The cancer genome atlas pan-cancer analysis project

OPEN

The Cancer Genome Atlas Pan-Cancer analysis project

The Cancer Genome Atlas Research Network[1], John N Weinstein[2,3], Eric A Collisson[4], Gordon B Mills[3], Kenna R Mills Shaw[5,6], Brad A Ozenberger[7], Kyle Ellrott[8,9], Ilya Shmulevich[10], Chris Sander[11] & Joshua M Stuart[8,9]

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. Although many genomic events with direct phenotypic impact have been identified, much of the complex molecular landscape remains incompletely charted for most cancer lineages.

[1]Full lists of members and affiliations appear at the end of the paper. [2]Department of Bioinformatics and Computational Biology, University of Texas MD Anderson Cancer Center, Houston, Texas, USA. [3]Department of Systems Biology, University of Texas MD Anderson Cancer Center, Houston, Texas, USA. [4]Department of Medicine, University of California, San Francisco, San Francisco, California, USA. [5]The Cancer Genome Atlas Program Office, Center for Cancer Genomics, National Cancer Institute, Bethesda, Maryland, USA. [6]University of Texas MD Anderson Cancer Center, Institute for Personalized Cancer Therapy, Houston, Texas, USA. [7]National Human Genome Research Institute, US National Institutes of Health, Bethesda, Maryland, USA. [8]Department of Biomolecular Engineering, University of California, Santa Cruz, Santa Cruz, California, USA. [9]Center for Biomolecular Science and Engineering, University of California, Santa Cruz, Santa Cruz, California, USA. [10]Institute for Systems Biology, Seattle, Washington, USA. [11]Computational Biology Center, Memorial Sloan-Kettering Cancer Center, New York, New York, USA. Correspondence should be addressed to J.M.S. (jstuart@ucsc.edu).

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[1–3], and a subset of tumor suppressor genes was identified by analyzing loss of heterozygosity[4]. 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[5–7], has established definitions of molecular subtypes[8–12] and has identified new biomarkers on the basis of genomic, transcriptomic, proteomic and epigenomic alterations. Some of these biomarkers have clinical implications[13,14]. 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 $BRAF^{V600}$ mutations[15]; and the fraction of lung cancers treated with targeted agents is increasing with the discovery of likely driver aberrations in most lung tumors[16,17]. Large-scale processes that shape cancer genomes have similarly been identified. Analyses of chromothripsis[18] and chromoplexy[19], which involve the breakage and rearrangement of chromosomes at multiple loci, and kataegis[20], which involves hypermutational processes associated with genomic rearrangements, are providing insights into tumor evolution (see Garraway and Lander[21] 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[22]. However, 'long tails' in the distributions of aberrations among samples have also been uncovered[23]. Indeed, a majority of the TCGA samples have distinct alterations not shared with other samples in their

BOX 1 TCGA: MISSION AND STRATEGY

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[8–12,14,27].

cohort. Despite the apparent uniqueness of each individual tumor in this regard, the set of molecular aberrations often integrates into known biological pathways that are shared by sets of tumor samples. In other cases, rare somatic mutations can be implicated as drivers by aggregating events across tumor types to improve the detection of patterns, for example, hotspot mutations in DNA segments that encode particular protein domains, leading to the identification of potential new drug targets.

Determining whether rare aberrations are drivers (oncogenic contributors) or just passengers (clonally propagated 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[22] 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[24,25], until recently, most research on the molecular, pathological and clinical natures of cancers has been 'siloed' by tumor type[26]. 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[10,27]. 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[12,27,28]. 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[29], skin[30] and cervix[31] but activated by mutation in leukemias[32].

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 (**Box 2**). 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 (**Fig. 1**). 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[33]. Further, the power to identify multiple signals of positive selection has increased the ability to distinguish 'driver' from 'passenger' aberrations[34].

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[35].

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[34] and (iv) aggregating genes using gene networks and pathways[36].

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[37].

