has also been applied to 316 ovarian cancer samples to identify 27 hot subnetworks whose genes were mutated in a significant number of patients. These analyses identified several networks of interest, including the iRef network, containing 11 subnetworks with three or more genes, and the HPRD network, which contained 20 subnetworks. Next steps include additional QC on mutation data, incorporating a background mutation model, incorporating other data types, and incorporating Dendrix to assess exclusive gene sets. These algorithms are available for download.

One participant asked how to transition from defining these pathways to validating them, to which Dr. Raphael replied that scaling presents some challenges, given that differences between mutation frequencies increase as data sets become larger.

Detection, Diagnosis, and Correction of Batch Effects in TCGA Data
Rehan Akbani, Ph.D.; The University of Texas M.D. Anderson Cancer Center

Dr. Akbani noted that there are many points at which batch effects can potentially be introduced into TCGA’s data pipeline. As such, it is critical to detect and quantify batch effects and to identify their sources. The Akbani laboratory has recently developed the MBatch R package (http://bioinformatics.mdanderson.org/tcgabatcheffects) that includes a suite of tools, both novel (e.g., PCA-Plus and BatchCorr) and standard (e.g., hierarchical clustering, box plots, clinical correlates, ANOVA/MANOVA), to identify batch effects. PCA-Plus plots points according to batch centroids, and BatchCorr provides a value that can be used with p-value to identify areas where batch effects are present. He noted that some batch effects will be inevitable, although the issue is not rampant in TCGA data. To correct batch effects, the source should be corrected when possible, although some corrective algorithms (e.g., ANOVA, ComBat, Median Polish) can also be employed. However, he cautioned that correction algorithms may go too far and "correct" the underlying biology as well. One participant asked if the data presented in these analyses were raw or normalized, to which Dr. Akbani replied that Level 3 data were used. Analysis of Level 1 data is impractical due to size.

Papillary Thyroid Carcinoma Analysis

Thomas Giordano, M.D., Ph.D.; The University of Michigan Medical School

Dr. Giordano began by noting that although the incidence of thyroid cancer is on the rise, mortality remains fairly constant. Papillary carcinoma consists of three main types--classical, follicular variant, and tall cell variant. While many genetic defects in thyroid cancer (e.g., *BRAF* mutation, *RET* rearrangement, *NTRK1* rearrangement) have been identified, approximately 25% of cases have no common driver mutations. Preliminary, non-validated analysis of TCGA papillary thyroid carcinoma specimens indicates that papillary carcinoma is highly differentiated with a low overall mutation rate. Genotype and histologic type appear to be strongly correlated. An interesting novel mutation was identified in *EIF1AX*, an X-linked translation initiation factor with no known role in any type of cancer. Recurrent fusions identified from integrating low-pass WGS and RNA Seq data include *CCDC-RET*, *ETV6-NTRK3*, and others. DNA methylation identified four subtypes that correlate with histologic type and mutational status. Dr. Giordano noted that thyroid-specific genes provide insight into progression and loss of response to radioactive iodine therapy. Methylation analyses have identified thyroid peroxidase (TPO), which is silenced by DNA hypermethylation in classical and tall cell variant tumors. miRNA analysis can cluster subtypes into either four or seven subtypes. These data analyses have also identified multiple associations with histologic type. In conclusion, Dr. Giordano noted that TCGA papillary thyroid carcinoma cohort is outstanding and representative of the disease. Overall, thyroid cancer has a low mutation rate with few CN changes. Preliminary analyses of TCGA data reveal strong associations between tumor morphology, genotype, gene expression, CN changes and methylation status, and many interesting leads for novel mutations and gene expression patterns have been identified. The first publication of these results is expected in mid-2013.

One participant asked if BRAF inhibitors could be a therapeutic strategy for these tumors, to which Dr. Giordano replied that radioactive iodine will likely remain the standard of care for the immediate future. Another attendee asked about similarities between thyroid carcinoma and lung adenocarcinoma, given shared thyroid transcription factor 1 and *RET* rearrangements. Dr. Giordano replied that thyroid cancer contains relatively few *EGFR* mutations; although a similarity with lung adenocarcinoma exists, it may be tenuous. The majority of deaths from thyroid carcinoma occur in poorly-differentiated and anaplastic cases. The role of heredity remains unclear, even for early-onset cases, because papillary carcinomas differ from medullar thyroid carcinomas, which are strongly familial.

Wednesday, November 28

Session III

Chair: John N. Weinstein, M.D., Ph.D.; The University of Texas M.D. Anderson Cancer Center

The Bladder Cancer Analysis Working Group (BLCA AWG): A Progress Report
John N. Weinstein, M.D., Ph.D.; The University of Texas M.D. Anderson Cancer Center

Dr. Weinstein began by stating that bladder cancer is four times more likely to occur in men than women and is the fourth most common cancer in men. The US spends 2.2 billion dollars per year in health care for bladder cancer. Bladder cancer has two predominant phenotypes, low-grade and high-grade. Low-grade tumors are superficial, less likely to metastasize, frequently reappear after resection, are amenable to therapy, and are associated with a low mortality. High-grade tumors have a propensity to invade and metastasize and have a high mortality when invasive, although they respond well to treatment if detected early. The Weinstein laboratory focuses on muscle-invasive forms of the disease, characterized as Stages 2-4 (15-20% of bladder cancer patients). He noted that 80% of patients with muscle-invasive disease present *de novo*, with distant metastases accounting for the most common cause of treatment failure. Cisplatin-based multi-agent chemotherapy is the standard of care for neoadjuvant therapy prior to cystectomy and for measurable metastatic diseases. The FDA has not approved a new agent to treat muscle-invasive bladder cancer in more than two decades.

TCGA analyses have focused on muscle-invasive urothelial cancer, for which 126 samples were included in the data freeze for the marker manuscript. A total of 153 samples have been qualified, with 138 in the analysis pipeline. Follow-up data are available on 126 cases, with a median time of 209 days. Thirty-five patients have died to date. Most of the patients studied are node-negative, Stage 3 cases, with a predominant percentage of reformed smokers. Dr. Weinstein noted that specimen accrual has been a factor, although the process has been recently accelerated. A data freeze has been established at 126 samples plus normals. Significantly mutated genes identified in preliminary analysis include *MLL2*, *TP53*, and *KDM6A*, among others. Unsupervised clustering of methylation data indicates four categories, and mRNA analysis identifies three subtypes. SuperCluster results that aggregate many types of data identify two to three clusters. Many fusion proteins, including FGFR3-TACC3, have been identified in these preliminary studies. Splice variation is a significant feature of these data sets. Heavy smokers show a CD44 exon skip event that is not present in non-smokers. Viral integration also appears to be significant. Four samples feature integration sites for four different viruses; three sites with viral sequences have no viral integration sites.

