Integration of -omic data to understand biological processes and traits

Xianfa Xie[1], Mary Jo Ondrechen[2], Joshua Urrutia[3]

1 Department of Biology, Virginia State University, Petersburg, VA 23806
2 Department of Chemistry and Chemical Biology, Northeastern University, Boston, MA 02115
3 Texas Advanced Computing Center, University of Texas at Austin, TX 78758

Summary: How to understand the molecular basis and mechanism of biological traits is a fundamental question in biology, as well in all the applied fields of biological sciences, such as medical science, agriculture, bioindustry, conservation biology, and ecology. To fully address this question would require the integration of biological data at different levels, including the genomic, epigenomic, transcriptomic, proteomic, metabolomic, and phenomic data. Here we discuss the importance of such research efforts and analyze some of the challenges in integration, as well as offer some future directions.

Introduction

Understanding the molecular basis and mechanisms of biological traits is a fundamental question in biology and in all the applied fields of biological sciences, including medical science, agriculture, bioindustry, conservation biology, and ecology. To answer this question fully would require the integration of biological data at different levels, including at the DNA sequence, gene expression, gene regulation including the epigenetic mechanisms, protein synthesis and modification, biochemical pathways/networks, metabolic, and phenotypic levels. However, the interpretation and integration of genomic, epigenomic, transcriptomic, proteomic, metabolomic, and phenomic data across these levels is a major challenge so far in biology. Here we analyze the importance and challenges in -omics data integration and propose new directions for further development in this interdisciplinary field.

The integration of these different types of data will provide a comprehensive understanding of biological processes and biological traits at the molecular level and, most importantly, the mechanistic link between the genotype and phenotype as revealed by some pioneering studies and the series of ENCODE projects (Li *et al.* 2012, Gerstein *et al.* 2010, The modENCODE Consortium *et al.* 2010, The ENCODE Project Consortium 2012, Yue *et al.* 2014). Novel information can be obtained from the integration of datasets. For example, integrating RNA-seq data with epigenomics data would allow us to characterize which epigenetic marks have a functional impact on transcription, which would not be possible using only RNA or only epigenomic data. Further integration of these data with genomic data could illuminate how the epigenetic regulation is dependent on the genomic composition versus influenced by the environment. For another example, most of the protein structures from structural genomics studies are of unknown or uncertain biochemical function, but integrating them with transcriptomic data will help identify their functions.

The biological sciences have reached a point where vast quantities of -omics data have been amassed. Technologies for generating -omics data are relatively mature and broadly accepted. Integrating these data sources is the next logical step to carry biological knowledge to the next level. Reliable methods for the analysis and interpretation of each type of -omics data are still being developed but can be enhanced in parallel with integration across levels.

Existing Tools

Sequencing technologies have allowed massive amounts of genomic data to be routinely produced. Illumina short-read sequencing has low error rates for individual base pairs, and its newest sequencer (the NovaSeq) can generate up to 3000 gigabases of sequence data in a single run. Nanopore/Pac-bio long-read sequencing technologies can achieve a continuous 2MB for a single read (albeit with relatively high error rates for individual base-pairs). A variety of wet-lab techniques to extract and purify DNA/RNA have been developed to take advantage of these new technologies: ATAC-seq (Assay for Transposase-Accessible Chromatin using sequencing), ChIP-seq, bisulfite-seq, FAIRE-seq, RNA-seq, and exome-seq, each of which provides unique insights into biological processes. Tandem Mass Tag mass spectrometry (TMT-MS) has added new multiplexing capability to proteomics.

Similarly, a variety of software tools have been developed to analyze these data-types: BWA, STAR, GATK, Tophat2, k-mer counting/pseudo-aligners (Kallisto, Sailfish), Subread, DESeq2, EdgeR, Cufflinks, Trinity, SAMTools, SALSA (Wang, 2013), GRASP-Func (Mills, 2018), FastQC, MultiQC, SPADES, VELVET, Falcon, MACS2, SICER. Nonetheless, there exists a deficit in tools that leverage multiple data types to elucidate biological processes. One example, PARADIGM (Vaske *et al.* 2010) incorporates genomic and functional genomics data to infer which pathways are affected in a particular sample. Other methods, like Level of Evidence Scoring and Network analysis represent different data types as networks, and calculate the connectivity between different nodes in these networks in order to connect phenotypes (“anchor nodes”) to the genes (“target nodes”) responsible for those phenotypes (Furches *et al.* 2019). Nonetheless, standardization of integrated data formats, widespread adoption of data integration tools, methods of integrated data visualization, and interactive data exploration remain to be realized.

