frontiers in
GENETICS

REVIEW ARTICLE
published: 03 July 2014
doi: 10.3389/fgene.2014.00194

Toward a systematic understanding of cancers: a survey of the pan-cancer study

Zhaoqi Liu and Shihua Zhang*

National Center for Mathematics and Interdisciplinary Sciences, Academy of Mathematics and Systems Science, Chinese Academy of Sciences, Beijing, China

Edited by:
Fengfeng Zhou, Shenzhen Institutes of Advanced Technology, China

Reviewed by:
An-Yuan Guo, Huazhong University of Science and Technology, China
Ao Li, University of Science and Technology of China, China

****Correspondence:***
Shihua Zhang, Academy of Mathematics and Systems Science, Chinese Academy of Sciences, 55. Zhongguancun East Road, Beijing 100190, China
e-mail: zsh@amss.ac.cn

Studies on molecular aberrations of cancer patients have increased unprecedentedly in scale and accessibility, allowing large-scale integrative cross-cancer analysis. Pan-cancer study is becoming a valuable paradigm for cancer genomics. Here, we review recent advances in this field and highlight the potential challenges and directions especially from the computational angle.

Keywords: cancer genomics, pan-cancer study, data integration, The Cancer Genome Atlas, bioinformatics

INTRODUCTION

Cancers have been believed as complex genomic diseases nowadays. They are largely caused by molecular aberrations including somatic mutations, copy number alterations, transcriptional expression changes, epigenetic variations, and so on. Great advances in high-throughput techniques and comprehensive efforts have revealed a systematic investigation of the genomic landscapes of human cancer.

The Cancer Genome Atlas (TCGA) project started in 2005 with the goal of profiling and analyzing more than 10,000 tumor samples from about 20 tumor types, provides an unprecedented opportunity to systematically analyze molecular aberrations of cancer through the application of genomic technologies. For each individual cancer type, the rich molecular data from six types of omics platforms were analyzed and integrated to identify novel oncogenic drivers, establish molecular subtypes and discover new biomarkers (McLendon et al., 2008; Bell et al., 2011; Muzny et al., 2012; Hammerman et al., 2012; Koboldt et al., 2012; Creighton et al., 2013; Kandoth et al., 2013b). These comprehensive analyses have identified many important genomic similarities among tumor types and subtypes, which present an opportunity to design tumor treatment strategies and enable therapeutic discoveries among tumors regardless of tissue or organ of origin. This suggests the potential importance of developing a comprehensive analysis across cancers to find the pan-cancer similarities and tumor-specific characteristics.

To this end, TCGA launched the Pan-Cancer analysis project to compare the molecular data of 12 tumor types (**Figure 1**). The pan-cancer study aims to examine the similarities and differences among the genomic and cellular alterations across diverse tumor types by analyzing multiple profiles of large number of human tumors. The first batch of research achievements of this promising project have been released very recently (Ciriello et al., 2013; Kandoth et al., 2013a; Lawrence et al., 2013; Weinstein et al., 2013; Zack et al., 2013). The recently published studies based on these comprehensive datasets provide more systematic understanding of human cancer on genomic, epigenomic, transcriptomic, proteomic, and clinical levels. The majority of these findings focus on cancer genomic variations such as somatic mutation, copy number alteration, and chromosomal aberrations. Although the pan-cancer project has made great advances, deciphering the complicated data in meaningful terms is still in its early stage. In this paper, we mainly review the progresses of Pan-cancer project, discuss high-related "pan-cancer" studies and highlight potential challenges and directions (**Table 1**).

SOMATIC MUTATIONS

Somatic mutations are essential for tumorigenesis and most human cancers are caused by a small number of driver gene mutations that develop over the course of about two decades (Vogelstein et al., 2013). Therefore, a comprehensive investigation of the mutational landscape of multiple cancer types would definitely be a critical basis for cancer diagnostics, therapeutics, and selection of rational combination of therapies. Finding the driver mutations from passengers will still be the major challenge in cancer genomics. Large-scale genomic analysis and approaches that use cross-tumor principles definitely enable the identification of validated and novel driver genes while dramatically improving the sensitivity and efficiency compared to the traditional calls by individual-tumor-type research (Lawrence et al., 2013).

Kandoth et al. reported 127 significantly mutated genes from diverse cellular processes across different tumor types, and identified cancer type-specific signatures of driver mutations in several dominant cancer types (Kandoth et al., 2013a). They found that kidney renal clear cell carcinoma (KIRC) has the strongest exclusivity from the other 11 cancer types with high mutation