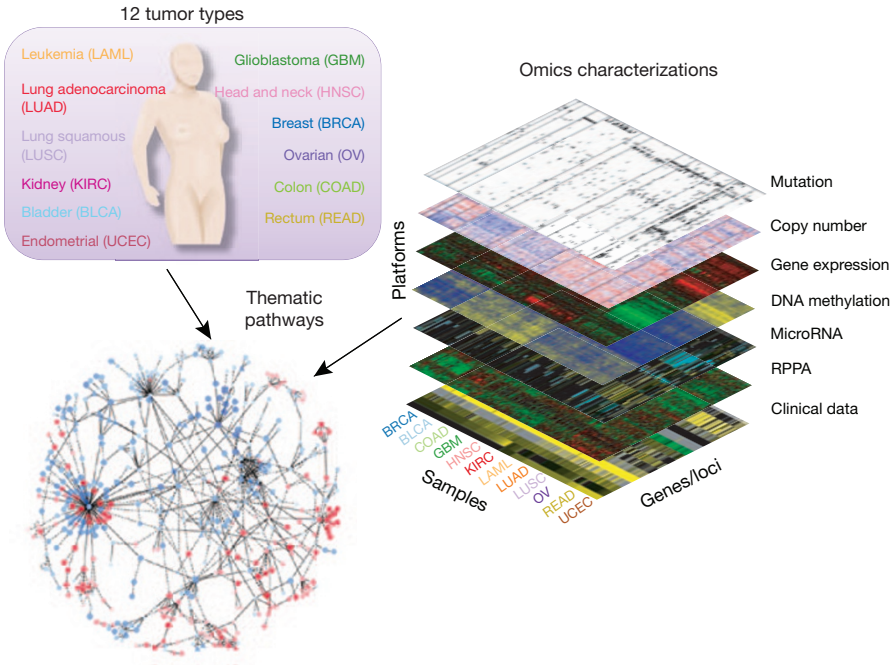

Figure 1 Integrated data set for comparing and contrasting multiple tumor types. The TCGA Pan-Cancer project assembled data from thousands of patients with primary tumors occurring in different sites of the body, covering 12 tumor types (top left) including glioblastoma multiformae (GBM), lymphoblastic acute myeloid leukemia (LAML), head and neck squamous carcinoma (HNSC), lung adenocarcinoma (LUAD), lung squamous carcinoma (LUSC), breast carcinoma (BRCA), kidney renal clear-cell carcinoma (KIRC), ovarian carcinoma (OV), bladder carcinoma (BLCA), colon adenocarcinoma (COAD), uterine cervical and endometrial carcinoma (UCEC) and rectal adenocarcinoma (READ). Six types of omics characterization were performed creating a 'data stack' (right) in which data elements across the platforms are linked by the fact that the same samples were used for each, thus maximizing the potential of integrative analysis. Use of the data enables the identification of general trends, including common pathways (bottom left), revealing master regulatory hubs activated (red) or deactivated (blue) across different tissue types.

Can molecular subtypes be delineated to disentangle 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[38]. 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[38]. 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

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 Firehose 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 (DOIs) 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.

characterization and increases in the quality and number of antibodies available for reverse-phase proteomic arrays (RPPAs). A series of analyses of batch effects has been carried out to assess systematic and platform-specific biases (R. Akbani, personal communication). However, more work is needed to establish best practices for minimizing unwanted batch effects while preserving biological signals.

The nature and quality of available clinical data vary widely by cancer type. Differences in these data limit the ability to establish one-size-fits-all norms for the comparison of demographic information, histopathologic characterization, behavioral context and clinical outcomes. For example, the Pan-Cancer survival data are relatively robust for serous ovarian cancer because of its poor prognosis but are still immature for breast and endometrial cancers, as (thankfully) most patients with these cancers do better for longer periods of time. Certain data elements are routinely collected only when they are anticipated to be relevant (for example, the smoking history of patients with lung, bladder and head and neck cancers). Clear viral etiologies have been identified in several solid tumor types, including head and neck cancer, cervical cancer, Kaposi's sarcoma and hepatocellular carcinoma. However, a pan-cancer analysis of the infectious etiologies of other cancers could not be conducted at present because infection status was recorded for only some tumors and tumor types (as an optional data element). Finally, tumor stage and grade are not easily comparable across different tumor types because, for good reason, each tumor type has its own system. This set of challenges to cross-tumor analysis highlights the fact that current clinical practice is largely conducted according to classification by tissue or organ.

