Meeting Summary

The Cancer Genome Atlas's (TCGA) 3rd Annual Scientific Symposium: Enabling Cancer Research through TCGA

Natcher Conference Center, NIH
Bethesda, MD

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

Table of Contents

Monday, May 12

Opening Remarks
Matthew Meyerson, M.D.; Dana-Farber Cancer Institute and Marco Marra, Ph.D.; British Columbia Cancer Agency

Dr. Meyerson welcomed attendees and thanked participants for their enthusiasm for the symposium. He noted that The Cancer Genome Atlas (TCGA) owes its success to a number of parties, including NCI and NHGRI leadership, the network of investigators who have participated in TCGA studies, and the vast network of physician collaborators and patients who have shared their specimens. He reported that TCGA's Biospecimen Core Resource (BCR) recently shipped its 10,000th specimen, and the initiative has produced numerous marker papers. Currently, manuscripts on lung and gastric adenocarcinoma are in press at *Nature*, and thyroid carcinoma and kidney chromophobe manuscripts are currently under review. Dr. Marra thanked members of the cancer community who have promoted the TCGA initiative, noting that the power of the data and the discoveries that they enable are just beginning to be realized.

Keynote Address: Big Data, the Community, and the Commons
Robert Grossman, Ph.D.; The University of Chicago

Dr. Grossman began by stating that issues relating to large data volumes can be conceptualized around four questions:

1. What are the similarities and differences between "big" biomedical data, "big" science data, and "big" commercial data?
2. What instrument should be used to make discoveries over big biomedical data?
3. Do we need new types of mathematical and statistical models for big biomedical data?
4. How should large biomedical datasets be organized to maximize discoveries and their impact on health care?

He noted that, while bioinformatics funding has remained constant, data volumes have increased exponentially. While TCGA investigators have analyzed nearly 10,000 genomes, what would be required to sequence one million genomes? Sequencing of one million genomes would likely change the understanding of genetic variation. The genomic data (including tumor and normal specimens) from a single patient is approximately one terabyte. Thus, one million genomes would produce approximately 1,000 petabtyes or one Exabyte of data. With compression, these data could perhaps be reduced to approximately 100 petabytes. At a cost of $1,000 per genome, the sequencing would cost approximately $1 billion dollars, which result, for example, from one hundred studies with 10,000 individuals each completed over three years.

Dr. Grossman stated that big data are disrupting the standard model of biomedical computing, which features public repositories, community software, and bioinformaticists who create local data. Under this structure, it may require several weeks to download one petabyte of data; three weeks would be required to download all TCGA data at a rate of ten gigabytes/second. Moreover, new data types continue to emerge, and there is a paucity of bioinformaticians relative to the amount of data being generated. For example, the smart phone is becoming a home for medical and environmental sensors, which will create a wealth of new data.

He noted that the field of computational advertising identifies the best match between a user in a given context (e.g., web-search, webpage content, social media and mobile contexts) and suitable advertisements. A modern advertising analytic platform will construct behavioral profiles on more than 100 million individuals, and refresh each of these models each day. These calculations are carried out at machine speed by operators with minimal training. Dr. Grossman noted that the biomedical community could borrow strategy and technology from this approach.

The new model of biomedical computing would include large scale biomedical data commons and biomedical clouds that interoperate with public data repositories, community software, community compute and storage resources, and private clouds at medical research centers, companies and other organizations In this model, large and interoperable biomedical data commons would store biomedical data and provide transparent access to data for those researchers authorized to access it. Cloud computing and automation would be used to manage these commons and to compute over it. He noted that this approach is just beginning to be realized.

During the last decade, the focus has been on creating and integrating bioinformatics tools, not on creating large scale bioinformatics infrastructures that is accessible to the research community.

He suggested viewing a data center as the “instrument” for computing with big data, broadly analogous to how a microscope is the instrument used for viewing small objects and the telescope is the instrument used for viewing distant objects. The challenge is developing software that scales to a data center and creating the automation required to operate software and hardware at this scale. However, setting up and operating large-scale, efficient, secure, and compliant racks of computing infrastructure at this scale is beyond the reach of most laboratories, even though it is essential for the community as a whole. He suggested that the biomedical informatics community develop smaller scale specialized data centers, say at the scale of 0.5-3 MW, versus the 15-25 MW size data center that a commercial cloud service provider (CSP) might use.

Many commercial CSPs, such as Amazon Web Services (AWS), are also available for use by the biomedical research community. CSPs have the advantages of scaling, offering a wide variety of services, and are available to anyone with a credit card. By contrast, community-based science and biomedical clouds can operate at lower cost (when operated at a large enough scale), can offer specialized computing infrastructure optimized for scientific workflows,, and can provide specialized security and compliance. These resources also support data that are too important to trust exclusively with a commercial provider. However, it is still critical to interoperate with CSPs whenever possible. However, when the amount of data grows beyond several petabytes, a medium-scale private/community cloud can be much less expensive than a commercial CSP.

He pointed out that we have recently entered an era of data center scale science (e.g., Bionimbus, CGHub, GenomeBridge, Cancer Collaboratory, etc.) and this technology will mature over the next decade. The big data genomics community has a large amount of data in aggregate but unlike disciplines, such as the big data physics and astronomy, it is produced by many small to medium size instruments, not one large instrument.

Dr. Grossman noted that Science CSPs must interoperate with commercial CSPs, although the two communities have different perspectives, requirements, and business models. The key question to ask is whether biomedical computing at the scale of a data center is sufficiently important for the research community or whether the community should exclusively outsource to commercial CSPs. The Open Science Data Cloud Consortium shows proof-of-principle of the model of biomedical computing at the scale of a data center.

Dr. Grossman then asked about the gains that could achieved when data are analyzed at the scale of the data center--do new phenomena emerge at scale in biomedical data or do we just see the same phenomena? In other words, is "more (data) different"?

When building models over big data, ensembles of models allowed models to scale to computer clusters. On the other hand, simple ensembles do not scale without change to the scale of a large data center. The challenge with data at this scale is often devising ways to decompose large and heterogeneous datasets into homogenous components that can be modeled. Drawing on the computational advertising paradigm, Dr. Grossman asked how research and healthcare would be impacted if we could analyze all of the biomedical data each evening.

He noted that we once we can successfully do science at the scale of a data center, the challenge is to develop new modeling techniques that can explore whether "more is different" at this scale.

Several genome clouds currently exist (e.g. the Bionimbus Protected Data Cloud), and TCGA data can be analyzed using them but it is not practical for every research organization to build their own. A "Cloud Condo" model is an organizational model in which several research organizations join together to build a common large scale cloud that provides each organization their own private cloud.

He noted that there is a societal benefit when biomedical data are also available in data commons operated by NFPs versus sold exclusively as data products by commercial entities or only offered for download by the USG. Large data commons should peer with each other and develop standards for interoperating, although these standards should not be developed ahead of open-source reference implementations. While it is expected that a period of experimentation will be necessary to develop optimal technology and practices, challenges remain when analyzing modeling data at scale and incorporating novel data types at scale.

One participant asked whether a commons model will support iterative method development, as is common in bioinformatics. Dr. Grossman replied that cloud computing in conjunction with data commons can easily support iterative method development.

In response to another question, he noted that, in many cases, models can often be improved more easily by doubling the amount of data as compared to doubling the complexity of the model.

Session I

Chair: Han Liang, M.D., Ph.D.; The University of Texas MD Anderson Cancer Center

Comprehensive and Integrative Genomic Characterization of Diffuse Lower Grade Gliomas

Daniel J. Brat, M.D., Ph.D.; Emory University School of Medicine

Dr. Brat began by noting that lower-grade gliomas (LGGs) comprise a family of infiltrative neoplasms. The World Health Organization (WHO) classifies LGGs into astrocytomas, oligodendrogliomas, and mixed oligoastrocytomas. Survival among these classifications varies greatly, with astrocytic-lineage tumors ultimately proceeding to glioblastoma. Astrocytomas of Grades II and III, which will progress to glioblastoma, are characterized by tumor cell infiltration and mutations in *IDH*, *TP53*, and *ATRX*. Oligodendrogliomas are characterized by 1p/19q co-deletion and mutations in *IDH*, *CIC*, *FUBP1*, and *TERT* promoter genes. These chemosensitive tumors are associated with better prognosis than astrocytomas. However, a large group of tumors contain multiple cell types and are morphologically ambiguous. The current LGG classification scheme uses a histogenic classification developed in 1926, which recognizes the difficulty of classifying these tumors. As such, identification of a reproducible and clinically meaningful molecular classifier will greatly enhance the understanding and treatment of these cancers. Dr. Brat noted that TCGA aims to identify 500 lower-grade gliomas, analyzed at different Characterization Centers for somatic mutations, DNA copy number (CN), mRNA expression (including fusions), DNA methylation, microRNA expression, protein levels and phosphorylation, and DNA copy-number/rearrangements. At the data freeze, 293 cases had been accrued, 254 of which featured overlapping data from major platforms. The initial manuscript for TCGA LGG analyses is expected to be submitted within the month. MutSigCV analysis identified significantly mutated genes and revealed mutation classes. *IDH* mutations occur in approximately 80% of lower-grade gliomas. Two classes of mutations are observed within this group: 1) tumors with *TP53*/*ATRX* mutations (mostly astrocytomas and oligoastrocytomas), and 2) tumors with *CIC*, *FUBP1*, *Notch1*, and *PIK3CA* mutations (mostly oligodendrogliomas). Wild-type tumors do not have these mutations but instead commonly feature mutations typical of glioblastomas, such as *EGFR*, *PIK3*, and *PTEN*. CN alterations by histology show characteristic 1p/19q losses in oligodendrocytomas. Unsupervised clustering shows tight clustering of specific gains and losses, such as chromosome 7 gain and chromosome 10 loss in *IDH* wild-type LGG. Using the OncoSign unsupervised clustering algorithm, three subgroups of LGGs were observed, based largely on *IDH* mutations and 1p/19q status. DNA methylation analysis showed five distinct subgroups. mRNA expression clustering yielded four subsets that correlate with *IDH* status and 1p/19q co-deletions. Clustering of molecular data (CN, mRNA, miRNA, methylation) identified 3-5 subtypes. “Clustering of clusters” analysis identifies three robust, non-overlapping LGG molecular classes, largely based on *IDH* and 1p/19q status. These molecular findings do not support a “mixed histology” classification. Tumor types also separate in terms of clinical outcomes. *IDH* wild-type LGGs have mutation frequencies, oncogenic gene fusions, and clinical outcomes similar to those observed in glioblastoma. Revere-phase protein array (RPPA) analysis shows upregulated proteins in the *IDH* wild-type group that could represent therapeutic targets. In summary, Dr. Brat noted that six histopathologic diagnoses can be distilled into three robust, clinically relevant molecular classes of LGGs. *IDH*-mutant, 1p/19q co-deleted gliomas feature mutations in *CIC*, *FUBP1*, *TERT* promoter, *Notch 1*, and *PIK3CA*. *IDH*-mutant, non-co-deleted gliomas are characterized by mutations in *TP53*, *ATRX*, and 8q24. *IDH* wild-type LGG have molecular alterations and clinical behavior similar to that observed in glioblastoma.

