The Cancer Genome Atlas Pan-Cancer analysis project

The MIT Faculty has made this article openly available. Please share how this access benefits you. Your story matters.

<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>As Published</td>
<td><a href="http://dx.doi.org/10.1038/ng.2764">http://dx.doi.org/10.1038/ng.2764</a></td>
</tr>
<tr>
<td>Publisher</td>
<td>Nature Publishing Group</td>
</tr>
<tr>
<td>Version</td>
<td>Final published version</td>
</tr>
<tr>
<td>Accessed</td>
<td>Sat Feb 09 18:20:41 EST 2019</td>
</tr>
<tr>
<td>Citable Link</td>
<td><a href="http://hdl.handle.net/1721.1/84662">http://hdl.handle.net/1721.1/84662</a></td>
</tr>
<tr>
<td>Terms of Use</td>
<td>Creative Commons Attribution-NonCommercial-No Derivative Works 3.0 Unported License</td>
</tr>
<tr>
<td>Detailed Terms</td>
<td><a href="http://creativecommons.org/licenses/by-nc-sa/3.0/">http://creativecommons.org/licenses/by-nc-sa/3.0/</a></td>
</tr>
</tbody>
</table>
The Cancer Genome Atlas Pan-Cancer analysis project


The Cancer Genome Atlas (TCGA) Research Network has profiled and analyzed large numbers of human tumors to discover molecular aberrations at the DNA, RNA, protein and epigenetic levels. The resulting rich data provide a major opportunity to develop an integrated picture of commonalities, differences and emergent themes across tumor lineages. The Pan-Cancer initiative compares the first 12 tumor types profiled by TCGA. Analysis of the molecular aberrations and their functional roles across tumor types will teach us how to extend therapies effective in one cancer type to others with a similar genomic profile.

Cancer can take hundreds of different forms depending on the location, cell of origin and spectrum of genomic alterations that promote oncogenesis and affect therapeutic response.

Molecular profiling of single tumor types

That cancer is fundamentally a genomic disease is now well established. Early on, large numbers of oncogenes were identified using functional assays on genetic material from tumors in positive-selection systems, and a subset of tumor suppressor genes was identified by analyzing loss of heterozygosity. More recently, systematic cancer genomics projects, including TCGA (Box 1), have applied emerging technologies to the analysis of specific tumor types. This disease-specific focus has identified novel oncogenic drivers and the genes contributing to functional change, has established definitions of molecular subtypes and has identified new biomarkers on the basis of genomic, transcriptomic, proteomic and epigenomic alterations. Some of these biomarkers have clinical implications. For example, we now view ductal breast cancer as a collection of distinct diseases whose major subtypes (for example, luminal A, luminal B, HER2 and basal-like) are managed differently in the clinic; the outcomes for metastatic melanoma have improved as a result of therapeutic targeting of BRAFV600E mutations, and the fraction of lung cancers treated with targeted agents is increasing with the discovery of likely driver aberrations in most lung tumors. Large-scale processes that shape cancer genomes have similarly been identified. Analyses of chromothripsis and chromoplexy, which involve the breakage and rearrangement of chromosomes at multiple loci, and kataegis, which involves hypermutational processes associated with genomic rearrangements, are providing insights into tumor evolution (see Garraway and Lander for a review).

Analysis across tumor types

Increased numbers of tumor sample data sets enhance the ability to detect and analyze molecular defects in cancers. For example, driver genes can be pinpointed more precisely by narrowing regions affected by amplification and deletion to smaller segments of the chromosome using data on recurrent events across tumor types. The use of large cohorts has enabled DNA sequencing to uncover a list of recurrent genomic aberrations (mutations, amplifications, deletions, translocations, fusions and other structural variants), both known and novel, as common events across tumor types. However, ‘long tails’ in the distributions of aberrations among samples have also been uncovered. Indeed, a majority of the TCGA samples have distinct alterations not shared with other samples in their...
Important information about the biological relevance of the molecular changes in cancer can be obtained through combined analysis of multiple different types of data. For that reason, TCGA's principal aims are to generate, quality control, merge, analyze and interpret molecular profiles at the DNA, RNA, protein and epigenetic levels for hundreds of clinical tumors representing various tumor types and their subtypes. Cases that meet quality assurance specifications are characterized using technologies that assess the sequence of the exome, copy number variation (measured by SNP arrays), DNA methylation, mRNA expression and sequence, microRNA expression and transcript splice variation. Additional platforms applied to a subset of the tumors, including whole-genome sequencing and RPPAs, provide additional layers of data to complement the core genomic data sets and clinical data. By the end of 2015, the TCGA Research Network plans to have achieved the ambitious goal of analyzing the genomic, epigenomic and gene expression profiles of more than 10,000 specimens from more than 25 different tumor types.