Statistically speaking, care must be taken to ensure that the increased sample size

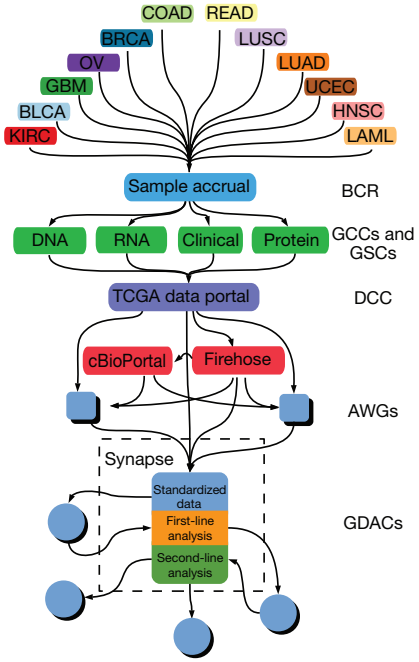

Figure 2 Data coordination for the Pan-Cancer TCGA project. Data were collected by the Biospecimen Collection Resource (BCR) from 12 different tumor types and characterized on 6 major platforms by the Genome Characterization Centers and Genomic Sequencing Centers (GCCs and GSCs). Data sets were deposited in the TCGA Data Coordination Center (DCC) from which they were then distributed to the Broad Institute's Firehose 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, Firehose 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.

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

Cancer	RPPA[a]	DNA methylation[b]	Copy number[c]	Mutation[d]	microRNA[e]	Expression[f]
LUSC	195	358	345	178	332	227
READ	130	162	164	69	143	71
GBM	214	405	578	290	501	495
LAML	NA	194	198	197	187	179
HNSC	212	310	310	277	309	303
BLCA	54	126	126	99	121	96
KIRC	423	457	457	417	442	431
UCEC	200	512	511	248	497	333
LUAD	237	431	357	229	365	355
OV	332	592	577	316	454	581
BRCA	408	888	887	772	870	817
COAD	269	420	422	155	407	192
Total	2,674	4,855	4,932	3,247	4,628	4,080

Tabulated are the numbers of unique tumor samples available for each tumor type (rows) and each measurement platform (columns). NA, not available.
[a]Reverse-phase protein arrays measuring protein and phosphoprotein abundance. [b]DNA methylation at CpG islands. [c]Microarray-based measurement of copy number. [d]Samples subjected to whole-exome sequencing to determine single-nucleotide and structural variants. [e]Sequencing of microRNAs. [f]RNA sequencing and microarray gene expression analysis.

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 compounding false positives known to result from current single-tumor investigations[33]).

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 biologies 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

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 cancer[43])?
- 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 to disruptions in promoter and enhancer sites and aberrations in noncoding RNAs, as well as the genomic integration processes at work in tumor evolution that result from mobile endogenous and exogenous DNA elements such as retrotransposons and viruses. Whole-genome sequencing will create a backdrop against which genome-wide association studies can relate inherited predispositions to particular forms of cancer. Systems-oriented approaches, based on relevant pathways and networks, will add to the therapeutic opportunities that arise from the wealth of data. Experimental follow-up will be critical to assess the functional consequences and therapeutic liabilities of these new findings.

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[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.

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.

