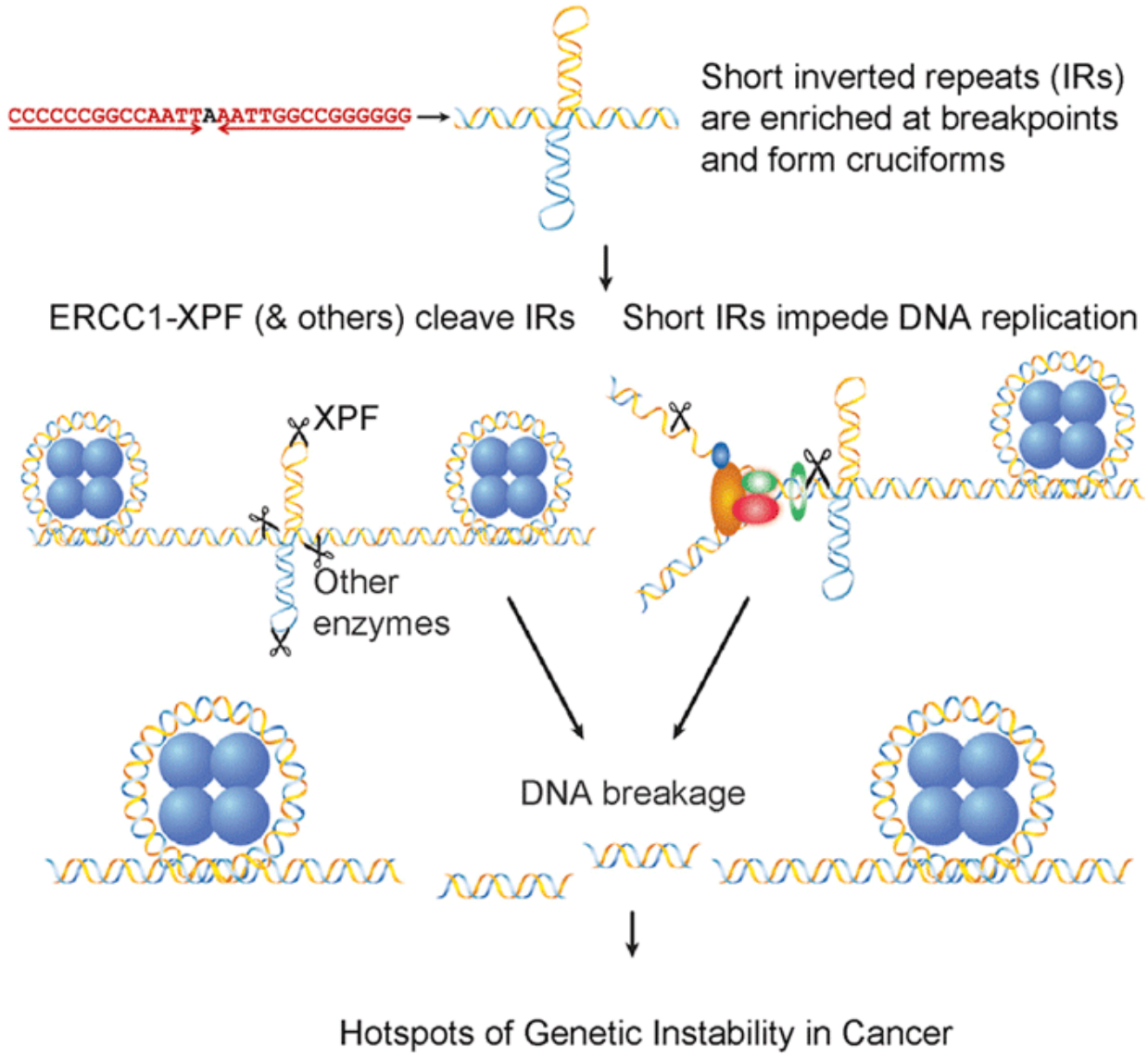
Supercomputers surprisingly link DNA crosses to cancer

June 19 2015



Short inverted repeat sequences of DNA nucleotides are enriched at human cancer breakpoints. Credit: Karen Vasquez, UT Austin

Supercomputers have helped scientists find a surprising link between cross-shaped (or cruciform) pieces of DNA and human cancer, according to a study at The University of Texas at Austin (UT Austin).