TCGA has other, complementary aims as well: to promote the development and application of new technologies, to detect cancer-specific molecular alterations, to make data and results freely available to the scientific community, to develop tools and standard operating procedures that can serve other large-scale profiling projects and to build cadres of individuals (including experimentalists, computational biologists, statistical analysts, computer scientists and administrative staff) with the expertise to carry out such large-scale, team science projects. As of 24 July 2013, TCGA had mapped molecular patterns across 7,992 total cases representing 27 tumor types. The data, along with tools for exploring them, are publicly available at http://www.cancergenome.nih.gov/. Eight ‘marker papers’ (comprehensive initial publications on each of the tumor types) have been published so far.

Determining whether rare aberrations are drivers (oncogenic contributors) or just passengers (clonally intertwined with neutral effect) and whether they are clinically actionable will require further functional evaluation as well as the analysis of additional tumors to increase power. The identification of more driver aberrations and acquired vulnerabilities for each individual tumor will undoubtedly boost personalized care. Developing treatments that target the ~140 drivers validated so far, however daunting, appears possible; devising one-off therapies for the thousands of aberrations in the long tails will be much more challenging.

Although important general principles have emerged from decades of study, until recently, most research on the molecular, pathological and clinical natures of cancers has been ‘siloed’ by tumor type. One has only to glance at the directory of oncology departments in any major cancer center to realize that medical and surgical cancer care are, for the most part, also divided by disease as defined by organ of origin. This framework has made sense for generations, but the results of molecular analysis are now calling this view into question; cancers of disparate organs have many shared features, whereas, conversely, cancers from the same organ are often quite distinct.

Important similarities among tumor subtypes from different organs have already been identified. For example, TP53 mutations drive high-grade serous ovarian, serous endometrial and basal-like breast carcinomas, all of which share a global transcriptional signature involving the activation of similar oncogenic pathways. Similarly, ERBB2/HER2 is mutated and/or amplified in subsets of glioblastoma, gastric, serous endometrial, bladder and lung cancers. The result, at least in some cases, is responsiveness to HER2-targeted therapy, analogous to that previously observed for HER2-amplified breast cancer. Other commonalities across tumor types include inherited and somatic inactivation of the BRCA1-BRCA2 pathway in both serous ovarian and basal-like breast cancers, microsatellite instability in colorectal and endometrial tumors, and the recently identified POLE-mediated ultramutator phenotype characterized by extremely high mutation rates, common to both colon and endometrial cancers. Conversely, there are important cases in which the same genetic aberrations have very different effects depending on the organ within which they arise. A prime example is provided by the NOTCH gene family, which is inactivated in some squamous cell cancers of the lung, head and neck, skin and cervix but activated by mutation in leukemias.

Such examples illustrate the importance of developing a comprehensive perspective across tumors, independent of histopathologic diagnosis; shared molecular patterns will enable etiologic and therapeutic discoveries in one disease that can be applied to another. Importantly, integrative interpretation of the data will help identify how the consequences of mutations vary across tissues, with important therapeutic implications. Relatively rare cancers, such as childhood malignancies, in particular stand to benefit from such an approach.

We know much more about the molecular details of major cancers than we did just a few years ago, but once a cancer is metastatic it remains incurable with few exceptions. Only time will tell whether the integration of molecular characteristics with data on histology, organ site and metastatic location will contribute to an improvement in patient outcomes. But the balance is shifting in this direction. Hence, the goal of the Pan-Cancer project is to identify and analyze aberrations in the tumor genome and phenotype that define cancer lineages as well as to identify aberrations that transcend particular lineages. This report outlines the scope of the project and introduces the first coordinated set of manuscripts to be published from the enterprise.