1. Soda, M. *et al. Nature* **448**, 561–566 (2007).
2. Parada, L.F., Tabin, C.J., Shih, C. & Weinberg, R.A. *Nature* **297**, 474–478 (1982).
3. Payne, G.S., Bishop, J.M. & Varmus, H.E. *Nature* **295**, 209–214 (1982).
4. Baker, S.J. *et al. Science* **244**, 217–221 (1989).
5. Tomlins, S.A. *et al. Science* **310**, 644–648 (2005).
6. Davies, H. *et al. Nature* **417**, 949–954 (2002).
7. Mardis, E.R. *et al. N. Engl. J. Med.* **361**, 1058–1066 (2009).
8. Cancer Genome Atlas Research Network. *Nature* **455**, 1061–1068 (2008).
9. Cancer Genome Atlas Research Network. *Nature* **474**, 609–615 (2011).
10. Cancer Genome Atlas Network. *Nature* **490**, 61–70 (2012).
11. Cancer Genome Atlas Research Network. *Nature* **489**, 519–525 (2012).
12. Cancer Genome Atlas Network. *Nature* **487**, 330–337 (2012).
13. Perou, C.M. *et al. Nature* **406**, 747–752 (2000).
14. Cancer Genome Atlas Research Network. *Nature* **499**, 43–49 (2013).
15. Chapman, P.B. *et al. N. Engl. J. Med.* **364**, 2507–2516 (2011).
16. Paez, J.G. *et al. Science* **304**, 1497–1500 (2004).
17. Takeuchi, K. *et al. Nat. Med.* **18**, 378–381 (2012).
18. Stephens, P.J. *et al. Cell* **144**, 27–40 (2011).
19. Baca, S.C. *et al. Cell* **153**, 666–677 (2013).
20. Alexandrov, L.B. *et al. Nature* **500**, 415–421 (2013).
21. Garraway, L.A. & Lander, E.S. *Cell* **153**, 17–37 (2013).
22. Vogelstein, B. *et al. Science* **339**, 1546–1558 (2013).
23. Wheeler, D.A. & Wang, L. *Genome Res.* **23**, 1054–1062 (2013).
24. Hanahan, D. & Weinberg, R.A. *Cell* **144**, 646–674 (2011).
25. Hanahan, D. & Weinberg, R.A. *Cell* **100**, 57–70 (2000).
26. McDermott, U. & Settleman, J. *J. Clin. Oncol.* **27**, 5650–5659 (2009).
27. Kandoth, C. *et al. Nature* **497**, 67–73 (2013).
28. Palles, C. *et al. Nat. Genet.* **45**, 136–144 (2013).
29. Stransky, N. *et al. Science* **333**, 1157–1160 (2011).
30. Wang, N.J. *et al. Proc. Natl. Acad. Sci. USA* **108**, 17761–17766 (2011).
31. Zagouras, P., Stifani, S., Blaumueller, C.M., Carcangiu, M.L. & Artavanis-Tsakonas, S. *Proc. Natl. Acad. Sci. USA* **92**, 6414–6418 (1995).
32. Weng, A.P. *et al. Science* **306**, 269–271 (2004).
33. Lawrence, M.S. *et al. Nature* **499**, 214–218 (2013).
34. Tamborero, D. *et al. Sci. Rep.* doi:10.1038/srep02650 (26 September 2013).
35. Zack, T.I. *et al. Nat. Genet.* doi:10.1038/ng.2760 (26 September 2013).
36. Hofree, M., Shen, J.P., Carter, H., Gross, A. & Ideker, T. *Nat. Methods* doi:10.1038/nmeth.2651 (15 September 2013).
37. Ciriello, G. *et al. Nat. Genet.* doi:10.1038/ng.2762 (26 September 2013).
38. Li, J. *et al. Nat. Methods* doi:10.1038/nmeth.2650 (15 September 2013).
39. Barretina, J. *et al. Nature* **483**, 603–607 (2012).
40. Garnett, M.J. *et al. Nature* **483**, 570–575 (2012).
41. Weinstein, J.N. *Nature* **483**, 544–545 (2012).
42. Heiser, L.M. *et al. Proc. Natl. Acad. Sci. USA* **109**, 2724–2729 (2012).
43. Kostic, A.D. *et al. Genome Res.* **22**, 292–298 (2012).