One participant asked about the major differences between *IDH* wild-type LGGs and glioblastomas. Dr. Brat replied that *IDH* wild-type LGGs have a slightly better survival, although the tumors are similar in terms of CN, mutation profiles, and epigenomic profiles.

Using TCGA Data to Inform on Precision Medicine in Late-Stage Cancer Settings
Andrew J. Mungall, Ph.D.; British Columbia Cancer Agency

Dr. Mungall began by noting that the British Columbia Cancer Agency (BCCA) offers a provincial, population-wide cancer control program that includes prevention, screening, diagnosis, and treatment. Within this setting a personalized oncogenomics (POG) program has been established that aims to bridge the divide between genomics research and clinical practice by identifying tumor-specific therapeutic targets in individuals with late-stage metastatic disease. Specimens collected include tumor biopsies, archival tumors, and peripheral blood. To date, this study has enrolled 83 individuals representing 28 tumor types. The average time from biopsy to reporting to clinicians is 38 days. The POG guides treatment decision-making by providing directed cytotoxic chemotherapy choices and targeted therapeutic options, providing data to complement or correct clinical tests, informing diagnosis, and identifying primary tumor sites. For example, targeted therapeutic options were provided for an individual with squamous cell carcinoma lesions on his chest that metastasized to his preauricular (ear) node. Biopsies of the chest and ear lesions identified single nucleotide variants (SNVs) and small insertions/deletions (in/dels) that rarely overlapped. Copy number profiles for the two tumor sites showed that the sites were essentially molecularly distinct, with few common breakpoints. Pathway analyses of integrated genome and transcriptome data showed numerous rearrangements with DNA repair defects in the chest lesion. These analyses suggested treatment options with erlotinib in the ear and everolimus for the chest lesion. Despite initial clinical response the preauricular tumor progressed nonetheless, with further *EGFR* amplification and overexpression evident from a re-biopsy. Another case study featured an individual diagnosed with non-small cell lung carcinoma (NSCLC) for whom an *EML4-ALK* fusion was identified from transcriptome data that was previously reported as negative in an approved clinical assay for this rearrangement. As a result of this oncogenic fusion, the patient highly expressed *ALK*. *ROS1* was also highly expressed relative to TCGA lung cancer tumours. Crizotinib was thus proposed, and the tumor responded dramatically within several months. In summary, for each POG patient, three or more genomes (e.g., normal, archival, and tumor) and one tumor transcriptome have been sequenced. To date, of 82 consented patients with advanced cancer, 74 biopsies were attempted. Full data sets were available for 50 of these individuals, and these data were clinically evaluated in 38 individuals. The POG was informative or actionable for treating 33 of the 38 patients; treatment was available and offered for 18 of the 33 individuals. Future POG studies include a Phase II trial with 5,000 individuals over the span of five years. This will include a rapid “oncopanel” to provide an initial assessment of each case. The emphasis in Phase II will expand beyond end-stage individuals.

One attendee asked about plans to make the data public, of which there are no immediate plans due to confidentiality issues. Tumor heterogeneity remains an issue with individual-based analyses; this study required a tumor content of 40% before sequencing was initiated.

The ICGC-TCGA DREAM Somatic Mutation Calling Challenge: Initial Results
Paul C. Boutros, Ph.D.; Ontario Institute for Cancer Research

Dr. Boutros began by noting that the overall goal of genomic research is translation to the clinic. Large-scale sequencing efforts have generated numerous genomic profiles, which are often

analyzed by algorithms whose comparability and reliability have not been established. The ways that data are preprocessed can greatly affect the conclusions drawn from their analysis; subtle differences in the analysis of a single data set (e.g., due to versioning differences in software packages) can affect conclusions dramatically. Dr. Boutros noted that this observation holds true for all tumor types. To address the output from different algorithms, SAGE Bionetworks, the University of California Santa Cruz, and the Ontario Institute for Cancer Research have devised a Challenge to determine the best whole genome sequencing (WGS) analysis methods, focusing on accuracy rather than speed, computational efficiency, or other considerations. This Challenge incorporates ten tumor/normal pairs representing tumor types that span a wide range of cellularity. For these specimens, raw and processed data are available, along with complete clinical protocols. To make this data set publicly available, a template was developed to expedite ethics committee approvals. As the Challenge was designed further, it was noted that a simulated-data component would be useful to encourage researchers beyond the cancer community. This *in silico* component was composed using a genome from a cell line or germline that had SNVs and SVs "burned in" using BAMSurgeon. A subset of reads was then introduced with additional SNVs and SVs to create a "tumor/normal" pair. Five releases of these data will be made available, with increasing complexity. These data may be accessed by registering for the Challenge at Synapse, downloading the data using GeneTorrent, or accessing the data directly in the Google Compute Engine cloud application.

The Challenge assesses SVs and SNVs in the human tumor data for balanced accuracy across the ten tumor/normal pairs. For the simulated tumor data, each individual tumor is assessed for SNVs and SVs. The Challenge is scored in several ways. For the ten real tumor/normal pairs, several thousand candidates will be validated using a minimum resequencing to approximately 300X coverage using AmpliSeq primers on an IonTorrent platform. For *in silico* data from the five synthetic tumor/normal pairs, a complete ground-truth is known for each dataset. The Challenge will calculate the sensitivity, specificity, and balanced-accuracy for each genome on a held-out piece of the genome. Final results from the Challenge are expected in November 2014. To date, 440 entries have been logged, with many users providing ongoing post-challenge submissions as they improve their calling algorithms. It is hoped that these continuing efforts will generate a "living" benchmark and improve the ability to simulate reads. Dr. Boutros noted that entries broadly reflect a single receiver operating characteristic (ROC) curve. There is also a surprisingly large chromosome bias; some chromosomes have very low calling rates for unknown reasons. Determinants of false positives included variant allele frequency, mapping ability, normal coverage, and tumor coverage. Determinants of false negatives included mapping quality, normal coverage, tumor coverage, and variant allele frequency. Efforts to date also reveal surprisingly strong trinucleotide effects, although coding regions appear to have substantially lower error rates than the rest of the genome. Parameterization was critical to users' success at improving calling. In summary, Dr. Boutros noted that Challenge results to date reveal surprising trends in error profiles, including chromosomal bias and trinucleotide bias. The best method for predicting SNVs is MuTect, whereas the best methods for predicting SVs are Delly (tumor data) and novoBreak (*in silico* data). The Challenge has established a community for rapid development and refinement of algorithms and benchmarking for next-generation sequencing platforms. The Challenge has also spurred users to improve their algorithms and led to improvements in tumor-read simulations. Dr. Boutros noted that calls generated by different callers will be made publicly available as part of an effort to make results complementary.

Multiple calls submitted by a single user are taken into account when assessing performance. Although several teams have overfit, all teams are encouraged to improve their algorithms.

Lessons Learned for the Genomic Characterization of Patient-Matched Frozen and Formalin-Fixed, Paraffin-Embedded Tissues: Progress Update
Erik Zmuda, Ph.D.; Nationwide Children's Hospital

Dr. Zmuda noted that massively parallel sequencing has enabled major advancements in the understanding of tumor biology, including the identification of novel drivers of tumor progression, elucidation of new therapeutic targets, and the development of a molecular-based cancer taxonomy. However, many seminal studies have drawn from and were optimized for frozen tissues. The concept of precision medicine involves the application of these advances to the clinic. One challenge to this translation is that available diagnostic specimens are primarily formalin-fixed, paraffin-embedded (FFPE) tissues, and FFPE fixation is known to introduce molecular artifacts. The goals for TCGA FFPE Pilot Study include identifying and optimizing the best practices for extracting, characterizing, and analyzing FFPE specimens, defining the patterns of artifactual alterations induced by formalin fixation and paraffin embedding (e.g., a "molecular signature" of FFPE), bridging the gap to diagnostic material, and facilitating the application of the merging cancer taxonomy to clinical testing environments. This study aims to develop procedures to co-isolate DNA and RNA from FFPE tissues optimally to maximize yield and integrity with consistent characterization using platforms beyond those used in the study. A survey of commercial FFPE extraction methods indicates that no single method provides optimal DNA and RNA integrity. However, by customizing these methods, an optimized co-isolation method has been developed. The technique uses scrolls rather than cores, and lysis generates a supernatant from which DNA and RNA are extracted. This method was assessed using FFPE-preserved tumor material and paired pathology-matched frozen specimens gathered from 38 qualified TCGA individuals spanning six tumor types. The average age of these tumor blocks was three years.

Paired fresh-frozen (FF) and FFPE-derived analytes were then distributed for characterization on five genomic platforms (exome sequencing, Broad SNP6 array, USC methylation, BCCA miRNA-seq, and UNC mRNA-seq). Perhaps unsurprisingly, none of the FFPE SNP arrays passed quality control (QC), in part due to a highly oversegmented CN profile. FFPE can validate the CN profile in fresh-frozen tissues, although segmentation artifacts resulted in a high false discovery rate. Exome sequencing results indicated that the overall mutation spectrum in lung adenocarcinoma reveals a shift towards C>T transitions in FFPE tissues. Binning by allele fraction illustrates that FFPE effects are limited to low-allele fraction components. These results support the use of FFPE tissues for exome sequencing, although additional tools are needed to compensate for low-allele fraction C>T SNV artifacts, which are consistent with effects of deamination caused by formalin fixation. mRNA sequencing using a pairwise Pearson correlation of transcript quantification between FF and FFPE tissues showed robust correlations between the results generated. Unsupervised clustering showed that the effect of FFPE varies among studies. Isolating differences between FF and FFPE revealed consistent trends in quantification. Overall, FF and FFPE expression signatures were highly concordant, although additional bioinformatics steps may be required to adjust for differences in the level of expression detected in FFPE samples. miRNA sequencing results suggested that FFPE has a weak effect on miRNA characterization. However, additional effort is needed to gain insight into

the cause/effect of increased diversity in miRNA species observed in FFPE tissues. DNA methylation analysis showed minimal effects from FFPE. An excellent concordance in methylation signatures was observed between FF and FFPE tissues, although an Illumina FFPE restoration protocol was required to achieve these results. Future efforts will include analyzing the FFPE signature in the context of multi-center calling, delineating the influence of tumor heterogeneity in the results of this study, and conducting a deeper analysis of the differences between FF and FFPE to identify potential bioinformatics mechanisms to correct for the artifacts caused by formalin fixation and paraffin embedding.

