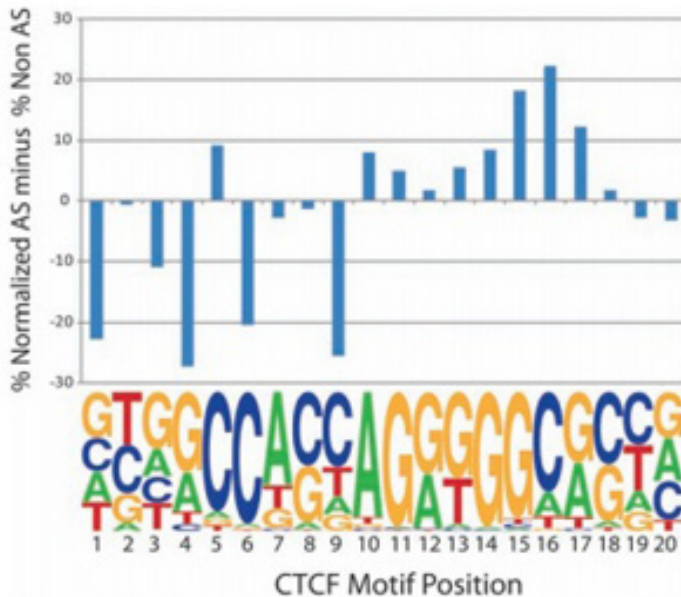


Placing Landmarks on the Genome Map

Bioinformatics researchers use Ranger to explore DNA and heredity



The schematic diagram shows human chromosome 21 with a small region outlined in red. The main rectangle below is a close-up of the outlined region, showing the binding locations of three transcription factors along the chromosome. [Courtesy of Vishy Iyer.]

We typically think of heredity — eye color, body type or susceptibility to a disease — as rooted in our genes. And it is. But scientists are finding that variants in the non-coding regions of the genome outside the genes, formerly considered junk, play a role in our personal traits as well. This insight is becoming clearer as biologists sequence more genomes and analyze their findings.

Just 10 years ago, scientists completed the sequencing of the first human genome. Since that time, the cost of DNA sequencing a human genome has dropped from billions of dollars to tens of thousands, enabling comparative studies and focused investigations of gene expression. These new sequencing technologies have greatly improved scientists' ability to understand the behavior of biological systems and are radically changing how diseases are understood and treated.

Many common diseases have a genetic component that predisposes a person to disease, though the connection is rarely simple. But genomic studies are beginning to lead to breakthroughs in the understanding and treatment of disease. These studies would not be possible without advances in the tools used to interpret the genetic blueprint, such as next-generation sequencers and the advanced computing systems that organize and give meaning to the terabytes of data they produce.

The combination of next-generation sequencers and high performance computers are enabling biologists to ask novel questions and glean new insights about DNA and heredity.

One example involves the role of transcription factor proteins in gene regulation, an important aspect of genetics that scientists are just beginning to explore on a genome-wide scale. These proteins bind to landing pads on the genome and act as control dials for gene regulation — turning genes on or off, or determining the amount of gene activity in a cell.

"If you're comparing normal cells to cancer cells, you want to know what happened in the cancer cell that makes it different from the normal cell," said Vishy Iyer, an associate professor in the Institute for Cellular and Molecular Biology at The University of Texas at Austin. "The gene expression patterns change, and we want to know which genes are up or down regulated, and how that came about."

About 2,000 of the transcription factor proteins have been identified, and some have been linked to breast and other cancers, Rett syndrome, and autoimmune diseases. However, little is known about the mechanism by which they work.