Genome Characterization Center: Kyle Chang[12], Chad J Creighton[12], Caleb Davis[12], Lawrence Donehower[12], Jennifer Drummond[12], David Wheeler[12], Adrian Ally[13], Miruna Balasundaram[13], Inanc Birol[13–15], Yaron S N Butterfield[13], Andy Chu[13], Eric Chuah[13], Hye-Jung E Chun[13], Noreen Dhalla[13], Ranabir Guin[13], Martin Hirst[13], Carrie Hirst[13], Robert A Holt[13], Steven J M Jones[13], Darlene Lee[13], Haiyan I Li[13], Marco A Marra[13], Michael Mayo[13], Richard A Moore[13], Andrew J Mungall[13], A Gordon Robertson[13], Jacqueline E Schein[13], Payal Sipahimalani[13], Angela Tam[13], Nina Thiessen[13], Richard J Varhol[13], Rameen Beroukhim[16], Ami S Bhatt[16,17], Angela N Brooks[16,18], Andrew D Cherniack[16], Samuel S Freeman[16], Stacey B Gabriel[16], Elena Helman[16,19], Joonil Jung[16], Matthew Meyerson[16,17], Akinyemi I Ojesina[16,17], Chandra Sekhar Pedamallu[16,17], Gordon Saksena[16], Steven E Schumacher[16,20], Barbara Tabak[16,20], Travis Zack[16,21], Eric S Lander[16], Christopher A Bristow[22], Angela Hadjipanayis[23], Psalm Haseley[24], Raju Kucherlapati[25], Semin Lee[24], Eunjung Lee[24], Lovelace J Luquette[24], Harshad S Mahadeshwar[22], Angeliki Pantazi[23], Michael Parfenov[23], Peter J Park[24,26,27], Alexei Protopopov[22], Xiaojia Ren[23], Netty Santoso[23], Jonathan Seidman[23], Sahil Seth[22], Xingzhi Song[22], Jiabin Tang[22], Ruibin Xi[24,28,29], Andrew W Xu[24], Lixing Yang[24], Dong Zeng[22], J Todd Auman[30], Saianand Balu[31], Elizabeth Buda[32], Cheng Fan[31], Katherine A Hoadley[31], Corbin D Jones[32,33], Shaowu Meng[31], Piotr A Mieczkowski[34], Joel S Parker[31,34], Charles M Perou[31,34,35], Jeffrey Roach[36], Yan Shi[31], Grace O Silva[31,34], Donghui Tan[34], Umadevi Veluvolu[34], Scot Waring[31,32], Matthew D Wilkerson[34], Junyuan Wu[31], Wei Zhao[31,34], Tom Bodenheimer[31], D Neil Hayes[31,37], Alan P Hoyle[31], Stuart R Jeffreys[31], Lisle E Mose[31], Janae V Simons[31], Mathew G Soloway[31], Stephen B Baylin[38], Benjamin P Berman[39], Moiz S Bootwalla[39], Ludmila Danilova[40], James G Herman[38], Toshinori Hinoue[39], Peter W Laird[39], Suhn K Rhie[39], Hui Shen[39], Timothy Triche Jr[39], Daniel J Weisenberger[39], Scott L Carter[16], Kristian Cibulskis[16], Lynda Chin[16,22], Jianhua Zhang[22], Gad Getz[16,41,42], Carrie Sougnez[16] & Min Wang[12]