Chromatin remodeling analysis showed that epigenetic modifiers were mutated in more than three samples out of 100 analyzed. Five significantly mutated chromatin-modifying genes, *KDM6A*, *ARID1A*, *MLL2*, *MLL3*, and *EP300*, have been identified in these analyses. Results indicate pathways involved in bladder cancer, including p53/Rb, histone modification, RTK/Ras/PI3K, and SWI/SNF.

One participant commented that the location of a virus is often linked to the mutation spectrum, to which Dr. Weinstein replied that the mutation spectrum is the factor that caused the working group to search for virus sites. Another attendee noted that *CDK1A* is a growth suppressor gene that has never been observed to be mutated in cancer and asked about the relationship between tumor suppressor mutations and *TP53* mutations. Dr. Weinstein replied that the tumor suppressor mutations seen in these analyses appear to be mostly truncating mutations; no mutual exclusivity with *TP53* has been noted.

Identification of Gene Fusions using RNA Sequencing Data

Siyuan Zheng, Ph.D.; The University of Texas M.D. Anderson Cancer Center

Dr. Zheng began by noting that gene fusion has been recognized as a driver and a target in cancer. Although the first fusion gene, *BCR-ABL,* was discovered in cancer more than 30 years ago, sequencing data were first used to identify fusions in 2005. To help use RNA sequencing data to identify fusion genes in cancer, the Zheng laboratory has developed the Pipeline for RNA Seq Data Analysis (PRADA; http://sourceforge.net/projects/prada). PRADA features four modules: processing, expression calculation and QC, gene fusion identification, and a supervised search model. PRADA aligns reads to the transcriptome and genome using a multi-tiered strategy. Mapping to the transcriptome captures all transcript variants, whereas mapping to the genome captures unannotated transcripts. PRADA uses many established tools as its infrastructure, including SamTools and Picard. Reads per kilobase per million mapped reads (RPKM), a gene expression estimate from RNA seq data, have been used to call GBM subtypes in previous TCGA work. PRADA uses two lines of evidence to detect fusions: discordant read pairs and fusion-spanning reads. Filters designed to exclude false positives include a homology filter (gene partners cannot have significant sequence homology), the ratio of fusion-spanning reads over discordant reads, and additional filters based on gene partners and junction patterns. PRADA has been applied to TCGA GBM and ccRCC data to identify gene fusions. Eighty gene fusions were identified in 416 ccRCCs, and 15% of the samples had at least one fusion. Recurrent fusions include *SPFQ-TFE3, TFG-GRP128,* and others. Thirteen fusions were selected, and RT-PCR was used to validate eleven of these. 232 fusions were identified in 164 GBM specimens, and 68% of these samples contained at least one gene fusion. The *TFG-GPR128* fusion was observed in both GBM and ccRCC cancers. CNV in these two genes has been previously annotated in a number of large human population cohorts and in healthy individuals, suggesting a germline event. The observed fusion was caused by focal amplification and inversion. The Zheng laboratory has also developed the GUESS-ft tool to support the supervised search for *TFG-GPR128* fusions in normal TCGA specimens. In all cases where this fusion was observed in the tumor, it also appeared in the corresponding normal sample, indicating a germline event. The fusion can activate *GPR128* expression. PRADA also features a module to identify intragenic rearrangements that has been applied to RNA seq data from GBM. In summary, Dr. Zheng noted that PRADA provides functions useful for processing RNA seq data and can be used as a standalone version or within portable batch systems or load-sharing facilities. Modular steps allow users the flexibility to pause or resume analysis, and the tool facilitates batch analysis at approximately 180 samples per two weeks.

One participant asked about deletions observed in exons 12, 13, and 14, to which Dr. Zheng replied that these will be investigated further as analysis proceeds.

Comparative Mutational Analysis in Frozen and FFPE Tumor Samples

Gaddy Getz, Ph.D.; Broad Institute

Dr. Getz began by noting that tissue banks and biorepositories contain many formalin-fixed, paraffin-embedded (FFPE) samples that are well-characterized clinically and histopathologically. FFPE samples could fill an accrual gap for TCGA, given that FFPE remains the standard clinical practice for biopsy. Moreover, use of FFPE specimens could enable connections with existing clinical trials and move genomic analyses toward standard clinical practice. However, challenges with using FFPE specimens include difficulty of extracting samples, poor yield of DNA, poor quality of extracted DNA due to warm ischemic time, and other variables. FFPE data sets analyzed to date include four TCGA prostate “trios” (e.g., four FFPE tumor samples plus four fresh-frozen tumor/normal pairs), 46 breast cancer trios, and 17 lung cancer “quartets” (e.g., 17 FFPE normal/tumor pairs and 17 fresh-frozen normal/tumor pairs). Results form these analyses have demonstrated that, while frozen tissues have approximately double the library size of FFPE tissues, the latter affords a sufficiently deep library for WES analysis. Most targets captured in frozen tissue are also captured in FFPE; the total mutation count is similar in prostate and lung samples. FFPE mutations are not swamped by artifacts; mutations have the same spectra. In addition, FFPE tissues support CN analysis. In these experiments, there was a 44% overlap between mutations identified in frozen and FFPE specimens. Dr. Getz noted that comparisons of frozen and FFPE tissues alter two variables at once because samples come from different parts of the tumor in terms of purity and subclonal composition. The sensitivity to detect a mutation depends on coverage and allelic fraction, and allelic fractions in frozen and FFPE tissues differ due to differences in purities. The ABSOLUTE algorithm was applied to 217 ovarian cancer specimens to distinguish between clonal and subclonal mutations.

To compare frozen and FFPE mutation sets, Dr. Getz noted that it is unnecessary to call the mutations in FFPE and frozen tissues independently. However, it is essential to validate the existence of the mutations found in FFPE in the frozen samples, which will require more than two reads. Afterwards, one must correct for the different allelic fractions in the two samples due to their different purities. Sites should then be stratified based on the power to validate a mutation, and a program such as ABSOLUTE must be used to distinguish clonal and subclonal mutations. Cancer genome projects can use FFPE samples to generate similar lists of significant genes to those found in frozen tissues using a program such as MutSig. In conclusion, Dr. Getz noted that WES on FFPE samples is robust. The overlap between FFPE and frozen samples can be calculated by controlling for relative coverage and adjusting for different allelic fractions. Mutation rates and categories are similar for the two tissue types, with subclonal mutations contributing to the differences. Thus, clinical FFPE tissues can be used for WES to support cancer genome projects. Ongoing challenges include low-yield samples, variability in FFPE blocks due to age and fixation methods, and sample preparation requirements for WGS

One participant asked whether it would be prudent to amplify DNA from small samples. Dr. Getz replied that such a strategy is not advisable, as it produces unpredictable artifacts.