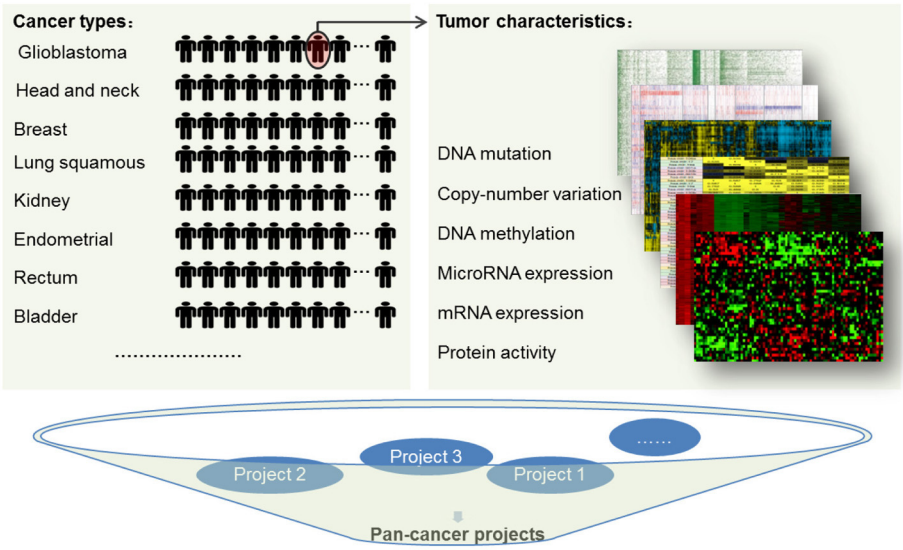

FIGURE 1 | Multiple omics datasets from diverse tumor types enable comprehensive analyses on numerous aspects of cancer genomics. The pan-cancer data includes rich biomolecular profiles on six types of platforms (mutation, copy number, methylation, gene expression, microRNA and reverse phase protein arrays) from tumors occurring in different sites of the body (glioblastoma, lymphoblastic acute myeloid leukemia, head, and neck squamous carcinoma, lung adenocarcinoma, lung squamous carcinoma, breast cancer, kidney renal, clear-cell carcinoma, ovarian carcinoma, bladder carcinoma, colon adenocarcinoma, uterine cervical and endometrial carcinoma, and rectal adenocarcinoma).

frequency of *VHL* and *PBRM1*. However, besides *TP53*, there was hardly any common mutation shared by multiple cancer types based on the observation of the reported genes, disenabling the discovery of potential extending of shared effective treatments among tumors. They pointed that the combination of drivers varies for individual patients in each cancer type and it was crucial for optimizing the treatment. Lawrence et al. conducted a large-scale genomic analysis of somatic point mutations in exome sequences from 4742 human cancers and identified nearly all known cancer driver genes and 33 novel candidates (Lawrence et al., 2014). However, more validations on these new drivers are required with experimentally follow-up studies.

Phosphorylation has been considered as an important factor in cancer which is involved in key processes such as the control of proliferation, oncogenic kinase signaling. It was recently reported that cancer may be driven by statistically significant and spatially specific mutations in protein sites involved in cellular phosphorylation signaling (Reimand and Bader, 2013). More recently, Reimand et al. extended their study to detect such mutations to 3185 tumor genomes across 12 cancer types, and predicted 54 additional cancer-specific drivers and 82 genes only seen in pan-cancer analysis (Reimand et al., 2013). However, this analysis only restricted known signaling alterations to protein-coding mutations which only comprise a minority of all cancer mutations, limiting the extent of mutated signaling in tumor cells caused by other mechanisms.

It has been demonstrated that computational analyses of sequence data for identifying driver mutations from large cohorts of tumor samples are not trivial due to the heterogeneous nature of cancer and all existing methods for the identification of genes exhibiting signals of positive selection show particular shortcomings and specific biases (Gonzalez-Perez et al., 2013a). Recently, Tamborero et al. proposed an integrative strategy to combine five complementary methods which enables the identification of a comprehensive and high-confident pan-cancer driver gene list (Tamborero et al., 2013). This analysis have shown that the combination of complementary methods are effective than individual methods. However, there is no gold-standard dataset of driver and passenger genes to assess the quality of such combination. Thus, it naturally introduces a computational issue that what the reasonable or optimal combination of different methods is. Practical exploration on the composition and structure of the investigated genomic dataset and detailed learning on the principle of each method would help to form a better combination analysis than traditional intuitive operation, e.g., combining the output *p*-values, or overlapping the top-ranking genes from diverse algorithms.

The investigation of temporal relationship of somatic genetic events would provide new insights into the discovery of driver oncogenes. It is reported that the timing of vital mutation is likely to be related to metastasis, which is responsible for the death of most patients with cancer. The genetic changes that occur early during malignant transformation may represent promising targets for therapeutic intervention (Vogelstein et al., 2013). Thus, a comprehensive analysis of determining the temporal sequence of somatic genetic events would help the identification of important mutations across 12 cancer types, which was untouched extensively by previous studies. This is probably because the lack of effective computational algorithms (Attolini et al., 2010). More efforts and techniques are needed in developing fast and accurate models to resolve this issue. Moreover, the identification of genetic alterations that leads to cancer metastasis is remarkably

Table 1 | Brief summary of recent pan-cancer studies.

	p/t	Resources	dt	Summary	References
1	3281/12	TCGA	S	Describe variable mutation frequencies and contexts and their links to environmental factors and defects in DNA repair, and identify 127 significantly mutated genes.	Kandoth et al., 2013a
2	7042/30	TCGA, ICGC, and others	S	Extract 21 distinct mutational signatures, find some present in many cancers and certain ones are associated with phenotypic features, and discover localized hypermutation "kataegis" in many cancers.	Alexandrov et al., 2013
3	3185/12	TCGA	S	Analyze known phosphorylation sites mutated by single nucleotide variants, predict signaling-specific cancer driver genes, and create a high-confidence collection of cellular signaling-related cancer mutations.	Reimand et al., 2013
4	4632/13	TCGA, ICGC, and others	S	Propose a platform for summarizing somatic mutations, genes and pathways involved in tumorigenesis, and identifying, ranking, and visualizing cancer drivers.	Gonzalez-Perez et al., 2013b
5	5277/19	TCGA	SE	APOBEC3B is the most likely cause of a large fraction of both dispersed and clustered cytosine mutations in six distinct cancers.	Burns et al., 2013
6	3205/12	TCGA	S(C)	Employ five complementary methods to search for mutational driver genes, demonstrate its advantage, and provide a list of 291 high-confidence cancer driver genes.	Tamborero et al., 2013
7	3083/27	TCGA	S	Demonstrate the false-positive cancer gene identification issue, provide a methodology MutSigCV to eliminate the artifactual findings and enable the identification of true cancer associated genes.	Lawrence et al., 2013
8	2680/14	TCGA, dbGaP, and others	S	Demonstrate a significant presence of the APOBEC mutation pattern in certain cancers.	Roberts et al., 2013
9	4742/21	TCGA, dbGaP	S	Find that large-scale genomic analysis can identify nearly all known cancer genes, report 33 novel genes, conduct down-sampling analysis and estimate the tumor number of samples for near-saturation.	Lawrence et al., 2014
10	4934/11	TCGA	C	Compare patterns of copy number change across cancer types, determine individual SCNA events and their temporal ordering from these profiles and identify functionally relevant correlations between SCNAs.	Zack et al., 2013
11	8227/19	GEO	C	Discover similarity of chromosomal arm-level alterations and co-occurring pairs of arm-level alterations, identify cancer-related gene enriched recurrent focal alterations, and tumor type-specific alterations with enriched functional categories.	Kim et al., 2013
12	3290/11	TCGA	RE(CM)	Infer recurrent cancer-associated miRNA-target relationships across multiple cancer types, which were highly consistent with published data from miRNA perturbation experiments and predictions based on sequencing technology.	Jacobsen et al., 2013
13	4186/11	TCGA, AGO-CLIP	MCRE	Describe a pan-cancer co-regulated oncogenic microRNA "superfamily," define mutations in microRNA target sites, and identify pan-cancer oncogenic co-targeting pathways by the miR-17-19-130 superfamily members.	Hamilton et al., 2013
14	82 cell lines	ENCODE	ME	Provide an atlas of DNA methylation across diverse samples, enable new discoveries about DNA methylation and its role in gene regulation and disease.	Varley et al., 2013

