Genomics 98 (2011) 242–252

Contents lists available at ScienceDirect

Genomics

journal homepage: www.elsevier.com/locate/ygeno

Perspective

Decoding the genome beyond sequencing: The new phase of genomic research

Henry H.Q. Heng [a,b,c,*], Guo Liu [a], Joshua B. Stevens [a], Steven W. Bremer [a], Karen J. Ye [a], Batoul Y. Abdallah [a], Steven D. Horne [a], Christine J. Ye [d]

[a] Center for Molecular Medicine and Genetics, Wayne State University School of Medicine, Detroit, Michigan 48201, USA
[b] Karmanos Cancer Institute, Wayne State University School of Medicine, Detroit, Michigan 48201, USA
[c] Department of Pathology, Wayne State University School of Medicine, Detroit, Michigan 48201, USA
[d] Department of Internal Medicine, Wayne State University School of Medicine, Detroit, Michigan 48201, USA

ARTICLE INFO

Article history:
Received 4 March 2011
Accepted 18 May 2011
Available online 26 May 2011

Keywords:
Genome theory
Genome context
Genome chaos
Genome instability
Non clonal chromosome aberrations (NCCAs)
Clonal chromosome aberrations (CCAs)
Genetic coding
Punctuated cancer evolution

ABSTRACT

While our understanding of gene-based biology has greatly improved, it is clear that the function of the genome and most diseases cannot be fully explained by genes and other regulatory elements. Genes and the genome represent distinct levels of genetic organization with their own coding systems; Genes code parts like protein and RNA, but the genome codes the structure of genetic networks, which are defined by the whole set of genes, chromosomes and their topological interactions within a cell. Accordingly, the genetic code of DNA offers limited understanding of genome functions. In this perspective, we introduce the genome theory which calls for the departure of gene-centric genomic research. To make this transition for the next phase of genomic research, it is essential to acknowledge the importance of new genome-based biological concepts and to establish new technology platforms to decode the genome beyond sequencing.

1. Introduction

Since the inception of the journal *Genomics* nearly 25 years ago, rapid development of various genomic technologies has greatly advanced the science of genomics. However, despite cutting edge technologies including whole genome scanning [1], global gene expression profiling [2], copy number variation analysis [3] and massive parallel sequencing [4], the understanding of the human genome and the mechanism of human diseases remains a challenging process [5–7]. These powerful technologies have generated scores of data, which paradoxically challenge the framework of current genomics and gene based concepts of common disease, including the rationale of analyzing large numbers of diverse samples with the highest resolution possible. Many diseases are in fact system diseases where sundry genetic variations can be involved in a seemingly stochastic fashion. Furthermore, heterogeneity occurring at multiple levels is a key feature of these diseases (rather than just "genetic noise"), which cannot be addressed simply by sequencing DNA and increasing the sample size [8]. This issue represents the very reason for the failure to identify major causative genes/variants in many common diseases including cancer despite extensive large scale sequencing and whole genome scanning [5].

There are two obvious but somewhat contrary options that can be undertaken to move the field forward. One popular option is to continuously push the limits of technology by increasing the resolution and speeds while lowering costs in order to analyze more samples [4,9,10]. It is believed that studying a larger number of samples will yield the long anticipated genetic patterns of disease by the elimination of 'noise'. Unfortunately, many initial reports of this approach have generated contradictory conclusions, revealing enormous diversity rather than the expected reduction in diversity, and that high levels of genetic heterogeneity seem to be the general rule [5,11,12]. Questions are now being raised about whether data from large scale genomic studies will ever prove to be of promised clinical value, even if each personal genome or "cancer genome" is sequenced [13,14]. The extreme complexity of disease heterogeneity, encompasses the following: low penetrance of specific gene mutations within patient populations; multiple genetic–epigenetic and environmental interactions; and the influences of stochastic evolutionary processes, render most individual molecular mechanisms less than useful for clinical prediction.

A second, new option requires a drastic change in our thinking, and is a departure from the traditional genetic framework and will provide answers from a different perspective or level of genetic organization rather than mainly focusing on DNA and RNA sequences. The basis for this option is that genome alterations are more common and profound than individual gene mutations in most human disease conditions. This new conceptual framework based on the genome

* Corresponding author at: Center for Molecular Medicine & Genetics, Wayne State University School of Medicine, 540 E. Canfield, 3226 Scott Hall, Detroit, MI 48201. Fax: +1 313 577 5218.
E-mail address: hheng@med.wayne.edu (H.H.Q. Heng).

0888-7543/$ – see front matter © 2011 Elsevier Inc. All rights reserved.
doi:10.1016/j.ygeno.2011.05.008

theory calls for a redirect of our efforts to systematically decode genetic information stored at the genome level [7,15,16]. Initial calls for just such a change have emanated from the field of cancer research [5,7,17–26].