Structural Variant Detection in Colorectal Cancer

Evert van den Broek, M.D., Ph.D.; VU University Medical Center Amsterdam

Dr. van den Broek began by stating that colorectal cancer (CRC) is the second cause of cancer-related death worldwide. Nearly 40% of CRC patients will die due to metastatic cancer. Nearly all CRC tumors exhibit chromosomal instability. Primary tumor and matched normal DNA from 356 patients enrolled in the Phase III Capecitabine, Irinotecan, and Oxaliplatin in Advanced Colorectal Cancer (CAIRO) and CAIRO2 trials was analyzed to identify recurrent somatic variants that cause CRC. Comparative genomic hybridization identified 5,737 genes with one or more breakpoints, and 482 candidate genes were identified with recurrent breakpoints. *MACROD2* was the most prominent gene identified. Limitations of array comparative genomic hybridization (CGH) data include the fact that the location of a breakpoint is an estimate, the DNA structure is unknown, and balanced events will be missed. As such, candidate validation is required. The van den Broek laboratory has developed a candidate-driven structural variant detection algorithm based on a read-pair approach. The read-pair approach is then combined with read depth to define breakpoint location and to determine tumor-specific events.

TCGA data were then used as a validation set for the candidate genes identified from the CAIRO data sets. Distribution of discordant pair groups per type showed an approximately five-fold higher number of translocation-discordant pair groups for candidate genes compared to control genes. In conclusion, Dr. van den Broek noted that 482 candidate genes with recurrent breakpoints were identified in 356 CRC samples based on array CGH analysis. TCGA provided an essential CRC reference dataset to validate structural variants in candidate genes with recurrent breakpoints. Identification of breakpoints based on array CGH is correlated with SV detection in TCGA CRC data. Further studies will be performed to investigate the clinical and functional significance of validated candidate genes. RNA Seq data will then be mined to assess the effect of these mutations on transcripts. In response to a participant’s query, Dr. van den Broek noted that low-coverage data were used for the validation studies.

Inhibitor-Sensitive Fibroblast Growth Factor Receptor Mutations in Lung Squamous Cell Carcinoma

Rachel G. Liao; Harvard University

Ms. Liao began by noting that adenocarcinoma of the lung has seen many targeted therapy advances in the past decade (e.g., *EGFR*, *ERBB2*, *EML4-ALK*), while squamous cell carcinoma has had few targets and no targeted therapies despite its great clinical burden. The FGFR receptor tyrosine kinase family has been implicated in many cancers. Analysis of TCGA lung squamous cell carcinoma specimens reveals that focal amplification of *FGFR1* occurs in approximately 10% of cases. *FGFR2* and *FGFR3* are each mutated in approximately 3% of cases. Neither gene is significantly mutated in this data set, possibly because these genes represent a family of receptors. *FGFR2* and *FGFR3* mutations are observed in lung SqCC but do not repeatedly co-occur with other events except for *TP53* mutation. *FGFR2/3* mutations are transforming (e.g., they can drive proliferation) in an anchorage-independent growth assay, although *FGFR2/3* transformation can be blocked by FGFR inhibitors. Loss of transformation correlates with loss of phosphorylation, and cells exhibiting dependency on the FGFR pathway are sensitive to FGFR inhibitors. One clinical case of head and neck squamous cell carcinoma with *FGFR2* mutation regressed upon treatment with a multi-kinase inhibitor known to inhibit

FGFR2. In conclusion, Ms. Liao stated that *FGFR2/3* mutations observed in lung SqCC are sufficient to drive transformation in the NIH-3T3 cell line model, and the transformation phenotype can be reversed by inhibiting *FGFR* with small molecules. Ba/F3 cells that depend on *FGFR2/3* signaling for proliferation can be growth-inhibited by small molecules. A clinical success confirms that these findings provide a rationale for further study of patients who have *FGFR* events in their tumors. TCGA data have been used effectively to identify novel targetable driving events in tumors, although these events do not always meet the threshold of statistical significance. One attendee asked about amplification events that could be driver mutations, to which Ms. Liao replied that a subset of tumors that present with amplifications are likely driven by *FGFR1* signaling events. Not all patients respond to therapies targeted to *FGFR*.

Analysis of 3,000 Cancer Exomes to Identify Novel Cancer Drivers and Therapeutic Opportunities
Nickolay Khazanov, Ph.D.; Compendia Bioscience

Dr. Khazanov stated that Compendia Bioscience seeks to use data from TCGA and other efforts to define a comprehensive catalog of driver mutations in cancer to support identification of novel therapeutic targets and appropriate patient populations. TCGA data, and WES analyses in particular, are rich in mutation and fusion information. Challenges to this effort include the scale of data to collect and process, the heterogeneity among data formats and analysis methods across data types and centers, the speed necessary to accommodate a rapidly growing dataset, and ways to identify true driver events using immature and evolving methods. Compendia uses data from TCGA, the Broad Institute's GDAC, and CGHub. To date, the company has processed 2,998 samples from 15 diseases (1.1 million mutations) using data from TCGA. Pan-cancer analyses have identified 107 potential gain-of-function (GOF) genes (e.g., Ras family genes) and 120 potential loss-of-function (LOF) genes (e.g., *PTEN*, *APC*). *PIK3C* and *RAF1* are mutations seen across several cancer types, and a possible central role of *RAF1* can be deduced only through pan-cancer analysis. Fusions have been assessed across 1,475 samples spanning six diseases with known gene fusions using an algorithm based on batch parallel processing with novel filtering and classification schemes. *RET* fusions were observed across fourteen samples representing breast, lung, and thyroid cancers. Future directions for these efforts include using TCGA data for integrative analysis (gene and pathway-level summarization, co-occurrence, mutual exclusivity, and clinical subtype), looking at cancer types beyond TCGA, using model systems to map drivers, and continuing partnership with the biopharmaceutical industry.

Session IV
Chair: Charles Perou, Ph.D.; The University of North Carolina at Chapel Hill

Integrated Genomic Characterization of Endometrial Carcinoma
Douglas A. Levine, M.D.; Memorial Sloan-Kettering Cancer Center

Dr. Levine began by noting that endometrial cancer can be thought of as essentially two diseases, given the differences between endometrioid and high-grade serous endometrial carcinomas. These types can be difficult to differentiate pathologically when tumors are high grade. Type 1 cancers are endometrioid and are treated with radiation, whereas Type 2 are more aggressive, metastatic tumors that are treated with chemotherapy. Many mutations have been identified in each Type. Approximately 50% of endometrial cancers that have spread beyond the uterus ultimately recur. This TCGA study (n=373) selected primary, newly-diagnosed, untreated, endometrial cancer with tissues collected from the endometrium or uterus. To date, analyses include 248 WES, 107 low-pass genomes, and 333 RNA Seq. Preliminary results indicate that mutation profiles become more complex as the histology becomes more aggressive. CN analysis reveals four groups, including one group that is ultramutated. Nearly one quarter of high-grade endometrioid tumors cluster with serous tumors. Common mutations include *PTEN* in low-grade endometrioid tumors and *TP53* in serous cases. The ultramutated group contains universal mutations in polymerase ε (*POLE*); 13 of 17 cases have one of two hotspot *POLE* mutations. These findings are similar those observed in colorectal cancer. In endometrial cancer, serous and high-grade cases are associated with worse outcomes. Groups with microsatellite stability (MSS and instability (MSI) show little difference in survival. Approximately 50 significantly mutated genes have been identified, although no events have been identified to date in the small *POLE* group. MicroRNA clustering identifies six subgroups, and methylation analysis identifies four subgroups. Gene expression clustering identified three clusters (e.g., mitotic, hormonal, and immunoresponsive). Supervised RPPA data were used to evaluate the clustering data, indicating that the mitotic cluster has increased expression of DNA repair and proliferative genes. The PARADIGM algorithm identified five clusters centered around pathways such as PTEN and PI3K/AKT. *PIK3CA* mutations show common hotspots but different mutation spectra from those of breast and colon cancers. Multiplatform analysis reveals molecular similarities among ovarian serous, uterine serous, and basal-like breast cancers, although mutation frequencies vary across these tumor types. Genomic similarities are likely due to shared *TP53* mutations, although it is possible that ovarian serous and uterine serous tumors share a common site. In summary, Dr. Levine noted that these efforts have identified recurrent *POLE* mutations that are associated with an altered mutation spectrum and a very high mutation rate in these tumors, with the PI3K/AKT pathway being the most activated in endometrial tumors.

