Making sense of the threads of ACGT

Rakesh Mishra, CCMB

How much we do understand regarding how information contained in DNA is stored and expressed? Genomes consists of genes and regulatory elements that control their expression. But these constitute of only a fraction of the genome. Most of the genome remains unexplored in terms of functional elements. A large amount of data in the form of genomic sequences and epigenome features have been accumulated using approaches like Next Generation Sequencing (NGS), sequencing using chromatin immunoprecipitation (ChIP-seq), high-throughput sequencing to capture chromosome conformation (HiC), etc. However, there is substantial gap in analysis and in making sense of this huge information content. New tools, approaches and theoretical perspectives/models may help us understand how genetic information is stored, contained and used by living forms.

A few basics

- Genetic information that is inherited from generation to generation and forms the basis of life processes is in the form of DNA.
- DNA is a polymer made of four units, viz., A, C, G and T. It is the sequence of these four units that forms the basis of information content in DNA.
- The complete set of DNA molecules present in the form of chromosomes is called the genome.
- Typically, a functional unit of DNA sequence is called gene which is transcribed as RNA. While some RNA molecules may be the final functional product, others are translated into proteins as the final functional product.
- Most of the known genes constitute only a small fraction of the genome in "higher" organisms. The remaining part of the genome is being actively researched for the presence of new functional elements.
- Lots of regulatory elements have emerged from such efforts that partly explain how genes are transcribed in a regulated manner.

What we see

- When genomes of different organisms are compared, it turns out that number of genes does not change much when we go from simple to complex organisms.
- The minimum size of the genome in relatively simpler organisms is smaller than that of more complex organisms, indicating that more complex organisms have more DNA or larger genome size.
- Developmental processes that lead to the formation of a complex organism starting from a single cell are based on the choice of the combination, level of expression aand temporal profiles of genes which eventually define a cell type.
- The more complex an organism, the larger the number of its cell types (humans have >200 cell types while flies have 50 cell types).
- A major question remaining to be answered is how such sets of genes are picked for expression and how the expression states are maintained during the entire life of the organism.
- This is a very important process, as each cell type has the entire set of genes but uses only a subset of those. For example, intestinal epithelial cells produce proteases that help digest proteins we eat, and not digest ourselves even though we are largely made of proteins. Brain cells also have the genes that make proteases, but here such genes are kept completely inactive to save the organ.

Potentials and possibilities

- The entire genome set is accommodated within the nucleus of the cell; this requires a great degree of compaction and packaging.
- In this process of genomic packaging, some parts may remain accessible and, therefore, available for expression while rest may go into an inaccessible form and, therefore, not be available for expression.
- Packaging of the genome, by this logic, offers a means to divide genome into accessible and inaccessible states and this may be a key to differentiation process which sets a subset of genes active in a cell type. It may even be causally linked to the cell type.
- According to this line of thought, complex organisms package their genomes in as many ways as the number of cell types they have.
- Much of the DNA that is not used for transcription and that has steadily increased during evolution of complexity, may code for this multiplicity of the packaging process.
- What is this packaging code? How to look for it?

New technologies and new insights

- Recent advances in technologies have started to yield insights by revealing the never seen before picture of nuclear interiors.
- We can now map at a whole-genome scale, specific DNA methylation patterns, histone modification patterns, precise binding sites of virtually any chromatin associated factor, accessible and inaccessible sites, etc.
- We can also estimate the proximity of any genomic region to any other part of the genome. This indirectly reveals the precise packaging state of the genome.

A large number of genomes have been sequenced and many features have emerged that remain largely to be understood

- Certain classes of DNA sequences have steadily accumulated with a concomitant increase in complexity, which includes a variety of repetitive elements including simple sequence repeats (SSRs), a.k.a., micro satellites.
- A number of observations suggest that accumulation of these repeats is not random. For example, type of repeat, their size, distribution across the genome, presence of proteins with specificity of binding to given repeat, etc., point to a functional relevance of this class of DNA.
- There is a pattern in the type of repeat composition and the evolutionary relatedness of organisms. Have the non-coding parts of the genome contributed to evolution of complexity by bringing in novel regulatory features?

Furthermore

- There are a large number of proteins that are primarily part of the process of genomic packaging. This includes histones, non-histone chromatin proteins, proteins that associate at specific sites and determine their functional status by either introducing epigenetic modifications or recognizing such modifications to deliver intended function.
- Many of these proteins are not only conserved across eukaryotes but often show expansion in higher organisms.
- There is a parallel between the accumulation of more, so-called non-coding parts of the genome and more chromatin-level regulatory factors with the evolution of complexity.

Can we put these facts together and theorize how information is accumulated, stored and expressed in the DNA-based information system of life?