RESEARCH PROFILE

Adam Siepel

In the 6 million years since humans diverged from chimps, the two species have remained astonishingly similar at the level of genes. Modern humans have only a handful of protein-encoding genes—from among the total set of about 21,000—that chimps don't have. The full set of proteins that humans and chimps make is almost identical.

What then makes people and chimps so different? It's one of several profound questions addressed in the wide-ranging research of Professor Adam Siepel, a quantitative biologist who uses advanced mathematics and analytical tools developed in disciplines ranging from computer science to physics and engineering to extract meaning from data collected in biological experiments. Using these tools, Siepel and colleagues have demonstrated, for instance, that it's not so much our genes, but the way they're regulated that distinguishes us from the great apes.

Siepel, 43, came to Cold Spring Harbor Laboratory from Cornell University in 2014 to lead the Simons Center for Quantitative Biology. The SCQB, launched with a $50 million donation by Jim and Marilyn Simons (see box, p. 12), had been gaining critical mass since 2009.

"I came to the Lab because it's a tremendous opportunity to do quantitative biology (QB) shoulder-to-shoulder with leaders in experimental biology," Siepel explains. Professor Mike Wigler, an early champion of a QB center, pointed out years ago that "we will all benefit from having very smart people at the Laboratory" like Siepel, with "deep insights into mathematics and the structure of things, including large data sets."

A hacker at age 12

Siepel is a member of the first generation raised with home computers. "I became a hacker at around age 12, and started writing video games." From his home town, West Valley, an hour from Buffalo in western New York, it was off to Cornell for engineering and then to Los Alamos National Laboratory, for his first chance to use advanced computers to answer biological questions—in this case, figuring out how HIV, the AIDS virus, evolved. It was 1994, and "I was immediately captivated."

He later earned a Ph.D. at the University of California, Santa Cruz, then returned to Cornell to teach, distinguishing himself in research focused on comparative genomics and the development of statistical methods and software tools to understand how genomes evolve.

Tracing the human genome's evolution is a fascinating intellectual and technical challenge. Siepel says it is also of practical value. "If you're going to try to associate mutations in DNA with diseases, like cancer, you need to understand—as background—the process by which mutations are propagated through populations in the *absence* of disease."

In other words: "In order to ask questions about 'what is surprising?' when you compare people who are sick and those who aren't, you have to have a really good model for what is not surprising." One needs to know what scientists call the "null case"—the kind of mutations you expect to see when DNA acts like DNA does in normal situations.

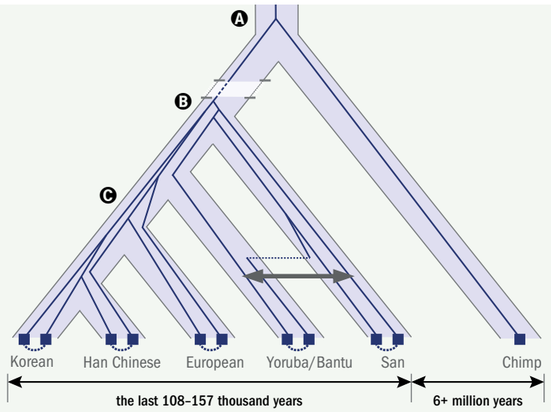
The average rate of human mutation is about 1 per 100 million DNA "letters," every generation. Since there are 3 billion letters in our genome, each of us has about 30 DNA letters that differ from the corresponding ones in our parents' genomes. Yet, Siepel and his team have asked: how can we know which of them, if any, actually matter?

Having DNA variants linked with a serious illness like autism or heart disease changes one's risk profile. But Siepel's inquiries have taken the question of a mutation's significance to an even deeper level. His research group invented a mathematical method, called INSIGHT, to predict which DNA letters in a given genome are important to evolution. By that, they mean which DNA mutations are likely to affect fitness. It involves comparing DNA changes among dozens of contemporary people with chimps, our closest relatives. Patterns of variation across these human and nonhuman individuals allow them to home in on the DNA letters that actually matter. "INSIGHT enables us to use these patterns to separate the important mutations from the ones that are likely not doing anything."

The analysis is complex, but the bottom line is simple and stunning: "Most of the mutations you see in present-day human populations have no impact whatsoever on fitness," Siepel says. Those that have an impact tend to disappear rapidly—they're either so advantageous that they are universally adopted and therefore lose their identity as mutations, or, much more often, they are harmful and vanish rapidly. Some human mutations result in a nonviable fetus, for example.

