It Turns Out There is Something Hidden in All That Junk (DNA)

High-throughput sequencing is revealing the ubiquity and significance of non-coding RNA.

Introduction

It has long been understood that only a small percentage of the human genome consists of protein coding sequence. These sequences, or genes, underlie the usual interpretation of the central dogma of molecular biology that genetic information is expressed as proteins, and that proteins transact most cellular functions. The remainder of the genome, the vast majority with no protein-coding information, has historically been disregarded as junk DNA, evolutionary debris with no critical function to the system. Any evidence to the contrary was largely ignored, or dismissed as noise associated with the analytical method. But that is all changing now. A mounting body of evidence unlocked by high-throughput sequencing is shedding light on these dark regions of the genome and generating insights about the functional significance of their transcripts. With heavy implications for the genetic architecture of traits and diseases, these findings are challenging long-held notions about the mechanisms that regulate biological processes. At the forefront of these discoveries is John Mattick Ph.D., a professor at the Institute for Molecular Bioscience at the University of Queensland in Australia.

Beyond Genes and Proteins

Dr. Mattick was the founding director of the institute, stewarding its growth from a small research center to a substantial institution of more than 400 scientists. After 20 rewarding years at the helm he decided to step aside to pursue his research interests. As he recalls, “I wanted to concentrate on my research because I thought that I had seen something that nobody else had.”

While at Baylor College of Medicine as a post-doctoral fellow in the late 1970’s he encountered the idea of introns. Large non-coding sequences spliced out of genes and apparently discarded, they were thought to have no functional importance. Thinking this was a strange conclusion, Dr. Mattick began to consider the alternative notion that perhaps this spliced RNA was functional; maybe its role was to send some type of communication into the system. If that were the case, it would mean that the system was configured much differently than previously understood and these RNAs must be involved with some sort of regulatory network. Looking back, he remembers “in an epiphany moment, I thought maybe the network is the thing that underpins the ontogeny of complex organisms. So there’s a whole layer of information here that we overlooked.” In the years that followed, this hypothesis became his intellectual hobby and the foundation of his research interests.

John Mattick, Ph.D., Professor at the Institute of Molecular Bioscience at the University of Queensland in Australia.

Before the arrival of next-generation sequencing technology, it was difficult for scientists to advance the idea that there might be some other significant element facilitating biological functions beyond the well-accepted notion of genes and proteins. As an enthusiast of scientific history, Dr. Mattick points out that as far back as the 1950’s, the great Barbara McClintock wrote a letter to her friend Marcus Rhoades questioning whether the increasingly established philosophy of the gene as primarily a protein-coding entity was valid. In light of her observations about transposition, for which she later won the Nobel Prize, she questioned whether the rigid notion of how genetic information is transacted was indeed correct. Dr. Mattick goes on to recite numerous other surprising observations over the past 60 years. From the discovery of abundant heterogeneous nuclear RNA to the observation that transposons were ubiquitous in eukaryotic genomes, there has been plenty of evidence over the years suggesting that the regions of the genome outside of protein coding sequences are very active indeed.

Sequencing Revels New RNA Insight

Throughout his career, Dr. Mattick continued to explore the eukaryotic genome from a different perspective, becoming more and more convinced that the system underlying complex organisms was largely misunderstood. However, in the absence of direct evidence, it was difficult for the ideas to get much traction in the scientific community. He recounts, "Apart from the conceptual idea that there's lots of intronic RNA being produced and that transposons were everywhere, we really couldn't say much more without sequencing." It wasn't until the first eukaryotic genome projects, enabled by capillary sequencing technology, that researchers were able to show there were transcripts everywhere—not just introns, but many antisense and intergenic transcripts as well. The arrival of next-generation sequencing technology brought forth more transcriptomic data and even more compelling evidence. He explains, "This was the foundation because now we could not only detect many of these transcripts through high-throughput sequencing, but equally important, we could determine their structure and actually start to study them as individual RNAs."

Despite these revelations, Dr. Mattick observes that many in the scientific community have ignored or dismissed the data and alternative interpretations. "Some of those who were aware or came across it said that's interesting, but it wasn't in their critical path for studying prostate cancer or whatever. So, many people ignored it, or more dismissively, would say 'it's probably just transcriptional noise.'" Yet, from his perspective, "noise" was only one possibility and lacked evidentiary support. As he explains, everything known about complex systems in other domains, such as computing, suggests that they are designed to minimize noise. As such, the same should be expected from the human body. Moreover, he adds, "The published literature on 'transcriptional noise' refers to the stochastic firing of genuine promoters, not illegitimate transcription from illegitimate promoters. It's examining, for example, the lac promoter and seeing how it fires, how it sort of pings off occasionally, and nobody has ever documented in bacteria or yeast a promoter that was not supposed to be a promoter."