One participant observed that almost all tumors with *PTEN* mutations also contain *PIC3K* mutations, suggesting that RPPA data could be mined to look for activation of downstream markers.

Assessing Tumor Heterogeneity and Tracking Clonal Evolution using Whole Genome or Exome Sequencing

Christopher A. Miller, Ph.D.; The Genome Institute at Washington University

Dr. Miller began by noting that tumors are heterogeneous, and evolution occurs at the cellular level. Clonal evolution in relapsed acute myelogenous leukemia (AML) suggests that hematopoietic stem cells gain mutations and may expand. At diagnosis, a cross-section of architecture (e.g., founding clone, subclones) is observed. Chemotherapy creates a bottleneck in which few cells pass through, although this process ultimately selects a clonal fraction that leads to relapse (approximately 5% of the original tumor). Thus, in AML, detecting minor subclones is critical. However, genomes are sequenced with low coverage (e.g., 30X), and algorithms are not designed to detect low-frequency events. The Miller laboratory has developed the Bayesian Scoring of Somatic Variant Read Counts (BASSOVAC) algorithm that incorporates purity, ploidy, base quality, allele frequency, and overall mutation rate, to obtain probabilities of heterozygous and homozygous somatic events in a set of data. BASSOVAC has been tested in primary breast tumors, matched normal tissues, and three metastases, all of which were whole-genome sequenced to 30X coverage. Mutation calls were made using the Somatic Sniper and

Varscan programs, and capture validation was performed for all variants. Deep read counts were obtained from validation sequencing for all variants in all samples. More than 50% of the variants present in the metastases were present at a detectable level in the tumor. BASSOVAC can be used to detect true variants at frequencies as low as 2%. To determine which variants are present in different subclones, an integrative approach is required that incorporates variant allele frequencies, CN calls, and purity and ploidy information. Clonality plots can be constructed, and clones can be inferred in an automated, unbiased manner, given that most tumors have a founding clone and one or more subclones. Dr. Miller noted that there is a lower bound on the number of clones. In summary, he stated that BASSOVAC can detect somatic mutations at very low frequencies and that robust automatic methods have been developed for inferring details about a tumor’s subclonal architecture. These efforts aim to characterize minor subclones at diagnosis rather than discovering their presence at relapse.

Dr. Miller noted raw data were fed into these algorithms. The fact that a mutation appears at a low frequency suggests more than one clump of cells; clumps may be a more accurate way than metastasis to conceptualize the spread of disease.

TCGA Benchmark 4: Evaluating SV and SNV Calls using Cell Line Genomes

Adam Ewing, Ph.D.; The University of California, Santa Cruz

Dr. Ewing described an open-invitation mutation calling benchmarking program using a set of BAM files for tumor/normal pairs, which will be distributed to participants. Mutations will be called and returned as VCF files, which will be collected and compared. Mutation types in this exercise include SNVs, in/dels, structural variants, and CNVs, called on whole genomes derived from cell lines. This exercise represents the fourth TCGA-sponsored benchmarking program; previous exercises have focused on calling SNVs on various sets of whole genomes and exomes. TCGA continues to benchmark because it must measure and set standards for accuracy of mutation calls and evaluate the performance of callers for in/dels, SNVs, and CNVs. Dr. Ewing noted that this controlled experiment will simulate normal contamination and subclonal expansion. The cell line data are publicly available, enabling wide participation. Although cancer genomics depends on the fidelity of somatic mutation calls, great discordance exists among centers and algorithms. This exercise will use two tumor/normal pairs derived from breast cancer tumor cell lines, HCC1143 and HCC1954, which are available from the American Type Culture Collection and sequenced at the Broad Institute. This mutation-calling exercise involves three components: comparison of tumor/normal pairs for each cell line and then simulating normal contamination and subclone expansion for both by mixing and spiking. BAM files for benchmarking will be distributed by CGHub. Although VCF is a successful standard, Benchmark 4 will also stimulate the creation of new tools, including Bamsurgeon, VCFcomparator, LeftShiftBreakends, and VCF to MAF Converter.

Polymerase Epsilon Mutations Accelerate Mutation Rates in Colorectal and Endometrial Cancer

David A. Wheeler, Ph.D.; Baylor College of Medicine

Dr. Wheeler began by noting that mutation rates classify patients with colorectal cancer. For example, one hypermutated group has MSI associated with a high rate of *MLH1* silencing, while another small group of patients presents very high mutation rates with no MSI and no apparent

MLH1 silencing. However, all of these patients have *POLE* mutations. When all mismatch repair systems are assessed, some genes are identified that do not increase in mutation rate with higher overall mutation frequency. *POLE* mutations in CRC are clustered mainly in the exonuclease domain and recur at several sites, even within the small data set. These polymerases, such as T4, are known to cause a high rate of mutation. The "mutator phenotype" has also been studied in bacteria, yeast, and mice. *POLE* and *POLD1* exonuclease knockout mice exhibit the mutator phenotype and die quickly from cancer.

Dr. Wheeler noted that ultramutated CRC patients show dramatic skewing in relative frequencies of mutations, which represent a combination of repair processes and polymerase actions. Results may indicate inefficient mutation repair, although this mechanism has not been established definitively. Similar phenomena related to *POLE* mutations are observed in CRC and endometrial cancers. The "polymerase B" domain family sequence has been relatively well conserved evolutionarily. The two major replicative DNA polymerases, POLD and POLE, show an asymmetry of function. POLE functions on replication of the leading strand, whereas POLD functions on lagging strands. The mutation profile is skewed at sites enriched for origins of replication. As such, no exonuclease domain *POLD1* mutations are observed in ultramutated patients. Mutation frequencies are anti-correlated with expression level in ultramutated patients. Progression-free survival in uterine corpus endometrioid carcinoma favors ultramutated patients as compared to hypermutated patients. In conclusion, Dr. Wheeler stated that a rare exonuclease mutation in *POLE* leads to an ultramutator phenotype in colorectal and endometrioid cancers. This phenotype defines a new subtype of these tumors that may have unique prognostic features and interesting biologic properties. Ultramutator patients exhibit a signature of transcription-coupled repair. The absence of POLD1 ultramutators suggests that POLD1 may perform an essential function in this new subtype of colorectal and endometrioid cancers. Whole genome sequencing should help to separate the effects of transcriptional repair from strand-specific mutation effects.