DNA naturally folds itself into cross-shaped structures called cruciforms that jut out along the sprawling length of its double helix. DNA cruciforms are abundant; scientists estimate as many as 500,000 cruciform-forming sequences may exist on average in a normal human genome. Over 80 percent of DNA cruciforms are considered small, i.e., under 100 base pairs of DNA. Small cruciforms enable DNA replication and gene expression, essential for human life. But scientists have also suspected these small cruciforms—a structure of DNA itself—to be linked to mutations that can elevate cancer risk.

DNA cruciforms are created by short inverted repeats of the nucleotides Adenine-Thymine-Cytosine-Guanine that form the bases of DNA structure. Inverted repeats are DNA nucleotide sequences followed by their reverse compliment. They're like a palindrome, a word phrase that reads the same backwards and forwards, such as 'Never a foot too far, even.'

The UT Austin study found that small DNA cruciforms are mutagenic, altering DNA in a way that can increase risk of cancer in yeast, monkeys, and in humans. High performance computing at UT Austin's Texas Advanced Computing Center (TACC) with the Stampede and Lonestar supercomputers helped the researchers find short inverted repeats of 30 base pairs and under in a reference database of mutations in human cancer that are somatic, meaning not inherited.

DNA strands commonly break in human cells. Repair proteins fuse the broken end of one DNA strand to the broken end of another. If formed in certain ways, these 'gene fusions, or translocations' can lead to cancer development.

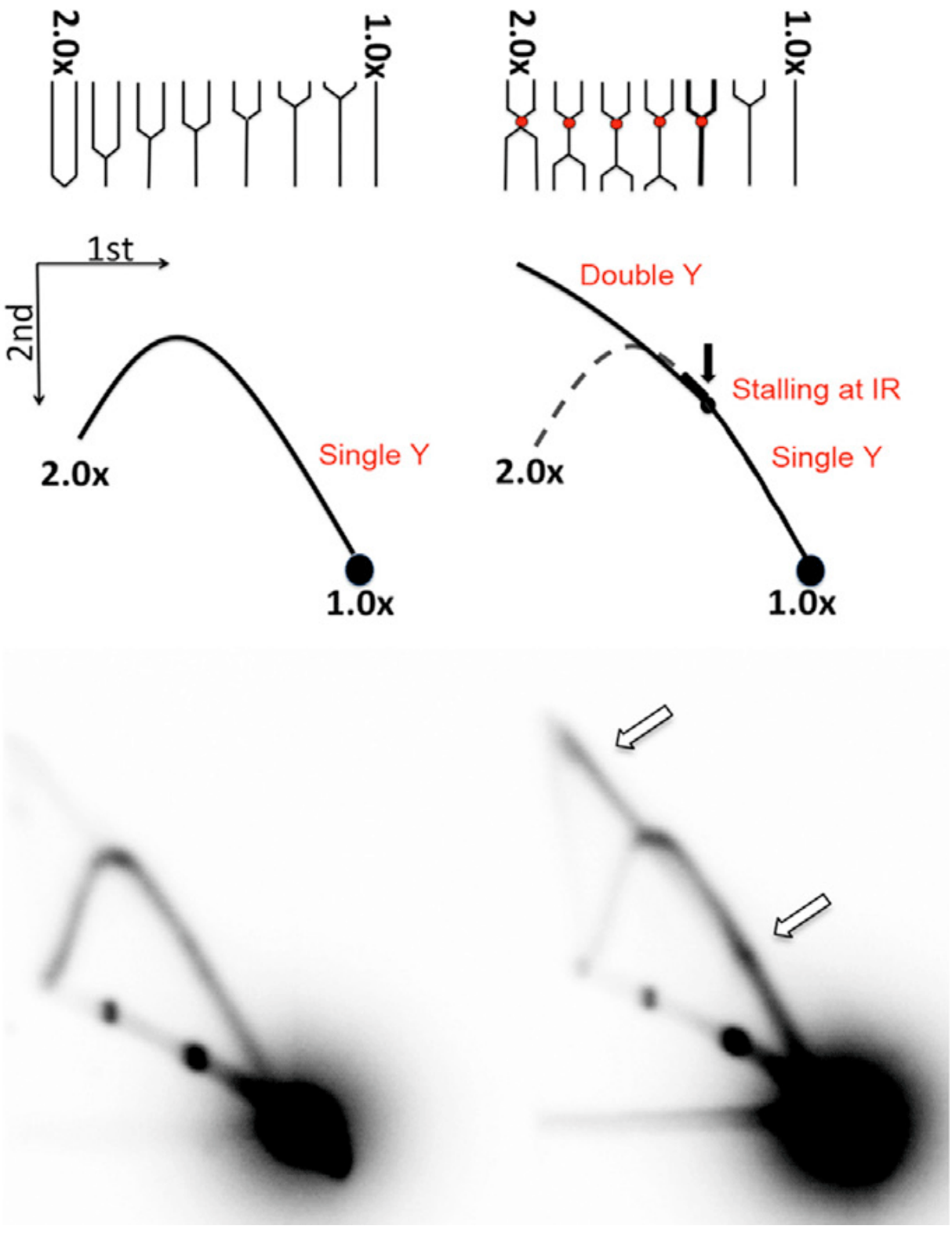
'We found that short inverted repeats are indeed enriched at translocation breakpoints in human cancer genomes,' lead author Karen Vazquez said. Vasquez is the James T. Delucio Regents Professor in the Division of Pharmacology and Toxicology at The University of Texas at Austin.