The Pan-Cancer project
To gain analytical breadth—defining commonalities, differences and emergent themes across cancer types and organs of origin—TCGA launched the Pan-Cancer analysis project at a meeting held on 26–27 October 2012 in Santa Cruz, California. The Pan-Cancer project is a coordinated initiative whose goals are to assemble coherent, consistent TCGA data sets across tumor types, as well as across platforms, and then to analyze and interpret these data. Within 2 months of the project’s launch, a ‘data freeze’ was declared on the first 12 TCGA tumor types, each profiled using 6 different genomic, epigenomic, transcriptional and proteomic platforms. Since that time, the aggregated data sets have been quality controlled, analyzed statistically and interpreted by a consortium of researchers, principally members of the TCGA Research Network.

The Pan-Cancer project lays the framework for an analytic process that, in the future, will include the integration of new tumor types and data from TCGA and other such enterprises. There are currently major consortium efforts in pediatric cancers (TARGET; Therapeutically
Applicable Research to Generate Effective Treatments) and adult cancers (ICGC; International Cancer Genomics Consortium), as well as smaller projects by research teams around the world. A critical component of such efforts will be the functional validation of aberrations in individual genes in team science efforts such as CTD\(^2\) (Cancer Target Discovery and Development) and the elucidation of pathway and network relationships in programs such as the US National Cancer Institute’s Integrative Cancer Biology Program.

A number of investigations that go beyond the single-tumor perspective are being addressed in the collection of Pan-Cancer manuscripts. Examples of the kinds of questions addressed by these investigations are given below.

Can increases in statistical power help to distinguish new driver mutations from the background of passenger mutations as the sample size is increased by aggregating the 12 tumor types? Assembled Pan-Cancer data have, in fact, enabled the identification of new patterns of genomic drivers. New computational approaches that leverage cross-tumor principles of replication timing and gene expression correlated with background mutation rates now enable the identification of frequently mutated genes while eliminating many false-positive and false-negative calls made in several single-tumor-type projects\(^3\). Further, the power to identify multiple signals of positive selection has increased the ability to distinguish ‘driver’ from ‘passenger’ aberrations\(^4\).

What tissue associations underlie the major genomic structural changes in cancer? Improved methods for the analysis of structural variation of large chromosome segments have refined the ability to identify genomic and epigenetic regulators in multiple peak regions seen only by collating data across different cancer types. Tissue-associated patterns have now been established for the rate and timing of whole-genome duplication events\(^5\).

What pathways emerge as critical and potentially actionable when all mutational events across many tissues are considered together? New classes of mutations, such as those in chromatin-remodeling genes, are emerging as cancer drivers identified only by (i) collecting less frequent events across tumor types, (ii) integrating event types such as mutations, copy number changes and epigenetic silencing, (iii) combining multiple algorithms to identify predicted drivers\(^6\) and (iv) aggregating genes using gene networks and pathways\(^7\).

Can an increase in the number of samples enhance analysis of the co-occurrence and mutual exclusivity of gene aberrations and improve the ability to distinguish driver aberrations from passengers? A bird’s-eye view of genomic and epigenomic events yields a ‘fate map’ of the alternative routes to carcinogenesis in a decision tree that spans tissue boundaries\(^8\).

Can molecular subtypes be delineated to discern tissue-specific from tissue-independent components of disease? Analyses of the epigenome, transcriptome and proteome show a strong influence of tissue on the state of altered pathways in tumor cells. For instance, analysis of the gene expression landscape reinforces the dominant tissue dependence of altered pathways and complements simultaneous profiling of over a hundred proteins important in cancer\(^9\). Using all of the tumor types together allows for any tumor-specific signals to be subtracted from the data sets. Intriguingly, subtracting tissue-specific signal from DNA microarray gene expression data sets identifies signatures of immune stromal influence that transcend tumor type boundaries (R. Verhaak, personal communication). Further, events that are common across lineages become apparent in a cross-tumor analysis\(^10\). Examples are the hormonal dependencies of breast, ovarian and endometrial cancers and a common ‘squamous cell’ signature across head and neck, lung, cervical and bladder cancers.

Which events actionable in one tumor lineage are also actionable in another tumor lineage, potentially increasing the range of indications for specific targeted therapeutics? A systematic evaluation of machine-learning approaches is needed to highlight methodological principles for predicting patient outcomes using integrated information across tissues (H. Liang, personal communication).