Besides the methodological challenges for integrating different types of -omic data, there are still some major challenges for integrating all the different levels of -omic data to better understand biological processes underlying important biological traits. First, some types of -omic data are still missing for many organisms, and complete -omic datasets for the same organisms are still rare. On the other hand, some datasets have significant error rates. For instance, in some databases the functional annotations within protein superfamilies have error rates as high as 80% (Schnoes *et al.* 2009). Furthermore, there is no universal agreement in the field for processes to correct errors in databases. In some cases, there does not exist a reasonably accessible process to correct an error in a public database, no matter how substantial the evidence. Secondly, different data formats have been used in various -omic studies, massaging and reformatting data takes an inordinate amount of time and energy. Developing a standardized format for working with integrated data types would be an incredibly valuable contribution to the community, and would allow researchers to devote more time to exploring (rather than reformatting) data. Third, while analytical tools for individual types of -omic data exist, the methods for integrated analysis of all the different types of -omic data still need to be further developed. Fourth, researchers with in-depth knowledge in genomic biology and analytic skills are rare and comprehensive training in both areas for the next-generation researchers is critically needed. Lastly but definitely not least, sufficient funding is needed from federal agencies to support the acquisition and analysis of the different types of -omic data for the same study system, as well as for training next-generation researchers with knowledge and skills in the multiple areas of genomic biology and computational analysis.

Integration of Biology: The integration of -omics data enables linkage of molecular-level properties with organism-level properties, and with all of the levels in between. This effort requires the dismantling of silos through collaborations and training across disciplines, including

molecular biology, genetics, epigenetics, cellular biology, biochemistry, organismal biology, ecology, evolutionary biology, statistics, computational biology, computer science, and bioinformatics.

Broader Impacts: The integration of the different levels of -omics data has important applications in every subfield of biological sciences, including agriculture, medical sciences, bioindustry, conservation biology, ecology, and evolutionary biology and will catalyze a major leap forward in the basic understanding of biology. Thus the subfields will each be impacted in a fundamental way.

References

Furches, A., Kainer, D., Weighill, D., Large, A., Jones, P., *et al.* 2019. Finding New Cell Wall Regulatory Genes in *Populus trichocarpa* Using Multiple Lines of Evidence. *Frontiers in Plant Science*, 10.

Gerstein, M.B., Z.J. Lu, E.L. Van Nostrand, C. Cheng, B.I. Arshinoff, *et al.* 2010. Integrative analysis of the Caenorhabditis elegans genome by the modENCODE project. Science 330, 1775-1787.

Li, X., J. Zhu, F. Hu, S. Ge, M. Ye, H. Xiang, G. Zhang, X. Zheng, H. Zhang, S. Zhang, Q. Li, R. Luo, C. Yu, J. Yu, J. Sun, X. Zou, X. Cao, X. Xie, J. Wang, W. Wang. 2012. Single-base resolution maps of cultivated and wild rice methylomes and regulatory roles of DNA methylation in plant gene expression. BMC Genomics 13:300.

Mills, C.L. R. Garg, J.S. Lee, L. Tian, A. Suciu, G. Cooperman, P.J. Beuning, M.J. Ondrechen. 2018. Functional classification of protein structures by local structure matching in graph representation. Protein Science 27, 1125-1135.

The ENCODE Project Consortium. 2012. An integrated encyclopedia of DNA elements in the human genome. Nature 489, 57-74.

The modENCODE Consortium, S. Roy, J. Ernst, P.V. Kharchenko, P. Kheradpour, *et al.* 2010. Identification of functional elements and regulatory circuits by *Drosophila* modENCODE. Science 330, 1787-1797.

Schnoes , A.M., S.D. Brown, I. Dodevski, and P.C. Babbitt. 2009. Annotation Error in Public Databases: Misannotation of Molecular Function in Enzyme Superfamilies. *PLoS. Comp. Biol.* 5(12): p. E1000605.

Vaske, C. J., Benz, S. C., Sanborn, J. Z., Earl, D., Szeto, C., Zhu, J., … Stuart, J. M. 2010. Inference of patient-specific pathway activities from multi-dimensional cancer genomics data using PARADIGM. *Bioinformatics* 26(12), i237–i245.

Wang, Z., P. Yin, J.S. Lee, R. Parasuram, S. Somarowthu, and M.J. Ondrechen. 2013. Protein Function Annotation with Structurally Aligned Local Sites of Activity (SALSAs). *BMC Bioinformatics* 14(Suppl 3): S13.

Yue, F., Y. Cheng, A. Breschi, J. Vierstra, W. Wu, *et al.* 2014. A comparative encyclopedia of DNA elements in the mouse genome. Nature 515, 355-364.