'In many cases, translocations are what turn a normal cell into a cancer cell,' co-author Albino Bacolla said. Bacolla is a research associate in the Vasquez Lab.'What we found in our study was that the sites of chromosome breaks are not random along the DNA double helix; instead, they occur preferentially at specific locations,' Bacolla said. 'Cruciforms structures in the DNA, built by the short inverted repeats, mark the spots for chromosome breaks, mutations, and potentially initiate cancer development.'

Vasquez said, 'We have also studied the potential mechanisms that are involved in the interplays among alternative DNA structures and cancer development. Our team has discovered at least two different mechanistic pathways: one involving DNA replication, where these unusual structures cause a roadblock to DNA replication; the other pathway is independent of that, where DNA repair proteins, we think, recognize these alternative DNA structures as damage, even though there is no damage per se. The cells try to process the structures as damage, but they are really processing naturally occurring unusual DNA formations and not actual damage. An abortive error prone repair process can then cause DNA double-strand breaks and lead to serious problems including neoplastic transformation.'

DNA double-strand breaks can increase the risk of cancer because they can result in translocations, deletions, and other mutagenic events that disrupt the coding properties of genes. 'These modifications of the DNA can lead to cancer,' Vasquez said.According to Paul Okano, program director at the Division of Cancer Biology of the National Cancer Institute, 'The focus of Dr. Vasquez' research on the mechanisms of alternate DNA structure-induced mutations, DNA breaks, and chromosome translocations is a novel and significant aspect of NCI grant supported studies on mechanisms of genomic instability. Dr. Vasquez' studies on the role of non-B DNA sequences in these mechanisms can contribute to our knowledge of the etiology of human cancer.'

Several studies went into the report Vasquez and her lab electronically published ahead of print in March 2015 in the journal *Cell Reports*. One study used reporter gene assays to confirm that the short inverted repeat sequences from COS-7 cells, derived from monkey kidney tissue, were mutagenic. 'We wanted to confirm that this was a biologically relevant finding,' Vasquez said. 'That's when we had to do some computational studies and insilico searching. We used the TACC supercomputers for that aspect of the work.'



Short inverted repeats (IR) stall replication forks in COS-7 cells. Upper panel:

schematic of a smooth Y-arc from 2D gel electrophoresis for the control plasmid (left), and the bulge on the Y-arc, as well as the double-Y-shaped replication intermediates caused by a replication barrier from the IR-containing plasmid. Lower panel: short IRs stall replication. The arrows designate the bulge on the Y-shaped arc, indicative of stalled replication intermediates, and the double Y-shaped replication intermediates containing IR-stalled replication forks colliding with forks progressing from the opposite direction. A representative image of four independent analyses is shown. Credit: Karen Vasquez, UT Austin.

