Sorry, But It's Not Your Father's Genome

No, I'm not referring to the startlingly high prevalence of non-paternity in our society, rather to the fact that our understanding of the genome is changing rapidly and drastically. This evolving understanding is leading to advances in therapeutic and diagnostic technology that even 10 years ago would have seemed more like something from Star Trek than a near reality. Read on.

The Human Genome Project revealed that humans are, on a numerical basis, genetically less complex than a mustard plant (Arabidopsis). In fact, our genome contains about 20-25,000 sequences suggestive of "genes" encoding proteins, while Arabidopsis contains about 27,000. Hmm. That shouldn't make sense to most of you, and didn't to many of the scientists working on the human genome. Empirically most of us are far more complicated than a mustard plant. This paradox harkens back to one of the teachings of distant high school and college biology and genetic courses, written by our parents' generation. To paraphrase many texts: "About 98-99% of the human genome appears to be junk, leftover from evolutionary dead ends." Any fair student of biology could spot a problem here: evolution tends to trim baggage and inefficiencies. Why would we use only one percent of our genetic material after a few billion years of trimming the excess?

Genome scientists have recently completed the first phase of a massive collaborative project, supported by the National Human Genome Research Institute (NHGRI) of the National Institutes of Health, called ENCODE (**Enc**yclopedia **O**f **D**NA **E**lements, see: http://www.genome.gov/10005107). ENCODE, initiated in 2003, was designed to enhance our understanding of the functional anatomy of the raw DNA sequence that the Human Genome Project revealed. What have we learned from the first phase, completed in June 2007, that looked in detail at about 1% of the human genome's structure? All that junk DNA? No surprise, it's not junk after all. Our genome is a complex ecosystem with a wide variety of different types of DNA elements, not simply genes encoding proteins, interacting across space and time. Much of it appears to play a role in determining what genes are expressed, in what order, and at what levels. Substantial portions of our DNA encode information for the production of small RNA molecules that never become translated to proteins, but rather fold up on themselves and act in concert with peptides as regulators of gene expression. Some of these molecules likely act as RNA-based enzymes. Very long stretches of DNA that don't seem to code for any proteins and therefore would not be predicted by previous models to be highly conserved by evolution are exceptionally highly conserved. In contrast, regions that previously would have been predicted to be conserved turn out not to be under as much evolutionary constraint as had been thought. Intriguingly, we don't really know what the function of these long highly conserved sequences is (are?).

How about the low number of genes in humans versus mustard plants? Well, the emerging model is that human genes don't ascribe to the old teaching of "one gene, one protein," but are much more akin to Russian stacking dolls. Genes are nested inside of genes, using alternate promoters, start sites, splicing sites and stop sites. Remember that

DNA is double stranded? There is evidence that overlapping genes occur which run on opposite strands encoding proteins of different functions. Recall that RNA molecules are translated to form proteins? It turns out that seemingly unrelated RNA molecules can be assembled (trans-spliced) to form templates for entirely new proteins. To further complicate this picture, elements in the DNA known as pseudogenes exist which, save for minor variations, look exactly like other functional genes. ENCODE and related results suggest that they are numerous, that they may affect transcription of neighboring genes, and that they may not be as transcriptionally silent as once believed. All in all, the first phase of ENCODE has demonstrated how little we really know about the human genome and its functions, and has provided tantalizing glimpses on novel strategies for thwarting human disease. To this end, NHGRI has recently announced over $80 million dollars in grant awards to flesh out our understanding of the functional elements in the 99% of the genome not covered in the pilot phase of ENCODE.

Unraveling the Gordian knot of the regulation of DNA function has clear implications for biomedicine. Many drugs in current use either directly or indirectly affect DNA transcription or translation, a classic example being steroids. A better understanding of the mechanism of action of drugs like steroids could permit the development of far more selective drugs that can target up-regulation or down-regulation of various genes important to human disease. It seems likely that at least some of the hereditary burden of disease will fall into these newly described regulatory regions that affect how and when genes are expressed. As well, these newly discovered regulatory mechanisms and pathways might themselves be co-opted to yield therapeutics. Proteins and peptides have a long history of being used therapeutically, so why not small RNA (or smRNA) molecules?

Ok, back to the start of this piece. Remember that "ribosome transplant" from Star Trek (if you are normal like my boss at the NHGRI and have absolutely no idea what I am talking about please see: http://en.wikipedia.org/wiki/The_Enemy_(Star_Trek:_The_Next_Generation)) ? We have arrived at the moment each of you has been waiting for since that episode was aired: "ribosome transplants" are here! For those readers who are sticklers for precision: RNA-based therapeutics are being used in human clinical trials in the United States. Difficult to believe? Search www.clinicaltrials.gov for siRNA (small interfering RNA). You will find a handful of active clinical trials that investigate the use of siRNA molecules for chronic myelogenous leukemia and wet age-related macular degeneration (AMD). In the late 90's Craig Mello and Andrew Fire described a mechanism by which double stranded RNA molecules occurring naturally can modulate gene function – (this landed them a Nobel Prize in 2006). More recently the RNA interference approach has been used to develop targeted therapeutic. At least in the case of wet AMD, preliminary data suggests a great deal of promise. Given what we are learning from ENCODE suggesting the pervasiveness of small RNA molecules in the day to day regulation of our genome, these trials may be the tip of an iceberg! For those wishing to learn more about the results of ENCODE see the free June, 2007 issue of *Genome Research* at www.genome.org/content/vol17/issue6/).