Domain-Specific PIK3CA Mutations Affect Different Pathway Activities across More than 3,000 TCGA Pan-Cancer-12 Tumors

Christopher C. Benz, M.D.; Buck Institute for Research on Aging

Dr. Benz began by noting that PI3 kinases are involved in proliferation, survival, growth, and metabolism. *PIK3CA* is mutated frequently in multiple cancer types, with hotspot mutations that are early and possibly initiating events in several cancers. Preclinical evidence in model systems indicates that *PIK3CA* mutations are activating and gain-of-function, but domain-specific mutations may have different AKT-activating properties and phenotypic consequences. TCGA's Pan-Can-12 dataset is sufficiently large to offer an opportunity to inquire about possible pathway differences linked to domain-specific *PIK3CA* mutations across different tumor types. These analyses used the PARADIGM algorithm to integrate RNA gene expression data and DNA CN data onto a superimposed pathway structure to infer the integrated activity of about thirteen thousand different pathway features. The TCGA Pan-Can-12 data set includes 3,531 specimens with PARADIGM-calculated integrated pathway levels (IPLs) and 3,277 specimens with exon sequencing data. 2,637 of the specimens featured both IPL and *PIK3CA* mutation data. Somatic *PIK3CA* mutations have an overall mutation frequency of 22% across the Pan-Can-12 dataset, although the frequency varies from 0% to greater than 50% in some cancer types. 447 cases (spanning eleven cancer types) had single-domain missense *PIK3CA* mutations. 83% of the cases featured either *helical* or *kinase* domain missense mutations, with a significantly varied domain distribution of mutations by cancer type. Visualization of these IPLs was carried out using Cytoscape. Results were consistent with preclinical evidence in that kinase domain mutations associated most strongly and positively with a superpathway hub showing PI3K catalytic subunit activation, while this hub was negatively associated with *helical* domain mutations. Cell cycle and proliferation activities (e.g., PKL1, FOXM1) appeared most strongly associated with *kinase* domain mutations and negatively associated with *helical* domain mutations. In contrast, cell motility and dissociation activities (e.g., RHO GTPases, Gap junction degradation) appeared enriched and strongly associated with *helical* domain mutations, and negatively associated with *kinase* domain mutations. In conclusion, Dr. Benz noted that missense *PIK3CA* mutations distribute quite differently with respect to their overall mutation frequency and domain specificity, although these mutations are common across cancer types represented in the TCGA Pan-Can-12 dataset. *Kinase* domain mutations appear to be linked more strongly with pathway features that enable cell proliferation, whereas *helical* domain mutations are more strongly linked with features that enable cell migration and dissemination. Functional studies are needed to confirm these findings, including the suggestion that breast cancers preferentially mutate *PIK3CA* to drive cell proliferation while lung and head and neck squamous cancers prefer *helical* domain mutations to drive their malignant cell motility. Additional comparisons will identify potential tumor type-specific differences between *kinase* and *helical* domain pathway

preferences. Dr. Benz noted that these studies did not examine or adjust for mutations in the p85α regulatory subunit and did not control for concurrent mutations (e.g. PTEN) also impacting the downstream consequences of this pathway.

Comprehensive Molecular Profiling of Adrenocortical Carcinoma
Siyuan Zheng, Ph.D.; The University of Texas MD Anderson Cancer Center

Dr. Zheng began by noting that adrenal cortical cancer (ACC) is relatively rare, with an annual incidence between 0.5-2 cases/1,000,000 individuals. The five-year survival for individuals with metastasis is less than 20%. Moreover, no standard staging system has been developed for this cancer type. ACC is an endocrine tumor; more than half of afflicted individuals also have hormonal excess. Only one drug treatment (mitotane) is FDA-approved, and targeted therapy has so far proved disappointing. TCGA collected 92 ACC cases for genomic characterization, with RPPA forthcoming for 48 of these cases. Analyses to date reveal that ACC specimens show relatively high tumor purity compared to other tumor types. A subset of ACC tumors may be related to infiltrative leukocytes, although the mutation frequency of ACC is relatively low compared to other tumor types. Analysis of 91 ACC exomes identified five significantly mutated genes (e.g., *TP53*, *CTNNB1*, *MEN1*, *PRKAR1A*, and *RPL22*). The mutation pattern suggests gene function in ACC; four of the five genes are inactivating mutations. *PRKAR1A* is a binding partner of PRKACA, and this complex induces cortisol production and proliferation. CN analysis identified significantly amplified and deleted DNA segments that include important genes, such as *TERT*, *ZNRF3*, and *TERF2*. However, one subset of ACCs (n=30) harbored no putative driver mutations, referred to as the dark matter group. These results correlate with other published findings. Methylation clustering identified three groups (normal-like, CIMP, and hypermethylated) based on methylation levels. miRNA clustering analysis revealed six clusters, and gene expression clustering identified two groups that recapitulate published analyses. A significant correlation was found between the dark matter group and the normal like subtype, indicating that the genomic landscape of ACC is strongly reflected by subtypes. RNA-seq data identified putative, albeit sporadic, cancer-related gene fusions (e.g., *EXOSC10-MTOR*, *PPP1CB-BRE*). The WNT pathway was the most frequently altered pathway in these specimens. Overall, approximately 45% of ACC patients harbored at least one altered gene in the WNT pathway. In summary, Dr. Zheng noted that these analyses identified potential new ACC driver genes, including *ZNRF3*, *TERT*, *TERF2*, and *PRKAR1A*. Integrative analysis indicates that approximately 30% of ACCs harbor no putative driver mutations. Infrequent alterations, such as gene fusions, may contribute to adrenal tumorigenesis. The WNT pathway is the most altered pathway in ACC, mostly resulting from *ZNRF3* deletion and *CTNNB1*-activating mutations. Dr. Zheng noted that WGS has not been carried out on these tumors. Known driver genes are infrequent, possibly reflecting the limitations of the cohort size. Additional studies will analyze germline data.

Session II
Chair: Josh Stuart, Ph.D.; University of California, Santa Cruz

Comprehensive Molecular Characterization of Chromophobe Renal Cell Carcinoma
Chad Creighton, Ph.D.; Baylor College of Medicine

Dr. Creighton began by noting that chromophobe renal cell carcinoma (ChRCC) represents approximately 5% of cancers that arise from the kidney nephron. Due in part to this relative rarity, this disease has been understudied at the molecular level. ChRCC has been comprehensively characterized by TCGA as the first of its set of Rare Tumor Projects. These analyses include 66 cases subjected to comprehensive TCGA profiling, 50 of which feature WGS and 61 of which have mitochondrial DNA sequencing. Somatic alteration patterns in ChRCC tumors differ from clear-cell renal carcinomas (CCRCCs). Somatically-mutated genes identified by WES included *TP53*, *PTEN*, *FLT4*, and many others that were mutated in one or two cases. DNA methylation revealed widespread differences between CCRCC and ChRCC, possibly reflecting the cells of origin of the two tumor types. Published orthogonal data (Cheval L, et.al. *PLoS One* 2012;7:e46876) suggest that gene expression patterns in ChRCC versus CCRCC reflect the distal versus proximal nephron, respectively. Mitochondrial DNA (mtDNA) analyses show that mtDNA mutations involve the electron transport chain; ChRCC relies upon oxidative phosphorylation. WGS identified kataegis in some ChRCC cases, usually in the vicinity of genomic rearrangements. Twenty-one genes (including *TERT*, *TBX22*, and others) were associated with kataegis. Structural breakpoints that were identified were associated with the *TERT* promoter region; these variants correlated with high levels of *TERT* expression. *TERT* promoter-associated structural variants were validated. Of the seven rearrangements identified by WGS, six were confirmed by PCR. In conclusion, Dr. Creighton stated that comprehensive molecular analysis of a rare cancer type can be used as a platform for discovery, and global molecular patterns may provide clues as to a cancer's cell of origin. This project incorporated mtDNA sequencing into a multi-platform molecular characterization of a cancer and identified recurrent genomic rearrangements involving the *TERT* promoter region. He noted that specimens used for these analyses were identified only as ChRCC. Some of the tumors accrued by TCGA to support the CCRCC analysis were actually characterized as ChRCC specimens. These specimens were removed from the clear-cell manuscript and excluded from the ChRCC analysis.

Prediction of Individualized Therapeutic Vulnerabilities in Cancer from Genomic Profiles
Bulent Arman Aksoy; Memorial Sloan-Kettering Cancer Center

Mr. Aksoy discussed Project Statius (http://cbio.mskcc.org/cancergenomics/statius/), a computational resource that describes individualized therapeutic vulnerabilities (Aksoy BA, et.al. *Bioinformatics* 2014; DOI:10.1093/bioinformatics/btu164). He noted that metabolic reactions within a cell are regulated by multiple catalytic isoenzymes. In a cancer cell, many factors affect this process, including instability-induced homozygous deletions in which one of the isoenzymes is lost by chance. Given that there is an intact copy of the extant enzyme, the cell can still catalyze the reaction. However, if cells are perturbed by introducing targeted, selective drugs, the remaining isoenzyme can be selectively inhibited. The drug thus exploits this vulnerability and selectively kills the cancer cells, which can no longer catalyze the given reaction. These drugs have reduced toxicity since they have non-catastrophic effects on normal cells. Mr. Aksoy and his colleagues have developed a computational pipeline to establish a systematic vulnerability screen using genomic profiles, metabolic pathways, and targeted drugs (e.g., from Pathway Commons 2, cBioPortal, PiHelper, and other resources) as input. This pipeline creates a list of metabolic vulnerabilities that are individualized and can be tested in cell lines. Analysis of 16 cancer studies identifies a number of vulnerabilities associated with each cancer type or data resource. To date, Project Statius has identified nearly 4,000 vulnerabilities across cancer specimens and cell lines. Statistics and additional information about the vulnerabilities identified

to date can be found at the URL listed previously. To allow users to prioritize these vulnerability results, a confidence score has been assigned to the individualized vulnerabilities based on the supporting evidence associated with each of these vulnerabilities. Mr. Aksoy noted that cancers with the highest rates of vulnerabilities in these experiments tend to correlate with the number of available specimens. However, cancers that have subtypes that are driven by CN alterations (e.g., ovarian, breast) are also more likely to contain vulnerabilities.

Recurrent Epistates Define Tumor Methylome Differences
Huy Q. Dinh, Ph.D.; University of Southern California

Dr. Dinh discussed whole genome bisulfite sequencing (WGBS), a DNA methylation analysis platform that has provided methylation profiles of 28 million CpGs with more sequence variation information than that obtained from microarray-based DNA methylation platforms used for other TCGA projects. To date, WGBS has been completed on 47 TCGA patient samples (nine cancer types) at greater than 15x sequence coverage per sample. Differentially-methylated regions (DMRs) have been identified across multiple tumors based on single-molecule analysis of WGBS. Epipolymorphism (Landan G, et.al. *Nat Genet* 2012;44:1207-1214) was used as a basis for epistates to explain different interpretations of methylation patterns across loci. Expectation maximization (EM) was used to assess the epistate mixture in these experiments. This algorithm can estimate epistate frequencies within individual samples, and epistate and its frequency can be used to identify DMR loci. The group led by Peter Laird and Ben Berman has developed a method to infer epistates from a pool of WGBS short reads in tumor and normal specimens. Epistate frequency can then be used to estimate tumor purity based on methylated epistates found in tumor specimens but not in normal specimens and leukocyte samples. The tumor purity is estimated based on epistate frequency density distribution. The result is highly correlated with estimation using ABSOLUTE method. Besides, the EM method enables epistates to be detected at low frequency. Dr. Dinh noted that sequencing-based DNA methylation approaches offer advantages over array-based platforms in that the former can identify epigenetically-silenced distal elements. In summary, Dr. Dinh noted that a single-molecule epistate method has been developed for analyzing multiple WGBS samples. Applications of epistate analysis include estimation of methylation-based tumor purity and application of DMR analysis to identify epigenetically-silenced regulatory regions especially those at low epistate frequencies. He noted that the EM algorithm must be re-run several times to achieve convergence and to assign significance to a region with presence of distinct epistates.