Limitations of analysis across tumor types
Several data integration challenges place unavoidable limitations on cross-tumor analysis at the current time. A key challenge is the integration of data that have been generated on different platforms or updates of the same platform, as technologies improve. In the Pan-Cancer studies, for example, there have been transitions to much higher density DNA methylation arrays, use of different exome capture technologies, addition of RNA sequencing to microarray-based RNA
**Commentary**

The first goal of the Pan-Cancer Analysis Working Group was to assemble data from the separate disease projects to build a well-coordinated joint data set spanning multiple tumor types. A data freeze (21 December 2012) based on six different genomic and epigenomic characterization platforms was made available as the pancan12 data set to all analysis groups. Twelve tumor types (GBM, OV, BRCA, LUSC, LUAD, COAD, READ, KIRC, UCEC, BLCA, HNSC and LAML; see Fig. 1 for definitions) were selected on the basis of data maturity, adequate sample size and publication or submission for publication of primary analyses. The Pan-Cancer 12 data set includes a total of 5,074 tumor samples, of which 93% had been assessed for genomic, epigenomic, and gene and protein expression data on at least one platform each (Table 1). The essential purpose of such a joint data set is twofold: to increase the statistical power to detect functional genomic determinants of disease and to identify both tissue-specific aspects of cancer and intrinsic molecular commonalities across tumor types.

The Pan-Cancer analysis project started as an informal collaboration among members of the TCGA Research Network but then quickly expanded to include many other interested researchers. Ensuring standardization and consistency of the data and annotations across multiple platforms and clinical data elements was a necessity for the project. To coordinate analyses across this large group of researchers, formal pipelines were created to establish a coherent working base of data and results. The process of TCGA data generation and Pan-Cancer analysis was as follows (Fig. 2). First, tumor and germline samples were obtained from a large number of tissue source sites and processed by the Biospecimen Core Resource (with sample selection according to criteria established for each tumor type and with extensive quality controls) to generate purified DNA, RNA and protein preparations. The preparations were sent to Genome Characterization Centers (GCCs) and Genome Sequencing Centers (GSCs) for molecular profiling, and the resulting data were deposited at the TCGA Data Coordinating Center (DCC) to provide a primary source of data at four levels of data processing. Seven Genome Data Analysis Centers (GDACs), along with analysts at the GCCs and GSCs and the external research community, shared analysis and interpretation of the data, coordinating activities through face-to-face meetings and regular (usually weekly) teleconferences.

A data freeze was created by pulling higher levels of interpreted data (Level 3) from the DCC into a coordinating repository called Synapse created by Sage Bionetworks. To create a coherent data set, a sample “white list” was created by synchronizing flagged samples with the DCC on the basis of annotations and criteria from the individual disease working groups. The Pan-Cancer project leveraged the TCGA infrastructure for sample acquisition, sample processing and data generation on individual tumor types, as well as for the production of derived data sets and a variety of analysis results assembled in the Broad Institute’s Firehoose system. Assembled robust, self-consistent data sets across all 12 Pan-Cancer tumor types were deposited into Synapse. The Synapse system implements mechanisms for tracking provenance and metadata, stable digital object identifiers (DOI) for data referencing and flexible methods for data access, either through a wiki-like web-based environment or programmatically through application programming interfaces (APIs). The pancan12 data sets and selected results are available under the Synapse accession doi:10.7303/syn300013.

The process of cross-tumor analysis was as follows (Fig. 2). First, a data freeze was created by pulling higher levels of interpreted data (Level 3) from the DCC into a coordinating repository called Synapse created by Sage Bionetworks. To create a coherent data set, a sample “white list” was created by synchronizing flagged samples with the DCC on the basis of annotations and criteria from the individual disease working groups. The Pan-Cancer project leveraged the TCGA infrastructure for sample acquisition, sample processing and data generation on individual tumor types, as well as for the production of derived data sets and a variety of analysis results assembled in the Broad Institute’s Firehoose system. Assembled robust, self-consistent data sets across all 12 Pan-Cancer tumor types were deposited into Synapse. The Synapse system implements mechanisms for tracking provenance and metadata, stable digital object identifiers (DOI) for data referencing and flexible methods for data access, either through a wiki-like web-based environment or programmatically through application programming interfaces (APIs). The pancan12 data sets and selected results are available under the Synapse accession doi:10.7303/syn300013.

