nature immunology

Resources for life

Understanding how gene expression is regulated requires much more than the identity of genomic sequences.

Sequencing the entire human genome by the year 2000 was an immense goal. Hailed as discovering the blueprint for life, that project faced many logistical and technological challenges. The scientific community engaged in open debates on whether its substantial cost would affect the funding of other investigator-initiated science. In hindsight, it is evident that the genome sequence has provided a framework with which to address more fundamental questions, such as how this 'handbook of genetic instructions' is dynamically executed to achieve appropriate developmental and tissue-specific gene expression in a person. However, assembling the complete sequence of the human genome was only the beginning of genomic science.

Even before the draft sequence was finished came the startling realization that protein-encoding exons represent only a small fraction of the genome. Without knowledge of its function, much of the genome was relegated to 'junk' status. Interpretation of the genome required another large-scale effort. This begat the ENCODE consortium, whose mission is to intensively annotate the human genome and to generate a community resource called the "Encyclopedia of DNA Elements." The ENCODE resource is derived from a multitiered systematic analysis of RNA expression, chromatin modification, transcription-factor binding and chromatin accessibility and connectivity compiled from 147 different human cell types, including various immortalized cell lines, primary cells and stem cells. The recent publication of 30 papers in *Nature*, *Genome Research* and *Genome Biology* represents the vanguard of insights that are emerging from this vast collaborative project. Reassuringly, one of the major insights of this first wave of analyses has dispelled the idea that much of the genome is junk. Instead, functional attributes have been assigned to over 80% of the genome. These studies have established rules for coregulated gene expression by identifying common patterns of *cis*-acting motifs through the discovery of new motifs and combinatorial interactions of transcription-factor binding, including those that require chromatin looping to bring multiple interacting regulatory sites into proximity.

The effort to understand genomes has proved to be a driver of technological innovation. Before the ENCODE goals could be achieved, new sequencing technologies and new analysis tools needed to be developed to ensure that the program would be more cost-effective and less time-consuming than the original genome project. Standardized methods of data acquisition had to be derived to generate robust data sets that would allow direct comparison of results generated from different laboratories. Advances such as the development of microarray platforms have allowed the systematic investigation of gene expression. Next-generation high-throughput sequencing technologies have vastly accelerated sequence acquisition by improving speed and lowering sequencing costs. Genome-wide association studies have led to the identification of single-nucleotide polymorphisms significantly associated with genetic disease traits. Specific histone-DNA and regulatory protein–DNA interactions can be identified by chromatin-immunoprecipitation and mapped to the genome through subsequent deep sequencing. Chromatin-conformation capture followed by high-throughput sequencing or related technologies (HiC, 5C and ChIA-PET) has identified chromatin architecture and higher-order protein-mediated interactions between noncontiguous genomic sequences. Comparative genomic analyses, made possible by sequencing of the genomes of other species, have also provided important insights. Such studies have helped to identify conserved sequences, including many in noncoding regions, involved in regulating gene expression. The ongoing development of new computational algorithms and statistical tools has likewise made it possible to manage and systematically investigate the vast amount of information that continues to be generated.

Hematopoietic cells, given the relative ease of their isolation from the blood, represent a tractable system for measuring changes in gene expression that accompany developmental progression and can be used to identify lineage relationships based on the hierarchical clustering of coregulated gene expression. However, gene expression by effector cells may prove to be more complex, as hinted at by the patterns of hypersensitivity to DNAse I observed by Stamatoyannopoulos and colleagues (*Nature* **489**, 75–82, 2012). In that study, genes involved in the immune response had the most diverse array of enhancer-promoter interactions, which perhaps reflects the need for tight control to elicit a robust but transient immune response to prevent collateral tissue damage or autoimmunity.

Nature Immunology has published systematic analyses of transcriptional profiling of cells of the mouse immune system done under the aegis of the Immunological Genome project consortium (including the Resource in this issue by Randolph and colleagues). Although these studies have not examined the genome-wide chromatin landscape associated with gene-expression patterns as described above, they have identified developmental relationships and common functional pathway modules based on gene ontogeny patterns. Previously unanticipated phenotypic molecules and transcription factors have been identified as unique markers of various stages of lineage development. These markers and nodal points can be modulated to assess their role in regulating the development or function of cells of the immune response and/or their contribution to immunity or tolerance. Moreover, how these factors influence gene expression, or are altered themselves, in response to overt immunization or infection can likewise be easily tested in various mouse models. Cross-comparison of mouse expression profiles with those of the human ENCODE project should also provide evidence of functional relevance.

The ENCODE and Immunological Genome consortia are providing rich resources to the biological science community. These framework data sets and associated analytical tools make it possible to formulate more specific hypothesis-driven questions and to identify common patterns of gene expression and regulation. The immunology community should embrace such efforts.