(Continued)

Table 1 | Continued

	p/t	Resources	dt	Summary	References
15	4379/11	TCGA	P	Develop a user-friendly data portal, The Cancer Proteome Atlas (TCPA) with six modules: Summary, My Protein, Download, Visualization, Analysis, and Cell Line.	Li et al., 2013
16	2920/11	TCGA and other 31 datasets	E(CS)	Describe a method called "ESTIMATE" that uses gene expression signatures to infer the fraction of stromal and immune cells in tumor samples.	Yoshihara et al., 2013
17	4433/19	TCGA	E	Screen for expressed viruses across diverse cancers, provide a large-scale virus–tumor association map, and confirm and extend current knowledge.	Tang et al., 2013
18	3299/12	TCGA	SCM(E)	Develop an algorithmic approach to hierarchically stratify tumors, divide tumors into two major classes, and reveal oncogenic signatures to characterize tissue-independent subclasses of tumors.	Ciriello et al., 2013

p/t, number of patients/number of tumor types; Resources, major data resources; dt, major molecular data types used for pan-cancer study and validation analysis in bracket including somatic mutation (S), copy number variation (C), DNA methylation (M), microRNA expression (R), mRNA expression (E) and reverse-phase protein arrays (P); Summary, summary of the key contributions.

limited still now and need to be further studied with the abundant pan-cancer data.

In order to reveal the causes of extensive somatic mutations accrued in cancers, a global analysis with the pan-cancer dataset found that APOBEC3B-catalyzed genomic uracil lesions are responsible for a large proportion of mutations in distinct cancer types (Burns et al., 2013). Cytidine deaminases, which convert cytosine bases to uracil during RNA editing, may contribute to DNA damage. A similar study showed a significant presence of the APOBEC mutation pattern in bladder, cervical, breast, head and neck, and lung cancers (Roberts et al., 2013). Meanwhile, a newly introduced concept of understanding the biological processes generating mutations, mutational processes, were explored on the TCGA, ICGC and other datasets using a previously developed computational framework. Finally, they extracted more than 20 distinct mutational signatures, one of which attributed to the former mentioned APOBEC family of cytidine deaminases (Alexandrov et al., 2013). In addition, hypermutation localized to small genomic regions called "kataegis" was found in many cancer types.

All these comprehensive analyses on the mutation profiles have proven the enhanced ability of detecting driver genes with the increase in the number of patients across 12 tumor types. However, cancer is a disease of pathways driven by underlying systematic alterations. The main subjects of alterations are not individual driver genes, but rather modules of functionally related proteins at pathway-level. With an increase in the number of mutational profiles across different tissues, critical and tumorigenesis-associated pathways would be discovered to enable physicians to select the best combination therapy for each patient. To provide an exhaustive description of potentially actionable pathway-level catalog of the driver mutations would be a challenge for specific targeted therapeutics across cancer types.

Computational methods for integrating, comparing and interpreting genome-scale molecular information are urgently needed in current stage and known algorithmic approaches may be adopted and adapted for such analysis. For example, to identify mutated driver pathways using somatic mutation data, Vandin et al. developed a method by considering mutual exclusive principle and high coverage (Vandin et al., 2012). Zhao et al. proposed a powerful mathematical programming model to solve it and suggested to incorporate gene expression data to prioritize the true functional ones (Zhao et al., 2012). Ciriello et al. proposed a network-based method to detect driver modules that obey the mutual exclusivity principle (Ciriello et al., 2012). Hofree et al. devised a network-based approach to integrate discrete somatic mutation data with known biological molecular networks to stratify patients into subtypes for individual tumor types (Hofree et al., 2013). We believe that the similar network-based framework can be adopted for cross-tumor comparative analysis and finding tumor-specific features (**Figure 2**).

SOMATIC COPY NUMBER ALTERATIONS

Somatic copy number alterations (SCNA) are ubiquitous and affect a majority of the genome in cancers. It has been comprehensively demonstrated that SCNAs play critical roles in activating oncogenes and in inactivating tumor suppressors. Distinguishing the driver events from the passenger SCNAs, identifying their gene targets and describing their functional roles are major challenges in current stage. The unprecedented large-scale data of copy number profiles across cancers will enable the identification of recurrent chromosomal alterations with potential clinical benefits, provide more systematic understanding of human cancer, and leads to substantial advances in cancer diagnostics and therapeutics.