RNA Capture-Seq

To address the perceptual gap in the scientific community, Dr. Mattick's team, led by Dr. Tim Mercer, collaborated with Dr. John Rinn at Harvard and Dr. Jeffrey Jeddeloh at Roche on a project to develop a new method, termed RNA Capture-Seq, which uses Illumina sequencing to provide a more detailed view of the transcriptome[1-2]. In principle, the idea is similar to exome sequencing. Researchers create a capture microarray to target an intergenic region or other locus for interrogation. The captured cDNAs are then eluted and sequenced. Using this method, they began to look at a series of samples to evaluate the scattered tags in a deep sequencing data that had previously been attributed to noise. For comparison, they first sequenced the samples using conventional RNA-Seq so that they had both types of data for each sample. What appeared as scattered tags in the conventional deep sequencing data set were revealed as full transcripts with detailed splice patterns when analyzed using the RNA Capture-Seq method. Dr. Mattick explains, "They just appeared out of the mist, and there's no way you can look at this data and say this is noise. They are transcripts that are difficult to see in conventional deep sequencing data sets. It's a huge advance built on a mixture of array technology and sequencing, and Illumina does both very well."

According to Dr. Mattick, the significance of this method is not limited to proving that these transcripts are real. It provides a powerful tool for endeavors like genome-wide association studies, allowing researchers to document all of the gene products at the RNA level, including those that go on to form proteins. He and his colleagues have uncovered a number of striking results using the RNA Capture-Seq method. They've found not only a plethora of new transcripts, most of which are non-coding, but also many novel splice variants, even among some of the most well-studied genes. In one example, they identified four new isoforms of p53, the "guardian of the genome." Three of these isoforms actually changed the domain structure of the protein. They've also identified new variants of hox genes, known to encode highly conserved transcription factors that regulate morphological development. The implications of these findings may greatly impact our understanding of the critical biological mechanisms governed by these genes.

> "Apart from the conceptual idea that there's lots of intronic RNA being produced and that transposons were everywhere, we couldn't say much more before sequencing."

Traditionally, many of these transcripts and splice variants were missed because they appeared at very low levels in a whole-transcriptome sequencing data set. What Dr. Mattick has pointed out is that in some cases they are expressed in just a few cells within a tissue. When the tissue is homogenized for a sample, the transcripts only appear at low levels within the cell population. Thus, very deep sequencing is needed to actually see them. He explains, "We have shown that to get the same depth of coverage of the transcriptome on the Genome Analyzer™ or any platform, you'd need at least 10 billion reads for the same region. With the HiSeq® system, we will be able to go in much greater depth and get those numbers. But if you're interested in a particular region or locus, then you need to use the RNA Capture-Seq method to interrogate the transcripts."

For Dr. Mattick, evidence for the ubiquity and functional significance of these non-coding transcripts is overwhelming, demonstrated in an increasing number of publications. He predicts, "I think it's going to become a very dynamic area of research. Consider the reverse as an analogy: if having worked on RNA all these years, we suddenly discovered proteins and said, wow, there's 20,000 of them. There are splice variants. It's like a new world." In light of these discoveries, he believes that many biological fields are on the precipice of a revolution in terms of how they view the functionality behind the system. This is especially true for developmentally complex organisms, where the architecture of the system requires a tremendous amount of information. Here the non-coding RNA can have a major impact on the epigenetic control of development. They can provide insight about the complex, sophisticated mechanisms that control human ontogeny, explaining

the reasons why each person looks different from another. Dr. Mattick illustrates this point by considering identical twins, each resulting from the split of an embryo into two organisms that are virtually identical aside from the differences imposed by life. The high similarity between them throughout adulthood suggests that the genetic program is very robust. He elaborates, "The differences between two persons are not because of some stochastic process, but because their genomic specifications and developmental program differ, controlling all the unique characteristics of an individual. Making a muscle cell is easy, but making all the fine muscles in your face with the different shapes and anchor points—that's a big job, and requires huge amounts of information."

"I didn't foresee the spectacular increase in sequencing output and decrease in cost. I knew it was going to come, I just didn't know what the pace would be. This is the only way you can learn about the system—It has to be through sequencing."

The Next Frontiers of RNA Research

Another observation Dr. Mattick makes is that this RNA regulatory system, put in place to control a precise ontogeny, is acted upon by evolution to confer plasticity, or modify the information. The mechanism through which this occurs appears to be RNA editing. He explains, "This fundamental advance gave the brain the epigenetic plasticity it needed to be able to learn. The evidence for that is compelling. It appears that the RNA-based system that evolved to control differentiation and development has been elaborated and made plastic in the brain as a precondition for the evolution of higher-order information processing." If that hypothesis is correct, it will revolutionize neuroscience, the last frontier in biological terms. His lab has now started a program to study RNA editing using deep sequencing. They have developed a protocol to enrich for certain editing signatures like adenosine-to-inosine transitions, which are common in primates. In terms of sequencing, inosine behaves like guanine, so they have developed a cleavage-based method to enrich for these edits in order to sequence them and explore the phenomenon more efficiently.

