Together, big data, bench science and genome-wide diagnostics predict genomic instability that can lead to disease

August 7 2018, by Ana María Rodríguez, Ph.d.

Dr. James R. Lupski. Credit: Baylor College of Medicine

They are the most common repeated elements in the human genome; more than a million copies are scattered among and between our genes.

Called Alu elements, these relatively short (approximately 300 Watson-Crick base pairs), repetitive non-coding sequences of DNA have been implicated in the rapid evolution of humans and non-human primate species. Unfortunately, these repeats also cause genomic structural variation that can lead to disease.

Disease-causing Alu elements do not work alone. To cause structural variations, pairs of elements (Alu/Alu) mediate genomic rearrangements that result in either gene copy number gains or losses, and these changes can have profound consequences for an individual's health.

For instance, the first Alu-mediated rearrangement was described 30 years ago in a patient with familial hypercholesterolemia or very high levels of cholesterol in the blood. The patient carried a small deletion—8-kilobase long—of the gene for the low-density lipoprotein (LDL) receptor that binds to low-density lipoprotein particles, which are the primary carriers of cholesterol in the blood. Alu/Alu-mediated rearrangements had resulted in the small deletion of the LDL receptor in this patient, rendering it unfit to capture LDL-cholesterol particles and remove them from the blood.

Years later, other similarly severe medical conditions were linked to Alu/Alu-mediated structural variations, such as spastic paraplegia 4 and Fanconi anemia. Scientists have estimated that Alu/Alu-associated copy number variants cause approximately 0.3 percent of human genetic diseases.

In their laboratories at Baylor College of Medicine, Dr. James R. Lupski and Dr. Chad A. Shaw have been studying the mechanisms mediating a number of structural variations for many years; Dr. Lupski's research interest in structural variant mutagenesis has spanned decades. Among other things, his lab and the findings from other labs pointed at Alu element-mediated variation as the cause of a significant portion of some

pediatric genetic diseases.

"The Alu elements we are talking about are thought to be completely inert, they are not actively producing proteins, but problems arise when the machinery that repairs broken DNA incorrectly replicates a genomic segment flanked by a pair of repetitive Alu elements. The machinery 'gets confused' by the repetitive Alu sequences and responds in a way that leads to either duplication or deletion of the sequence between the Alu elements, and this can lead to disease," said Shaw, who is a statistician, a computational scientist and an associate professor of molecular and human genetics at Baylor College of Medicine, as well as senior director of bioinformatics at Baylor Genetics.

The situation would be analogous to reading a text that has the same sentence repeated twice at intervals. In this analogy, the gene is represented by a paragraph of text flanked by the two same short phrase of words. The reader would see the repetition, get confused and probably skip that section, possibly missing important information between the repeats. Conversely, the reader would read the same sentences multiple times by returning to the first sentence. In the genome, 'missing' a section that includes important genes—a deletion copy number variant—or repeating a segment—causing a duplication or copy gain—can both have serious health consequences.

Given the relevance of Alu elements in human genetic diseases as well as genome evolution, the researchers wanted to find a way to predict which genes are susceptible to Alu/Alu-mediated rearrangements. Current clinically applied methods for measuring genome variation have limitations to achieve this goal, such as insufficient resolution or great cost, so the researchers developed a novel approach.

"We began by conducting a comprehensive statistical study to identify the characteristics of the Alu pairs known to cause diseases," said

Xiaofei Song, a graduate student in the Lupski lab. "This would enable us to build a machine-learning model to predict genes that would likely be susceptible to changes due to Alu/Alu-mediated rearrangements."

How to build and test a machine-learning model to predict disease-causing genes

The researchers applied a comprehensive and unbiased computational approach to identify the features of the Alu pairs that make genes susceptible to copy number gain or loss.