The Somatic Genomic Landscape of Glioblastoma Multiforme

Roel Verhaak, Ph.D.; The University of Texas M.D. Anderson Cancer Center

Dr. Verhaak stated that his presentation would focus on marker analysis in GBM tumors using a more substantial data set than that supplied for previous TCGA publications. The current analysis included 291 WES analyses, 17 WGS profiles, 544 mRNA expression profiles, 491 miRNA expression profiles, and 545 DNA methylation analyses. Results show that 71 genes are significantly mutated in 291 GBMs, including some novel genes (e.g., *SPTA1*, *ATRX*, *TCHH*) mutated at frequencies above 3%. No significant gene mutation was observed in approximately 10% of samples. Five cases presented *BRAF* V600E mutations, which are sensitive to vemurafenib in melanoma. Mutations in chromatin-modifying genes were detected in 41% of GBM samples. Permutations of similarly-sized gene sets suggest significance for chromatin-remodeling mutations. Analysis of more than 540 samples allowed the precise definition of CNA target regions. Focal CN loss targets tumor suppressor genes. WGS identifies complex rearrangements between chromosomal sections, such as the formation of double minutes. RNA Seq identified 84 in-frame fusion transcripts, which frequently result from local inversions. Genome breakpoints are associated with CN difference. 6.4% of GBM tumors analyzed harbor transcript fusions involving *EGFR*. Intragenic rearrangements in *EGFR* can be detected through RNA sequencing, with the majority of point mutations occurring in the extracellular domain.

Three different C-terminal deletions were identified, and two relatively unknown variants (exon 12-13 and exon 14-15) were detected. Approximately 45% of GBM tumors harbor an *EGFR*-associated point mutation or genomic rearrangement. GBM expression subtypes are related to abnormalities in *MYC*, *EGFR*, and *IDH1*, among others. These analyses suggest that G-CIMP hypermethylators associate with better outcome, and the proneural class performs more poorly than other subtypes when taking out G-CIMP. Moreover, protein expression levels associate with transcriptomal class. In summary, Dr. Verhaak noted that approximately 600 GBM specimens have been profiled comprehensively to characterize the somatic alteration landscape of GBM. Novel significantly mutated genes detected include *SPTA1*, *LZTR1*, *KEL*, and *TCHH*. Whole-genome and mRNA sequencing detected genomic rearrangements, most notably involving *EGFR*. The proneural subclass may perform worse than other subtypes.

In response to participants' queries, Dr. Verhaak noted that both double-minute samples contained *MDM2*, and one contained *EGFR*. Moreover, exome 8 and 9 deletions were observed in 1-2% of samples. The novel mutated genes may be biologically significant, although supporting functional data will be required.

Network-Based Stratification of Tumor Mutations
Matan Hofree, M.S.; The University of California, San Diego

Mr. Hofree noted that stratification of cancer subtypes will improve prognostics, enhance understanding of tumor biology, define new subtype-specific drug targets, and improve tailored treatment. Efforts to stratify subtypes using gene expression have identified GBM subtypes linked to survival. However, recapitulating a clinical phenotype using expression-defined subtypes has been less successful for ovarian cancer (TCGA Network. *Nature* 2011;474:609-615). Analysis of somatic mutations in high-grade serous ovarian cancer tumors suggests that these could possibly be used for subtyping, although data are sparse and often difficult to cluster according to genotype. Network-based stratification begins with bootstrapping and network-smoothing steps, after which a network clustering approach is applied. This process is then repeated, and data are aggregated. An intuition for network smoothing forms a more nuanced network that can identify areas of overlap between genotypes. Network-based stratification using somatic mutations from TCGA ovarian cancer specimens has provided a biologically relevant signal in which clusters associate with patient survival. Four subtypes have been identified, one of which performs poorly in terms of survival. However, these clusters do not always match those identified using other data types. Clinical translation of subtypes using expression signatures can be accomplished by defining subtypes using somatic mutations that predict a clinical phenotype (survival, drug response), training a model on matched gene expression to predict subtypes within the same set of patients, and then predicting subtypes using expression in new patients. These efforts have identified several pathways in TCGA ovarian cancer samples, including caspase and FGFR. In conclusion, Mr. Hofree noted that network-based stratification recovers biologically relevant subtypes of ovarian cancer. Somatic mutation subtypes differ from those recovered using other molecular profiles, and these subtypes can be recapitulated using gene expression. Each cancer subtype appears to have specifically-effected networks.

One participant asked if clinical variables provide additional predictive value. Mr. Hofree replied that clinical variables have not been included because doing so will render the approach more supervised. Clinical variables such as stage, grade, and age appear subtype-independent.

Session V

Chair: Louis M. Staudt, M.D., Ph.D.; National Cancer Institute

Comprehensive Analysis of Lung Adenocarcinoma

Matthew Meyerson, M.D.; Dana-Farber Cancer Institute

Dr. Meyerson began by stating that lung cancers cause more than 25% of cancer deaths in the US each year and are the leading cause of cancer deaths among men and women in America. Lung cancer kills more than one million people each year worldwide. Major lung cancer histologies include lung adenocarcinoma, squamous cell lung carcinoma, and small cell lung carcinoma. The most common form of lung cancer, lung adenocarcinoma, accounts for approximately 40% of lung cancer diagnoses and approximately 65,000 deaths each year in the US. Lung adenocarcinoma uniquely often occurs in non-smokers. As such, the disease has become a paradigm for molecular subtyping, as treatments have moved toward molecular-based strategies. Targeted inhibitors of *EGFR* (e.g., gefitinib, erlotinib) and *ALK* (e.g., crizotinib) have been made possible by genomic discoveries. Several previous comprehensive genomic studies of lung adenocarcinoma have identified many disease-associated genes. Despite the identification of molecular subsets, more than half of all lung adenocarcinomas lack an identifiable driver mutation.

