Ancient DNA

Looking back in time with evolutionary genetics

Tony Capra is an Associate Professor of Biological Sciences at Vanderbilt University, and one of the leaders in the field of evolutionary genetics.

Tony, many thanks for agreeing to contribute to *The Biochemist*, we are delighted to be able to speak to you about your work in the field of evolutionary genetics.

Tell us about yourself – you originally studied mathematics, what drew you to the molecular biosciences?

I studied math and computer science in college and then went on to a PhD program in computer science. However, through a personal connection, I got a job as a programmer in a mouse neurogenetics lab for several summers during college. It exposed me to the fact that many fields of biology (genetics, in particular) were beginning to generate more data than they could analyse using their traditional methods. During my first year of grad school, I tried out a few small computational biology projects and realized that it was fun to be both a scientist and a computer scientist at the same time. I haven’t looked back since!

Can you tell us about evolutionary genetics? What is the aim of the field?

The goal of evolutionary genetics is to understand how evolutionary processes have produced the astonishing diversity of form and function we see in the living world and to identify the specific genetic differences responsible for differences between individuals and species.

You use computational approaches to analyse DNA – could you explain that for our readers?

Genomes are very big. For example, one copy of the human genome contains more than 3 billion DNA base pairs (the As, Ts, Cs, Gs). Thanks to advances in DNA sequencing technology, whole genome sequences from more than 100,000 people and thousands of different species are available. Similarities and differences between these genomes can reveal how different individuals and species are evolutionarily related.

There’s no way we could analyse these trillions of data points without the help of computer programs and statistical models. Computational modelling helps us to identify patterns in the DNA that would be impossible to detect otherwise. For example, my group is interested in how DNA sequence patterns influence when and where different genes are expressed. We use machine learning to extract sequence motifs that are predictive of gene regulatory activity in different cellular contexts.

How do you apply algorithms and statistical models to ancient and modern genomes?

Thanks to recent advances in the ability to identify, extract, and sequence DNA from ancient material (e.g., bones and teeth), DNA sequence information is now available from thousands of ancient humans and several of our close relatives like the Neanderthals. These data are revolutionizing how we think about the recent past of our species and model evolution. Ancient DNA is enabling us to identify differences between humans and our close extinct relatives and to try to infer traits of these groups that we cannot directly observe. For example, my group recently developed an algorithmic approach that uses modern genomes paired with gene expression data to predict differences in gene regulation in different cellular contexts between ancient and modern humans. When I started in this field, I never thought we’d have data like this available. It is like having a time machine!
One of the aims of your research is to explore how our increasing knowledge of genomic variation can be translated into the treatment and prevention of disease. Could you tell us about that?

Given the importance of our genomes to how our bodies develop and function, knowing an individual's genome sequence has great potential to inform how they should be treated. For example, some people have genetic variants that make the enzymes that process drugs more or less efficient. Failure to account for these differences can result in improper dosage. Many of these genetic variants have different frequencies in different human populations, so it is very important to base such predictions on data sets that match the ancestry of the patient. Furthermore, for many common diseases we do not fully understand how genetic variants influence risk, so we cannot (yet) use genetic information to guide treatment or prevention.

Previous work has explored the impact of Neanderthal genes on the immune systems of modern humans. Can you tell us about it?

Small amounts of Neanderthal DNA are present in modern non-African individuals. My group pioneered an approach that uses biobanks with DNA samples from patients linked to de-identified electronic health records to test how remaining Neanderthal DNA influences traits. We identified the Neanderthal DNA in tens of thousands of people and then tested for traits that associated with different Neanderthal alleles. We've found many associations, including influences on many genes involved in the innate immune system and risk for many autoimmune diseases. Other groups have now shown that cells from humans with Neanderthal DNA at certain positions in their genomes are more likely to respond to different viral and bacterial challenges than human cells without Neanderthal ancestry. Overall, Neanderthal DNA influences many systems in the human body, but there is a particularly strong influence on the immune system. However, I want to be clear that we cannot blame Neanderthals for these diseases; Neanderthal DNA only contributes a small fraction of the overall genetic risk.

Your recent paper described analysing Neanderthal DNA to find differences in how genes are controlled between modern humans and Neanderthals.

What did you find?

We found thousands of genes that were likely active at different levels between humans and Neanderthals. By looking at what we know about the functions of these genes, we found many traits that were likely divergent. Some of these agree with known differences in skeletal structure from the fossil record, e.g. they were shorter and had a more pronounced brow ridge. Others point us to differences in their immune, reproductive, and cardiovascular systems. Our work also suggests that some of these differences in gene regulation may have been detrimental in human–Neanderthal hybrids.

Our October issue explored the use of machine learning and artificial intelligence in research, including work using deep learning programs to predict protein structures. How do you think the use of AI will affect research in the future?

AI is already commonly used in genomics and biomedicine. Its role will only increase over the next decade. The development and application of these powerful computational methods will be the easy part. In my view, the two major challenges posed by the increasing use of machine learning in research are: 1) their interpretation and 2) ensuring that they do not introduce subtle biases into our knowledge. For example, we've developed deep learning algorithms that can accurately identify parts of our genome that influence gene regulation, but the DNA patterns 'learned' by these algorithms to make their predictions often do not fully capture all the DNA motifs required for the function of these regions in the cell. Sometimes these algorithms can make very accurate predictions without reflecting biological reality. Furthermore, these algorithms are only as good as their training data. Since most of our genomic databases have strong biases towards individuals of European ancestry, algorithms trained on these data have the potential to be much less accurate on individuals with other backgrounds.

Finally, if you could bring back one extinct species (human or otherwise), what would it be and why?

Rather than bringing back an extinct species, I think we should use our growing understanding of evolutionary genetics to better take care of the humans and species that remain. This kind of stewardship poses many challenges but is vitally important to our world.