



Genomics Proteomics Bioinformatics

www.elsevier.com/locate/gpb
www.sciencedirect.com



PERSPECTIVE

Precision Medicine: What Do We Expect in the Scope of Basic Biomedical Sciences?



Jun Yu^a

CAS Key Laboratory of Genome Sciences and Information, Beijing Institute of Genomics, Chinese Academy of Sciences, Beijing 100101, China

Received 20 January 2016; accepted 3 February 2016
Available online 12 February 2016

Handled by Hongxing Lei

The era of Precision Medicine (PM) has entered its execution phase where acquisition of human genome variation data is already leading the charge [1,2]. PM has come a long way to this phase from the Human Genome Project started some 25 years ago [3,4]; it marks the beginning of disease-centric research in the field of biomedical sciences. Other than delivering high-coverage human genomes by millions—of course tailored to patients and common diseases—in the next decade or so, what else do we expect to have and what should we be doing to make the best out of PM projects beyond the current expectation of genetics?

Believe or not, the PM bandwagon is still largely bannered with slogans promoting large-scale, data-driven (or discovery-driven) approaches, such as in-depth discovery of cancer-causing mutations, drug targets, and biomarkers, and may be coupled with hypotheses to balance the currently-debated environment-centric *vs.* mutation-centric arguments. Nevertheless, we are still in time to speculate how to grasp disease-centric PM tasks, largely focusing on various common diseases, which are clear in priority and complex in nature, including but not limited to cancers, as well as cardiovascular, infectious, metabolic, and neurological diseases or disorders. Other than disease-oriented objectives, we should think more

and deeper about other useful data and well-thought tasks in addition to genome sequences, as well as new synthesis as to how to acquire novel data and to interpret them with wisdom [5–7].

First of all, once we have an enormous amount of high-quality sequences, understanding the human population structures and defining haplotypes within and between populations, as well as their disease relevance, are of essence. As the nature of this particular venture is largely informational, population-based sequence variation databases [8] are long-awaited, since the raw data accumulation may exceed the current computation and storage capacities. An integrated database hosts all sequence variations and functional annotation is highly desirable. In addition, mutation biases can also be further used to define function-selected sequence elements beyond protein-coding sequences [9,10]. For instance, when human genes are partitioned into house-keeping and tissue-specific genes, the mutation rate of tissue-specific genes appear 33% higher than that of house-keeping genes [11,12]. This observation suggests that the germline associated genes and chromosome organization hold keys to a variable mutation rate. Another example is the fact that the size of universal introns (size-invariable within lineages) are found—based on population data—to be functionally selected toward size optima, albeit lineage-associated [9,13]. After all, mutations are neither created randomly nor equally in an operational sense for overall DNA sequences and their carriers—chromosomes, let alone selections that are largely attributable to function or phenotype and

E-mail: junyu@big.ac.cn (Yu J).

^a ORCID: 0000-0002-2702-055X.

Peer review under responsibility of Beijing Institute of Genomics, Chinese Academy of Sciences and Genetics Society of China.

<http://dx.doi.org/10.1016/j.gpb.2016.02.001>

1672-0229 © 2016 The Author. Production and hosting by Elsevier B.V. on behalf of Beijing Institute of Genomics, Chinese Academy of Sciences and Genetics Society of China.

This is an open access article under the CC BY license (<http://creativecommons.org/licenses/by/4.0/>).

poorly defined in structural terms of genes and intergenic sequences [14].

Second, studying cellular gene expression and regulation in precision presents another challenge. As a compartmental unit of life, cellular heterogeneity provides functional diversity, even when spoken about asexually divided bacteria. Although transcriptomics is well-respected as a critical paradigm for gene expression study tailored to cells [15], not yet has a single standard human transcriptome been produced, claimed, validated, and hosted in an authorized database. The difficulties are enormous at present time [16]. We have not yet been able to sequence RNA directly in a resolution of copies per cell and to define chemically-modified RNA sequences quantitatively at single-molecule resolution [17–19]. We have not yet been able to separate cellular RNAs into appropriate classes, ribosomal *vs.* total, messenger *vs.* non-coding, small *vs.* large, *etc.* All these call upon a genome-wide large-scale project world-wide: the Human Transcriptomes Project. The Human Transcriptomes Project will certainly come after the sequencing effort in the early phase of the PM project, or being a sequel of it or maybe even sooner, as the current sequencing capacity and tasks will have to be redirected after the genome sequencing effort reaches a peak.

