Human genetic variation in Tartu, Estonia

Re-shaping Europe one SNP at a Time

Andres Metspalu and his group, while exploring the genetic background of their fellow Estonians, confirmed the genetic gradients running across Europe from south to north and from east to west.

Curiosity has always played a major role in human progress. Will we fall off the earth if we wander too far? What pulls the apple towards the ground? Why, when green peas are crossed with yellow peas, are their offspring yellow? The questions were seemingly simple, but the answers, when they came, were brilliant. Humankind still has questions that remain unanswered, and though finding a solution appears more complex, one quest that appears to be drawing Europe together ironically aims to identify what makes Europeans genetically different.

"As it turns out, we are very similar in Europe," says Andres Metspalu, director of the Estonian Genome Project and the Institute of Cell and Molecular Biology at the University of Tartu. "Previous population studies suggest a genetic gradient running across Europe from south to north and from east to west, implying that the greatest genetic difference occurs between northwest and southeast populations."

Metspalu and his group have now confirmed this gradient in a recent publication about population genetics (*PLoS ONE* 4(5):e5472). Entitled "Genetic Structure of Europeans: A view from the north east", the paper truly reflects the willingness of scientists and institutions to share and be part of the discovery process with contributions from more than 20 scientists from all over Europe. For the first time, information was available for Eastern Europe and the Baltic region, in particular Estonia, which was important because it provided insight into the validity and relevance of genetic data coming from small countries like Estonia to the rest of Europe.

Getting an idea off the ground

Any two unrelated people share approximately 99.5 % of their DNA, but in certain regions of the chromosome one individual may have the nucleotide adenine, whereas the other may have guanine. This change of a single nucleotide in the genome sequence is called a single nucleotide polymorphism or SNP (pronounced *snip*) and must occur in more than 1% of the population to be classified a SNP. In most cases, SNPs occur outside the "coding sequence", posing no threat to cell function, but SNPs have also been found to be associated with different diseases, and affect how people respond to chemicals, drugs and environmental factors.

Currently, several countries are mapping SNPs within their populations. Large international collaborations such as the International HapMap Project, which looks for common patterns of variation in the human genome, and the Genome Wide Association Study (GWAS), which tries to associate SNPs with specific diseases, are underway.

Metspalu knew that countries from Eastern Europe and the Baltic region were underrepresented in these larger population genetic studies. Therefore, at the beginning of 2008, the former president of the European Society of Human Genetics began contacting friends and colleagues all over Europe to share his idea of performing a European-wide population genetic analysis, requesting samples or, where available, SNP data. The response from his peers was favourable and he credits the good relationship they have with one another, including a high level of trust, as essential for conducting such a large collaboration. He jokes that it took more time to get the final version of the paper approved by the more than 20 authors than it took to receive the samples and perform the analyses.

Crunching the numbers

Obtaining samples from Estonia was relatively easy because of the Estonian biobank, which is a part of the Estonian Genome Project that Metspalu was instrumental in setting up. Currently, the biobank contains samples from more than 36,000 people or approximately 4% of the population across Estonia, making it a true representation of the Estonian population.

Once data were received and samples prepared from the different European populations, it was the overwhelming task of doctoral student Tõnu Esko to make sense of it all by employing complex statistical methods, which he kindly explained. "We used the Principal Component (PC) analysis because it is the most elegant and fastest method available to analyse 270,000 markers in the more than 3,000 samples we received. It is a mathematical method which transforms large amounts of complex information (3,000 x 270,000 matrix or table) into simple components."

This particular analysis, according to Esko "looks for SNP or SNP alleles that differ most between populations and based on the most different SNP allele, provides a PC score which is used to draw genetic maps." EIGENSOFT software analysis produced several components but, as Esko explained, only the first two components are usually used because they represent the most diversity. For this analysis, the first two PCs illustrated the genetic diversity along the northwest to southeast gradient of Europe and separated populations along similar lines as their geographic origins. The genetic map showed populations distributed into regions in the form of a triangle with Finland at one apex, the Baltic countries (Estonia, Latvia and Lithuania) and Russia and Poland at the second, and Italy at the third. Central and western European countries remained at the center. This analysis was also able to pinpoint population differences within Italy, Finland and Germany.