**Figure 2** Data coordination for the Pan-Cancer TCGA project. Data were collected by the Biospecimen Collection Resource (BCR) from 12 different tumor types and characterized on 6 major platforms by the Genome Characterization Centers and Genomic Sequencing Centers (GCCs and GSCs). Data sets were deposited in the TCGA Data Coordination Center (DCC) from which they were then distributed to the Broad Institute’s Firehoose and the Memorial Sloan-Kettering Cancer Center’s cBioPortal for various automated processing pipelines. Analysis Working Groups (AWGs) conducted focused analyses on individual tumor types. Results from the DCC, Firehoose and AWGs were collected and stored in Sage Bionetworks’ Synapse database system to create a data freeze. Genome data analysis centers (GDACs) accessed and deposited both data and results through Synapse to coordinate distributed analyses.

**Box 2. Coordination of Data and Results**

The first goal of the Pan-Cancer Analysis Working Group was to assemble data from the separate disease projects to build a well-coordinated joint data set spanning multiple tumor types. A data freeze (21 December 2012) based on six different genomic and epigenomic characterization platforms was made available as the pancan12 data set to all analysis groups. Twelve tumor types (GBM, OV, BRCA, LUSC, LUAD, COAD, READ, KIRC, UCEC, BLCA, HNSC and LAML; see Fig. 1 for definitions) were selected on the basis of data maturity, adequate sample size and publication or submission for publication of primary analyses. The Pan-Cancer 12 data set includes a total of 5,074 tumor samples, of which 93% had been assessed for genomic, epigenomic, and gene and protein expression data on at least one platform each (Table 1). The essential purpose of such a joint data set is twofold: to increase the statistical power to detect functional genomic determinants of disease and to identify both tissue-specific aspects of cancer and intrinsic molecular commonalities across tumor types.

The Pan-Cancer analysis project started as an informal collaboration among members of the TCGA Research Network but then quickly expanded to include many other interested researchers. Ensuring standardization and consistency of the data and annotations across multiple platforms and clinical data elements was a necessity for the project. To coordinate analyses across this large group of researchers, formal pipelines were created to establish a coherent working base of data and results. The process of TCGA data generation and Pan-Cancer analysis was as follows (Fig. 2). First, tumor and germline samples were obtained from a large number of tissue source sites and processed by the Biospecimen Core Resource (with sample selection according to criteria established for each tumor type and with extensive quality controls) to generate purified DNA, RNA and protein preparations. The preparations were sent to Genome Characterization Centers (GCCs) and Genome Sequencing Centers (GSCs) for molecular profiling, and the resulting data were deposited at the TCGA Data Coordinating Center (DCC) to provide a primary source of data at four levels of data processing. Seven Genome Data Analysis Centers (GDACs), along with analysts at the GCCs and GSCs and the external research community, shared analysis and interpretation of the data, coordinating activities through face-to-face meetings and regular (usually weekly) teleconferences.

A data freeze was created by pulling higher levels of interpreted data (Level 3) from the DCC into a coordinating repository called Synapse created by Sage Bionetworks. To create a coherent data set, a sample ‘white list’ was created by synchronizing flagged samples with the DCC on the basis of annotations and criteria from the individual disease working groups. The Pan-Cancer project leveraged the TCGA infrastructure for sample acquisition, sample processing and data generation on individual tumor types, as well as for the production of derived data sets and a variety of analysis results assembled in the Broad Institute’s Firehoose system. Assembled robust, self-consistent data sets across all 12 Pan-Cancer tumor types were deposited into Synapse. The Synapse system implements mechanisms for tracking provenance and metadata, stable digital object identifiers (DOI) for data referencing and flexible methods for data access, either through a wiki-like web-based environment or programmatically through application programming interfaces (APIs). The pancan12 data sets and selected results are available under the Synapse accession doi:10.7303/syn300013.
achieved by cross-cancer comparison does not lead to increased false-negative rates for discovery (for example, by ‘diluting out’ an important mutation specific to one disease) or false-positive rates (for example, by confounding false positives known to result from current single-tumor investigations33).