In a recent study, Zack et al. conducted a comprehensive analysis of high-resolution copy number profiles of the TCGA data and identified common patterns of SCNA across cancer types (Zack et al., 2013). They found that whole-genome doubling observed in 37% of cancers was associated with higher rates

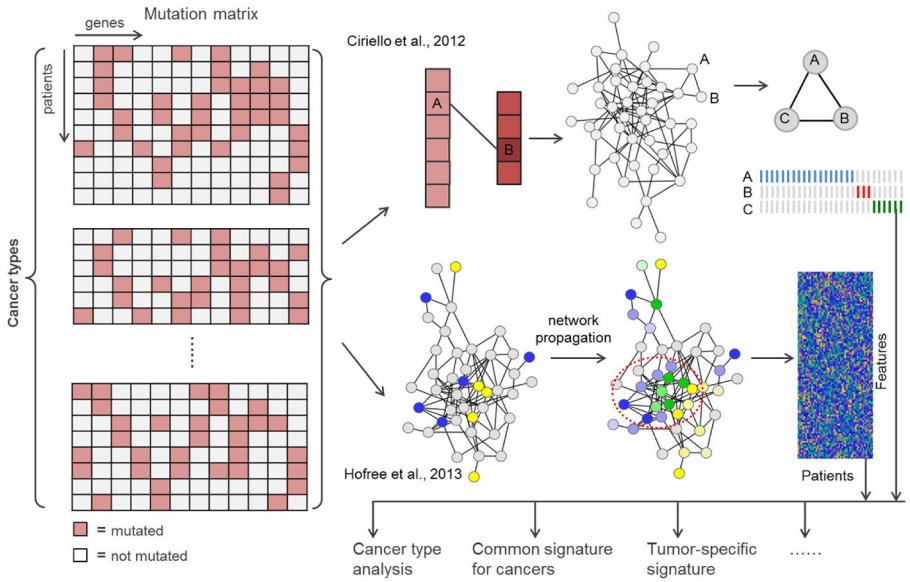

FIGURE 2 | Two potential network-based frameworks for cross-tumor comparative analysis and tumor-specific feature discovery.

of every other type of SCNA. They suggested that the diverse lengths of SCNAs in the middle of chromosomes and those of the telomere-bounded ones indicate different mechanisms of generation. They reported a number of significantly recurrent focal SCNAs in 140 regions, some of which were enriched for genes involved in epigenetic regulation, or encompassed genes tending to generating interacting protein product. In another study, Kim et al. found that chromosomal arm-level alterations among developmentally related tumor types tend to be similar (Kim et al., 2013). They also reported a number of co-occurring pairs of arm-level alterations, and found that recurrent pan-cancer focal alterations are enriched with known cancer related genes. Both of these two studies explored the rare cataclysmic event that occurs in a small fraction of chromosomes called chromothripsis. Specifically, a number of localized chromothripsis events associated with known cancer-related genes were revealed on chromosome 2 in some neuroblastoma cases (Kim et al., 2013). Similarly, Zack et al. revealed that chromothripsis was detected in 16% of glioblastomas, reaching the highest rate across cancers (Zack et al., 2013). The fact that arm-level alterations tend to be shared among diverse cancer types was also reported in a previous study (Beroukhim et al., 2010), where 158 regions of focal somatic copy-number alterations were identified and a large majority of them were present at significant frequency in multiple cancer types. Such large-scale analysis provides insights into mechanisms of generation and functional consequences of cancer-related SCNAs, which cannot be revealed directly in individual cancer.

However, due to the intrinsic complexity of cancer genomes, powerful algorithmic approaches are still needed for deep exploration of the large-scale copy number alteration profiles for driver events and the chromothripsis based on the integration of new data resources. Statistical analysis of co-occurrence and mutual exclusivity of genomic aberrations were needed to be further explored. Akavia et al. once developed a computational framework to integrate chromosomal copy number and gene expression data for identifying cancer drivers based on the hypothesis that copy number aberrations often influence the expression of genes in a module via changes in expression of the driver (Akavia et al., 2010). Extending of such computational framework for identifying robust common drivers and cancer-specific ones will be promising in the near future.

Recently, a novel algorithmic approach that uses 479 selected functional events obtained from significance analysis on the mutation, copy number variation and DNA methylation profiles derived a hierarchical classification of 3299 TCGA tumors from 12 cancer types (Ciriello et al., 2013). The top two classes of the clusters are dominated by mutations (M class) and copy number changes (C class) respectively, which would be due to the treatment of the mutation and copy number features equally and separately. The M class of tumors contained almost all the samples in kidney clear-cell carcinoma, while almost all ovarian cancer fell into the C class. Patients within the same subclass may be benefited from the observed cross-cancer distribution of targetable events. Integrating these three kinds of significant features with different weights on gene-levels would provide diverse findings and the ready-processed data matrix of selected functional events will be valuable for other related analysis.

DNA METHYLATION ALTERATIONS

DNA methylation is a key determinant of regulatory chromatin complexes that has a complex relationship with gene expression and was found to be dysregulated in many cancers. Recently, a large-scale DNA methylation study on 82 human cell lines and tissues provides an atlas of DNA methylation across diverse and well-characterized samples and enables new discoveries about DNA methylation and its role in gene regulation and disease

(Varley et al., 2013). The comparisons of methylation profiles across different cancer cell lines identified cancer-associated and cell-type specific methylation signatures. The relationship between DNA methylation and gene expression levels were well observed across the genome; however, evidences of its directed or indirected associations with other molecular and phenotypic characteristics across multiple cancers are of potential interest and are worth further exploring with the aid of pan-cancer data.