Genome Data Analysis Center: Gordon Saksena[16], Scott L Carter[16], Kristian Cibulskis[16], Lynda Chin[16,22], Jianhua Zhang[22], Gad Getz[16,41,42], Huyen Dinh[12], Harsha Vardhan Doddapaneni[12], Richard Gibbs[12], Preethi Gunaratne[12,43], Yi Han[12], Divya Kalra[12], Christie Kovar[12], Lora Lewis[12], Margaret Morgan[12], Donna Morton[12], Donna Muzny[12], Jeffrey Reid[12], Liu Xi[12], Juok Cho[16], Daniel DiCara[16], Scott Frazer[16], Nils Gehlenborg[16,24], David I Heiman[16], Jaegil Kim[16], Michael S Lawrence[16], Pei Lin[16], Yingchun Liu[16], Michael S Noble[16], Petar Stojanov[16,17], Doug Voet[16], Hailei Zhang[16], Lihua Zou[16], Chip Stewart[16], Brady Bernard[10], Ryan Bressler[10], Andrea Eakin[10], Lisa Iype[10], Theo Knijnenburg[10], Roger Kramer[10], Richard Kreisberg[10], Kalle Leinonen[10], Jake Lin[10], Yuexin Liu[48], Michael Miller[10], Sheila M Reynolds[10], Hector Rovira[10], Ilya Shmulevich[10], Vesteinn Thorsson[10], Da Yang[44], Wei Zhang[44], Samirkumar Amin[45], Chang-Jiun Wu[22], Chia-Chin Wu[22], Rehan Akbani[2], Kenneth Aldape[46], Keith A Baggerly[2], Bradley Broom[2], Tod D Casasent[2], James Cleland[2,46], Chad Creighton[47], Deepti Dodda[2], Mary Edgerton[48], Leng Han[2], Shelley M Herbrich[2], Zhenlin Ju[2], Hoon Kim[2], Seth Lerner[49], Jun Li[2], Han Liang[2], Wenbin Liu[2], Philip L Lorenzi[2], Yiling Lu[3], James Melott[2], Gordon B Mills[3], Lam Nguyen[2,46], Xiaoping Su[2], Roeland Verhaak[2], Wenyi Wang[2], John N Weinstein[2,3], Andrew Wong[2,46], Yang Yang[2,50], Jun Yao[51], Rong Yao[2], Kosuke Yoshihara[2], Yuan Yuan[2,52], Alfred K Yung[51], Nianxiang Zhang[2], Siyuan Zheng[2], Michael Ryan[3,46], David W Kane[2,53], B Arman Aksoy[11], Giovanni Ciriello[11], Gideon Dresdner[11], Jianjiong Gao[11], Benjamin Gross[11], Anders Jacobsen[11], Andre Kahles[11], Marc Ladanyi[54,55], William Lee[11], Kjong-Van Lehmann[11], Martin L Miller[11], Ricardo Ramirez[11], Gunnar Rätsch[11], Boris Reva[11], Chris Sander[11], Nikolaus Schultz[11], Yasin Senbabaoglu[11], Ronglai Shen[56], Rileen Sinha[11], S Onur Sumer[11], Yichao Sun[11], Barry S Taylor[4,56,57], Nils Weinhold[11], Suzanne Fei[58], Paul Spellman[58], Christopher Benz[59], Daniel Carlin[8,9], Melisssa Cline[8,9], Brian Craft[8,9], Kyle Ellrott[8,9], Mary Goldman[8,9], David Haussler[8,9,60], Singer Ma[8,9], Sam Ng[8,9], Evan Paull[8,9], Amie Radenbaugh[8,9], Sofie Salama[8,9,60], Artem Sokolov[8,9], Joshua M Stuart[8,9], Teresa Swatloski[8,9], Vladislav Uzunangelov[8,9], Peter Waltman[8,9], Christina Yau[59], Jing Zhu[8,9] & Stanley R Hamilton[44]

Sequencing Center: Gad Getz[16,41,42], Carrie Sougnez[16], Scott Abbott[61], Rachel Abbott[61], Nathan D Dees[61], Kim Delehaunty[61], Li Ding[61–63], David J Dooling[61], Jim M Eldred[61], Catrina C Fronick[61], Robert Fulton[61], Lucinda L Fulton[61], Joelle Kalicki-Veizer[61], Krishna-Latha Kanchi[61], Cyriac Kandoth[61], Daniel C Koboldt[61], David E Larson[61], Timothy J Ley[61,64], Ling Lin[61], Charles Lu[61], Vincent J Magrini[61], Elaine R Mardis[61,63,65], Michael D McLellan[61], Joshua F McMichael[61], Christopher A Miller[61], Michelle O'Laughlin[61], Craig Pohl[61], Heather Schmidt[61], Scott M Smith[61], Jason Walker[61], John W Wallis[61], Michael C Wendl[61,65,66], Richard K Wilson[61,63,65], Todd Wylie[61] & Qunyuan Zhang[61,65]

Data Coordinating Center: Robert Burton[67], Mark A Jensen[53], Ari Kahn[53], Todd Pihl[53], David Pot[53] & Yunhu Wan[53]

Tissue Source Site: Douglas A Levine[68]

Biospecimen Core Resource Center: Aaron D Black[69], Jay Bowen[69], Jessica Frick[69], Julie M Gastier-Foster[69,70], Hollie A Harper[69], Carmen Helsel[69], Kristen M Leraas[69], Tara M Lichtenberg[69], Cynthia McAllister[69], Nilsa C Ramirez[69,70], Samantha Sharpe[69], Lisa Wise[69] & Erik Zmuda[69]