As of October 2012, TCGA's lung adenocarcinoma project has accrued 303 samples with comprehensive molecular data, 230 of which were included within the data freeze. The majority of samples excluded were due to pathology review, and these cases will be included in a subsequent pan-NSCLC report. The goal is to submit a manuscript in Spring 2013. Copy number analysis of lung adenocarcinoma indicates that chromosome 3q is frequently gained in SqCC but not lung adenocarcinomas. Focal CN alterations in lung adenocarcinoma include *TERT*, *TERC*, *EGFR*, *MET*, and *CCDN3*, among others. Exome and RNA Seq analyses of 230 tumor/normal pairs and 230 tumor RNAs indicate that lung adenocarcinoma has a very high rate of somatic mutations, which can pose a problem when identifying significantly mutated genes. For example, known recurrently-mutated genes (e.g., *ERBB2*, *CTNNB1*) are not classified as significant regardless of the method used. While expression filtering enriches for real genes, a variety of alternate approaches, including two-stage statistical analysis and functional significance analysis, must be considered. In the end, a much larger sample size may be required to elucidate the full population of causative mutations in lung adenocarcinoma. The most prevalent mutated genes in lung adenocarcinoma include *TP53*, *KEAP1*, *STK11*, and *NF1*; novel candidate genes identified include *BCL9L*, *MGA*, and *MKI67IP*. Recurrent mutations are observed in SWI/SNF chromatin remodeling genes such as *ARID1A* and *SMARCA4*. Expression-based clustering (n=230) shows reproducible classes of lung adenocarcinoma (e.g., bronchioid, squamoid, magnoid). Low-pass WGA of lung adenocarcinoma (n=133 tumor/normal DNA pairs for low-pass; n=230 for RNA Seq analysis) identify fusions with known fusion partners, including *ALK*, *ROS1*, and *RET*. A *VMP1-RPS6KB1* fusion was detected in approximately 6% of cases, and peptidase fusions were also identified that bear further exploration. DNA methylation analysis (n=181 tumors/181 normal) showed that *CDKN2A* is inactivated by multiple genomic mechanisms. miRNA clustering (n=352 tumors) has identified signatures for five groups that can be discriminated by miR-10a/183, 143, 375, 148a, and 21, with miR-21 expression defining a large subset of lung adenocarcinoma. MutSigCV analysis of "oncogene-positive" and "oncogene-negative" samples examined mutational events in tumors lacking *RTK* activation and other defining events.

Integrative cross-platform analysis showed major deregulation of RTK/RAS/RAF and PI3K/AKT pathways in lung adenocarcinoma. RPPA analysis (n=183) showed that lung adenocarcinoma clusters into distinct groups, including *RTK* activation, *MEK* activation, and DNA repair groups, that are independent of smoking status. In conclusion, Dr. Meyerson stated that lung adenocarcinoma and SqCC have similar CN profiles. These cancers feature very high mutation rates, thus challenging the identification of novel mutated genes. RNA sequencing data identified three distinct expression subtypes in lung adenocarcinoma, and multiple fusions and mechanisms for *CDKN2A* inactivation are also expressed in this disease. Distinct clusters are observed from RPPA and miRNA analysis. Mutational differences between “oncogene-positive” and “oncogene-negative” subtypes include enrichment of *NF1* mutations in the oncogene-negative group.

One participant asked about the possible estrogen receptor (ER) positivity in lung adenocarcinoma subsets, to which Dr. Meyerson replied that there is some evidence to support the role of ER signaling in lung adenocarcinoma. Another attendee asked whether the expression subgroups indicate their cells of origin. Dr. Meyerson noted that the bronchioid subgroup likely arises from alveolar type 2 pneumocytes. Another participant asked whether smokers and non-smokers showed differences in expression subtype. Dr. Meyerson replied that approximately 15% of these cases represent persons who never smoked. However, comparative analyses of smokers and non-smokers in this cohort are underway.

Predicting Time to Ovarian Carcinoma Recurrence using Protein Markers

Ji-Yeon Yang, Ph.D.; The University of Texas M.D. Anderson Cancer Center

Dr. Yang began by observing that the standard treatment for ovarian cancer is surgery followed by platinum-based chemotherapy. However, approximately 25% of patients do not respond to chemotherapy. As such, identifying platinum-resistant patients at the time of diagnosis could alter the approach to therapy for ovarian cancer. The Classification of Ovarian Cancer (CLOVAR) model predicts outcome in ovarian cancer by using gene expression signatures to predict survival. Using TCGA data, CLOVAR identified disease subtypes and survival gene-expression signatures that could inform a prognostic model. In initial analyses, CLOVAR was less effective at predicting progression-free survival than overall survival. Efforts are underway to develop a predictor of platinum resistance based on protein markers. RPPA was carried out on 172 proteins and phosphoproteins across 412 TCGA serous ovarian cancer samples (222 cases in the model set). The Least Absolute Shrinkage and Selection operator (LASSO) was applied to TCGA training set (n=222) to select the nine proteins most associated with progression-free survival in ovarian cancer. The Protein-Driven Index of Ovarian Cancer (PROVAR), a weighted, linear combination of the nine markers, was predictive of overall and progression-free survival in TCGA data set. PROVAR was validated in an independent set of 229 high-grade serous samples, in which the index predicted time to tumor recurrence and overall survival. PROVAR was then compared to CLOVAR using a set of 130 samples with available gene expression data, indicating that PROVAR improves survival prediction. The robustness of the nine protein markers was assessed to identify five proteins from the validation samples (one of which overlapped with those identified in the training set). Hierarchical clustering of 172 TCGA proteins identified four clusters. In conclusions, Dr. Yang noted that PROVAR is predictive of progression-free and overall survival in high-grade serous ovarian cancers. Unlike genetic signatures in previous studies that often contained a large number of genes, PROVAR is sufficiently simple and predictive to be used in clinical practice.

One attendee commented that androgens are statistically significant predictors of ovarian cancer. Another participant suggested removing the nine identified markers and re-analyzing the data. Dr. Yang noted that ElasticNet could be used.

The Molecular Diversity of Luminal A Breast Tumors
Giovanni Ciriello, Ph.D.; Memorial Sloan-Kettering Cancer Center

Dr. Ciriello began by stating that breast cancer represents a set of diseases characterized by HER2 and ER receptors but with distinct molecular traits, prognosis, and therapeutic options. Luminal breast cancer tumors, the most heterogeneous subtype, are HER2-negative and estrogen receptor (ER)-positive. These tumors are classified either as luminal A or luminal B. Using multiple datasets, approximately 1,500 luminal tumors have been analyzed to compare CNA, somatic mutations, and mRNA expression. While HER2-positive and basal breast cancers feature the highest number of mutations per sample, luminal A tumors are clinically and molecularly heterogeneous. These results suggest that molecular diversity could possibly be linked with outcome variability. Copy number alteration identified five major subgroups of luminal A tumors, some of which are characterized by low-level (or no) copy number alteration. In these analyses, genomic instability correlated with poor prognosis. Copy number-high luminal A tumors show high levels of expression of *TP53*, *MYC*, chromosome 23 gain and 3p loss but downregulate *PIK3CA*. To assess the pathways most deregulated in other luminal A subtypes, the unbiased Mutual Exclusivity Models (MeMo; www.cbio.mskcc.org/memo) program was used to identify alterations to components of Ncor and SMRT, co-repressor complexes important to ER+ tumors. Ncor/SMRT complexes are required to mediate tamoxifen effects. In conclusion, Dr. Ciriello noted that the genomics of luminal A tumors have been dissected using multiple datasets to explain their molecular and clinical heterogeneity. Five major subtypes of luminal A tumors were characterized by CNA and somatic mutations. The atypical luminal subtype is characterized by high genomic instability, high levels of *TP53* mutation, upregulation of Aurora kinases, and poor prognosis. Luminal A hallmark mutations are prevalent in tumors characterized by low copy number alterations. Multiple rare but mutually exclusive associated with the loss of Ncor/SMRT were identified that may predict lack of response to endocrine therapy.