MicroRNA AND GENE EXPRESSIONS

MicroRNAs (miRNA) have been demonstrated to play key roles in gene regulation by binding target mRNAs in a sequence complementary manner. Previous studies have shown that dysregulation of microRNAs can contribute to tumor formation and progression. Recently, Jacobsen et al. explored the common processes of tumor biology regulated by microRNAs across 11 diverse cancer types (Jacobsen et al., 2013). They adopted a multivariate linear regression model to evaluate a causal relationship score of each pair of miRNA and mRNA in individual cancer types and employed a rank-based statistical method to integrate scores obtained from multiple cancer types to infer recurrent pan cancer-associated miRNA-mRNA relationships from miRNA and mRNA expression profiles. The predicted miRNA-target interactions were shown to be highly consistent with published experimental data and computational predictions, and form a high-confidence pan-cancer network of 143 recurrent target relationships for further analysis. Computationally, this current analysis didn't address the potential nonlinear effect between miRNA and mRNA which need to be addressed further.

In another study, Hamilton et al. explored the microRNA regulatory landscape and identified pan-cancer microRNA drivers of cancer by integrating the TCGA Pan-Cancer microRNA, mRNA, copy number variation (CNV) and exome-sequencing data sets from 12 tumor types with a miRNA target atlas composed of publicly available Argonaute Crosslinking Immunoprecipitation (AGO-CLIP) data (Hamilton et al., 2013). They showed a pan-cancer, coregulated oncogenic microRNA "superfamily," which cotargets critical tumor suppressors via a central GUGC core motif. Through these two integrative pan-cancer analysis frameworks, we were able to understand microRNA regulatory architectures across multiple tumor types. Both studies have shown new examples of miRNAs that coordinately regulate cancer pathways across many cancer types, demonstrating the potential roles of miRNA-target co-modules. In the future studies, computational techniques for identifying such co-modules (Zhang et al., 2011) can be developed to pan-cancer data set for exploring common ones across diverse cancer types directly.

In addition, other pan-cancer related works include a method that uses gene expression signatures to infer the fraction of stromal and immune cells in tumor samples (Yoshihara et al., 2013), and a landscape of virus–tumor map generated using transcriptome sequencing data (Tang et al., 2013). Actually, 10 year ago, Segal et al. have conducted a study to address the commonalities and variations between different types of tumor using DNA microarrays (Segal et al., 2004). They implemented an integrated analysis of 1975 published microarrays spanning 22 tumor types. They defined co-expression modules based on expression profiles in different tumors and employed a unified analysis to characterize gene-expression profiles in tumors with activated and deactivated modules. They have found that activation of some modules is specific to particular types of tumor and other modules are shared across a diverse set of tumors. We believe that the revisit of the large-scale pan-cancer study in terms of expression profiles and integration analysis with other genomics data will improve the understanding for diagnostic, prognostic and therapeutic studies.

WEB TOOLS FOR PAN-CANCER STUDY

Several useful web tools have been developed to interactively visualize and explore the large-scale TCGA pan-cancer data (**Table 2**). Specifically, Gonzalez-Perez et al. developed a web platform called IntOGen-mutations to identify and visualize cancer drivers across tumor types, which provides convenience for better clinical decision-making (Gonzalez-Perez et al., 2013b). Moreover,

Table 2 | Brief summary of useful webserver or database for pan-cancer study.

Name	Website	Key purposes	References
IntOGen-mutations	http://www.intogen.org/mutations	Identify and visualize cancer drivers across tumor types.	Gonzalez-Perez et al., 2013b
CancerMiner	http://cancerminer.org	Search recurring microRNA-mRNA associations across cancer types.	Jacobsen et al., 2013
Synapse	https://www.synapse.org/	Collaborate with the TCGA pan-cancer group to share and update data, results and methodologies.	Omberg et al., 2013
TCGA	http://cancergenome.nih.gov/	Provide a platform for researchers to search, download, and analyze data sets generated by TCGA.	Weinstein et al., 2013
TCPA	http://bioinformatics.mdanderson.org/main/TCPA:Overview	Facilitate access of the broader research community to cancer proteomics datasets.	Li et al., 2013
UCSC Cancer Genomics Browser	https://genome-cancer.ucsc.edu	Offer interactive visualization and exploration of TCGA genomic, phenotypic, and clinical data.	Cline et al., 2013

Li et al. developed a user-friendly data portal with six modules, The Cancer Proteome Atlas (TCPA), which provides comprehensive, and unique cancer proteomic data and powerful visualizing and analysis modules for exploring such data (Li et al., 2013). Jacobsen et al. have presented all predictions of miRNA-target relationships in their study on an online resource, which allows exploration, prioritization and visualization of novel miRNA-target interactions in TCGA data (Jacobsen et al., 2013). The University of California Santa Cruz (UCSC) Genome Browser has become a very important tool which offers online public access to a growing database of genomic sequence and annotations for a large collection of organisms and provides an integrated environment for visualizing, comparing, analyzing and sharing both publicly available and user-generated genomic data sets with various web-based tools. Cline et al. has extended this powerful Browser to explore the impact of genomic alterations on phenotypes by visualizing data of different platforms and levels, performing cancer classifications and conducting patient survival analysis (Cline et al., 2013). The Synapse web server developed by Sage Bionetworks is an informatics platform of public resources for the scientific community and encourages scientists to discover and share data, models, and analysis methods. The TCGA pan-cancer group has collaborated on this system to share and evolve data, results, and methodologies throughout the full duration of the project (Omberg et al., 2013). More importantly, updates of new datasets and discoveries will be immediately available based on this system. In summary, all these resources and tools will provide great convenience and promote pan-cancer type of study.

DISCUSSION AND CONCLUSION

Although several previous pan-cancer studies focusing on multiple tumor types or cell lines have been reported before (Segal et al., 2004; Lee et al., 2008; Sahin et al., 2008; Wu et al., 2010; Beroukhim et al., 2010), the ongoing pan-cancer project has provided an unequaled resource for the integrative analysis of multiple cancer types, and achieved remarkable discoveries. Generally, the main investigation and observations are attributable to two fundamental aspects: intra-cancer heterogeneity and cross-cancer similarity reflected in different levels of molecular properties. However, along with these progresses, new challenges are emerging and pressing to be resolved.