The years of research ahead will determine where these discoveries will lead. What is clear is that availability of high-throughput RNA sequencing has enabled forward thinkers like Dr. Mattick to open a window to exciting new possibilities. He reflects, "I didn't foresee the spectacular increase in sequencing output and decrease in cost. I knew it was going to come, I just didn't know what the pace would be. This is the only way you can learn about the system—it has to be through sequencing. There is a whole world that we need to discover to understand human evolution, development, and cognition properly—we're going to have to go there." For Dr. Mattick, we are going there, revealing a new world of RNA-based regulatory systems—the hidden layer in human genetic programming. The alternative notion he considered many years ago logically led him to the discovery of something very significant in what most had considered junk. If his foresight is correct, these discoveries have just scratched the surface of breakthroughs that will transform our understanding of the system that underlies human life.

References

1. Mercer TR, Gerhardt DJ, Dinger ME, Crawford J, Trapnell C, Jeddeloh JA, Mattick JS, and Rinn JL (2011) Targeted RNA sequencing reveals the deep complexity of the human transcriptome. Nature Biotechnology, epub ahead of print, doi:10.1038/nbt.2024.
2. Roberts A and Pachter L (2011) RNA-Seq and find: entering the RNA deep field. Genome Medicine, 3:74.

Illumina, Inc. • 9885 Towne Centre Drive, San Diego, CA 92121 USA • 1.800.809.4566 toll-free • 1.858.202.4566 tel • techsupport@illumina.com • illumina.com

FOR RESEARCH USE ONLY

© 2011 Illumina, Inc. All rights reserved.
Illumina, illuminaDx, BaseSpace, BeadArray, BeadXpress, cBot, CSPro, DASL, DesignStudio, Eco, GAIIx, Genetic Energy, Genome Analyzer, GenomeStudio, GoldenGate, HiScan, HiSeq, Infinium, iSelect, MiSeq, Nextera, Sentrix, SeqMonitor, Solexa, TruSeq, VeraCode, the pumpkin orange color, and the Genetic Energy streaming bases design are trademarks or registered trademarks of Illumina, Inc.
All other brands and names contained herein are the property of their respective owners.
Pub. No. 070-2011-034 Current as of 12 December 2011

illumina®

AGAATGATAACAGTAACACACTTCTGTTAACCTTAAGATTACTTGATCCACTGATTCAACGTACCGTAACGAACGTATCAATTGAGACTAAATATTAACGTACCATTAAGAGCTACCGTCTTCTGTTAACCTTAAGATTACTTGATCCACTGATTCAAC
TCAACGTACCGTAACGAACGTATCATTAAGATTACTTGATCCACTGATTCAACGTACCGTAACGAACGTATCAATTGAGACTAAATATTAACGTACCATTAAGAGCTACCGTGCAACGACGAAAAGAATGATAACAGTAACACACTTCTGTTAACCTT
CGACGAAAAGAATGATAACAGTAACACACTTCTGTTAACCTTAAGATTACTTGATCCACTGATTCAACGTACCGTAAAGATTACTTGATCCACTGATTCAACGTACCGTAACGAACGTATCAATTGAGACTAAATATTAACGTACCATTAAGAGCTACC
AACGTACCATTAAGAGCTACCGTGCAACAGTAACACACTTCTGTTAACCTTAAGATTACTTGATCCACTGATTCAACGTACCGTAACGAACGTATCAATTGAGACTAAATATTAACGTACCATTAAGAGCTACCGTGCAACGACGAAAAGAATGATAA
AGAATGATAACAGTAACACACTTCTGTTAACCTTAAGATTACTTGATCCACTGATTCAACGTACCGTAACGAACGTATCAATTGAGACTAAATATTAACGTACCATTAAGAGCTACCGTCTTCTGTTAACCTTAAGATTACTTGATCCACTGATTCAAC
GATTACTTGATCCACTGATTCAACGTTAAGATTACTTGATCCACTGATTCAACGTACCGTAACGAACGTATCAATTGAGCTTCTGTTAACCTTAAGATTACTTGATCCACTGATTCAACGTACCGTAACGAACGTATCAATTGAGACTAGCAACGAGC
CGTATCAATTGAGACTAAATATTAACGTACCATTAAGAGTCTGTTAACCTTAAGATTACTTGATCCACTGATTCAACGTACCGTAACGAACGTATCAATTGAGACTAAATATTAACGTACCATTAAGAGCTACCGTGCAACGAAAAGAATGATAACAGTA