Cancer Cell

CellPress

Commentary

Next-Generation Analytics for Omics Data

Jun Li,[1,4] Hu Chen,[1,2,4] Yumeng Wang,[1,2,4] Mei-Ju May Chen,[1,4] and Han Liang[1,2,3,*]
[1]Department of Bioinformatics and Computational Biology, The University of Texas MD Anderson Cancer Center, Houston, TX 77030, USA
[2]Graduate Program in Quantitative and Computational Biosciences, Baylor College of Medicine, Houston, TX 77030, USA
[3]Department of Systems Biology, The University of Texas MD Anderson Cancer Center, Houston, TX 77030, USA
[4]These authors contributed equally
*Correspondence: hliang1@mdanderson.org
https://doi.org/10.1016/j.ccell.2020.09.002

The increasing omics data present a daunting informatics challenge. DrBioRight, a natural language-oriented and artificial intelligence-driven analytics platform, enables the broad research community to perform analysis in an intuitive, efficient, transparent, and collaborative way. The emerging next-generation analytics will maximize the utility of omics data and lead to a new paradigm for biomedical research.

The Challenge in Omics Data Analysis

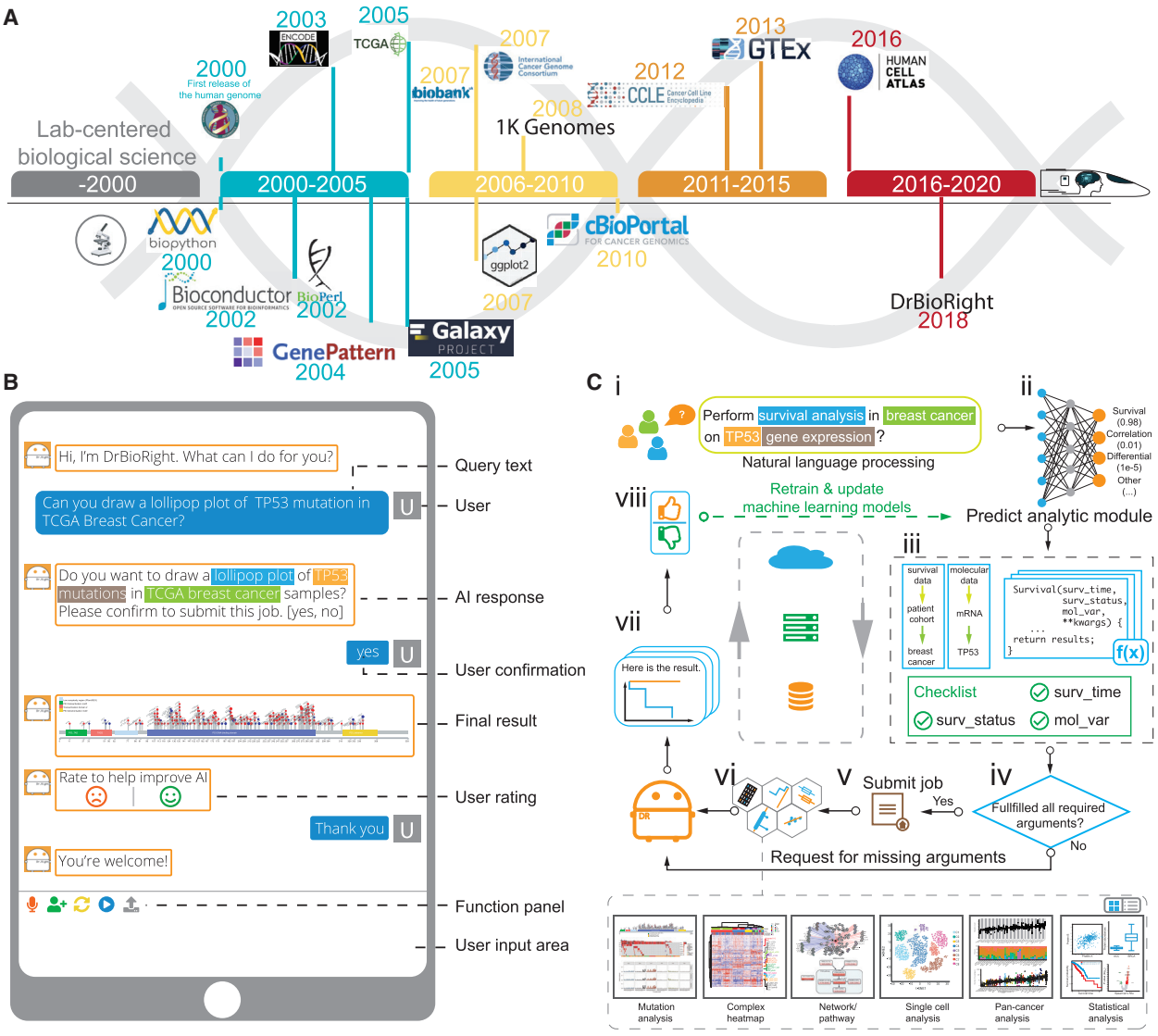
Over the last two decades, high-throughput molecular profiling technologies have revolutionized biomedical sciences. Various omics data (e.g., genomic, transcriptomic, proteomic, epigenomic, and metabolic data) generated from thousands of patients, animal models, and cell lines are accumulating at an increasing speed, especially through large consortium projects such as ENCODE (ENCODE Project Consortium, 2012), Genotype-Tissue Expression (GTEx) (GTEx Consortium, 2013), and The Cancer Genome Atlas (TCGA) (Hutter and Zenklusen, 2018) (Figure 1A). These rich omics data have provided unprecedented opportunities to systematically characterize molecular mechanisms and develop related biomedical applications. The data surge also presents a major challenge for researchers in data analysis to obtain meaningful insights.