Tumor lineage has an important role in the observed patterns of co-aberrations and gene expression profiles that indicate different consequences of seemingly similar events, for example, involving the same gene(s) or amplicon(s). Likewise, new methods for accurately probing cross-tumor trends will need to account explicitly for differences across tissues in mutation rates, copy number changes on the focal and arm-level scales, and the prevalence of other co-occurring events in the genetic and epigenetic backgrounds.

Despite these challenges, the collection of Pan-Cancer publications presented here represents a landmark in the continuing effort to understand the common and contrasting biologicals of cancers from a molecular perspective. Still, major questions amenable to further cross-tumor investigations remain (Box 3), and the techniques used to compare different tumors will undoubtedly improve with use, time and further collaborative efforts.

Future directions
The Pan-Cancer project represents one of the first of what will surely be many efforts to coordinate analysis across the molecular landscape of cancer, especially as additional tumor types are investigated in large numbers. Further increasing the number of samples per tumor type and the variety of tumor types will improve our ability to detect rare driver events in heterogeneous tumor samples. But the true power will come from a detailed analysis across tumor types—with links to high-quality clinical outcomes and eventual experimental validation and clinical trials to test the hypotheses that emerge. Technologies such as laser capture microdissection and cell sorting will improve the ability to distinguish whether omic signals arise from malignant or stromal cells. Histone profiling, protein analysis based on mass spectrometry and deconvolution of tumor heterogeneity through single-cell sequencing are examples of technologies expected to add important new dimensions of information. Continued efforts to identify the progenitor

Table 1 Data used by the Pan-Cancer project as defined on 21 December 2012

<table>
<thead>
<tr>
<th>Cancer</th>
<th>RPPAa</th>
<th>DNA methylationb</th>
<th>Copy numberc</th>
<th>Mutationd</th>
<th>microRNAs</th>
<th>Expressionf</th>
</tr>
</thead>
<tbody>
<tr>
<td>LUSC</td>
<td>195</td>
<td>358</td>
<td>345</td>
<td>178</td>
<td>332</td>
<td>227</td>
</tr>
<tr>
<td>READ</td>
<td>130</td>
<td>162</td>
<td>164</td>
<td>69</td>
<td>143</td>
<td>71</td>
</tr>
<tr>
<td>GBM</td>
<td>214</td>
<td>405</td>
<td>578</td>
<td>290</td>
<td>501</td>
<td>495</td>
</tr>
<tr>
<td>LAML</td>
<td>NA</td>
<td>194</td>
<td>198</td>
<td>197</td>
<td>187</td>
<td>179</td>
</tr>
<tr>
<td>HNSC</td>
<td>212</td>
<td>310</td>
<td>310</td>
<td>277</td>
<td>309</td>
<td>303</td>
</tr>
<tr>
<td>BLCA</td>
<td>54</td>
<td>126</td>
<td>126</td>
<td>99</td>
<td>121</td>
<td>96</td>
</tr>
<tr>
<td>KIRC</td>
<td>423</td>
<td>457</td>
<td>457</td>
<td>417</td>
<td>442</td>
<td>431</td>
</tr>
<tr>
<td>UCEC</td>
<td>200</td>
<td>512</td>
<td>511</td>
<td>248</td>
<td>497</td>
<td>333</td>
</tr>
<tr>
<td>LUAD</td>
<td>237</td>
<td>431</td>
<td>357</td>
<td>229</td>
<td>365</td>
<td>355</td>
</tr>
<tr>
<td>OV</td>
<td>332</td>
<td>592</td>
<td>577</td>
<td>316</td>
<td>454</td>
<td>581</td>
</tr>
<tr>
<td>BRCA</td>
<td>408</td>
<td>888</td>
<td>887</td>
<td>772</td>
<td>870</td>
<td>817</td>
</tr>
<tr>
<td>COAD</td>
<td>269</td>
<td>420</td>
<td>422</td>
<td>155</td>
<td>407</td>
<td>192</td>
</tr>
<tr>
<td>Total</td>
<td>2,674</td>
<td>4,855</td>
<td>4,932</td>
<td>3,247</td>
<td>4,628</td>
<td>4,080</td>
</tr>
</tbody>
</table>

Tabulated are the numbers of unique tumor samples available for each tumor type (rows) and each measurement platform (columns). NA, not available.

aReverse-phase protein arrays measuring protein and phosphoprotein abundance. bDNA methylation at CpG islands. cMicroarray-based measurement of copy number. dSamples subjected to whole-exome sequencing to determine single-nucleotide and structural variants. eSequencing of microRNAs. fRNA sequencing and microarray gene expression analysis.