Siepel's team has also used an evolutionary perspective to shed new light on fundamental biological mechanisms. For instance, the regulation of gene transcription—the process by which a gene's coded message is copied into RNA. The process was described in great detail decades ago, from the activation of genes, to the "reading" of DNA by protein machines called DNA polymerases, to the generation of RNA messages, called transcripts. Yet research published last year by Siepel and colleagues, including his longtime collaborator John Lis at Cornell, gave evidence of several things not previously suspected.

First, it turns out that not only are the long DNA passages that "spell out" genes being "read" and "transcribed" into messages; so are shorter DNA regions that *regulate* genes, called enhancers and promoters. (Different combinations



Our demographic history

Where do we come from? Our genomes are the product of many ancestral genomes, shuffled through generations of genetic recombination. Siepel's team devised methods to estimate when various current populations branched off from one another, long after the great human-chimp divergence about 6.5 million years ago (Ⓐ in chart). Genomes of 6 living people, representing, San, Yoruba, Bantu, European, Han Chinese and Korean populations, were compared with one another and a chimp genome. The San occupy the oldest existing branch of our common family, having diverged 108–157 thousand years ago (Ⓑ). Europeans and Asians diverged from common African ancestors (represented here by Yoruba and Bantu peoples) 38–65 thousand years ago (Ⓒ). The data also suggest the original human ancestral group numbered ~ 9000 people. (Adaptation of figure by Siepel and colleagues)

Our closest relatives, in evolutionary terms.

of promoter and enhancer activity help explain why some genes are active only in specific cell types.)

The second revelation is that this process of reading, copying and writing RNA messages proceeds in opposite directions, simultaneously, on the twin DNA strands—at genes, enhancers and promoters alike. All this message-making raises a problem: how does the cell end up with stable RNA messages that tell the cell how to make proteins? What happens to all those additional messages being generated at enhancers and promoters? These messages fall away from the double helix, the team recognized, and are quickly destroyed.

This adds up to what Siepel calls a "unified model" for how DNA transcription is initiated. Why does it matter? It simplifies: "We found that the same process of RNA message making gets applied not only at genes but also at the regulatory elements." Moreover, the mechanism is useful—for the well-being of the individual *and* the species. RNA messages that regulate gene expression are made and then destroyed. The only surviving message is that of the gene. This is what the cell needs to make a protein. The entire machinery tends to ensure that proteins are made, made properly, and only when they are supposed to be made.

Another significance of the work concerns how these facts were ascertained. It involved some ingenious tagging of RNAs that Dr. Lis invented. It also depended upon a massive compilation and sorting of data, drawing upon the vast data set of the Encyclopedia of DNA Elements, or ENCODE (a consortium in which CSHL Professor Thomas Gingeras plays a lead role). It also pivoted upon the ability of Siepel's team to build a model that explained how to distinguish between long- and short-lasting RNA messages.

Both examples of Siepel's recent research shed light on gene regulation and help answer many questions, including the mystery of what makes men and chimps different. The work shows that while only a small part of today's genome has been under evolutionary "selection pressure," it is changes in factors like enhancers and promoters that *regulate* genes, and not in genes themselves, that appear to account for much of the difference.

Peter Tarr

A transformative gift

The 2014 gift to the Lab of $50 million from Marilyn and Jim Simons to establish the Simons Center for Quantitative Biology (SCQB) was only the most recent in a long and fruitful series of the couple's philanthropic acts, many focusing directly on support of math and basic research.

"I became convinced some time ago that quantitative methods were going to get more and more important in biology," Mr. Simons says of the inspiration for the Simons Center gift. "After the genome was sequenced and deep study of its structure got under way, it was evident that we were going to need more and better mathematical and statistical analysis." The Simonses had previously supported quantitative biology at the Institute for Advanced Study at Princeton. "Well, Marilyn [who is Vice Chairman of CSHL's Board] and I like Cold Spring Harbor, too, and so we thought some intensification of the quantitative effort could be made—that a real concentration of quantitative people would amplify efforts across the Lab." The final step in launching the Simons Center was finding a leader. "They only interviewed first-class people, and they were lucky to find Adam [Siepel]," Mr. Simons says. "He's first-class."