'We have used both the Stampede and the Lonestar Linux clusters. We usually back up our data on Corral,' Bacolla said.

The challenge and need for using HPC, said Bacolla, is that the time required to find all combinations of inverted repeats, given a DNA sequence is enormous. The Vasquez team designed their algorithm to take a string of letters corresponding to the DNA bases A-T-C-G and check if the nearby strings of letters match the reverse compliment of the first string. They then varied the string length and the distance between the strings.

'For every position along the DNA, the program has to perform several hundred iterations. Then the number of these iterations needs to be multiplied by the length of the DNA, then by the number of the translocations in our cancer patients,' Bacolla said.'We developed mostly our own scripts, which we wrote in Oak in the shell,' Bacolla said. He used the scripts to generate about 20,000 random chromosomal breaks. 'We needed to compare the frequencies of inverted repeats found in the COSMIC data set with those that we would find in control, by chance.' COSMIC is a database maintained by the Sanger Institute in the U.K. of mutations found in human somatic, or non-inheritable cancer.

'We had 20,000 translocations from human cancers from the COSMIC

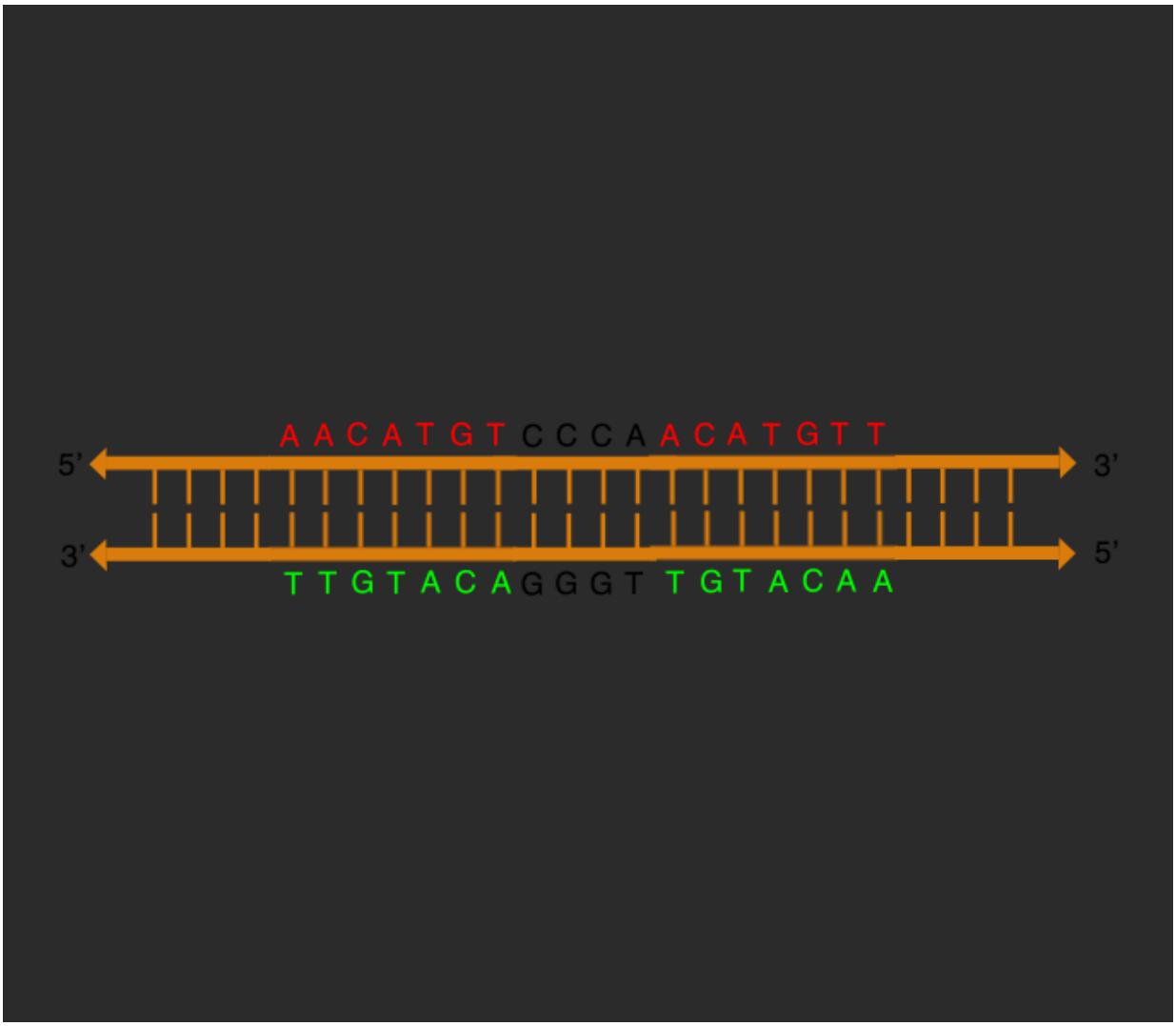
database; 200 bases of DNA for each translocation; and about 200-400 iterations at each position,' Bacolla said. The number of iterations totaled about two billion.