BOX 3 EXAMPLES OF ADDITIONAL MAJOR QUESTIONS AMENABLE TO FURTHER PAN-CANCER ANALYSES

- What is the spectrum of nucleotide- and dinucleotide-level changes associated with different carcinogenic etiologies (for example, tobacco, pathogens or inflammation) operating in different parts of the body?
- Will integration of additional data sources, including additional tumor types from TCGA and other projects, increase the power of analysis in useful ways?
- How can characterization based on molecular changes complement pathological analysis for classification of cancers into tumor lineages with potentially different clinical management?
- Can molecular profiles effectively categorize cancers for therapeutic decision-making?
- Are there predictive expression-based signatures for genomic events that transcend tissues, reflecting pathways disrupted by the alterations?
- Will comprehensive protein analysis through emerging mass spectrometry approaches in the Clinical Proteomic Tumor Analysis Consortium (CPTAC) and other efforts extend the power of the genomic, transcriptomic and proteomic analyses in TCGA?
- Will whole genome analysis demonstrate the influence of mobile elements, mutations in non-coding regions and connections to constitutional risk-associated loci?
- How are changes in protein families distributed across different tumor types?
- Are aberrations in specific protein domains or pathways distributed differentially across tumor lineages?
- Beyond the known examples, including in cervical, head and neck, esophageal and hepatocellular cancers, can we identify other cancer types that show virally mediated initiation?
- Are bacteria associated with different cancer lineages (as Fusobacteria are in colorectal cancer33)?
- Can the answers to any of these questions help in the design of novel therapies and clinical trials, with the ultimate goal of improving patient outcomes?
cells of tumors will enable universal properties to be distinguished from parochial ones. Clone-level and other types of studies may identify even more connections among tumor types. Longitudinal genomic studies on primary resected tumors paired with their local recurrences and/or metastases will be undertaken by large consortium efforts, which have heretofore been restricted to primary disease and have lacked information about response to treatment. The characteristics of primary tumors may change markedly when they metastasize to distant sites, particularly bone and brain. Analysis of metastasis across tumor types will therefore be highly informative.

The power of cross-tumor analysis will increase as technologies for monitoring individual tumor cells at high resolution come into play. Now that the price of genome sequencing has fallen, the next Pan-Cancer enterprise will be able to analyze large numbers of whole-genome sequences across tumor types. Whole-genome analysis will complement the current studies by shedding light on mutational processes in the noncoding parts of the genome, which have not been as well explored so far. This expanded analysis will bring focus only over time and with broader efforts. Still, the collection of TCGA Pan-Cancer publications represents a significant contribution to a new period of discovery in cancer research.

ACKNOWLEDGMENTS
We thank J. Zhang for the administrative coordination of TCGA Pan-Cancer Analysis Working Group activities, C. Perou and K. Hoadley for contributions to Figure 1, and A. Margolin, D. Wheeler, M. Meyerson and L. Ding for comments on early drafts of the manuscript. The study was funded by the National Cancer Institute and the National Human Genome Research Institute.

COMPETING FINANCIAL INTERESTS
The authors declare no competing financial interests.

From many tumors to the individual
The hope is that investigations across tumor type such as the Pan-Cancer project will ultimately inform clinical decision-making. We hope such studies will enable the discovery of novel therapeutic agents that can be tested clinically—perhaps in novel adaptive, biomarker-based clinical trials that cross boundaries between tumor types. Toward this end, TCGA Pan-Cancer data sets have been made available publicly in one location. Although coordination remains a challenge, the data sets comprise an unequalled resource for the integrative analysis of cancer in its many forms.