Virus Analysis in Head and Neck and Bladder Cancers
Michael Parfenov, M.D., Ph.D., M.S.; Harvard Medical School

Dr. Parfenov began by noting that viral infection is an important risk factor for many cancer types. Head and neck squamous cell carcinomas are the sixth most common cancers worldwide. Between sixty and eighty percent of oropharyngeal cancer and approximately twenty percent of oral and laryngeal cancers are caused by HPV. HPV-mediated cancers have significantly improved outcomes relative to non-HPV-mediated tumors. Bladder cancer is the second most common genitourinary cancer in adults. There is a moderate association between HPV and *BK polyomavirus* infection and bladder tumors. To date, genome sequencing data have been analyzed from 115 TCGA head and neck and 105 TCGA bladder cancer tumor/control pairs. For many HPV-positive samples, the entire viral genome presents in infected cells. For some tumors, less than 80% of the viral genomes are present in the cells. Coverage varies across samples and within the same sample. Viral integration disrupts the gene that stops viral oncogenes E6 and E7; HPV integrates into the *TRPC4AP* gene. Among HPV- or *polyomavirus*-positive samples, integration events were detected in fourteen cancer samples and two controls. Some cases showed the formation and amplification of chimeric episomes. The current model suggests that viruses integrate into the *TRPC4AP* gene, although the chromosome scar is not visible. Several alternate models based on segregation and re-replication have also been proposed. In conclusion, Dr. Parfenov stated that viral sequences and their cellular status can be detected effectively from low-pass WGS data. Eight percent of head and neck cancers and four percent of bladder tumors are HPV-positive. These results suggest that integration events may contribute directly to carcinogenesis through viral gene expression and modification of cellular tumor suppressors or oncogenes. Approximately 25% of all HPV integration events are followed by excision of the fused host and viral regions that form the circular mini-chromosomes that present in multiple copies within the cancer cells.

One participant asked if multiple events were observed in tumors for which the virus integrated into 100% of malignant cells. Dr. Parfenov replied that one tumor showed two such events. Cases where this phenomenon is not observed would suggest that the virus is not picking up the driver gene.

GeneSpot: A Portal for Interactive Gene-Centric Exploration of The Cancer Genome Atlas

Brady Bernard, Ph.D.; Institute for Systems Biology

Dr. Bernard began by noting that cancer researchers want to know TCGA mutation profiles, significant CN aberrations, and the data-derived statistical annotations for a given gene. However, there is no consensus method by which data repositories are organized. The typical workflow for obtaining data includes downloading, parsing, processing, and merging all data to extract features of interest. This process requires significant time, resources, and expertise. Of the billions of data points generated, only a few thousand are relevant. To circumvent the need to download data or software, the Bernard laboratory has developed GeneSpot (http://genespot.cancerregulome.org), a web-based controllable canvas with numerous gene-centric views of all cancer types supported by TCGA. GeneSpot allows any data file stored in a TCGA repository to be browsed, retrieved, and filtered, either by gene lists or row and column labels. Based on an HTML5/JavaScript architecture with a cloud application, GeneSpot allows users to establish, save, and share sessions, thereby enabling research collaborations. Dr. Bernard noted that the challenge of designing mining tools is to display information usefully. With GeneSpot, users can dynamically control the layout and visualization. Future directions include adding more views and making content publishable to social media applications.

Functional Characterization of **KEAP1** *TCGA Mutants in Lung Squamous Cell Carcinoma*

Bridgid E. Hast; The University of North Carolina at Chapel Hill

Ms. Hast began by noting that *KEAP1/NRF2* signaling is the major cellular mechanism to control reactive oxygen species (ROS) and regulate intracellular oxidation-reduction (redox) homeostasis. *NRF2* activity modulates survival via redox homeostasis. *NRF2* target genes such as *HMOX1*, *GCS*, *NQO1*, and *MRP* mitigate acute spikes in ROS, assist with chemotherapeutic/xenobiotic clearance, and control metabolically-derived ROS. Analysis of 178 squamous cell lung carcinomas reveals that the disease features pathway mutations in *KEAP1/NRF2* signaling. Mutations in *KEAP1* and *NRF2* are mutually exclusive; collectively, *KEAP1*, *NRF2*, and *CUL3* mutations were altered in 34% of analyzed tumors. *KEAP1* mutations exhibit differential suppression of *NRF2*-mediated transcription, and *KEAP1* mutants differentially bind to interacting proteins. *KEAP1* mutants cluster into four classes based on *NRF2*-binding ability. Certain mutants consistently bind more *NRF2* than does wild-type *KEAP1* yet cannot suppress *NRF2*-mediated transcription. These “superbinders” may represent subpar structures that are slow at turning over *NRF2* because their structure is perturbed. These studies show that *KEAP1* cysteine residues are stress-specific. There are several ways to induce oxidative stress; for example, cysteine 151 forms adducts with electrophiles that compromise the microenvironment and reactivity of C151. These results suggest that cysteine reactivity may be altered in *KEAP1* in cancer. Analysis of TCGA data reveals that *KEAP1* mutations cluster, and these clusters could point toward important regions of *KEAP1*. *KEAP1* mutations are hypomorphic and can be further inactivated by interacting proteins. Overexpression of the ETGE-containing protein, DPP3, in lung squamous cell carcinomas further activates *NRF2* signaling in a *KEAP1* mutant background. In conclusion, Ms. Hast noted that mutations in *KEAP1* in lung squamous cell carcinoma can be grouped into four phenotypic classes. The “superbinder” class exhibits enhanced *NRF2* activity and stability and likely results from structural changes in the *KEAP1* homodimer. *KEAP1* mutations in cancer cluster around cysteines with reactivity to electrophilic compounds, and overexpression of ETGE-containing proteins can further activate *NRF2* activity in a *KEAP1* mutant background.

Dr. Harold Varmus then briefly took the podium to express his admiration for TCGA as a paradigm of team effort to create a program that sets a high standard for examining genomes and using the subsequent information to understand cancer at a molecular level.

Keynote Talk: The Genetic Basis for Cancer Therapeutics—The Opportunity and the Obstacles

William Sellers, M.D.; Novartis Institutes for BioMedical Research, Inc.

