

VIEWPOINT

Whole genomes may hold clues to autism, but patience is key

BY BERNIE DEVLIN, MICHAEL TALKOWSKI

27 JUNE 2017

It's been 14 years since scientists spelled out most of the more than 3 billion letters of the human genome. The feat, which took 13 years and cost just under \$3 billion to complete, signaled a new era in biomedical research.

Much of human genetic research has focused on the roughly 2 percent of the genome that makes up genes, called the **exome**. Amino acids, the building blocks of proteins, are encoded in three-letter 'triplets' throughout the exome. This triplet code has allowed us to predict which mutations are likely to alter the function of a protein, and which are likely to be silent.

The most severe mutations are those that disrupt the proteins critical to health and development. Natural selection acts against these changes. Some of the mutations seen in people with autism are severe and rarely seen in the general population. We have used this information to identify genes that are likely relevant to the condition.

We know relatively little, however, about the 98 percent of the genome that does not code for genes. These sweeping swaths of DNA, once blown off as 'junk,' are now known to contain important sequences that switch genes on and off and fine-tune their expression.

It's reasonable to assume that a small subset of the mutations that occur in the noncoding genome contribute to autism. And now that the cost of sequencing a genome has dropped to about \$1,500, we can finally test that assumption.

One of the enticing things about mutations in the noncoding genome is their frequency in all of us: Each of our exomes carries perhaps 1 new mutation, whereas our noncoding genomes carry around 100. But most of these mutations are surely benign, and we lack a 'decoder' that allows us to predict which mutations are harmful.

If finding mutations tied to autism in the exome is like finding a needle in a haystack, then finding mutations in the noncoding genome is like finding a peculiar piece of hay in that stack without knowing the properties that distinguish it from the rest. If we are going to be successful in our search, we need to understand what we're looking for.

Hidden hits:

It is possible that some noncoding mutations are as damaging as those in the exome. For instance, they might disrupt a stretch of DNA that regulates the expression of a key gene for brain development. But we have no way to interpret which DNA letters are crucial for the function of these regulatory regions and may therefore affect gene function when mutated.

So how can we approach this daunting problem? History suggests that we must scour the noncoding genome for mutations tied to autism agnostically, without any preconceived notions about where these mutations may be hiding. This unbiased approach has served us well in previous efforts to analyze the genome.

We expect our initial results using this approach to be lean, but we will avoid the pitfalls of a past era of human genetics when many investigators focused on 'candidate' genes they assumed played a role in a particular condition. The record of replication from the candidate-gene approach was abysmal, and in the end very little was learned about the conditions at all. Indeed, several decades of research have taught us that scientists as a whole are not terribly prescient when it comes to predicting the genetic causes of human conditions.

We have begun the search using whole-genome sequencing data from 519 families that have one child with autism but unaffected parents and siblings. To explore these data, we have assembled a consortium of scientists with extensive expertise in many facets of human genetics, genomics, statistical genetics and computer science. Perhaps we can best liken our initial analysis to Alfred Tennyson's poem "The Charge of the Light Brigade," in which a confluence of circumstances led a British light cavalry unit into a battle against impossible odds.

Forging ahead:

Figuratively, like the plight of Light Brigade, the outcome of our initial advance into the noncoding genome was likely predetermined. The data from only 519 families are no match for the complexity of the noncoding genome and the sheer number of tests required to properly evaluate it. Only a strong and focused noncoding signal could overcome this testing burden, and if such a signal were present, it's likely we would have seen it with other methods.

We detected a small increase in the burden of noncoding variation in individuals with autism compared with their unaffected siblings, but the risk associated with these regulatory variants does not approach the risk associated with protein-coding mutations.

We plan to continue to develop new statistical and bioinformatics methods to interpret the impact of mutations that alter gene regulation. As we amass additional whole-genome sequences, we will continue our unbiased search, and eventually, reliable insights will emerge.

It is not reasonable to expect breakthroughs at this early stage. Instead, we expect to learn much about the nature of the noncoding genome and how to analyze it. As sample sizes and knowledge increases, we will soon transition from this era of initial exploration to one of true biological discovery.

When that transition will occur is impossible to say at this point. Our proverbial haystack will not change in size, content or complexity. However, with many scientists committed to searching together, we will eventually discover the peculiar features of those pieces of hay we seek.

Bernie Devlin is professor of psychiatry at the University of Pittsburgh. Michael Talkowski is associate professor of neurology at the Center for Genomic Medicine at Massachusetts General Hospital.