How to integrate the data generated on different platforms or different versions of the same platform is an unavoidable challenge which doesn't account for the challenge in the integration of data across cancer-types. Consensus and reliable standardization of the input data will be a key step to obtain robust and reliable results from the true biological signals and conquer the unwanted batch effects. Large-scale collaborative analysis and open community-based competition has been suggested to be one possible solution to establish best practices for overcoming this challenge (Omberg et al., 2013).

To our knowledge, there were no well-established and unified approaches to integrate different molecule data in pan-cancer studies. There were only some general routine techniques such as robust quantile normalization or z-score transformation for conquering the data scale issue of different cancer datasets. Currently, most published pan-cancer studies prefer to rerun the same algorithm on each cancer type individually and compare or combine the results to derive the pan-cancer similarities in statistical fashion or meta-analysis.

The multi-dimensional genomic profiling data provide unique opportunities to study the coordination between regulatory mechanisms on multiple levels. Recently, we have developed methods for the integrative analysis of multi-dimensional genomics data and the discovery of underlying combinatorial patterns (Li et al., 2012; Zhang et al., 2012). The discovered multi-dimensional modules have been demonstrated to reveal perturbed pathways that would have been overlooked with only a single type of data, uncover associations between different layers of cellular activities and allow the identification of clinically distinct patient subgroups. It will be valuable to adopt such study to uncover hidden patterns of multi-dimensional "omics" data across tumor types.

Due to the shared molecular dimension, the pan-cancer studies are focusing on the molecular properties. However, unlike molecular profiles, most clinical features are incomparable across tumor types due to the nature and availability of such data (Weinstein et al., 2013). For example, tumor stage and grade are not comparable as each tumor has its own system. Furthermore, some clinical features are collected according to the classification by tissue or organ, making them vary widely across tumors. Thus, how to effectively employ the clinical features in performing comparative analysis involving multiple cancers will be an important but challenging issue.

Almost all the pan-cancer studies are involved in direct use of existing computing techniques, or previously well-developed approach with an extending analysis on the new dataset. However, the uniqueness and complexity of the pan-cancer data may require more specific and modified approaches for novel discoveries of underlying principles in tumor evolution. Moreover, in contrast to well-studied phenotypic heterogeneity in tumors, the genetic heterogeneity among the cells of an individual tumor or tumors of different patients set an obstacle to effective response to uniformly designed therapeutics. This issue could not be simply resolved by numerical calculation. More efforts on personalized medicine and development of treatment should take advantage of detection of this heterogeneity. For example, Liu et al. evaluated the patient survival prediction performance of genomic and clinical data on the five intrinsic breast cancer subtypes and revealed that molecular gene profiles and clinical features have different prognostic power (Liu et al., 2014). How to extend this kind of studies to make a pan-cancer type of analysis will be an interesting and meaningful problem.

Distinguishing and interpreting the functional role of variants in the noncoding parts of the sequences is an open frontier which has not been as well explored so far. In addition, the prediction of the functional consequences of chromosomal- scale structural variation are also challenging (Weinstein et al., 2013; Wheeler and Wang, 2013).

The ultimate goal of almost all cross cancer studies is to affect clinical decision-making, accelerate the discovery of novel therapeutic agents applied for tumors rising from different organs with similar genomic characteristics. The number of commercially available targeted cancer drugs is still limited nowadays. New

computational findings require rapid and effective methods for functional validation. Experimental follow-ups are always critical to assess the hypotheses and consequences. Therefore, a great challenge is how to speed up the process of translating novel discoveries into treatments based on experimental measurements.

As we known that, the process of tumor usually take decades to develop but cancer metastasis occurs only a few years before death. Thus, the investigation on molecular aberrations account for cancer metastasis should be highly informative. Moreover, the knowledge learned from cancer genomics can also be exploited to develop methods for prevention and early detection of cancer, which will be essential to reduce cancer morbidity and mortality (Vogelstein et al., 2013). Besides, the causal relationships of several carcinogenic etiologies with multiple cancer types are also worth exploration (Weinstein et al., 2013). All these challenges enable the pan-cancer study to be a hot topic.

AUTHOR CONTRIBUTIONS

Shihua Zhang and Zhaoqi Liu conceived this study and wrote the paper.

ACKNOWLEDGMENTS

This work was supported by the National Natural Science Foundation of China, No. [61379092, 11001256], the Foundation for Members of Youth Innovation Promotion Association, CAS, and the Scientific Research Foundation for ROCS, SEM.

REFERENCES

Akavia, U. D., Litvin, O., Kim, J., Sanchez-Garcia, F., Kotliar, D., Causton, H. C., et al. (2010). An integrated approach to uncover drivers of cancer. *Cell* 143, 1005–1017. doi: 10.1016/j.cell.2010.11.013

Alexandrov, L. B., Nik-Zainal, S., Wedge, D. C., Aparicio, S. A., Behjati, S., Biankin, A. V., et al. (2013). Signatures of mutational processes in human cancer. *Nature* 500, 415–421. doi: 10.1038/nature12477

Attolini, C. S. O., Cheng, Y. K., Beroukhim, R., Getz, G., Abdel-Wahab, O., Levine, R. L., et al. (2010). A mathematical framework to determine the temporal sequence of somatic genetic events in cancer. *Proc. Natl. Acad. Sci. U.S.A.* 107, 17604–17609. doi: 10.1073/pnas.1009117107

Bell, D., Berchuck, A., Birrer, M., Chien, J., Cramer, D. W., Dao, F., et al. (2011). Integrated genomic analyses of ovarian carcinoma. *Nature* 474, 609–615. doi: 10.1038/nature10166

Beroukhim, R., Mermel, C. H., Porter, D., Wei, G., Raychaudhuri, S., Donovan, J., et al. (2010). The landscape of somatic copy-number alteration across human cancers. *Nature* 463, 899–905. doi: 10.1038/nature08822