Dr. Sellers began by observing that oncogenes and tumor suppressors represent an outstanding model for understanding and treating cancer. In the previous twenty years, a paradigm shift has taken place, beginning with the development of the targeted agent, imatinib, for the treatment of CML. However, translation from discovery to clinic has been historically slow. In 1960, the first somatic lesion in cancer was discovered. Fifty years later, imatinib, which inhibits the output of this oncogene, was developed. Since the introduction of imatinib in 2001, CML mortality has been greatly reduced. As a result of imatinib’s success, a set of second-generation *ABL* inhibitors has reached the market. Critics initially voiced concern that the cancer stem cell population would progress and that imatinib would be a temporary cure. However, the ratio of mortality per prevalent case continues to decline suggesting that an inexorable progression of a stem cell population is not inevitable. Dr. Sellers then asked whether this paradigm could translate from a relatively simple genetic disease to a more complex solid tumor.

He noted that the imatinib paradigm is translatable, as mutations in the kinase domain of *EGFR* can be treated by inhibitors in solid tumors. Thus drugs can be developed by either an empirical approach in unselected populations or increasingly by using an understanding of pathogenic mechanisms to drive patient selection. To facilitate the latter approach, five key hurdles were discussed. First, genetic-based therapy of cancer requires that the complete compendium of somatic genetic alterations in cancer is known for all cancer types, subtypes, and cancer stages. This catalog must be sufficiently complete to define functionally redundant, cooperating, and antagonistic events. Second, most known genetic alterations have not led directly to drug candidates. Most oncogenes are not druggable, and known entities such as *RAS*, *ERG1/ETV1*, and *BCL2* have been refractory to drug discovery. Identifying viable cancer treatments will also involve understanding tumor suppressor pathways and the effects of a gene's absence. For example, germline mutation of *PTCH* leads to the development of hereditary basal cell carcinoma, and acquired mutations in *PTCH* have been identified in medulloblastoma and in sporadic basal cell carcinoma. Such discoveries can inform efforts to design drugs that target tumor suppressor pathways. For example, LDE225 is a CNS-penetrating SMO antagonist that leads to tumor regression for up to 28 days from the initiation of treatment in an established *PTCH*-deficient model of medulloblastoma. Recent work established a medulloblastoma gene expression signature for the Hedgehog activated subtype using 40 FFPE tissues; five genes have been selected for clinical evaluation. To date, five of five patients with the pathway signature have responded to the Smoothened antagonist LDE225, while 16 patients with signature negative cancers have not responded. This example, illustrates the concept of synthetic lethality as applied to cancer. In this framework, a somatic mutation in the cancer enhances its viability, yet at the same time the combination the mutation and the drug yields a selective therapeutic response.

Efforts to further identify synthetic lethal relationships have explored common cancer-related pathways, such as PI3K. In glioma, *PTEN* loss leads to PIK3Cβ dependence. This discovery has informed the development of BKM120, a pan-Type 1 PI3K inhibitor that can penetrate the blood-brain barrier that is currently in clinical trials. shRNA screening can support efforts to prospectively identify instances of synthetic lethality. For instance, a pooled shRNA screen identifies the β-catenin dependence of the WNT pathway, which is commonly mutated in cancer cell lines.

The third problem facing the developing of genetic-based therapeutics is that tumors develop resistance to targeted agents, suggesting that successful approaches will require combinations of agents. For example, tumors resistant to *BCR-ABL* inhibition contain mutations that lead to insensitivity/resistance to imatinib—these cancers must mutate *ABL* to survive. Dr. Sellers noted that more potent kinase inhibition increases the rate of response. In some cases, improved or enhanced targeted inhibition can overcome resistance. For example, *ALK*-driven lung cancers show a high response rate to crizotinib, but resistance develops rapidly, suggesting that a more potent *ALK* inhibitor should work in crizotinib-refractory settings. Dr. Sellers noted that targeting key oncogenes with potent inhibitors is the key to preventing resistance. In the *BRAF* setting, efforts have sought to understand the mechanisms that bypass target dependence. Emerging data here suggests that combinations of inhibitors targeting different nodes in the RAS pathway may be a way to prevent the emergence of resistance.

These emerging data highlight the fourth problem, namely the need to identify highly active combinations directed at genetic subtypes of cancer. Experiments are underway to test the capabilities of large-scale systematic screening to explore the space of possible combinations.

Finally, the lack of a robust pre-clinical translational infrastructure has limited the ability to preclinically profile the same number of patients that will be treated clinically. To support efforts to explore the pre-clinical therapeutic efficacy Novartis and the Broad Institute have created a so-called Cancer Cell Line Encyclopedia (CCLE). These resources aim to support efforts to enroll only the most appropriate patients into a clinical trial such as selecting patients with appropriate PIK3CA mutations in a trial of the PIK3CA inhibitor, BYL719. However, studies using cell lines do not relinquish the need for primary tumor models, and Novartis has developed a set of 410 primary human tumors that can be propagated in immunodeficient mice and used as an in vivo model systems.

In summary, Dr. Sellers noted that the cancer genome must be mapped in depth, and efforts must include validated but difficult-to-drug targets and synthetic lethal drug targets. Resistance should be studied preclinically, and novel highly-active combinations must be identified and tested. Finally, a robust preclinical infrastructure should be built to support these efforts.

Discussion:

One participant inquired about using assays with one gene per well to support synthetic lethality investigations. Dr. Sellers replied that Novartis has moved toward pooled screens due to transfection heterogeneity at a large scale. He noted that recurrence specimens are being accrued for TCGA efforts to compare primary versus metastatic cancers. TCGA collaborates with the Broad Institute and Novartis on the CCLE, which is expected to become available through CGHub in 2013. Although it is not known how many cell lines will be required to represent human cancer, efforts are underway to convert primary human tumors into cell lines. Cells may be manipulated with exonucleases to create isogenic cell lines, although the use of cancer cells for this purpose may generate cell lines that are not isogenic.

Another participant asked how to develop a therapeutic armamentarium for a future in which cancer is approached as a chronic disease. Dr. Sellers replied that ideally cancer therapy should not be thought of as such, because ultimately, curative therapy will likely require patients' doses of medications that may have significant side effects. Targets must be inhibited potently, with the recognition that side effects will be present (although these can be moderated). Efforts must aim to cure disease completely rather than to administer sub-curative and sub-therapeutic doses. Another attendee asked if the pharmaceutical industry supports sequencing studies at the time of disease progression, to which Dr. Sellers replied that Novartis is interested in building an infrastructure around such clinical samples. One participant asked if mutations that drive resistance exist in a small fraction of cells from the beginning or whether they arise during progression. Dr. Sellers replied that this is not known. However, efforts must aim to pressure a tumor at several points so that no one or two mutations are sufficient to overcome the therapy. He noted that a druggable driver mutation identified in 20% of cancer cells from one patient would be insufficient to merit further resources, as efforts should identify the earliest set of mutations persistently required for disease.

Closing Remarks

Matthew Meyerson, M.D.; Dana-Farber Cancer Institute and Ilya Shmulevich, Ph.D.; Institute for Systems Biology

Dr. Meyerson thanked all contributors, noting that TCGA investigators have set aside personal interests to advance the project for the community, NCI and NHGRI leadership, and for cancer patients who have shared their tissues for TCGA studies.

The meeting was then adjourned.