Editorial

From genome data to biological “understanding”

Volume 2 Issue 2 - 2015

Pedro Martinez
Department of Genetica, University of Barcelona, Spain

Correspondence: Pedro Martinez, Department of Genetica, University of Barcelona and Catalan Institution for Research and Advanced Studies, Barcelona, Spain, Tel 34934035302,
Email pedro.martinez@ub.edu

Received: January 21, 2015 | **Published:** February 13, 2015

Abbreviations: DNA, deoxyribonucleic acid; RNA, ribonucleic acid

Editorial

Over the last few years, a flurry of papers dealing with the generation and analysis of huge amounts of genome and transcriptome data has for the first time provided us with a privileged view of the structure and function of the genomes of many organisms. It is important to point out though that the data is still patchy. While some taxonomic groups, including all vertebrates and insects, are represented well, a large part of the Earth’s biodiversity is still very poorly represented. Our exponentially expanding databases are clearly anthropocentric and reflect our (mostly practical) interests. Nevertheless, we should acknowledge the enormous insights that analysis of these molecular sequences provides us with. Our understanding of the structural organization of genomes, the dynamic nature of genome evolution, the encoding of regulatory information, and the global changes in transcription during development and disease, is all informed by the rich data provided by sequencing projects. Over recent years our view of biology has definitely changed: from a classical “gene-centric” view to the current “genome-centric” one. Gone are the times when analysis of developmental or pathological processes was performed “one gene at a time”. The accumulated data allows us routinely to look at processes on a more global scale (a “systems” perspective, if you like).

Moreover, the complementary datasets-including genomic, transcriptomic and even proteomic data-collected on different biological systems allows us to correlate activity patterns at different levels. Needless to say, those patterns can also be analysed at an unprecedented level of resolution, with the aid of technology that ranges from single-cell sequencing to all kinds of high-resolution microscopies. At any level of research, from the generation and analysis of sequence data to increasing our understanding of cellular (or organismal) processes, bioinformatics has become the crucial toolbox. The complexity of the new datasets has prompted the development of a vast array of informatics tools and the development of bioinformatics as a solid new discipline. In fact, this is a time when many scientists are warning us that there is more data accumulated in databases than manpower (plus the appropriate software) to analyse it!

It is important though that we understand a basic fact of biology: having data, as comprehensive as they might be, does not mean that we understand the processes from which those data are derived. A clear idea of how development or disease proceeds does not naturally stem from the accumulation of huge amounts of data. The integration of different data sources and the correlation of gene patterns are not equivalent to “understanding”, however close it may bring us to it. Understanding can only be derived from a “reconstruction” of the process as it happens “in nature”. This is certainly difficult because it ultimately implies the use of a “reverse engineering” approach. We may know which all the pieces that compose an engine are, and perhaps even know how most of them are arranged with respect to each other, but that does not necessarily mean that we understand how the engine works. This is probably not something that needs to be stated here, though there is still an undercurrent within all modern biology that assumes that more data translates directly into clear understanding.

How could we gain further understanding? One promising way of doing just that when it comes to the analysis of biological processes, is to decipher their controlling gene regulatory networks. These GRNs provide us with a causal description of a process based on the interactions that a set of transcription factors may enter into with their binding targets in the genome of the organism analysed. The transcription factors represent the actors in the process, with mutations in any of them affecting the outcome of the process itself. How these transcription factors interact with each other is encoded in the regulatory apparatus of each gene. Our representation of those interactions, whether activatory or inhibitory, can take the form of a network, with nodes (genes) and connections (interactions). Only a comprehensive network in which all the actors in a process are included will provide a fair understanding of the process itself. Ultimately, this complete representation should function as a predictive tool; a necessary condition for proper understanding of any process. Hence, we should emphasize the need to study GRNs in more detail experimentally (and not just computationally) in different biological contexts. The genomic and transcriptomic data that populate today’s databases should prove to be especially useful when it comes to suggesting candidates involved in the GRNs that control many processes (and in fact, they are). These data are now being complemented with those related to transcription binding sites, enzymatic hypersensitive chromatin sites, in situ screens, mutant phenotype databases, etc. We just need to mine the databases and go back to the bench to develop appropriate tests. This will constitute a most fruitful use of the impressive amounts of DNA/RNA data currently stored in our databases. My aim here is that we should move from the genome data back to the biological processes, while focusing on how the regulatory genome builds cells, tissues, organisms and perhaps even ecosystems. I would like to end with an exercise in restraint, by stating that although GRNs are particularly informative if what is needed is knowledge of how genomes control biological processes; but obviously this is just the vision from the nuclei. Other relevant levels of analysis are needed to gain a fuller understanding; and those range from the cellular to the tissular and the organismal.

These are real challenges we biologists will be facing over the next few years.

Acknowledgements

None.

Conflict of interest

Author declares that there is no conflict of interest.

Citation: Martinez P. From genome data to biological "understanding" *J Investig Genomics*. 2015;2(2):30–31. DOI: 10.15406/jig.2015.02.00018