"We analyzed a training data set composed of 219 Alu pairs that are known to contribute to diseases by affecting specific genes," Song said. 'First, we identified the sequence features of the Alu elements in those 219 pairs; then, we looked on the entire human genome, using the current human genome reference sequence to which the Baylor Human Genome Sequencing Center (HGSC) contributed significantly, for other Alu pairs with similar characteristics. So, if we found a region including a number of Alu pairs with these specific features, then we would consider it to be a 'hotspot' of genomic instability associated with Alu pairs."

"We also looked at other features, such as the characteristics of the DNA section surrounding two Alu elements," said Shaw, who also is adjunct associate professor of statistics at Rice University. "If the pairs are at a certain distance from each other and are oriented in a certain way, then this is a risk factor. Having a high similarity level on the DNA sequence is another clue that an Alu pair may confuse the replication machinery and mediate rearrangements."

The researchers conducted an extensive computational analysis of the human genome and approximately 78 million Alu pairs using the BlueGene supercomputer at Rice University that integrated all these data

and built a comprehensive model. They used the model to evaluate the whole genome, characterizing the risk of Alu/Alu-mediated rearrangement for each gene.

"In addition, we carried out computational work to test our model in real human genome data—more than 54 thousand personal genome samples. For each of these samples, the copy number variation has been determined and is available as anonymized genomic variation information at the Baylor Genetics diagnostic laboratory," Song said. "This analysis predicted that a number of known disease genes were at risk of Alu/Alu mediated copy number gain or loss."

The researchers selected 89 of the predicted cases and, using PCR and genomic sequencing in the Lupski lab, tested for the presence of Alu-mediated rearrangements, confirming the prediction in 94 percent of the cases.

"These are all new discoveries of copy number variations caused by Alu-mediated rearrangements," Shaw said. "We also identified the junction, the piece of DNA between Alu elements, which may include one or more genes that have been rearranged."

The work also enabled Song to produce an AluAluCNVpredictor, a web-based tool that allows researchers around the world to predict the risk of Alu/Alu-mediated rearrangements for the genes of their interest. This tool can be accessed at http://alualucnvpredictor.research.bcm.edu:3838/ .

Interdisciplinary collaboration uncovers hidden clues in the DNA

This work shows the power of collaboration between experimental

geneticists, genomicists and computational scientists. Years of research have produced extensive knowledge of the genetic basis of disease as well as vast amounts of genomic data that, thanks to the computational teams that built sophisticated computational tools, can now be analyzed to uncover hidden clues in the DNA. The results are a deeper understanding of the structure of the genome, the ability to elucidate novel disease-gene associations, improved molecular diagnosis and the revelation of further insights into genomic instability, human gene structure and human genome evolution.

"Our approach allows us to visualize evidence for genomic rearrangements at very high resolution," Shaw said. "One of the things Song's work has helped us learn is that a large portion of human variation, including both variants associated and not associated with disease, is driven by small scale Alu/Alu-mediated events."

This research marks another important chapter in more than a decade of collaboration between wet-bench science in the Lupski laboratory, genomics in the Baylor HGSC and computational science in the Shaw laboratory, as well as the rich data for research provided by Baylor Genetics. This work highlights the unparalleled environment for interdisciplinary research at Baylor College of Medicine.

"The power of our study is the marriage of computational and statistical analysis of 'BigData' with wet-bench experimental science, as well as real human personal genome variation data from the diagnostic laboratory. In the process, we gained insights into genomic stability/instability and structural variation of the human genome responsible for disease," said Lupski, Cullen Professor of Molecular and Human Genetics and professor of pediatrics at Baylor. Lupski also is an attending physician at Texas Children's Hospital, a member of the HGSC, principal investigator at the Baylor-Hopkins Center for Mendelian Genomics and faculty with the Baylor Genetics and

Genomics graduate training program.

More information: Xiaofei Song et al, Predicting human genes susceptible to genomic instability associated withAlu/Alu-mediated rearrangements, *Genome Research* (2018). DOI: 10.1101/gr.229401.117

Provided by Baylor College of Medicine