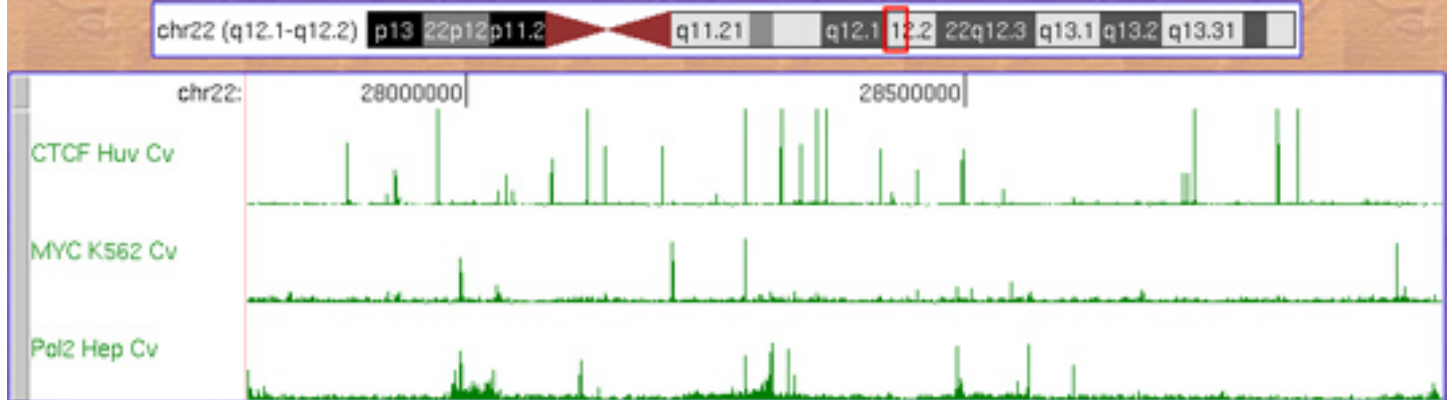
Vishy Iyer, associate professor in the Institute for Cellular and Molecular Biology at The University of Texas at Austin

Iyer, along with colleagues at Duke, UNC-Chapel Hill and Hinxton, UK, are trying to change that. Their published work is one of the first studies that used next-gen sequencing and HPC analysis to explore the expression of genes related to a specific regulatory transcription factor (called CTCF), and to investigate the role of heredity in the transcription binding process.

Their proof of principle study, “Heritable Individual-Specific and Allele-Specific Chromatin Signatures in Humans,” published in *Science* in April 2010, determined that distinctions in transcription binding can be studied using their method, and that transcription factor binding appears to be a heritable trait.

“We showed for the first time that some of the differences in DNA between individuals can affect the binding of transcription factors,” said Iyer, “and more importantly, that those differences could be inherited.”

The group used a relatively new sequencing technology, ChIP-Seq, which combines chromatin immunoprecipitation (ChIP) with massively parallel DNA sequencing to identify the DNA-associated proteins and to precisely map binding sites for proteins



The schematic diagram shows human chromosome 21 with a small region outlined in red. The main rectangle below is a close-up of the outlined region, showing the binding locations of three transcription factors along the chromosome. [Courtesy of Vishy Iyer.]

of interest. The technique enables researchers to pull out only the regions of DNA to which the proteins of interest are bound. These base pairs are then sequenced to determine the order of nucleotides and to count how many molecules of the promoter are bound to the protein.

Sounds simple enough, until you try to sequence millions of these regions of DNA and locate their exact position among the approximately three billion base pairs in the human genome.

“The genome is a vast area with many features,” explained Iyer. “You can think of the proteins as landmarks that we’re trying to place on the genome map.”

The Ranger system at the Texas Advanced Computing Center takes the short sequence reads generated by ChIP-Seq and aligns them to the reference genome. “It’s like a text search. Though if you try to run it in Microsoft Word, it will never finish,” joked Iyer. Using several thousand cores in parallel on Ranger, the alignment runs take several hours for each of the data sets the group investigated.

In total, the project has used more than 175,000 processor hours, or the equivalent of 20 years on a single processor.

The experiment that Iyer and his colleagues’ conducted explored the gene expression of six individuals: a mother, father and child from two different populations. The single base resolution offered by next-gen sequencing enabled the researchers to look at single, known differences in the DNA and to use those differences to examine how genes on each chromosome bind transcription factors.

“We’re able to tell the difference in binding from the gene that you inherited from your father and mother—that was the big advance,” said Iyer. “We’re now using this technology to look at those differences and apply it to cases where you know that the gene from one of your parents has a mutation that pre-disposes you to some disease.”

More generally, the findings bring science one step closer to personalized medicine based on a close reading of an individual’s genome.

The sequencing technology is young, and consequently, Iyer and his team had to create many of the software tools used to analyze the results.

“There’s not a standard pipeline where you can take next-generation sequencing data, push a button, and get some result or some insight. You really have to work at it, and part of that process is writing code,” said Iyer. “But the hardest thing, as with all science, is asking the right question, doing the right experiment, and using the right tools.”

Iyer intends to expand the experiment in the future to compare gene regulation in healthy and sick individuals. Despite the tremendous complexity of the genome, Iyer is optimistic that his group’s research will have an impact on human health.

“There are lots of diseases, lots of variants that are associated with them, and for a subset, they’ve got to be affecting gene expression by impacting transcription factors,” he opined. “If we pick the diseases and the factors smartly, I think we’ll find them.”

Published January 19, 2011