Burns, M. B., Temiz, N. A., and Harris, R. S. (2013). Evidence for APOBEC3B mutagenesis in multiple human cancers. *Nat. Genet.* 45, 977–983. doi: 10.1038/ng.2701

Ciriello, G., Cerami, E., Sander, C., and Schultz, N. (2012). Mutual exclusivity analysis identifies oncogenic network modules. *Genome Res.* 22, 398–406. doi: 10.1101/gr.125567.111

Ciriello, G., Miller, M. L., Aksoy, B. A., Senbabaoglu, Y., Schultz, N., and Sander, C. (2013). Emerging landscape of oncogenic signatures across human cancers. *Nat. Genet.* 45, 1127–1133. doi: 10.1038/ng.2762

Cline, M. S., Craft, B., Swatloski, T., Goldman, M., Ma, S., Haussler, D., et al. (2013). Exploring TCGA pan-cancer data at the UCSC cancer genomics browser. *Sci. Rep.* 3:2652. doi: 10.1038/srep02652

Creighton, C. J., Morgan, M., Gunaratne, P. H., Wheeler, D. A., Gibbs, R. A., Muzny, D., et al. (2013). Comprehensive molecular characterization of clear cell renal cell carcinoma. *Nature* 499, 43–49. doi: 10.1038/nature12222

Gonzalez-Perez, A., Mustonen, V., Reva, B., Ritchie, G. R., Creixell, P., Karchin, R., et al. (2013a). Computational approaches to identify functional genetic variants in cancer genomes. *Nat. Methods* 10, 723–729. doi: 10.1038/nmeth.2562

Gonzalez-Perez, A., Perez-Llamas, C., Deu-Pons, J., Tamborero, D., Schroeder, M. P., Jene-Sanz, A., et al. (2013b). IntOGen-mutations identifies cancer drivers across tumor types. *Nat. Methods* 10, 1081–1082. doi: 10.1038/nmeth.2642

Hamilton, M. P., Rajapakshe, K., Hartig, S. M., Reva, B., McLellan, M. D., Kandoth, C., et al. (2013). Identification of a pan-cancer oncogenic microRNA superfamily anchored by a central core seed motif. *Nat. Commun.* 4, 2730. doi: 10.1038/ncomms3730

Hammerman, P. S., Lawrence, M. S., Voet, D., Jing, R., Cibulskis, K., Sivachenko, A., et al. (2012). Comprehensive genomic characterization of squamous cell lung cancers. *Nature* 489, 519–525. doi: 10.1038/nature11404

Hofree, M., Shen, J. P., Carter, H., Gross, A., and Ideker, T. (2013). Network-based stratification of tumor mutations. *Nat. Methods* 10, 1108–1115. doi: 10.1038/nmeth.2651

Jacobsen, A., Silber, J., Harinath, G., Huse, J. T., Schultz, N., and Sander, C. (2013). Analysis of microRNA-target interactions across diverse cancer types. *Nat. Struct. Mol. Biol.* 20, 1325–1332. doi: 10.1038/nsmb.2678

Kandoth, C., McLellan, M. D., Vandin, F., Ye, K., Niu, B., Lu, C., et al. (2013a). Mutational landscape and significance across 12 major cancer types. *Nature* 502, 333–339. doi: 10.1038/nature12634

Kandoth, C., Schultz, N., Cherniack, A. D., Akbani, R., Liu, Y., Shen, H., et al. (2013b). Integrated genomic characterization of endometrial carcinoma. *Nature* 497, 67–73. doi: 10.1038/nature12113

Kim, T. M., Xi, R., Luquette, L. J., Park, R. W., Johnson, M. D., and Park, P. J. (2013). Functional genomic analysis of chromosomal aberrations in a compendium of 8000 cancer genomes. *Genome Res.* 23, 217–227. doi: 10.1101/gr.140301.112

Koboldt, D. C., Fulton, R. S., McLellan, M. D., Schmidt, H., Kalicki-Veizer, J., McMichael, J. F., et al. (2012). Comprehensive molecular portraits of human breast tumors. *Nature* 490, 61–70. doi: 10.1038/nature11412

Lawrence, M. S., Stojanov, P., Mermel, C. H., Robinson, J. T., Garraway, L. A., Golub, T. R., et al. (2014). Discovery and saturation analysis of cancer genes across 21 tumor types. *Nature* 505, 495–501. doi: 10.1038/nature12912

Lawrence, M. S., Stojanov, P., Polak, P., Kryukov, G. V., Cibulskis, K., Sivachenko, A., et al. (2013). Mutational heterogeneity in cancer and the search for new cancer-associated genes. *Nature* 499, 214–218. doi: 10.1038/nature12213

Lee, G., Laflamme, E., Chien, C. H., and Ting, H. H. (2008). Molecular identity of a pan cancer marker, CA215. *Cancer Biol. Ther.* 7, 2007–2014. doi: 10.4161/cbt.7.12.6984

Li, J., Lu, Y., Akbani, R., Ju, Z., Roebuck, P. L., Liu, W., et al. (2013). TCPA: a resource for cancer functional proteomics data. *Nat. Methods.* 10, 1046–1047. doi: 10.1038/nmeth.2650

Li, W., Zhang, S., Liu, C. C., and Zhou, X. J. (2012). Identifying multi-layer gene regulatory modules from multi-dimensional genomic data. *Bioinformatics* 28, 2458–2466. doi: 10.1093/bioinformatics/bts476

Liu, Z., Zhang, X. S., and Zhang, S. (2014). Breast tumor subgroups reveal diverse clinical prognostic power. *Sci. Rep.* 4:4002. doi: 10.1038/srep04002

McLendon, R., Friedman, A., Bigner, D., Van Meir, E. G., Brat, D. J., Mastrogianakis, G. M., et al. (2008). Comprehensive genomic characterization defines human glioblastoma genes and core pathways. *Nature* 455, 1061–1068. doi: 10.1038/nature07385