Significant progresses have been made over the years to overcome this challenge (Figure 1A). Initially, omics data are usually analyzed using in-house scripts written in general-purpose programming languages, such as Python, R, and Perl, by bioinformaticians or computational biologists. Several collections of specialized bioinformatic programming modules, such as Biopython (Chapman, 2000), BioPerl (Stajich et al., 2002), Bioconductor (Gentleman et al., 2004), and ggplot (Wickham, 2009), allow easier analysis and visualization of omics data. However, these tools still require users to have some programming expertise, which many experimental researchers do not possess. Many web-based or stand-alone bioinformatics tools then enable users to perform various analyses or visualization of omics data without extensive programming skills. These tools, however, are of limited use, as they only support a predefined set of analyses.

Recently, two types of more generalized bioinformatics platforms for omics data exploration have gained popularity. One type is "module hubs," such as Galaxy (Giardine et al., 2005) and GenePattern (Reich et al., 2006), which provide graphic infrastructure for users to assemble bioinformatics pipelines and perform user-defined tasks. The other type is "interactive data portals," such as cBioPortal (Cerami et al., 2012) and GTEx portal (GTEx Consortium, 2013), which focus on easy analysis and visualization of preloaded datasets. Despite these impressive efforts, users still have to spend considerable time identifying appropriate tools and learning distinct user interfaces and procedures, in addition to keeping track of the status and updates for the quickly evolving tools and datasets. As a result, there is still a substantial barrier preventing most researchers (especially those with no or limited bioinformatics and statistical expertise) from making full use of omics data in a straightforward manner.

DrBioRight, a Prototype Natural Language-Oriented, Intelligent Analytics

We hypothesized that most of the commonly used standard analyses of omics data could be conducted effectively using natural languages. To test the feasibility of this idea, we developed "DrBioRight," a natural language-oriented, artificial intelligence (AI)-driven omics data analysis platform (https://drbioright.org). DrBioRight consists of two subsystems: a user-friendly web interface and a backend compute server. Compared to other bioinformatics tools, DrBioRight employs a simple online chat interface with only one input area and one output area, and all the interactions with users are based on human languages (Figure 1B). Users can simply type an omics data analysis question in the input area. For example, a user can type "perform survival analysis in breast cancer on TP53 gene expression" to test if there is a correlation between *TP53* gene expression level and overall survival in breast cancer patients. After receiving an input text (Figure 1C), DrBioRight will run its natural language processing (NLP) module to tag the recognized entities, and based on the features identified in the input, the backend AI module will calculate scores to predict the best-matched analytic task. The program will then call the specific analytic module, identify the related dataset, and check whether all required parameters are filled. Before submitting the compute task, DrBioRight will ask the user to confirm if the detected task is indeed the intended analysis. If confirmed, a job scheduler will submit the task to a job queue and use cloud-based compute nodes to process it. Once the job is complete, DrBioRight will call an appropriate visualization module and send the results (usually an interactive table or plot) to the user in the output area. Last but not least, DrBioRight will ask for a rating for each successfully executed job. The feedback thus collected will be used to further improve the performance of the NLP and AI modules.



(legend on next page)

Importantly, DrBioRight has a flexible modularized framework, based on which a new computational analysis can be added with just two simple steps: adding the necessary modules and training the modules using natural human languages.