Third, defining DNA structural elements and gene organization as landmarks of chromosomes is undoubtedly a major endeavor. The ENCODE project has been paving ways for thorough definition of operational DNA elements for each cell type and tissue (<http://www.genome.gov/encode/>). There are quite a few unanswered questions along the line. For instance, most human genes are organized into clusters but circadian-regulated genes are cluster-avoiding. How are they synchronized in expression and organized into chromosome territories [14]? How transcript-rich cells, such as testis, brain, and stem cells, are organized to express most of their genes [20,21]? How chromosomes are organized, inherited, and regulated in step-wise changes precisely in germline cells to ensure body development from a zygote? Obviously, a human chromosome-based gene organization map becomes important, which may include experimental data and information, such as sites of chemical modifications, gene clustering and regulation (such as antisense transcription-based regulation), nucleosome occupancy (density *vs.* expression levels), non-transcribed regulatory elements, and organizer-anchorage sequences. We still have a long way toward three-dimensional modeling of the human chromosomes in a dynamic way for development and differentiation.

Fourth, other than informational and operational (structural and interactive functions) systems, rules and nature of various homeostatic processes are also critical and unique, including generation and control of energy, material, and signal transduction. The leptin-adipocyte signal control system represents an excellent example; the discovery of *obese* (*ob*) gene and its mutation in mouse has not yet led to an ultimate cure for obesity [22]. In this particular regime, cellular processes involved in cross-physiological systems are to be deciphered and large PM projects to categorize components and metabolites in circulating and excreting body fluids are to be expected. To measure everything in precision, novel assays and instrumentation are both essential.

Fifth, some PM projects have to go longitudinal as life is after all governed by time. In the dimension of time, we have so many mysteries to be solved; in addition to normal

development and aging, there are more than enough symptoms to be reduced and maybe even cured, including menopause syndrome, Alzheimer's disease, osteoporosis, osteoarthritis, diabetes, just to name a few. In this regime, plasticity or cellular responsiveness to stress signals and materials comes to the center stage, and the degree and timing are both to be measured in precision. Whether the relevant PM projects are named exposomes, stressomes, or platisomes may not be important but time-lapse records and measures are the keys. The connectome projects for neurology of several model organisms together with the Human Connectome Project have been pioneering on cognitive plasticity (<http://www.human-connectomeproject.org/>). Similar projects on the lymphatic system have also come to their time.

It is clear that the stratification of biology into distinct systems is equally important to that of diseases, as we are increasingly capable of exploiting new territories of research fields [23,24]. The fields of genetics, epigenetics, and environment have not provided enough conceptual freedom to allow precise description of genotype–phenotype relationship, where complexity, phenotypic plasticity, and cellular heterogeneity are conceptions frequently used. On the one hand, multi-track biology takes a divide-and-conquer approach to define useful data for understanding disease mechanisms at molecular levels. On the other hand, multi-track biology also takes a systematic approach to integrate data into the information commons that can be synthesized into knowledge on physiological systems as well as diseases, the pathological states of the physiological systems. The medicine side of the PM projects is also disease-centric, and their success largely depends on the organization of cohorts; we also expect the data from this side merges into the same information commons, where mechanisms of diseases are deciphered and strategies are designed to make humans healthier more than ever.

Competing interests

The author declared that there is no competing interest.

Acknowledgments

This work was supported by the “Strategic Priority Research Program” of the Chinese Academy of Sciences (Grant No. XDA08010304) awarded to JY.