What Do We Learn from Pan-Cancer Subtyping?
Josh Stuart, Ph.D.; University of California, Santa Cruz

Dr. Stuart described results from Pan-Cancer analysis of twelve tumor types (~3,500 tumors). Once these tumors were contextualized into molecular classifications, analysis aimed to determine whether the cancers segregate along boundaries imposed by tissue of origin or by pathways/molecular features. A TumorMap tool has been developed to display samples that are in similar “zip codes” from other samples. Results indicate that disease-specific subtypes are recapped in TumorMap. Maps were developed for each of six platforms (mRNA, microRNA, protein, DNA CN, DNA methylation, and exome mutations). Each platform produced 8-19 clusters, with DNA methylation providing the greatest number. All subtypes showed a strong correlation with tissue of origin. TumorMap recapitulated subtypes identified by the Working

Group. Single-platform subtypes correlated with tissue of origin. Exome mutation clusters showed the least amount of tissue correlation (approximately 70%; Kandoth C, et.al. *Nature* 2013;502:333-339). Using TumorMap, single platform maps are tissue-driven, with each layout driven by a different data platform. Cluster of cluster assignments (COCA) subtyping defined thirteen subtypes, eleven of which were analyzed. Twelve tissue-of-origin sites translated into eleven COCA subtypes. Mutation frequencies were associated with COCA subtypes; only three genes were present at a frequency greater than 10%. DNA CN analysis according to COCA subtypes showed distinct patterns that define subtypes. HotNet2 was used to identify interconnected mutated networks that revealed subtype and preferred tissue. Dr. Stuart noted that integrated subtypes provide new prognostic information, improving predictive ability over clinical data and tissue type. The bladder cancer cases diverged into three subtypes (bladder-enriched, squamous, and LUAD-enriched islands) on TumorMap. Integrated subtyping of bladder cancers distinguished patient outcomes. However, genomic determinants can define these bladder subclasses, with squamous-like bladder cancers showing significant genomic differences compared to the other types. HER2 and Rab25 proteins were expressed to a greater extent in non-squamous cases, with epithelial-mesenchymal transition (EMT) markers expressed in squamous-like bladder cancers. The Gene Programs algorithm identified 22 sets of genes known to be related functionally across Pan-Can-12 tumors. Gene Programs also recapitulated the integrated subtypes. These tools also revealed that oncogenic *Tp63* forms are more active in squamous tumors versus those with BRCA/basal *TP53* mutations. These results are currently in press at *Cell*. In summary, Dr. Stuart noted that analysis of twelve tumor types using six platforms displayed tissue-of-origin as dominant. Integrated analysis revealed eleven major groups, with some tumor types merging (e.g., HNSCC, lung squamous, some bladder cancers) and others separating (e.g., breast luminal versus basal-like). Intriguing subtype-specific differences in TP53 pathway activity were observed between ovarian, BRCA/basal breast, and squamous bladder tumors. Classification adds prognostic information that is independent of tissue and stage; COCA clusters defined clear prognostic groups for bladder cancer. Additional information, especially that generated from approaches that subtract tissue-of-origin signals, will be useful to these analyses. Dr. Stuart noted that principal component analysis was used to assess batch effects. Some batch effects were observed when mRNA platforms were changed in the course of these analyses.

Profiling Long Intergenic Non-Coding RNA Interactions in the Cancer Genome
Samir B. Amin, M.D.; The University of Texas MD Anderson Cancer Center

Dr. Amin began by noting that these efforts primarily aim to 1) catalog the expression levels of long intergenic non-coding RNA (lincRNAs) across TCGA cancer types, extending the initial profiling efforts from investigators at MSKCC and MD Anderson (Akrami R, et.al. *PLoS One* 2013; Han L, et.al. *Nat Commun* 2014) and 2) use integrative analysis to understand the emerging gene-regulatory role of lincRNAs in cancer progression. LincRNAs are greater than 200 base pairs RNA with no protein-coding potential. Most have a poly-A tail and epigenetic marks consistent with that of a transcribed gene. LincRNA interactions can facilitate oncogene-driven downstream gene regulation, acting as a molecular scaffold to mediate RNA-protein interactions at regulatory regions of oncogene-targeted genes. However, the mechanism for these interactions is not known; it has been hypothesized that a sequence-specific motif in the lincRNA structure may form the molecular scaffold. These analyses aim to quantify lincRNA expression across TCGA tumor types, to correlate lincRNA expression with gene expression, mutation, and

methylation subtypes, and to identify enrichment of sequence-specific lincRNA-DNA interactions at regulatory domains of cancer genes. Two sources of annotation (ENCODE and the Broad Institute) were used to support lincRNA quantification. Intragenic and overlapping transcripts were excluded from these analyses. Dr. Amin noted that most lincRNAs are poly-A–enriched and show very low expression in comparison to mRNA expression. Unsupervised analysis of 189 lincRNAs across 327 melanoma samples showed differential lincRNA expression based on CIMP subtype and mutation signatures. Several of the identified lincRNAs have roles in cancer-specific modulations. Differentially-expressed lincRNAs show proximity to several oncogene-regulated genes. These lincRNAs are more enriched in distal regulatory sites than in proximal sites, possibly indicating enhancer-like activity. Motif discovery analysis of 189 variably-expressed lincRNAs indicated a potential role of lincRNAs in transcription regulation and cancer growth signaling pathways. *ALU* elements are preferentially enriched in lincRNA exonic regions; approximately 23% of lincRNA transcripts have at least one *ALU* sequence in their coding region. Preferential hits from specific *ALU* subfamilies correspond to the most recent expansion of *ALU* elements within the context of primate evolution, suggesting a possible regulatory role. Ongoing tasks in these studies include determining the functional relevance of differentially-expressed lincRNAs by co-expression network analysis and integrating CN alterations involving differentially-expressed lincRNA regions, and making lincRNA data and analyses accessible via a Synapse portal and through Firehose portal. Dr. Amin noted that WGS analyses are available for 39 of the melanoma cases, which will enable the exploration of how mutations correlate with expression.

Multi-omics Classification of Head and Neck Cancer Ties TP53 Mutation to 3p Loss
Andrew M. Gross; University of California, San Diego

Mr. Gross began by stating that there are 560,000 cases of HNSCC annually, with 300,000 deaths attributed to the disease. HNSCC has a five-year survival rate of approximately 40%. These analyses aim to understand the molecular composition of individuals with HNSCC, to identify molecular subtypes within the patient cohort, to develop methods to integrate data across diverse measurement platforms, and to isolate genetic interactions in a cancer cohort. These analyses featured a cohort of individuals who were not infected with human papillomavirus (HPV). The study design included defining a set of candidate biomarkers, identifying biomarkers that stratify the patient cohort with respect to outcomes, and searching for associations among pairs of the prognostic biomarkers. Mr. Gross noted that *TP53* mutations and loss of the 3p arm co-occur in approximately 70% of HNSCC patients. The adverse prognostic effect of *TP53* is mediated by 3p loss. Multivariate analysis suggests that this observation is not an artifact of the relationship between *TP53* and chromosomal instability; no additional effect of instability was observed when 3p loss was taken into account. HPV-positive status confers a relatively large negative prognostic effect for individuals with HNSCC. To stratify this cohort further, a secondary prognostic screen was carried out for 179 patients with *TP53* mutation and 3p loss, associating mir-548k expression with poor prognosis. Individuals with amplification but not expression of mir-548k have better prognosis. *MIR548K* was recently implicated in esophageal cancer (Song Y, et.al. *Nature* 2014;509:91-95). Patients without the co-occurring *TP53*-3p loss have relatively good outcomes. A secondary association screen to assess event status revealed

that CASP8 and RAS signaling are important drivers in individuals who lack the *TP53*-3p loss. In conclusion, Dr. Gross noted that *TP53* mutation frequently co-occurs with 3p loss in HNSCC patients. In *TP53*-3p individuals, mir-548k leads to a worse prognosis. Mr. Gross noted that individuals who harbor disruptive *TP53* mutations have the worst prognoses, followed by individuals with non-disruptive *TP53* mutations. Individuals with non-disruptive mutations show significantly different prognoses from wild-type individuals. Given the most of the 3p arm is lost in these patients, it is not known which individual genes on 3p are implicated in these processes.

Tuesday, May 13

Session III

Chair: Peter Laird, Ph.D.; University of Southern California

Integrated Genomic Characterization of Papillary Thyroid Carcinoma
Thomas N. Giordano, M.D., Ph.D.; University of Michigan

Dr. Giordano began by presenting a model of thyroid cancer progression, noting that a gradual loss of differentiation is a hallmark of progression and the foundation of the classification scheme presented in this manuscript. He noted that 85% of TCGA thyroid cancer cases are papillary thyroid cancers (PTCs), which comprise three histologic types. The classical type features *BRAF*-V600E mutations and *RET* fusions. The follicular variant type is characterized by *RAS* mutations and recapitulates normal thyroid architecture. The tall-cell variant type also features *BRAF*-V600E mutations. A strong genotype-phenotype correlation is observed in these tumors. Prior to TCGA, *BRAF* and *RAS* were known mutations in PTC, with infrequent mutations in *PI3K* genes also observed. This study analyzed 496 primary PTCs, 391 of which featured analyses on all major platforms. In addition, 49 WGS analyses were completed on PTCs that lacked apparent driver mutations. Dr. Giordano noted that PTC has a relatively low mutation density, especially among solid tumors. PTC has a relatively quiet genome, with *BRAF* alterations observed in approximately 60% of cases. A diverse set of fusions, including *BRAF* fusions, was observed in 15% of the cohort. Fusions observed include *RET* fusions, diverse *BRAF* fusions, *ALK* fusions, and *ETV6-NTRK3* fusions. Many *BRAF*-mutated tumors had few CN-mediated changes, and a few cases had no putative drivers. *EIF1AX* was identified as one putative driver. The 14 (of 402 cases) that had no apparent drivers have some possibly explained mutations. The five drivers that were identified in these analyses were clonal, with implications for targeted therapy. Dr. Giordano noted that one challenge associated with this project is its focus on papillary carcinoma, which is an indolent cancer type with a 95% cure rate and little follow-up data. These tumors have a relatively low mutation density compared to other carcinomas. The manuscript focuses on the mutual exclusivity of *BRAF* and *RAS* mutations, the quiet nature of the PTC genome, and the availability of multidimensional data. To aid these analyses, the working group developed a $BRAF^{V600E}$-*RAS* score (BRS) that defines a gradient between two classes of PTCs—$BRAF^{V600E}$-like (BVL) and *RAS*-like (RL). The score explores how other mutations fall along this gradient. When a tumor acquires a *BRAF* mutation, its iodine metabolism machinery is silenced. These analyses indicate that *BRAF* mutations do not represent a homogeneous tumor group. *BRAF*-like tumors signal through MAP kinase pathways. These analyses also identify a set of tall-cell tumors with distinct molecular profiles. By leveraging the BRS, thyroid differentiation score, histologic type, and tumor grade, differences between clusters can be identified, including various miR markers. Overarching conclusions include the

observation that RL-PTCs and BVL-PTCs differ fundamentally in their genomic, epigenomic, and proteomic profiles. These studies have identified clinically-relevant subgroups of BVL-PTCs, with a potential role of miRs. As such, Dr. Giordano proposed a reclassification of thyroid cancer that reflects more accurately the genotypic and phenotypic differences of *RAS*- and *BRAF*^{V600E}-driven tumors. One participant noted that genomically-silent tumors are enriched in the follicular variant. The practical implications of mutations in the thyroglobulin gene in PTC are not known at present.