We strongly support the option of refocusing on the entire genome (not just the sequence of), not only because this approach has been overlooked, but also because of its ultimate importance in understanding how the entire genome functions and in the underpinnings of the general mechanism of many common complex diseases [7]. In this perspective, we will briefly discuss the key differences between genes and the genome by walking readers through our own experience of making the transition from the gene centric view to the genome theory [27]. In particular, to convince readers of the importance of this issue, we would like to point out that current genome study efforts only decode parts of the genome and do not address the key issue of decoding the genome as a whole system. Even for the ENCODE project (the encyclopedia of DNA elements) and Human Epigenome project (as well as many other 'omics' projects) [28,29], the conceptual framework is still the gene theory. At a fundamental level, DNA sequences (including their chemical modifications) and the genome represent distinctively different levels of coding and system control, and future genomic research must directly address the issue of genome coding, genome system control, and how interactions work across different genetic and epigenetic levels [8]. Equally important, new technical platforms are urgently needed to synthesize information at the higher levels and integrate them with the genome system. Emergent properties of the genome suggest that system information at the genome level is not a simple summary of gene sequence information. New technologies also need to integrate other key features of normal and diseased genomes such as: heterogeneity at multiple levels; differences between the system status (such physiological and pathological conditions where pathological conditions often involve genome alterations) and stochasticity of somatic evolution.

2. Gene vs. genome

2.1. The genome is not equal to the sum of all genes or its entire sequence

The genome is the entity containing an organism's hereditary information and the main evolutionary selection platform [27,30]. Traditionally, heritable information has been thought to be encoded exclusively in DNA and RNA sequences. The current, popular concept of the genome where the collection of all genes and non-coding sequences explain a given species has been influenced by the gene centric concept. A key unique feature of the genome however, is the genomic topology (a multi-dimensional interactive relationship that exists between genes and is the physical basis of genome architecture) and the emergent properties that exist at this higher level which have been largely ignored. As a result, the terminology, the Human Genome Project, as used by sequencing consortiums, implies that decoding DNA is equal to decoding the genome. This has spawned many popular but incorrect analogies, including considering the genome to be a book, where each chromosome represents a chapter. The problem with this metaphor is that one cannot simply read the basics of each chapter and comprehend it without including the multi-dimensional interactions within the system. A chromosome does not stand on its own as a biological entity and therefore there are no meaningful messages based on individual chromosomes. To put it succinctly, a simple parts list does not give a clue as to the assembly instructions. Similarly, the conventional statement that the sequencing of the human genome has provided a roadmap (or the foundation) of modern biomedical research is flawed. It is flawed particularly with regard to nonlinear systems as most complex systems have multiple levels of organization and the nonlinear relationship between them is connected through emergent properties, which is difficult to understand by only summarizing information of the lower parts. The reductionist tradition of understanding the "parts" first before understanding the "whole" system is only effective in a linear system [15,31].

Clearly, considering the relationship of the parts (genes) versus the whole (genome), where the whole is more than the sum of all the parts, and includes the 3-D interactive structure of the genome within nuclei, the current reductionist approach of treating the genome as a "bag of genes" or a collection of linear DNA structures does not mirror the complexity of the genome system. To illustrate this point, we have introduced the term, 'genome context', to differentiate the gene and the genome [21,27]. Genome context (the DNA sequence plus the genomic topology), rather than gene content, defines the structure of a genetic network and is the total interactive package that functions in organismal and somatic cell evolution. The revelation of the genome context results in the need for additional, crucial questions to be asked within the genomic community. These include: When a genome is altered, does the same gene mutation have the same biological meaning as it does in the original, unaltered genome? If the genome's key properties are emergent from the DNA level and are fundamentally different from DNA itself, is DNA sequencing crucial to understand the function of a eukaryotic genome?

In the real world the difference between a parts list and the blueprint of parts assembly is clear. In genome research, however, this key difference is often forgotten. It is thus necessary to separate genes and the genome as these are two different entities in genomic research. The difference between them is not just a difference of quantity but also a difference of the level of organization. One interesting analogy is to consider genes as building material, parts or even tools, and the genome as the architecture. The same materials can be used to construct different architectures with distinct functions [21]. As pointed out by Barbara McClintock, the genome is the organization that is responsible for activating and reconstructing in response to environmental challenges [32].

It is true that increasing attention has been given to the "gene context" rather than genes alone. However, without the conceptual framework of the genome theory, isolated approaches will not solve most paradoxes that the gene theory created [27].

2.2. Different genetic coding and control systems

If the above analogy that the gene/genome relationship represents a part versus the whole relationship is correct, then it is obvious that genes and genomes represent different levels of the system with different mechanisms for coding genetic information. Understanding how DNA codes RNA and proteins has been a major achievement of molecular biology [33], however, the codes for producing the parts using DNA and the codes for assembling parts using the package of the genome are very different. Current systems biology suggests that the function of the genome depends on the genetic network, without knowing what defines the genetic network structure in the first place. In contrast, genome theory states that the genome context defines the structure of a genetic network by creating the structural or topological basis of genetic interactions [21,27], thus a specific network is the emergent property of a given genome (Fig. 1).

According to the genome theory, both genes and the genomic topology are the key to maintaining the genome system. Therefore, new systems can be effectively created naturally either by creating many novel genes, or by re-organizing the existing genome through changing the genomic topology rather than gene content. Many less complex species like sponge and water fleas have comparable or even more genes than humans (comprised of a simple body plan without organs, muscles or nerve cells, the sea sponge has 18,000 genes compared to 21,000 genes for humans, while Daphinia water fleas have over 30,000 genes) [34,35], and most mammals share similar genes but have very different chromosomal arrangement and