With the natural language-oriented interactions and AI-driven modules, DrBioRight has immense potential to increase the efficiency and reproducibility of omics data analysis. We have curated and loaded some widely used cancer omics datasets, including TCGA, International Cancer Genome Consortium (ICGC/TCGA Pan-Cancer Analysis of Whole Genomes Consortium, 2020), and Cancer Cell Line Encyclopedia (Ghandi et al., 2019) (>20,000 samples in total). As an initial effort, we have built 10 analytic modules that cover the most common omics analyses and the related visualizations. Using these modules, users can easily get answers to questions like "What is the mRNA expression correlation of gene x and gene y in liver cancer?" and "Is there a correlation between TP53 mutations and the overall survival in patients with lung cancer?" and visualize the results using scatter, Kaplan-Meier, or boxplots. Moreover, DrBioRight supports bioinformatics analysis from raw next-generation sequencing reads. For example, a user can start an analysis by simply asking, "Could you do an RNA-seq analysis?" and then provide the source or location of the raw data (e.g., an SRA ID). Through a dialog with DrBioRight, the user can finish the entire analysis step by step, including quality control, read mapping, gene expression quantification, differential expression analysis, and gene set enrichment analysis. Finally, DrBioRight enables users to conveniently check the reproducibility of published results. To demonstrate this aspect, we focus on a classic cancer genomics paper (Nik-Zainal et al., 2016) in which the mutation patterns of 560 breast cancer whole genomes were analyzed. After loading the published dataset from the paper, the key results in the main figures can easily be reproduced through a quick dialog with DrBioRight (Figure S1). This side-by-side comparison not only validates the results using our platform but also highlights its potential for improving research reproducibility.

Key Features of the Next-Generation Data Analytics

With the successful development of DrBioRight, and having demonstrated its capability and utility, we propose five key features that next-generation data analytics should possess that will empower a board biomedical research community to explore omics data in an intuitive, efficient, reliable, and collaborative manner (Figure 1D).

Natural Language Understanding (NLU)

Human language is the most natural and intuitive system for communication among people. To serve the broadest research community, it is essential to employ natural human languages (text or voice) as the direct input to bridge users' thoughts with next-generation analytics. By integrating NLU, the analytics reduces the communication barrier for data analysis to a minimum, including identifying and confirming user intentions, translating them into executable bioinformatics analysis tasks, and interpreting and discussing the results in the context of current literature.

Artificial Intelligence (AI)

The next-generation analytics should use data-driven predictive models to correctly translate users' intention, identify appropriate datasets and algorithms, and select informative visualization. Importantly, with users' preference and feedback, the analytics system can, proverbially, "learn on the job" and use those lessons to improve its performance over time through flexible adaption.

Transparency

Reproducibility is a major concern in biomedical research nowadays. Instead of being a "black box," the next-generation analytics should be able to generate detailed analysis reports in real time. The analysis reports will contain detailed information on the dataset, processing procedures, and algorithms, ensuring that the executed analyses are transparent and that the obtained results are reproducible. It is also important to provide functionalities that allow users to check the reproducibility of omics results from published studies.

Mobile and Social Media Friendliness

As the most convenient communication tool, smartphones provide an excellent platform for researchers to perform omics data analysis without the restriction of place and time. Mobile-friendly next-generation analytics will allow greater flexibility in performing data analysis and visualization through smartphone devices. Another desirable feature will be to enable social media functions. Like Facebook Messenger or Slack, through an online chat interface, a user can not only start a one-on-one conversation with the analytics but can also invite collaborators to join a "group discussion" and explore the results together.

Crowdsourcing

To harness the wisdom of crowds, next-generation analytics should actively support open development by the entire research community, including inputs from algorithm developers, data scientists, biologists, and clinicians. This requires building an open-development user center that will allow software dissemination and contributions to and from other bioinformaticians and software developers (e.g., through Docker and GitHub), and a data-sharing system that allows users to share their private data for third-party use.

Toward a New Research Paradigm in Omics Science

Armed with the aforementioned features, next-generation analytics will essentially become an intelligent partner, rather than a tool, that works with human researchers to explore, analyze, and interpret omics data. In such an analytics platform, the AI module is the agile and powerful "brain" that is capable of various cutting-edge bioinformatics analyses and

Figure 1. The Next-Generation Analytics for Omics Data

(A) A timeline showing the major omics data resources and bioinformatics tools in the last two decades.
(B) A snapshot of an online chat interface of DrBioRight,
(C) An overview of the analytic flow of DrBioRight.
(D) Key features of the next-generation data analytics.
(E) A new research paradigm in omics science.

is always kept informed of the latest knowledge and resources; the NLU module allows researchers to efficiently communicate with the "brain" in the convenient format of a dialog, akin to talking to a bioinformatics collaborator, and the social media function promotes teamwork by facilitating the exchange of ideas, tool and data sharing, and team management. With these advances, we envision a new and exciting research paradigm (Figure 1E): a researcher can start a project by directly "talking" to the data analytics and can obtain the desired omics analyses in a timely manner; they can then interpret the obtained results in the context of available literature and even conduct reproducibility checks on published results; during the analysis process, the analytics also helps leverage various resources (data, tools, and expertise) in the community to increase the quality and impact of the researcher's findings; and finally, through possible integration with lab automation and a self-governing system, the analytics can direct lab robots to generate new experimental data that can be used by the analytics to perform further analyses and test new hypotheses.

