Cell

Human Genome: What's Been Most Surprising?

Let's Remember the Chromosomes

David Page
Massachusetts Institute of Technology

What surprises me most about the state of the human genome in 2011 is how much of the sequence remains to be assembled accurately and how important that achievement would be for fulfilling the Human Genome Project's original goal: a comprehensive reading of the book of life. Achieving this aim will require conquering the most structurally ornate and dynamic regions of the genome, including the essential but elusive elements (such as centromeres and telomeres) that are responsible for faithful transmission of the genome from one generation to the next. Students of human biology and medicine would finally be able to see the genome as an orchestra of chromosomes—not merely a "parts list" of genes. We would, at last, be positioned to address some of the longstanding mysteries of human biology and medicine, such as the fragile and tenuous nature of human reproduction, and we would understand why a sizeable proportion of all human conceptions—and half of spontaneously lost pregnancies—display dramatic anomalies of one or more chromosomes. Just as high-resolution crystal structures of macromolecular complexes enable unforeseen insights into function in health and disease, a complete and accurate assembly of the human genome will answer questions that we do not even know to ask.

Variation and Complexity

Vivian Cheung
University of Pennsylvania

The simplicity of A, C, G, and Ts as building blocks of the human genome is deceptive. Although various genome-scale projects seek to identify functional units based on DNA sequences, it is surprisingly difficult to find genes and regulatory elements, and such information is critical for determining how sequence variants affect disease risks.

DNA sequences are highly processed and modified during transcription and translation. The same DNA sequence can code for different proteins by alternate promoter usage or splicing, and base modification such as methylation and chromatin modifications also affect how DNA sequences are converted into functional units. In addition, DNA sequences are not always copied exactly into RNA and proteins; processes such as RNA editing lead to proteins that are not encoded by the underlying DNA sequences. Even after the proteins are synthesized, different modifications affect their functions.

Although DNA is viewed as the template for transcripts and proteins, there is a lack of a direct relationship between DNA sequence and functional elements. A deeper understanding of the fundamental yet complex relationships between DNA and RNA (as well as proteins) is necessary to assess the functional significance of genome variation. An approach may involve studying individual variation in the transcriptome and proteome in addition to DNA polymorphisms to determine which and how genetic variants affect disease susceptibility and response to therapy.

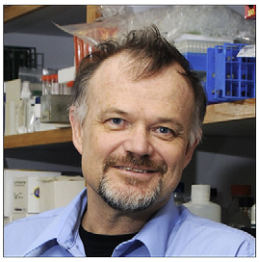
A Hidden Ecosystem

David Haussler
University of California, Santa Cruz

The role of functional, noncoding DNA has been simultaneously the biggest surprise and the biggest mystery emerging from our first glimpse of the human genome. Comparative sequencing of the genomes of other species revealed that, since the inception of placental mammals, more than 5% of the human genome has been under selective constraint to preserve some presumably fitness-enhancing function. Careful gene structure analysis proved that only about 1.2% codes for protein. Many of the most dramatically constrained genome segments are noncoding. What does this "dark matter" do, and how did it evolve? Ample experimental evidence suggests that the bulk of the evolutionarily constrained segments will have gene regulatory functions either as DNA elements or as parts of noncoding RNA genes. It has become clear that these elements turn over and are reinvented anew on a much faster timescale than the protein-coding genes.

The other big surprise is the mounting evidence that transposons—mobile elements within a cell's DNA that are viewed by most as almost exclusively parasitic—play a critical role in this turnover. Previously viewed as a library of cellular information infested with a few parasites, the genome has turned out to be a lively and complex ecosystem unto itself, with many unusual denizens partaking in complex evolutionary alliances and battles that remain to be deciphered.

Huge Heterogeneity



Richard Gibbs
Baylor College of Medicine

It is remarkable how much of our perception of the human genome is shaped by the nature of its "accessible" portions. The regions that are unique, more easily finished, and accessible to high-throughput genotyping are naturally the best studied. This excludes many repetitive regions and segments that are otherwise recalcitrant, such as those with high GC content. Many interesting and disease-related alleles fall in these regions. It is a shock to hear from colleagues involved in medical diagnostics that, for some regions, "we ignore the reference sequence." This is because there is enormous population heterogeneity, and the different alleles involve complex rearrangements. The assessment in the diagnostic laboratory is bootstrapped by all of the data from the patients that are tested with standard arrays and only modest dependence on the reference.

Another surprise is that there are many more private, even "unique," and potentially functionally significant variants in populations than we expected. This means that it will be a long time before all of the important alleles that contribute to human health have been discovered. And in general, when new sequences are examined, there will be newly discovered alleles for which the greatest challenge will be assessing their functional significance.