Muzny, D. M., Bainbridge, M. N., Chang, K., Dinh, H. H., Drummond, J. A., Fowler, G., et al. (2012). Comprehensive molecular characterization of human colon and rectal cancer. *Nature* 487, 330–337 doi: 10.1038/nature11252

Omberg, L., Ellrott, K., Yuan, Y., Kandoth, C., Wong, C., Kellen, M. R., et al. (2013). Enabling transparent and collaborative computational analysis of 12 tumor types within the cancer genome atlas. *Nat. Genet.* 45, 1121–1126. doi: 10.1038/ng.2761

Reimand, J., and Bader, G. D. (2013). Systematic analysis of somatic mutations in phosphorylation signaling predicts novel cancer drivers. *Mol. Syst. Biol.* 9, 637. doi: 10.1038/msb.2012.68

Reimand, J., Wagih, O., and Bader, G. D. (2013). The mutational landscape of phosphorylation signaling in cancer. *Sci Rep.* 3:2651. doi: 10.1038/srep02651

Roberts, S. A., Lawrence, M. S., Klimczak, L. J., Grimm, S. A., Fargo, D., Stojanov, P., et al. (2013). An APOBEC cytidine deaminase mutagenesis pattern is widespread in human cancers. *Nat. Genet.* 45, 970–976. doi: 10.1038/ng.2702

Sahin, U., Koslowski, M., Dhaene, K., Usener, D., Brandenburg, G., Seitz, G., et al. (2008). Claudin-18 splice variant 2 is a pan-cancer target suitable for therapeutic antibody development. *Clin. Cancer Res.* 14, 7624–7634. doi: 10.1158/1078-0432.CCR-08-1547

Segal, E., Friedman, N., Koller, D., and Regev, A. (2004). A module map showing conditional activity of expression modules in cancer. *Nat. Genet*. 36, 1090–1098. doi: 10.1038/ng1434

Tamborero, D., Gonzalez-Perez, A., Perez-Llamas, C., Deu-Pons, J., Kandoth, C., Reimand, J., et al. (2013). Comprehensive identification of mutational cancer driver genes across 12 tumor types. *Sci. Rep*. 3:2650. doi: 10.1038/srep02650

Tang, K. W., Alaei-Mahabadi, B., Samuelsson, T., Lindh, M., and Larsson, E. (2013). The landscape of viral expression and host gene fusion and adaptation in human cancer. *Nat. Commun*. 4, 2513. doi: 10.1038/ncomms3513

Vandin, F., Upfal, E., and Raphael, B. J. (2012). De novo discovery of mutated driver pathways in cancer. *Genome Res.* 22, 375–385. doi: 10.1101/gr.120477.111

Varley, K. E., Gertz, J., Bowling, K. M., Parker, S. L., Reddy, T. E., Pauli-Behn, F., et al. (2013). Dynamic DNA methylation across diverse human cell lines and tissues. *Genome Res*. 23, 555–567. doi: 10.1101/gr.147942.112

Vogelstein, B., Papadopoulos, N., Velculescu, V. E., Zhou, S., Diaz, L. A., and Kinzler, K. W. (2013). Cancer genome landscapes. *Science* 339, 1546–1558. doi: 10.1126/science.1235122

Weinstein, J. N., Collisson, E. A., Mills, G. B., Shaw, K. R. M., Ozenberger, B. A., Ellrott, K., et al. (2013). The cancer genome atlas pan-cancer analysis project. *Nat. Genet*. 45, 1113–1120. doi: 10.1038/ng.2764

Wheeler, D. A., and Wang, L. (2013). From human genome to cancer genome: the first decade. *Genome Res*. 23, 1054–1062. doi: 10.1101/gr.157602.113

Wu, C. C., Hsu, C. W., Chen, C. D., Yu, C. J., Chang, K. P., Tai, D. I., et al. (2010). Candidate serological biomarkers for cancer identified from the secretomes of 23 cancer cell lines and the human protein atlas. *Mol. Cell. Proteomics* 9, 1100–1117. doi: 10.1074/mcp.M900398-MCP200

Yoshihara, K., Shahmoradgoli, M., Martínez, E., Vegesna, R., Kim, H., Torres-Garcia, W., et al. (2013). Inferring tumor purity and stromal and immune cell admixture from expression data. *Nat. Commun*. 4, 2612. doi: 10.1038/ncomms3612

Zack, T. I., Schumacher, S. E., Carter, S. L., Cherniack, A. D., Saksena, G., Tabak, B., et al. (2013). Pan-cancer patterns of somatic copy number alteration. *Nat. Genet*. 45, 1134–1140. doi: 10.1038/ng.2760

Zhang, S., Li, Q., Liu, J., and Zhou, X. J. (2011). A novel computational framework for simultaneous integration of multiple types of genomic data to identify microRNA-gene regulatory modules. *Bioinformatics* 27, i401–i409. doi: 10.1093/bioinformatics/btr206

Zhang, S., Liu, C. C., Li, W., Shen, H., Laird, P. W., and Zhou, X. J. (2012). Discovery of multi-dimensional modules by integrative analysis of cancer genomic data. *Nucleic Acids Res*. 40, 9379–9391. doi: 10.1093/nar/gks725

Zhao, J., Zhang, S., Wu, L. Y., and Zhang, X. S. (2012). Efficient methods for identifying mutated driver pathways in cancer. *Bioinformatics* 28, 2940–2947. doi: 10.1093/bioinformatics/bts564

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Received: 20 March 2014; accepted: 12 June 2014; published online: 03 July 2014.
*Citation: Liu Z and Zhang S (2014) Toward a systematic understanding of cancers: a survey of the pan-cancer study. Front. Genet. **5**:194. doi: 10.3389/fgene.2014.00194*
This article was submitted to Bioinformatics and Computational Biology, a section of the journal Frontiers in Genetics.