With continued advances in high-throughput omics technologies, the tremendous amount of omics data that have and will be generated have ushered in a golden era for biomedical research while at the same time presenting us with unprecedented challenges in digesting these data and formulating new hypotheses. Powered by a self-improving AI module, DrBioRight represents an initial attempt at conducting bioinformatics tasks directly through natural languages. Such an analytics platform with the aforementioned features will generate a new research paradigm that maximizes the utility of omics data, accelerates biomedical research, and ultimately leads to better health for everyone.

SUPPLEMENTAL INFORMATION

Supplemental Information can be found online at https://doi.org/10.1016/j.ccell.2020.09.002.

ACKNOWLEDGMENTS

We thank K. Mojumdar for editorial assistance. This work was supported by the US NIH (U24CA209851, U01CA217842, and P50CA221703, H.L.; and P30CA016672), an MD Anderson Faculty Scholar Award (H.L.), and the Lorraine Dell Program in Bioinformatics for Personalization of Cancer Medicine (H.L.).

DECLARATION OF INTERESTS

H.L. is an advisor and shareholder for Precision Scientific.

REFERENCES

Cerami, E., Gao, J., Dogrusoz, U., Gross, B.E., Sumer, S.O., Aksoy, B.A., Jacobsen, A., Byrne, C.J., Heuer, M.L., Larsson, E., et al. (2012). The cBio cancer genomics portal: an open platform for exploring multidimensional cancer genomics data. Cancer Discov. *2*, 401–404.

Chapman, B.A.C.J. (2000). Biopython: python tools for computational biology. ACM SIG-BIO Newsletter *20*, 15–19.

ENCODE Project Consortium (2012). An integrated encyclopedia of DNA elements in the human genome. Nature *489*, 57–74.

Gentleman, R.C., Carey, V.J., Bates, D.M., Bolstad, B., Dettling, M., Dudoit, S., Ellis, B., Gautier, L., Ge, Y., Gentry, J., et al. (2004). Bioconductor: open software development for computational biology and bioinformatics. Genome Biol. *5*, R80.

Ghandi, M., Huang, F.W., Jané-Valbuena, J., Kryukov, G.V., Lo, C.C., McDonald, E.R., 3rd, Barretina, J., Gelfand, E.T., Bielski, C.M., Li, H., et al. (2019). Next-generation characterization of the Cancer Cell Line Encyclopedia. Nature *569*, 503–508.

Giardine, B., Riemer, C., Hardison, R.C., Burhans, R., Elnitski, L., Shah, P., Zhang, Y., Blankenberg, D., Albert, I., Taylor, J., et al. (2005). Galaxy: a platform for interactive large-scale genome analysis. Genome Res. *15*, 1451–1455.

GTEx Consortium (2013). The Genotype-Tissue Expression (GTEx) project. Nat. Genet. *45*, 580–585.

Hutter, C., and Zenklusen, J.C. (2018). The Cancer Genome Atlas: Creating Lasting Value beyond Its Data. Cell *173*, 283–285.

ICGC/TCGA Pan-Cancer Analysis of Whole Genomes Consortium (2020). Pan-cancer analysis of whole genomes. Nature *578*, 82–93.

Nik-Zainal, S., Davies, H., Staaf, J., Ramakrishna, M., Glodzik, D., Zou, X., Martincorena, I., Alexandrov, L.B., Martin, S., Wedge, D.C., et al. (2016). Landscape of somatic mutations in 560 breast cancer whole-genome sequences. Nature *534*, 47–54.

Reich, M., Liefeld, T., Gould, J., Lerner, J., Tamayo, P., and Mesirov, J.P. (2006). GenePattern 2.0. Nat. Genet. *38*, 500–501.

Stajich, J.E., Block, D., Boulez, K., Brenner, S.E., Chervitz, S.A., Dagdigian, C., Fuellen, G., Gilbert, J.G., Korf, I., Lapp, H., et al. (2002). The Bioperl toolkit: Perl modules for the life sciences. Genome Res. *12*, 1611–1618.

Wickham, H. (2009). ggplot2: elegant graphics for data analysis (Springer New York).