In the form of a triangle

The first and second PC analyses accounted for 8% and 4% of the genetic variation observed, respectively. The overall 12% genetic variation, which is twice as much as previously reported in European studies, was used to create genetic maps. The maps provide a good visual image of the variation between different populations but as Metspalu insists, "most of the genes and markers we have in Europe are not different. The reason we are different, as shown in these maps, is that there are genes which are under pressure to be changed."

Metspalu goes on to explain some of the genetic variation observed in the population, such as genes involved in the immune response. "That's quite logical, because the environment used to be different in the north than in the south, including the animals and insects and food people were exposed to, and the people developed different immunities to their environment," the molecular biologist explains. Other genes identified included pigmentation genes due to the difference in sunlight exposure from north to south and genes involved in lactose tolerance.

"But looking at it from the other side," says Metspalu, "about 88% of genes are so similar or the differences so small that we don't even have to look at them. Therefore, I believe that most of the disease-causing genes, immune system apart, are very similar." He goes on to conclude that, "this makes Europe unique because studies can be done across different populations which is relevant all over Europe and North America including the US and Canada."

How low migration rates help science

Other goals were to determine whether different European populations could be combined for larger scale studies; which populations could best be combined and how much power is lost when these different populations are combined. This inter-population genetic distance was described by the inflation factor lambda (λ). Esko described λ as a score which indicates how much two populations differ by chance. Therefore the higher the λ value, the greater the variation. λ values were determined between two populations resulting in a number ranging from 1 (samples from the same country) to 4.21 (most genetically diverse countries). A graph comparing λ to the p values of the tests provides information on the loss of power when populations are combined. As Esko explains, "if λ between two populations is 2, then the p value of the statistical test must not be 0.05 but 0.01 to be statistically significant after correction, suggesting that when λ is 2, we lose 5 times power, and we must have 5 times lower p value.

Different views on research in Estonia: Andres Metspalu (above) and Tõnu Esko

Estonia, with its 1.3 million inhabitants, is also geographically small, bordered by the Baltic Sea and the Gulf of Finland to the north and Russia to its east. The country is divided into 14 counties, and, interestingly, PC analysis was able to measure genetic variation within these counties, surprising even Metspalu. He claims they can predict the county a person comes from in about 80% of cases. His explanation is simple. "People used to stay put in early times and families lived in one place for generations." The father of four uses his own family as an example, which he can trace back to the same county until the 16th century. He believes that this low migration rate is advantageous for Estonia because as they increase the sample size in their biobank and screen for more markers, they can create large family structures. These structures can then be used to tell who is related to whom and the diseases prevalent within that family structure.

For the good of Estonia

Idealistic as it sounds, Metspalu rationalizes that, "association studies today are plagued with false positives as control populations are polluted with cases, would-be cases or hidden cases." He sees the need for larger studies with at least 10,000 people as cases and another 10,000 as controls to get the statistical power needed to conclude that the findings are true. He reasons, "I think that the family structure will help us to reduce the number of cases and controls needed to perform a good study which means that such studies will not be as expensive to perform and more labs can afford to carry out these studies. This will allow science to advance much faster."

In Tartu, work continues on the Estonian Genome Project as Metspalu and his team carry on exploring the genetic background of their fellow Estonians in relation to the rest of Europe and the world. A true academic with an enviable publication record, Metspalu has travelled the world throughout his career, and is now gearing up to bring Estonia onto the global scientific map. And though he is a strong proponent of experience gained by studying abroad, young scientists like Esko are excited to be in Tartu, "I chose to stay here because of our biobank and the genome project; I don't need to go abroad, we have everything right here."

ROSEMARIE MARCHAN