Somatic Alterations in Clinically-Relevant Cancer Genes among 12 TCGA Tumor Types
Ali Amin-Mansour; Broad Institute

Mr. Amin-Mansour began by noting that TCGA has greatly facilitated the study of alterations in clinically relevant genes. However, a clinical lens should be applied to the findings, as an understanding of clinically-actionable somatic mutations will drive clinical care. These analyses assessed somatic mutations and in/dels called from exomes of 3,276 TCGA cases and curated clinically relevant target genes (e.g., those that were diagnostic, prognostic, or predictive of response or resistance to therapies). Users may add genes to this list, which is available at www.braodinstitute.org/cancer/cga/target. More than 500,000 non-synonymous mutations were identified across twelve TCGA tumor types. Some of these mutations will not be clinically relevant, although clinical relevance of a given mutation may become apparent only after being assessed across many tumor types. The frequency of alterations across tumor types shows a "long tail" of alterations, with *TP53* being the most frequently mutated. These analyses suggest that shifting from hotspot to exome sequencing across tumor types will generate much more clinically actionable data. An exclusive focus on targeted gene panels will necessarily exclude some of the ever-increasing number of targeted genes, and the cost of WES continues to decrease. In addition, well-known clinically relevant genes (e.g., *BRCA*) may be rarely altered in unexpected tumor types. Moreover, genes that are rarely mutated in any given tumor type (e.g., *TSC1*, *MTOR*) may occur frequently across aggregated tumor types. However, clinical annotations must be linked to these variants. In conclusion, Mr. Amin-Mansour noted that there is a "long tail" of alterations observed in clinically relevant genes. Upgrading from hotspot profiling to exome sequencing will yield a more complete and clinically useful patient tumor profile. Clinically relevant alterations in well-known genes occur rarely in unexpected tumor types, and genes that are rarely mutated in any given tumor type are more regularly altered when considered in the context of aggregate studies. Future efforts of these studies include integrating CNV and fusion data to provide a clinical focus to omics studies. Mr. Amin-Mansour noted that *BRCA* mutations observed in unexpected cancer types occurred in various positions and were not specific to the tumors.

Inferring Intra-Tumor Heterogeneity from Whole Genome/Exome Sequencing Data
Layla Oesper, M.S.; Brown University

Ms. Oesper began by stating that many tumors are highly heterogeneous, containing multiple populations each with its own complement of somatic mutations. Tumors evolve over time and pass mutations to their progeny, and tumors also feature admixture with normal cells. Given the proliferation of sequencing data with modest coverage, it can be challenging to infer tumor composition from a single, mixed tumor sample. SNV-based methods (e.g., PyClone, SciClone, etc.), which cluster variant allele frequencies, rely on examining individual points in the genome

and often require higher coverage. Copy number aberration (CNA)-based methods such as ABSOLUTE and ASCAT, which were originally designed for SNP arrays, examine large regions. CNA gives a strong signal in sequencing data that can be combined across aberrations to infer tumor composition. If a tumor population contains multiple CNAs, one can expect a consistent shift in read depth. The Raphael group (of which Oesper is a member) has developed a probabilistic model to infer tumor composition from sequencing data when different tumor populations contain different CNAs. This algorithm incorporates two parameters that represent the number of copies of genomic intervals in each sample subpopulation and the proportion of the subpopulation in a given mixture. The multinomial Tumor Heterogeneity Analysis (THetA) algorithm was developed to identify the most likely tumor composition from a measured read depth. THetA is efficient for mixtures that contain normal cells and a single tumor subpopulation, and it can infer the composition of a mixture that contains normal cells and any number of tumor subpopulations. This algorithm explicitly considers multiple subclonal populations when assessing tumor purity. The next-generation of THetA includes improved optimization for multiple tumor subpopulations, extension to WES and low-pass WGS data, and analysis of highly-rearranged genomes. When THetA is applied to WES data, a similar purity estimate to that calculated by ABSOLUTE is observed for many samples. In contrast to ABSOLUTE, however, THetA inferred several subclonal aberrations. In summary, Ms. Oesper noted that THetA infers tumor sample purity and cancer subpopulations. Improvements to this algorithm have enabled it to be applied to a range of data types, including WES and WGS. She noted that while read-depth information has been incorporated into THetA, mutation data has not yet been incorporated, but B-allele frequencies can be used to validate predictions.

Genomic Characterization of Invasive Lobular Breast Carcinoma

Michael L. Gatza, Ph.D.; The University of North Carolina at Chapel Hill

Dr. Gatza began by observing that invasive breast cancers include invasive ductal carcinoma (IDC; 50-80% of cases), invasive lobular carcinoma (ILC; 10-15% of cases), and mixed IDC/ILC (4-5% of cases). These analyses aim to identify the genomic differences between ILC and IDC. At the time of the data freeze, 817 invasive breast cancer specimens were accrued, 490 of which were IDC, 127 of which were ILC, and 88 of which were mixed IDC/ILC. Dr. Gatza noted that ductal tumors tend to be more diverse than lobular tumors. Lobular tumors have upregulated ATM network, immune signaling, and MAPK signaling pathway genes, whereas ductal tumors show upregulation of MYC targets and E-cadherin stabilization. Ductal tumors have lower cellularity than lobular tumors. The Working Group also developed an integrative MAF using data from WES, UNCeeR (which combines DNA-seq and mRNA-seq), and ABRA (which allows the detection of small variants in in/dels). These integrated MAFs identify more species of interest than do DNA-based MAFs and were used to determine the frequency of mutations in ductal or lobular samples and to assess IDC Luminal A- and ILC Luminal B-specific alterations. PARADIGM analysis identified signaling pathways associated with IDC and ILC. Genes that were downregulated in ILC tumors included *CDH1*, *MYC*, and *XBP1*. Genes upregulated in these tumors included *TP53*/DNA damage response genes and immune-related genes. Lobular tumors appeared sufficiently variable to possibly constitute multiple diseases. RNA-seq analysis of ILC tumors identified three mRNA-based classes. A two-class sequence/alignment map (SAM) analysis identified 988 differentially expressed genes among ILC classes. ILC class mRNA/miRNA expression patterns corresponded with IDC and adjacent

normal tissues. Only the ILC Class I corresponded with the RPPA-reactive subtype, and ILC Class I tumors exhibited altered PDGFR/STAT3 and FOXM1 signaling. ILC Class II was defined by high immune signaling and proliferation genes. PARADIGM analysis identified IFNγ and FOXM1 signaling as key pathways in Class II tumors. In summary, the Working Group developed a unique integrated MAF that utilizes both DNA exome and mRNA sequencing. Comparison of ILC versus IDC revealed that *FOXA1* and *CDH1* mutations are associated with ILC, whereas *GATA3* mutation is associated with IDC. Altered signaling is observed in ILC versus IDC, along with differentially expressed miRNA and methylation patterns. ILC Class I tumors were associated with the reactive stromal subtype, whereas Class II tumors featured an immune component and were highly proliferative. Dr. Gatza noted that the lobular tumor classes featured similar amounts of tumor purity, although ILC is an infiltrative disease. Background subtraction was used to correct for variant proportions of stroma and normal tissue.

LineUp: Identifying Deleterious Mutations using Protein Domain Alignment
Brady Bernard, Ph.D.; Institute for Systems Biology

Dr. Bernard discussed LineUp, an algorithm to identify deleterious mutations using sequence-based protein domain alignment. Normalization of mutation frequencies across tumor types identifies a few recurrent significantly mutated genes (e.g., *TP53*), but most genes feature a low frequency of mutations, many of which may turn out to be deleterious. When considering data from multiple tumor types, interesting profiles may emerge. For example, structural alignment of RAS domain elements shows that functional sites (e.g., binding, catalytic) are conserved, even if their functions differ. LineUp aligns sequences from matching domains across all tumor types and evaluates their missense mutation frequencies. This approach applies to approximately 40% of missense mutations but excludes nonsense and frameshift mutations in certain genes. For example, aligning several hundred homeobox domains across twenty TCGA tumor types identifies positions of interest. For genes with low mutation frequency, variants observed in TCGA data can be compared to those identified in 1000 Genomes to identify genes contributing to a point and to assess whether these variants are deleterious. The functional impact of these variants may also be interpreted structurally. In summary, Dr. Bernard noted that mutations have been comprehensively evaluated at all positions within all domains to identify low-frequency but likely deleterious mutations. Hotspots that are outside of domains and mutations that broadly disrupt structure and function are not addressed in this method, making integration with other methods essential. Functional validation of low-frequency events in such data sets remains a challenge. However, as data volumes for tumor and normal genomes increase, more robust normalization per position per domain can be achieved. Dr. Bernard noted that LineUp can complement Mutation Assessor, which is based on sequence conservation within subfamilies. One participant recommended considering structure-based domain alignments as a complementary method to LineUp's sequence-based alignments.

Integrative Analysis of TCGA Data Reveals that Wilms' Tumor 1 Mutation is a Driver of DNA Methylation in Acute Myeloid Leukemia
Subarna Sinha, Ph.D.; Stanford University

Dr. Sinha began by stating that acute myeloid leukemia (AML) is characterized by the accumulation of myeloid precursor cells in the bone marrow that are blocked with respect to their ability to differentiate into mature blood cells. As with other tumor types, AML is

associated with widespread deregulation of DNA methylation. Regions of aberrant hypo- and hypermethylation could possibly arise from stochastic processes, cells of origin, or genetic mutations. The aims of these studies were to identify the genetic drivers of aberrant methylation in AML and to identify leads for a mutation-specific therapy. The Boolean Implications tool can be used to assess pairs of attributes (e.g., methylation, mutations, CNA). This method assesses a set of four "if-then" implications in which Attributes A and B relate (e.g., If A is high, then B is high [HIHI implication]; if A is high, then B is low [HILO implication], and so forth). The Sinha group has developed a computational pipeline using 191 TCGA AML specimens that feature both mutation and methylation data. Boolean Implications were generated by assessing the mutation/methylation pairs, with the number of methylation HIHI and HILO Boolean implications counted for each mutation. These analyses revealed that *WT1*, *IDH2*, and *CEBPα* mutations in AML are linked to hypermethylation. These mutations are almost mutually exclusive; distinct CpG sites and associated genes are linked to hypermethylating mutations. However, it was not known previously whether *WT1* plays an active role in hypermethylation in AML tumors. Their experiments found that *WT1* induces hypermethylation in AML cell lines, and the mutant *WT1* methylation signature is enriched for Polycomb repressor complex 2 (PRC2) target genes in cell lines and TCGA specimens. *WT1*-mutant AML shows aberrant repression of PRC2 targets. Inhibition of PRC2 promotes differentiation in AML tumors that feature *WT1* mutation. In summary, Dr. Sinha noted that mutation in *WT1* is strongly linked to DNA hypermethylation in AML. Introduction of mutant *WT1* into wild-type cells induces the same pattern of DNA hypermethylation as in primary tumors. The pattern of methylation and gene expression is consistent with a differentiation block caused by *WT1*-mutants through dysregulated silencing of PRC2 targets. This differentiation block in *WT1*-mutant AML can be overcome by inhibiting EZH2, and EZH2 inhibitors have activity in *WT1*-mutant AML. Boolean implications are thus useful tools for mining large, heterogeneous cancer data sets.