A key challenge is the development of clinical trial strategies for connecting subsets of tumors from different tissues in terms of molecular signatures. Recent analyses of pharmacological profiling experiments across a diverse panel of cancer cell lines has suggested that common genetic alterations can sometimes predict response to therapy across multiple cell lineages.4,39–42. Biomarker-based design of clinical trials can increase statistical power, greatly decreasing the size, expense and duration of clinical trials.

The number and size of omic data sets on cancer available to the research community for mining and exploring continue to expand rapidly, and computational tools to derive insights into the fundamental causes of cancer are becoming more powerful. It is important to note that the full potential of the enterprise will be realized only over time and with broader efforts. Still, the collection of TCGA Pan-Cancer publications represents a significant contribution to a new period of discovery in cancer research.


Genome Data Analysis Center: Gordon Sakse16, Scott L Carter16, Kristian Cibulskis16, Lynda Chinn16,22, Jianhua Zhang22,24, Gad Getz16,41,42, Huyen Dinh12, Harsha Vardhan Diddaparen12, Richard Gibbs12, Preethi Gunaratne12,13, Yi Han12, Divya Kalra12, Christine Kovar12,13, Lora Lewis12, Margaret Morgan12, Donna Morton12, Donna Muzzy12, Jeffrey Reid12, Liu Xi12, Juok Cho16, Daniel DiCara16, Scott Frazer16, Nils Gehlenborg16,24, David J Heiman16,17, Jaegil Kim16, Michael S Lawrence16,17, Pei Lin16, Yingchun Liu16, Michael S Noble16, Petar Stojanov16,24,17, Doug Voet16, Hailei Zhang16, Lihua Zou16,17, Chip Stewart16, Brady Bernard10, Ryan Bressler10, Andrea Eakin10, Lisa Iype10, Theo Knijnenburg10, Roger Kramer10, Richard Kreisberg10, Kalle Leinonen10, Jake Lin10, Yuexin Liu18, Michael Miller10, Sheila M Reynolds10, Hector Rovira10, Ilya Shmulevich10, Vesteinn Thorro10, Da Yang44, Wei Zhang44, Samir Kumar Am14, Chang-Jian Wu22, Chia-Chin Wu22, Rehan Akbani2, Kenneth Aldape16,17, Keith A Baggerly2, Bradley Broom2, Tod D Casasent2, James Cleland2,4, Chad Creighton7, Deepi Dedota4, Mary Edgerton4, Leng Han2, Shelley M Herbrich2, Zhenlin Ju2, Hoon Kim2, Seth Lerner4, Jun Li2, Han Liang2, Wenbin Liu2, Philip L Lorenzi2, Yiling Lu3, James Melott2, Gordon B Mills2,17, Ilam Nguyen2,4, Xiaoping Su2, Roeland Verhaak2,17, Wenyi Wang2, John N Weinstein2,17, Andrew Wong2,4,6, Yang Yang2,5, Jun Yao2,17, Rong Yao2, Kosuke Yoshihara2, Yuan Yuan2,17, Alfred K Yung1, Nianxiang Zhang2, Siyuan Zheng2, Michael Ryan3,4,6, David W Kane2,53, B Arman Aksoy11, Giovanni Cirillo11, Gideon Dresdner11, Jianjiong Gao11, Benjamin Gross11, Anders Jacobsen11, Andre Kahles11, Marc Ladanyi15,44, William Lee11, Kjong-Van Lehmann11, Martin L Miller11, Ricardo Ramirez11, Gunnars Rätsch11, Boris Revia11, Chris Sanders11, Nikolaus Schultz11, Yasin Senbabaoglu11, Ronglai Shen56, Rileen Sinha11, S Onur Sumer11, Yichao Sun11, Barry S Taylor4,56,57, Nils Weinhold11, Suzanne Fei58, Paul Spellman58, Christopher Benz59, Daniel Carlin8,9, Melissa Cline8,9, Brian Craft8,9, Kyle Ellrott8,9, Mary Goldman8,9, David Haussler8,9,60, Singer Ma8,9, Sam Ng8,9, Evan Paull8,9, Amie Radenbaugh8,9, Sofie Salama8,9,60, Artem Sokolov8,9, Joshua M Stuart8,9, Teresa Swatloski8,9, Vladislav Uzunangelov8,9, Peter Waltman8,9, Christina Yau59, Jing Zhu8,9 & Stanley R Hamilton44