References

- [1] National Research Council (US) Committee on A Framework for Developing a New Taxonomy of Disease. Toward precision medicine: building a knowledge network for biomedical research and a new taxonomy of disease. National Academies Press (US); 2011. <http://dx.doi.org/10.17226/13284>.
- [2] Holst L. The precision medicine initiative: data-driven treatments as unique as your own body, 2015, <https://www.whitehouse.gov/blog/2015/01/30/precision-medicine-initiative-data-driven-treatments-unique-your-own-body>.
- [3] National Research Council (US) Committee on Mapping and Sequencing the Human Genome. Mapping and sequencing the

- human genome. National Academies Press (US); 1988. <http://dx.doi.org/10.17226/1097>.
- [4] DeLisi C. The Human Genome Project: the ambitious proposal to map and decipher the complete sequence of human DNA. *Am Sci* 1988;76:488–93.
- [5] Yu J, Wong GK. Genome biology: the second modern synthesis. *Genomics Proteomics Bioinformatics* 2005;3:3–4.
- [6] Yu J. Challenges to the common dogma. *Genomics Proteomics Bioinformatics* 2012;10:55–7.
- [7] Yu J. Life on two tracks. *Genomics Proteomics Bioinformatics* 2012;10:123–6.
- [8] Ling Y, Jin Z, Su M, Zhong J, Zhao Y, Yu J, et al. VCGDB: a dynamic genome database of Chinese population. *BMC Genomics* 2014;15:265.
- [9] Yu J, Yang Z, Kibukawa M, Paddock M, Passey DA, Wong GK. Minimal introns are not “junk”. *Genome Res* 2002;12:1185–9.
- [10] Yang L, Yu J. A comparative analysis of divergently-paired genes (DPGs) of *Drosophila* and vertebrate genomes. *BMC Evol Biol* 2009;9:55.
- [11] Cui P, Ding F, Lin Q, Zhang L, Li A, Zhang Z, et al. Distinct contributions of replication and transcription to mutation rate variation of human genomes. *Genomics Proteomics Bioinformatics* 2012;10:4–10.
- [12] Cui P, Lin Q, Ding F, Hu S, Yu J. The transcript-centric mutations in human genomes. *Genomics Proteomics Bioinformatics* 2012;10:11–22.
- [13] Wang D, Yu J. Both size and GC-content of minimal introns are selected in human population. *PLoS One* 2011;6:e17945.
- [14] Wu G, Zhu J, He F, Wang W, Hu S, Yu J. Gene and genome parameters of mammalian liver circadian genes (LCG). *PLoS One* 2012;7:e46961.
- [15] Wu J, Xiao J, Zhang Z, Wang X, Hu S, Yu J. Ribogenomics: the science and knowledge of RNA. *Genomics Proteomics Bioinformatics* 2014;12:57–63.
- [16] Yu J. The human transcriptomes project: is it hard? *Next Gener Sequenc Appl* 2015;2:e104.
- [17] Marinov GK, Williams BA, McCue K, Schroth GP, Gertz J, Myers RM, et al. From single-cell to cell-pool transcriptomes: stochasticity in gene expression and RNA splicing. *Genome Res* 2014;24:496–510.
- [18] Macaulay IC, Voet T. Single cell genomics: advances and future perspectives. *PLoS Genet* 2014;10:e1004126.
- [19] Ozsolak F, Milos PM. Single-molecule direct RNA sequencing without cDNA synthesis. *Wiley Interdiscip Rev RNA* 2011;2:565–70.
- [20] Cui P, Liu W, Zhao Y, Lin Q, Zhang D, Ding F, et al. Comparative analyses of H3K4 and H3K27 trimethylations between the mouse cerebrum and testis. *Genomics Proteomics Bioinformatics* 2012;10:82–93.
- [21] Cui P, Liu W, Zhao Y, Lin Q, Ding F, Xin C, et al. The association between H3K4me3 and antisense transcription. *Genomics Proteomics Bioinformatics* 2012;10:74–81.
- [22] Li MD. Leptin and beyond: an Odyssey to the central control of body weight. *Yale J Biol Med* 2011;84:1–7.
- [23] Niu Y, Zhao X, Wu YS, Li MM, Wang XJ, Yang YG. *N*⁶-methyladenosine (m⁶A) in RNA: an old modification with a novel epigenetic function. *Genomics Proteomics Bioinformatics* 2013;11:8–17.
- [24] Dominissini D, He C, Rechavi G. RNA epigenetics: DNA isn't the only decorated nucleic acid in the cell. *Scientist* 2016, <http://www.the-scientist.com/?articles.view/articleNo/44873/title/RNA-Epigenetics/>.