Session IV

Chair: Neil Hayes, Ph.D.; The University of North Carolina School of Medicine

Comprehensive Genomic Characterization of Cutaneous Melanoma

Ian R. Watson, Ph.D.; The University of Texas MD Anderson Cancer Center

Dr. Watson began by stating that this project focuses on analysis of TCGA melanoma specimens, noting that the melanoma TCGA project differs from other TCGA efforts in that 80% of specimens are of metastatic origin. To reduce heterogeneity, tumors that originate from non-glabrous skin primary tumors were selected. Patients were selected who had no prior systemic treatment, with the exception of interferon treatment for less than 90 days prior to obtaining the specimen. The rationale for these criteria is the scarcity of frozen primary tumor tissue in sufficient quantityand if discovered at an early stage, melanoma is highly curable—however, the five-year survival rate decreases to 62% for regional disease. Most of the tissues included in these studies were excised from lymph nodes and skin. Only two cases had matched primary and metastatic specimens. Dr. Watson noted that melanoma features the highest mutation rate of cancers sequenced to date, with UV playing a role. The mutation rate was 17 mutations per megabase, most of which were C>T transitions. As such, it can be challenging to identify driver mutations in this disease. MutSig analysis identified 42 significantly mutated genes. InVEX, which permutes called mutations within a gene's exons and introns and assesses whether more "non-silent" mutations are observed in the real data versus the permutations, identified 13

significantly mutated genes, all of which were also identified by MutSig. Many *BRAF* V600 and *NRAS* mutations were observed in these analyses, and *NF1* was mutated in 14% of specimens. Melanoma was characterized into four subgroups (*BRAF* hotspot [n=122], *NRAS* hotspot [n=70], *NF1*-mutant [n=38], and triple wild-type [n=39]). Fusion analysis, which incorporated RNA-seq, DNA deep-seq, and low-pass DNA-seq, identified 221 potential drivers, including *BRAF* fusions with *ATG7* and *TAX1PDB1*. A pattern of high DNA methylation at CpG islands was associated with *IDH1* mutations. Dr. Watson noted that the observed clustering may reflect tissue of origin. Hierarchical clustering identified melanoma subgroups with elevated epithelioid, lymphocytic, and MITF expression signatures that are not associated with mutation status. The high-lymphocytic group was associated with the best prognosis. However, tumors from lymph nodes have higher degrees of lymphocyte infiltration, and specimens with high levels of infiltration are associated with better prognosis. *TERT* promoter mutations were observed in melanoma at a similar frequency (65%) as previously reported. Only the C228T *TERT* promoter mutation was associated with increased expression, and a higher fraction of metastatic samples contain this mutation. In summary, these analyses have identified four genetically distinct melanoma subgroups (e.g., *BRAF* mutant hotspot, *RAS* mutant hotspot, *NF1* loss-of-function mutations with UV signature, and triple wild-type specimens that lack the UV signature but are driven by CNAs of known oncogenes). Integration with other data platforms identifies differential MAPK signaling pathways that are altered in these genetic subtypes. Methylation clustering analysis has identified a CIMP subtype that is enriched for *IDH1* R132 mutations. Ongoing analyses include the incorporation of Oncosign analysis, miRNA clustering analysis, comparative analyses of primary versus metastatic tumors, and analysis of genetic determinants of lymphocytic infiltration. Degree of lymphocytic infiltration did not differ between the genetic subtypes. Dr. Watson noted that chromothripsis could possibly explain the genetic signatures observed in triple wild-type samples.

The Pan-Cancer Proteomic Landscape of TCGA Projects

Rehan Akbani, Ph.D.; The University of Texas MD Anderson Cancer Center

Dr. Akbani began by stating that his group analyzed 3,467 TCGA samples across eleven tumor types (equivalent to TCGA's Pan-Can-12 with the exception of AML). In these studies, 181 proteins were analyzed, including 128 total proteins, one cleaved, one acetylated, and 51 phosphorylated forms. RPPA data were produced in six batches that were controlled for batch effects. Correlations were established between protein and other platforms. The correlation between mRNA and protein was approximately 0.3, which is common for patient samples but less robust than the correlation observed in cell lines. The individual mean correlations between mRNA and protein were not uniform across the cancer types, with colon cancers showing the least degree of correlation. Protein:CNV correlation studies showed that amplification results in a 5% increase in protein expression levels, whereas deletions produce a 5% decrease in protein expression. Dr. Akbani noted that these observations may be affected by large numbers of passenger CNVs. Elevating mutations increased protein expression levels by 20%, whereas suppressive mutations decreased protein expression levels by 10%. miRNA and protein level correlated with a coefficient of ±0.07, whereas protein:protein mean correlation was approximately ±0.15. When measuring a specific entity, such as ERBB2, using CNV, mRNA, or protein level, differences are readily apparent in some tumor types (CRC, BLCA, LUAD, UCEC), with protein being high in a much larger percentage of samples than those measured by CNV or mRNA. While existing ERBB2-based therapies target cell-surface proteins, many of the

tests for ERBB2 rely on CN- or mRNA-based platforms, suggesting that a test that measures protein directly may identify a greater number of individuals who will benefit from therapy. Unsupervised clustering of the 3,467 samples identified eight major clusters, most of which cluster by tumor type with a few exceptions. Bladder cancers that cluster with endometrial cancers have worse prognosis than other bladder cancers. Squamous-like kidney cancers are associated with worse prognosis than other kidney cancers. Marker proteins that could serve as potential targets for therapy were identified across different tumor types. These included ER-alpha and AR (women's cancers), AR, BCL2, FASN, ACC1, and pACC (luminal breast cancers), CYMC (ovarian cancer), pSRC (all except for women's cancers), HER2 (bladder, endometrial, breast, colorectal cancers), HER3 (kidney cancers), and pEGFR with NOTCH1 and HER3 activation (GBM). Adjustment for tissue-specific effects yielded seven clusters driven by pathway activations/deactivations across tumor types. A reactive cluster that features predominantly breast cancers was characterized by protein markers that included CAVEOLIN1, MYH11, RICTOR, and COLLAGENVI. This reactive cluster was associated with good prognosis in breast, colon, and kidney tumors. In lung squamous cancers and bladder tumors, however, the reactive cluster was associated with poor prognosis for reasons unknown. Suppression of AKT pathways improves prognosis (as expected) in ovarian, but not kidney, cancers. Activation of the AKT pathway improves the prognosis for kidney tumors but not for ovarian cancers. Pathway activity thus appears to depend on tumor type. In summary, Dr. Akbani noted that mutations have greater mean-fold changes than CNV on average. Furthermore the correlations between mRNA and protein can vary widely by disease, and HER2 protein levels are not well predicted by CNV or mRNA in certain diseases, such as colorectal cancer, lung adenocarcinoma, and bladder cancer. In addition, these analyses identified several novel markers. Outcome differences were observed across clusters, possibly driven by pathway differences. Finally, pathway effects were not equal by disease; certain pathway activations affect prognosis positively or negatively in a disease-dependent fashion. Protein: protein correlations also varied by disease in these analyses.

Data Mining TCGA Breast and Ovarian Exomes for Novel Susceptibility Markers
John Martignetti, M.D., Ph.D.; Icahn School of Medicine at Mount Sinai

Dr. Martignetti began by noting that his research has mined TCGA in a different manner than that of other presenters, with a focus on risk genes using family-based studies to identify ovarian and breast cancer susceptibility genes. Family history is the strongest single predictor of a woman's chance of developing breast and/or ovarian cancers. While *BRCA1* and *BRCA2* mutations still represent the strongest known genetic predictors, they are responsible for less than 50% of all families that contain two or more cases in first-degree relatives and explain less than 50% of the excess familial cancer risk. Genetic studies that seek to identify breast and ovarian cancer susceptibility genes have therefore focused on those families with a high incidence of cancer across multiple generations. This approach avoids many of the technical, clinical, and statistical issues associated with genome-wide association studies (GWAS). The approach also overcomes issues associated with "rare" alleles by eliminating bias against identification of rare disease-causing alleles. Some families will harbor a "private" mutation, whereas others may share a gene. This family-centric approach can identify events at the population level that are not apparent when viewed in isolation. In these analyses, 21 exomes were sequenced with coverage depths ranging from 80X to 250X. Studies focused on families with wild-type *BRCA1/2* and identified other susceptibility genes. A filtering approach was used to reduce the number of

genes to validate; potential candidates were then curated manually to identify those most likely to validate. This approach identified 24 candidate genes shared among three individuals with family pedigrees of breast cancer. The mutation frequency in germlines of individuals with ovarian cancer was then assessed using TCGA data. Of the eight genes selected from this analysis, five contained the specific variant at frequencies of 1-2% in the wild-type *BRCA* subpopulation, and two of these were enriched from the general population. Mutation Assessor was applied to the 24 genes to score their functional impact. Seven variants were subsequently assessed as functional; three of these variants affected genes that are involved in cancer. Fifteen genes were chosen between TCGA and Mutation Assessor gene sets, and their distribution was analyzed in TCGA Pan-Cancer data. A candidate gene with biologic potency affecting long-term survival in breast cancer was identified. In summary, Dr. Martignetti noted that a number of high-interest "candidate" ovarian and breast cancer susceptibility mutations were discovered within and between families. The use of germline TCGA data allowed this candidate list to be refined. Planned validation studies include generating mutation-specific ovarian cancer cell lines and "humanized" mouse lines to understand the biology of one of the ovarian cancer susceptibility mutations and sequencing of a breast cancer candidate in an independent breast cancer family cohort followed by functional studies in cultured cell lines. Dr. Martignetti noted that three of the eight candidates had no associated SNPs.