'This simply takes a lot of time to perform. We certainly cannot do this kind of work on our laptop or anything like a normal system in our laboratories; we need a very powerful computing system to accomplish our gene sequence searches.'

Right away Bacolla found that the scripts he wrote would stall out when he tried to scale up to 100 or more sequences on one processor.

'To solve this, we had to get in touch with TACC support staff,' Bacolla said. 'They checked our script and analyzed the error log that we got. Finally, we ended up with a solution by giving to each processor fewer sequences.' This allowed Bacolla to scale his code up and use over 1,000 processors at once.

'It would not have been possible to do this job without the TACC resources,' Bacolla said. 'The center is an incredible resource in terms of its capacity and support. We have been using the resources and staff support for some time now. It's a wonderful opportunity for researchers at UT Austin.'



Animation showing transition of linear DNA to cruciform state. The nucleotide sequence (red, left) of A-A-C-A-T-G-T is followed by the gap sequence (black) C-C-C-A and the inverted repeat A-C-A-T-G-T-T (red, right). Scientists call the inverted sequence palindromic, in that it reads the same way from 5' to 3' (A-A-C-A-T-G-T on top strand, red) or 5' to 3' on the complimentary strand (A-A-C-A-T-G-T on bottom strand, green).

'With TACC's support, we were able to see that this is at least one plausible explanation in human cancer etiology, because these sequences

are enriched at translocation breakpoints,' Vasquez said. 'That gives us hope, inspiration, and enthusiasm to move forward.'

Vasquez sees that the next step for her lab is to apply these findings to improve human health. 'Our overarching interest is to understand how DNA structure can influence cancer development. With access to TACC, we are more confident that DNA sequences capable of forming particular unusual structures present a plausible explanation for how DNA breaks can lead to translocations in cancer,' Vasquez said.

'Our next steps are to go forward with a mouse model that can detect mutations and translocations in the mouse genome using human sequences from these cancer breakpoints,' Vasquez said.Does this really occur now in the context of chromosomes in living organisms? Is it tissue specific? Does aging make a difference? These are the types of questions that the researchers will ask.

'The long term goal for these studies is to develop better prevention or treatment strategies for cancer patients,' Vasquez said.

It's important to realize, stressed Vasquez, that short inverted repeats and the cruciform structures they create also do good for the body. They facilitate replication origin firing, initiating human DNA replication. 'They have both positive and negative functions, I would argue,' Vasquez said. ' it's not really something necessarily that we want to try and change, to remove these sequences from our DNA, but to better understand what they're doing for life processes and to attenuate any negative events that might occur because of their 'ubiquitous' presence'

'If we can help clinical scientists apply mechanistic information such as we hope will be gained from our research to better cancer treatment and a cancer prevention strategies, we are benefiting all of us.'

Vasquez sees a bright future in the intertwining of computation and the life sciences. 'I think the potential of the computational analysis is mind-blowing. Bioinformatics and computational centers like TACC are critical for the next steps in science. It's an exciting time,' she said.

Provided by University of Texas at Austin