National Cancer Institute/National Human Genome Research Institute Project Team: Stephen J Chanock[71], Tanja Davidsen[71], John A Demchok[71], Greg Eley[72], Ina Felau[71], Brad A Ozenberger[7], Margi Sheth[71], Heidi Sofia[7], Louis Staudt[71], Roy Tarnuzzer[71], Zhining Wang[71], Liming Yang[71] & Jiashan Zhang[71]

Collaborators: Larsson Omberg[73], Adam Margolin[73], Benjamin J Raphael[74], Fabio Vandin[74], Hsin-Ta Wu[74], Mark D M Leiserson[74], Stephen C Benz[75], Charles J Vaske[75], Houtan Noushmehr[76,77], Theo Knijnenburg[10], Denise Wolf[78], Laura Van 't Veer[78], Eric A Collisson[4], Dimitris Anastassiou[79], Tai-Hsien Ou Yang[79], Nuria Lopez-Bigas[80,81], Abel Gonzalez-Perez[80], David Tamborero[80], Zheng Xia[82], Wei Li[82], Dong-Yeon Cho[83], Teresa Przytycka[83], Mark Hamilton[84], Sean McGuire[84], Sven Nelander[85,86], Patrik Johansson[85,86], Rebecka Jörnsten[87], Teresia Kling[85,86] & Jose Sanchez[87]