Discovery and Functional Characterization of Recurrent Gene Fusions from 4,932 Primary Tumor Transcriptomes across 19 Human Cancers
Chai Bandlamudi, M.S.; The University of Chicago

Mr. Bandlamudi began by stating that COSMIC maintains a database of known fusion genes, with only 48 recurrent fusions appearing at a frequency of 1% or greater. Some of these genes are tissue-specific. To make a fusion call, one needs discordant reads and anchor reads. The specificity of the anchor read is determined by the anchor length, and a higher number of anchor reads is associated with increased specificity. However, alignment errors can generate false-positive fusion calls. To control for artifacts, the Bandlamudi group has developed Minimum Overlap Junction Optimizer (MOJO) to identify canonical fusions from paired-end transcriptome data. The sensitivity of MOJO was assessed using 18 cell line transcriptomes with previously-validated fusions. MOJO's performance was compared against those of other published methods (e.g., deFuse, Tophat, and others) across 7,175 tumor transcriptomes. Because the size of this sample set will engender some false-positive calls, MOJO analysis was run in its most sensitive mode, followed by filtering of calls from normal transcriptomes. This approach identified 15,582 high-confidence somatic fusions and captured most of the known fusions from TCGA marker papers. Mr. Bandlamudi noted that breast and ovarian cancers feature the highest numbers of fusion calls, with nearly all of the fusion calls featuring a known fusion gene. More than 75% of thyroid tumors had no fusions identified in these analyses. Many novel recurrent fusions were identified, some of which appeared across multiple tumor types. Functional validation used a workflow that includes synthesizing fusion constructs, packaging them into viral particles, creating stable cell lines that express the fusion construct, and assaying these for three hallmarks of cancer (proliferation, invasiveness, and evasion of apoptosis). This workflow was applied to eleven fusion constructs from nine fusion genes in MCF10A cells. Future directions for this work include carrying out functional validations using NIH3T3 cells, overexpressing full-length individual genes as controls, and characterizing functional fusions. Recurrently-fused genes are of special interest in these studies. Integrated analysis will also be carried out to identify

associations between CNAs, mutations, and fusion events. Mr. Bandlamudi noted that increases in RNA-seq depth correlate with the identification of additional clonal events. If the clonality or the purity of a tumor specimen can be estimated, then the required degree of RNA-seq depth can be modeled for discovery applications. The fusions identified in these studies have not yet been validated by PCR and Sanger sequencing.

Widespread Genetic Epistasis among Cancer Genes
Audrey Q. Fu, Ph.D.; The University of Chicago

Dr. Fu began by noting that epistasis or genetic interaction can be thought of as masking and determined using a multiplicative model. Phenotype outcome for double-mutant is product of the two phenotypes. [The video for this presentation was truncated].

Understanding the Evolution of the Melanoma Epigenome
Kadir C. Akdemir, Ph.D.; The University of Texas MD Anderson Cancer Center

Dr. Akdemir presented on efforts to understand the evolution of the epigenome in melanoma progression. Melanomas resist therapy through dedifferentiation, indicating that the cellular state associated with the disease is reversible. Epigenetic signatures play a known role in several human cancers, including melanoma and colon cancers. But how does the epigenome contribute to melanoma progression? In these experiments, a primary human melanocyte-based cell-line system that incorporates tumorigenic, non-tumorigenic, and metastatic lines, has been used to identify epigenetic changes. High-throughput chromatin immunoprecipitation (ChIP) sequencing was used to profile chromatin marks. Results indicate that melanoma features a loss of histone acetylations at observed pro-tumorigenic melanocytes and that de-acetylated enhancers contain putative tumor-suppressor motifs. Dr. Akdemir noted that the preexisting chromatin landscape could determine tumor suppressor-based regulations; a high degree of chromatin acetylation promotes binding of transcription factors. Analysis of ten human melanomas has provided insight into the evolution of the epigenome during development of human melanoma. These ten melanoma tumor samples from primary and metastatic lesions were comprehensively profiled by TCGA. Changes in chromatin states were evaluated using ChIP-seq, which requires approximately 3 mg of tissue (1,000-10,000 cells). To date, 36 histone marks, two forms of RNA polymerase, and three histone variants, and CCCTC-binding factor (CTCF) have been characterized across eight of these tumors. Dr. Akdemir stated that cancers show retrograde remodeling of their regulatory landscape. Cancer cells acquire regulatory regions, and new oncogenic sites may actually represent “older” developmental pathways. Data from Roadmap Epigenomics was used to investigate the developmental pathways that are corrupted during melanogenesis. These analyses revealed potential reorganization in the regulatory landscape during melanoma formation, including evolution at the H3K4me1 site. Integrating epigenomic data sets can determine the functionality of non-coding variants. One identified melanoma GWAS site loses its active histone marks in pro-tumorigenic melanocytes and in human melanoma. In summary, comprehensive epigenomic characterization in a primary melanocyte-based melanoma model has revealed loss of histone acetylation around genes involved in carcinogenesis. De-acetylated enhancers could hinder binding of key transcription factors to DNA. Preliminary epigenome profiling of human melanoma tumors suggests epigenomic reorientation during melanomagenesis. These studies have demonstrated that epigenomic profiles

are useful to annotate non-coding variants in cancer. These sequencing results will be made publicly available.

Session V
Chair: Ilya Shmulevich, Ph.D.; Institute for Systems Biology

Comprehensive Molecular Characterization of Gastric Adenocarcinoma
Adam Bass, M.D.; Dana-Farber Cancer Institute

Dr. Bass began by stating that gastric cancers cause approximately 723,000 deaths annually. From a pathology perspective, the two main categories of gastric cancers are intestinal and diffuse. The former, which are more typical adenocarcinomas with a glandular structure, commonly metastasize to the liver. Diffuse tumors are poorly cohesive, invasive, and frequently metastasize to the ovary and peritoneum. These tumors are associated with much worse survival than similar lobular breast cancers. Gastric adenocarcinoma may represent many different diseases, possibly due to anatomic site, histologic features, geographic parameters, and mutational profiles. In clinical trials, however, these nuances generally become ignored. The goals of TCGA analysis of gastric tumors included classifying the tumors more accurately with a scheme that can applied to practical settings, identifying key pathways in distinct tumor types, and identifying targets/biomarkers for distinct tumors and tumor types. To develop a classification scheme, these studies used the organization inherent in the data in an agnostic manner. Cluster of Cluster Assignments (COCA) and iCluster were used to identify key features of molecular clusters of tumors. These identifying features were then used to categorize tumors via a decision tree, with analysis based on assignments from the decision tree rather than on initial clustering. COCA identified four tumor subtypes (MSI, diffuse, Epstein-Barr virus [EBV]-positive, and aneuploidy). iCluster identified five groups that recapitulated the COCA results, with the aneuploidy clusters separating into two subclusters. Dr. Bass noted that the diffuse tumors were overwhelmingly present in the genomically stable cluster, although the anatomic site affected the classification to an extent. Distinct CIMP profiles differentiate EBV-positive and MSI-positive gastric cancers, with methylation patterns playing a large role. Dramatic rates of *PIK3CA* mutation were observed in EBV-positive gastric tumors. CNA identified 9p amplification, which was enriched in EBV-positive tumors (15%). *JAK2*, *CD274*, and *PDCLILG2* are genes within this area that code for PD-L1 and PD-L2, which are proteins that are targets of emerging inhibitors. Expression of PD-L1 and PD-L2 is elevated in EBV-positive gastric cancers; these EBV tumors feature a strong immune cell signaling signature. Aneuploid groups show highly-recurrent amplification of oncogenes, including numerous targetable genes. MSI tumors feature more mutations, including recurrent *ERBB2* and *ERBB3* mutations, but fewer CNAs. The diffuse cancers are relatively stable genomically. Significantly mutated genes in gastric cancer have been identified in these analyses. Highly recurrent *RHOA* GTPAse mutations are enriched in diffuse-type gastric tumors. *RHOA* has roles in invasion and migration, which could contribute to the diffuse growth phenotype. A novel recurrentclaudin-18 (*CLDN18)* -*ARHGAP26* fusion gene impacts the RHOA pathway and adhesion. Both *RHOA* and *ARHGAP* fusions were enriched in diffuse gastric cancers. Claudin-18 is a component of tight junctions and the cellular adhesion complex, whereas ARHGAP26 is a RHO-GAP, GTPase-activating protein that should reduce RHOA activity. Dr. Bass noted that *H. pylori* annotation is sparse for these tumors, as *H. pylori* does not enter the cancer cells. Moreover, relatively few clinical data

are currently available on these tumors. Upcoming efforts will merge gastric and esophageal tumor data.

Integrated Analysis of Metastatic Disease in Clear Cell Renal Cell Carcinoma: A Collaborative TCGA Analysis

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

Dr. Hakimi discussed the clinical TCGA (cTCGA) effort to gather information associated with TCGA specimens that is of greatest use to clinicians. He noted that clinical information collected at the time of accrual of TCGA specimens is often limited. Moreover, these data are often not reviewed in advance by disease experts, and cancer-specific outcomes are often not collected. Data of specific interest to the clinical community include risk factors, post-surgery treatment information, and detailed metastatic information. The cTCGA effort collected additional data from source sites on 342 of 389 (82%) TCGA kidney cancer cases. Acquired data included history, comorbidities, laboratory values, metastatic disease, systemic therapy, and longer-term recurrence data. Genomic insights from TCGA can be used to study epidemiologic phenomena. Many risk factors have been established for kidney cancer, including smoking and body mass index (BMI). Although BMI is a risk factor for acquiring kidney cancer, meta-analyses suggest that it is a protective factor at the time of surgery. To understand such observations, a study was conducted using data from 2,119 individuals who underwent renal mass surgery at MSKCC between 1995 and 2012. Logistic regression models identified associations between BMI and advanced disease overall, as well as in subgroups defined by comorbidities, presentation, and albumin level. Multi-variable competing risk regression models estimated associations between BMI and cancer-specific mortality. Somatic mutation, CN, DNA methylation, and expression data were examined by BMI among a subset of 126 individuals who participated in TCGA for CCRCC. These analyses indicated that more obese patients tended to have lower-stage tumors. Higher BMI was thus associated with better outcomes, although this relationship disappeared when analyses were controlled for tumor stage and grade. The protective effect of BMI was maintained even when nutritional status was poor, suggesting that BMI is an independent predictor in this context. Genomic interrogation was then carried out on 126 individuals from the same cohort that were analyzed as part of TCGA. These studies assessed the impact of BMI classes on mutations, CN events, DNA promoter methylation, and mRNA expression. Pathway analysis was performed on genes that were differentially expressed in the obese and normal-weight cohorts. While CN, DNA methylation, and the top fifteen mutated genes were not different among BMI cohorts (Hakimi AA, et.al. *JNCI* 2013;105:1862-1870), mRNA expression identified genes in fatty acid metabolism and beta oxidation-enriched pathways in obese individuals. *FASN* is downregulated in obese individuals with kidney cancer. However, lower *FASN* levels are associated with poor outcome. *FASN* allows for *de novo* lipid synthesis and promotes cell survival. *FASN* overexpression assessed by immunohistochemistry has been associated with aggressive RCC and shorter cancer-specific survival, and the pharmacologic inhibition of *FASN* can reduce RCC tumor growth *in vitro*. Lower expression of *FASN* was observed among obese colorectal cancer patients from the Nurses' Health Study, and other studies of colorectal and prostate cancer patients suggest that the adverse impact of *FASN* overexpression is limited to obese patients. High *FASN* levels are associated with worse survival, and obese patients downregulate this pathway. Ongoing efforts are investigating the number and timing of metastatic cases. In this cohort, 75% patients presented with evidence of metastatic disease at the time of surgery. These analyses also aim to combine genomic and clinical data to

correlate treatment response with underlying genomic differences and to assess why certain advanced-stage tumors metastasize. RNA-seq data will also be used to assess components of immune response as potential biomarkers. In summary, Dr. Hakimi noted that cTCGA Consortia can provide powerful insights into clinical and epidemiologic phenomena. The rich genomic information can serve as discovery sets for targeted validation in larger clinical cohorts. However, collaborative infrastructures are critical to make significant advances in these studies. Dr. Hakimi noted that these experiments will provide insight into whether BMI is the cause or the effect in these observations. It was also noted that TCGA continues to accept clinical data from source sites, and contributors are encouraged to provide as much follow-up data as possible. The data collected for the experiments described in this presentation will be curated and uploaded to TCGA.

Multi-Center Mutation Calling in TCGA
David Wheeler, Ph.D.; Baylor College of Medicine

Dr. Wheeler noted that mutation calls underpin many of the analyses in TCGA and that the field of mutation calling has evolved during the course of the initiative. Sources of error in sequencing data include randomly-distributed base-calling error, which is largely reflected in Q-values, and mapping error/alignment ambiguities. The latter category is systematic and depends on the details of the repeat structure of the genome (which differs between the tumor and normal) and on the sequencing chemistry. These errors generate a high-quality variation. Dr. Wheeler stated that all first-generation callers have a “truth engine” that distinguishes true calls from error. The downstream events thus generated must then be filtered; initial variation calls must be filtered by heuristic criteria. Dr. Wheeler noted that the best example of documenting such heuristic criteria is that for the Broad Institute’s MuTect platform (Cibulskis K, et.al. *Nat Biotechnol* 2013;31:213-219). Mutations provided by a single caller can yield sensible genomic profiles that align to pathways. However, if one compares calls from several different callers, relatively little overlap is observed. Early attempts to compare a set of callers indicated a high discordance between callers. High-quality calls as defined by one caller were missed by other callers. However, multi-center mutation calling may ameliorate these issues. When various callers are applied to diploid SNP analysis, the agreement is approximately 57% (O’Rawe J, et.al. *Genome Med* 2013;5:28), suggesting that these difficulties extend beyond somatic mutation calling. Heuristic callers used to filter data represent one component of this issue. Dr. Wheeler then described three-Center calling on TCGA cancers. Among genes called by all three centers, entities were highly accurate based on validation rate. Validation rates were lower for in/dels than for SNVs. These results led to the development of a meta-caller based on multiple mutation callers calibrated by validation data (Kim S-Y and Speed TP. *BMC Bioinformatics* 2013;14:189). With this approach, validation data are necessary to calibrate each caller. This effort also formalized multi-Center calling with the recognition that mutation callers continue to improve and that different callers detect different events. Moreover, validation cycles can be lengthy and cause delays in data submission. Three-Center calling can be streamlined using a multiple caller to stratify the calls by quality, with a timeline of approximately three weeks. He noted that the publication of a marker paper includes validation, which requires a second independent sequencing event. Multi-Center calling also enables other researchers to add their callers and therefore play a role in the marker paper. The TCGA adrenocortical carcinoma effort to call SNVs in 91 individuals included five Centers. Second-generation mutation callers (e.g., MuTect v2, VIPER, and MuSE) feature increasingly sophisticated heuristic filters and underlying genetic

models. These tools are being evaluated in the ICGC-TCGA DREAM Mutation Calling challenge. In conclusion, Dr. Wheeler noted that TCGA paradigm for mutation discovery is improved by multi-center calling, as evidenced through reduced false-negative rates and delivery of a set of somatic SNVs of calibrated accuracy. TCGA's multi-center mutation calling paradigm accelerates the submission of marker manuscripts and stimulates the development of new mutation callers by providing relatively rapid "benchmarking." A formal "meta-caller" has been developed that may be useful for retrospectively refining mutation calls from TCGA tumor sets. However, multi-center mutation calling has yet to be applied to other mutation modalities. These callers all use the same set of BAM files.

Extensive trans- and cis-QTLs Revealed by Large-Scale Cancer Genome Analysis
Kjong-Van Lehmann, Ph.D.; Memorial Sloan-Kettering Cancer Center

Dr. Lehmann began by stating that alternative splicing events can include exon skips and alternate 3' or 5' ends and can dramatically change the function of the genes in question. These experiments aimed to identify cancer-specific splicing patterns, variants that regulate splicing within a gene (e.g., *cis* associations), and variants that regulate splicing in other cancer genes (e.g., *trans* associations). TCGA provides RNA-seq and matching exome data. RNA-seq data can be used to identify and quantify splicing events, whereas exome data can be mined to identify variants in exons and in flanking intronic regions. However, TCGA data have not been processed uniformly. To correct for this issue in these analyses, all raw TCGA RNA-seq and WES data have been reanalyzed using a re-mapping pipeline, followed by variant calling with MuTect and GATK U.G. and splice-variant quantification using SplAdder. Analysis of splicing events using RNA-seq data has identified new splicing events that occur frequently in specific cancer types. Many of these represent potential targets for treatment, although independent confirmation is required. Quantitative trait locus (QTL) analyses have been applied to homogeneous data sets, and application to TCGA data offers the opportunity to understand tissue- and cancer-specificity of splicing and the chance to identify *trans*-associations. However, sample purity and heterogeneity and the presence of multiple rare events present challenges. Common variant association analysis must account for the role of patient populations in variant expression. *Cis*-associations in 45 genes have been identified across multiple cancer types to identify those associations replicated across the various tumors. Splicing *trans*-associations have also been identified in various genes. In conclusion, Dr. Lehmann stated that his group has developed a resource of novel and known alternative splice events that has been used to identify cancer-specific isoforms that appear to be rarely expressed in normal samples. A common variant association study was performed to map splicing phenotypes. The sample size available from TCGA data also enables the detection of *trans* associations. However, all of the identified associations require validation. He noted that efforts are underway to filter these events to determine which are tissue-specific and which are cancer-specific. Efforts are also ongoing to associate all common somatic variants in the data set, with rare-variant analysis underway. Somatic variant calls can be made publicly available.

Pan-Cancer Analysis of APOBEC Mutagenesis
Dmitry A. Gordenin, Ph.D.; National Institute of Environmental Health Sciences, NIH

Dr. Gordenin began by noting that single-strand DNA (ssDNA)-specific apolipoprotein B mRNA editing enzyme, catalytic polypeptide-like (APOBEC) cytidine deaminases are strong

endogenous mutagens in human cancers, with APOBEC3B and APOBEC3A believed to be the most likely responsible for mutagenesis. APOBECs can cause C>T and C>G substitutions *in vivo*. APOBEC is the only known strong endogenous carcinogen. APOBEC mutagenesis plays a large role in cancers such as bladder cancers. Recent research has indicated that error-prone trans-lesion synthesis in damaged long ssSNA can be a source of localized hypermutation and strand-coordinated mutation clusters. Mutation motifs are observed in C-coordinated clusters (Roberts SA, et.al. *Mol Cell* 2012;46:424-435), which are similar to kataegis events reported in breast cancer. This approach can be used at a sample-level analysis. Dr. Gordenin noted that APOBEC mutation patterns, as assessed using 3,103 TCGA cancer specimens, are abundant in cervical, bladder, head and neck, breast, and lung cancers, and APOBEC mutagenesis likely occurs in the background of all cancer types (although it is more abundant in certain types). Endometrial cancers rarely contain APOBEC-mutated samples, although most of those that do occur in the serous subtype. The four expression-based breast cancer subtypes vary in APOBEC-hypermutation, with the HER2-enriched subtype featuring the most pronounced fraction of samples enriched with APOBEC mutagenesis. These analyses also enable the separation of ABOPEC-driven from non-APOBEC-driven mutagenesis. For example, the impact of *ERCC2*-mutations in bladder cancers became evident after ABOPEC signature mutations were removed from analysis. Dr. Gordenin noted that TCGA-related efforts in this area include providing input into cancer-specific working groups, integrating analysis of APOBEC mutagenesis in cancer exome MAFs into Firehose, and analyzing updated and new TCGA exome MAFs. Furthermore, mechanistic and bioinformatic approaches can be combined to understand the mutation processes in cancer. Mechanistic knowledge can be used to establish stringent statistical hypotheses, which can then be applied to bioinformatic exploration of large databases of clinical mutations (e.g., TCGA) to understand disease-relevant mutagenic mechanisms. He also noted that mutual exclusivity and co-occurrence analyses can be carried out in the context of APOBEC studies, and ABOPEC mutations appear to co-localize with breakpoints of structural rearrangements.

Integration of Multiple Data Types for Genomic Characterization of Virus-Associated Tumors
Matthew A. Wyczalkowski, Ph.D.; Washington University School of Medicine

Dr. Wyczalkowski began by noting that viruses cause approximately 10-15% of cancers worldwide. Several virus types have been associated with human cancers, although the roles of certain viruses remain controversial. Some viruses integrate into the host genome, whereas other remain episomal. However, the role of integration in gene expression and the ways that integration is associated with disease initiation and progression are poorly understood. Moreover, viruses are ubiquitous, yet cancer is not, thus suggesting specific mechanisms. TCGA data provide a rich source to address such issues. Dr. Wyczalkowski then noted that this presentation focuses on four virus-associated cancers—bladder, cervical, head and neck, and gastric cancers. A viral integration pipeline has been developed that begins by extracting unmatched reads from RNA-seq data. These reads are then Blasted to an NT database to discover viruses and to select a virus reference. WGS data are referenced against this virus reference. Exome and WGS data are then queried for viral integration. This pipeline enables a per-tumor analysis of viral integration frequencies and patterns. For example, approximately 5% of bladder cancers were associated with viruses, whereas nearly all cervical cancers were associated with HPV. A small fraction of head and neck and gastric cancers were associated with EBV. To analyze viral integration, a reference containing both human genome and virus was used. By identifying discordant read pairs in which one read maps to the genome and the other maps to the virus, it was revealed that

EBV is involved in no viral integration events, whereas 100% of HPV18 viruses are integrated. For other viruses, approximately one-half to one-third are integrated. Detailed analysis of a single head and neck tumor shows that *RAD51B* is associated with viral integration. CN increases are commonly seen in viral integration sites. In *RAD51B* exons, viral integration upregulates the expression of the downstream exon. For CN increases, within an integration event, per-exon expression is upregulated. For CN decreases, within an integration site, per-exon expression is downregulated. In summary, Dr. Wyczalkowski noted that a pipeline for multi-modal integration of TCGA data (e.g., RNA-seq for virus discovery and expression analysis, WGS and WES for integration analysis) with unified visual representation of integration events has been developed. These analyses have illustrated the close association between viral integration, CN domains, and expression levels. He noted that these analyses have yet to reveal different mutational spectra in tumors that have integrated viruses, although these analyses are forthcoming. He noted that this pipeline will be made publicly available in the near future.

Closing Remarks
Matthew Meyerson, M.D., Ph.D. and Marco Marra, Ph.D.

Dr. Marra thanked participants for their efforts to date, noting that the quality of work has been energizing. He noted that great progress is expected in the future from these explorations. Dr. Meyerson noted that these efforts will affect and improve the care of cancer patients.

The meeting was then adjourned.