[12]Human Genome Sequencing Center, Baylor College of Medicine, Houston, Texas, USA. [13]Canada's Michael Smith Genome Sciences Centre, BC Cancer Agency, Vancouver, British Columbia, Canada. [14]Department of Medical Genetics, University of British Columbia, Vancouver, British Columbia, Canada. [15]School of Computing Science, Simon Fraser University, Burnaby, British Columbia, Canada. [16]The Eli and Edythe L. Broad Institute of the Massachusetts Institute of Technology and Harvard University, Cambridge, Massachusetts, USA. [17]Department of Medical Oncology, Dana-Farber Cancer Institute, Boston, Massachusetts, USA. [18]Dana-Farber Cancer Institute, Boston, Massachusetts, USA. [19]Harvard-MIT Division of Health Sciences & Technology, Cambridge, Massachusetts, USA. [20]Department of Cancer Biology, Dana-Farber Cancer Institute, Boston, Massachusetts, USA. [21]Biophysics Program, Harvard University, Boston, Massachusetts, USA. [22]Institute for Applied Cancer Science, Department of Genomic Medicine, University of Texas MD Anderson Cancer Center, Houston, Texas, USA. [23]Department of Genetics, Harvard Medical School, Boston, Massachusetts, USA. [24]The Center for Biomedical Informatics, Harvard Medical School, Boston, Massachusetts, USA. [25]Harvard Medical School–Partners HealthCare Center for Genetics and Genomics, Boston, Massachusetts, USA. [26]Division of Genetics, Brigham and Women's Hospital, Boston, Massachusetts, USA. [27]Informatics Program, Children's Hospital, Boston, Massachusetts, USA. [28]School of Mathematical Sciences, Peking University, Beijing, China. [29]Center for Statistical Science, Peking University, Beijing, China. [30]Eshelman School of Pharmacy, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina, USA. [31]Lineberger Comprehensive Cancer Center, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina, USA. [32]Carolina Center for Genome Sciences, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina, USA. [33]Department of Biology, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina, USA. [34]Department of Genetics, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina, USA. [35]Department of Pathology and Laboratory Medicine, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina, USA. [36]Research Computing Center, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina, USA. [37]Department of Internal Medicine, Division of Medical Oncology, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina, USA. [38]Cancer Biology Division, The Sidney Kimmel Comprehensive Cancer Center at Johns Hopkins University, Baltimore, Maryland, USA. [39]USC Epigenome Center, University of Southern California Keck School of Medicine, Los Angeles, California, USA. [40]The Sidney Kimmel Comprehensive Cancer Center at Johns Hopkins University, Baltimore, Maryland, USA. [41]Cancer Center, Massachusetts General Hospital, Boston, Massachusetts, USA. [42]Department of Pathology, Massachusetts General Hospital, Boston, Massachusetts, USA. [43]Department of Biology and Biochemistry, University of Houston, Houston, Texas, USA. [44]University of Texas MD Anderson Cancer Center, Houston, Texas, USA. [45]Department of Genomic Medicine, University of Texas MD Anderson Cancer Center, Houston, Texas, USA. [46]In Silico Solutions, Fairfax, Virginia, USA. [47]Baylor College of Medicine, Houston, Texas, USA. [48]Department of Pathology, University of Texas MD Anderson Cancer Center, Houston, Texas, USA. [49]Department of Urology, Baylor College of Medicine, Houston, Texas, USA. [50]Division of Biostatistics, University of Texas Health Science Center at Houston, School of Public Health, Houston, Texas, USA. [51]Neuro-Oncology Department, University of Texas MD Anderson Cancer Center, Houston, Texas, USA. [52]Graduate Program in Structural and Computational Biology and Molecular Biophysics, Baylor College of Medicine, Houston, Texas, USA. [53]SRA International, Inc., Fairfax, Virginia, USA. [54]Department of Pathology and Human Oncology, Memorial Sloan-Kettering Cancer Center, New York, New York, USA. [55]Pathogenesis Program, Memorial Sloan-Kettering Cancer Center, New York, New York, USA. [56]Department of Epidemiology and Biostatistics, Memorial Sloan-Kettering Cancer Center, New York, New York, USA. [57]Helen Diller Family Comprehensive Cancer Center, University of California, San Francisco, San Francisco, California, USA. [58]Molecular & Medical Genetics, Oregon Health & Science University, Portland, Oregon, USA. [59]Buck Institute for Research on Aging, Novato, California, USA. [60]Howard Hughes Medical Institute, University of California, Santa Cruz, Santa Cruz, California, USA. [61]The Genome Institute, Washington University, St. Louis, Missouri, USA. [62]Department of Medicine, Washington University, St. Louis, Missouri, USA. [63]Siteman Cancer Center, Washington University, St. Louis, Missouri, USA. [64]Division of Oncology, Washington University, St. Louis, Missouri, USA. [65]Department of Genetics, Washington University, St. Louis, Missouri, USA. [66]Department of Mathematics, Washington University, St. Louis, Missouri, USA. [67]SAIC-Frederick, Inc., Frederick, Maryland, USA. [68]Gynecology Service, Department of Surgery, Memorial Sloan-Kettering Cancer Center, New York, New York, USA. [69]The Research Institute at Nationwide Children's Hospital, Columbus, Ohio, USA. [70]The Ohio State University, Columbus, Ohio, USA. [71]National Cancer Institute, US National Institutes of Health, Bethesda, Maryland, USA. [72]Scimentis, Statham, Georgia, USA. [73]Sage Bionetworks, Seattle, Washington, USA. [74]Department of Computer Science, Brown University, Providence, Rhode Island, USA. [75]Five3 Genomics, LLC, Santa Cruz, California, USA. [76]Department of Genetics, Ribeirão Preto Medical School, University of São Paulo, São Paulo, Brazil. [77]Center for Integrative Systems Biology (CISBi), NAP/USP, São Paulo, Brazil. [78]Department of Laboratory Medicine, University of California, San Francisco, San Francisco, California, USA. [79]Department of Electrical Engineering, Columbia University, New York, New York, USA. [80]Research Unit on Biomedical Informatics, Department of Experimental and Health Sciences, Universitat Pompeu Fabra, Barcelona, Spain. [81]Catalan Institution for Research and Advanced Studies (ICREA), Barcelona, Spain. [82]Dan L. Duncan Cancer Center, Baylor College of Medicine, Houston, Texas, USA. [83]National Center for Biotechnology Information, National Library of Medicine, US National Institutes of Health, Bethesda, Maryland, USA. [84]Department of Molecular and Cellular Biology, Baylor College of Medicine, Houston, Texas, USA. [85]Institution for Immunology, Genetics and Pathology, Uppsala University, Uppsala, Sweden. [86]Science for Life Laboratory, Uppsala University, Uppsala, Sweden. [87]Mathematical Sciences, University of Gothenburg/Chalmers, Gothenburg, Sweden.