Strength in numbers

Genomic and epigenomic datasets are meaningless without statistical analysis. Professor Andrew Teschendorff leads a group that uses novel statistical tools to help interpret DNA methylation data, which could lead to new knowledge in cell biology and clinical advances in cancer diagnostics.

Your research is at the interface between statistics and genomics, and specifically epigenomics. Could you elaborate further on this?

The purpose of my research is to develop novel statistical methods to help us draw meaningful conclusions from a large variety of genomic and epigenomic datasets. The complexity and high-dimensional nature of these data have grown substantially over the last decade; so much so, that statistical analysis of data now represents one of the greatest bottlenecks in translating data into clinical advances. A lot of my work is conducted in close collaboration with clinicians and biologists who generate the data but lack the skills to analyse and interpret them. Generating data may soon take only a fraction of the time and cost required to properly and fully analyse them.

How has your early research career shaped your later direction?

I was actually trained as a theoretical particle physicist. However, after my doctoral work I became disheartened with the field due to a lack of data. Although I have now worked in statistical genomics for over 10 years, I remain a physicist at heart, approaching current problems and challenges in biology as physicists would. For instance, physicists always try to find the simplest solution to a problem, because this is often how nature works. Likewise, although molecular biology is undoubtedly complex, as a physicist I firmly believe that the principles underlying cellular biology and carcinogenesis will turn out to be relatively simple – although this does not mean that curing cancer will be easy once these principles have been elucidated.

In one of your studies you made a breakthrough using the statistical notion of differential variability. What exactly is differential variability?

In statistical genomics, the common approach is to identify features (these could be gene activity levels or epigenetic markers like DNA methylation) that differ in their average levels between two phenotypes; for instance, between healthy and cancerous cells. Such an approach could be dubbed ‘differential mean/average’. In contrast to this, differential variability aims to identify those features where the variance is widely different between two phenotypes.

Using differential variability, we were able to identify epigenetic markers in normal cells that could predict their future risk of becoming cancerous, whilst using the common differential average approach we could not; we also understand biologically why this is the case. The algorithm implementing these ideas and methods is called EVORA (Epigenetic Variable Outliers for Risk prediction Analysis) and is discussed in two 2012 papers published in Genome Medicine and Bioinformatics.

Why are advanced statistical methods necessary to unmask the true value of data in some datasets?

One area of my work concerns how to deal with confounding factors which, unfortunately, are omnipresent in large genomic datasets and can seriously bias analysis if not correctly accounted for. An example of a confounding factor could be the date or the lab in which an experiment has been performed. Laboratory conditions can vary between labs and dates, and this will cause some technical variation in the data that could mask or obscure the true biological signals. Statistical methods are therefore needed to perform this unmasking, and this is particularly challenging when the confounding factors are unknown – a not too uncommon scenario in large studies.

Who are your main collaborators and how do they help advance your studies?

Much of my career has been shaped by very fruitful collaborations with Professors Carlos Caldas and Martin Widschwendter. Collaboration with clinicians and biologists is very important to my work. Very often it is the novel type of data being generated, or the novel biological/clinical questions being asked, which drive the development of new statistical methods. For instance, without having access to DNA methylation data collected years in advance of cancer diagnosis, I would never have developed the EVORA algorithm.

Where will your research take you next?

From the perspective of statistical genomics/epigenomics, there is much need to further develop methods and models that incorporate the latest biological information and insights. For instance, we have recently learned that a lot of the dynamic DNA methylation changes seen in normal cell development occur in genomic regions that are far away from genes, but which can still affect the activity of these genes. Incorporating this type of information in our statistical models will be key to making further progress in understanding the principles underlying diseases like cancer.

From the biological and clinical perspective, we are currently studying DNA methylation profiles in easily accessible tissues such as blood or buccal cells and assessing whether there are specific DNA methylation fingerprints in these tissues that could indicate the future risk of cancer.
Delving into data

Understanding large epigenomic datasets requires the development and application of statistical techniques. A group at the CAS-MPG Partner Institute in Computational Biology, Shanghai, China, and University College London, UK, is using novel methods to unlock information about how epigenetic factors can influence ageing and disease.

**DURING THE LAST** decade, improved technologies have led to an increase in the volume, complexity and variety of genomic and epigenomic data. Without the ability to properly analyse these datasets, the clinicians and biologists who generate them are unlikely to draw statistically and biologically valid conclusions, meaning that the use of novel statistical methods can often be the key determinant underpinning the successful outcome of a genomic or epigenomic study.

**EVORA**

Based at both the Chinese Academy of Sciences-Max-Planck Gesellschaft (CAS-MPG) Partner Institute for Computational Biology in Shanghai, China, and University College London (UCL) Cancer Institute in the UK, Professor Andrew Teschendorff is leading research into statistical methods to analyse data obtained from studies of the epigenome – the record of chemical changes to DNA and histone proteins. His team’s research delves into how epigenetic factors can influence ageing and disease. The data they work with show that epigenetic changes, and specifically changes to the DNA methylome, occur throughout life in normal tissues and are among the earliest events in carcinogenesis. Although acquired genetic mutations can also trigger changes in DNA methylation, the evidence the group has produced points towards genetic mutations being secondary or representing later events in carcinogenesis.

For example, in one study that Teschendorff recently worked on, his group used a novel statistical paradigm called differential variability, and a statistical algorithm based on this concept known as EVORA (Epigenetic Variable Outliers for Risk prediction Analysis). By employing EVORA, they were able to demonstrate, for the first time, that aberrant DNA methylation profiles in normal cervical epithelial cells, collected three years in advance of cancer diagnosis, can predict the risk of such transformation. Although accuracy (64 per cent) was not high enough for clinical application, the study is important because it demonstrates that DNA methylation changes are among the earliest to occur in carcinogenesis. Future studies that measure more epigenetic marks in larger numbers of cellular specimens offer the promise to improve risk prediction accuracy further, so as to reach levels that might be of clinical interest.

**AGEING**

Teschendorff is also investigating the epigenomics of ageing, an area he was introduced to while collaborating with Professor Martin Widschwendter, who heads the Department of Women’s Cancer at UCL. In 2008, whilst analysing DNA methylation profiles in blood cells from a large cohort of women, the group’s statistical analysis revealed a component of variation that significantly correlated with the age of the women. The team observed that a particular subset of this age-DNA methylation signature...
was affecting transcription factors, which play a key role in normal tissue development. “What excited us especially was that this age-DNA methylation signature was seen across all normal tissue types, so not only blood but also in normal epithelial cells, skin and even in stem cell populations,” Teschendorff recalls. Moreover, the same class of transcription factors which were altered as a function of age were also seen to be altered in cancer tissue. “So, in effect, we had identified a common age-cancer signature, which could explain why for many cancers their incidence increases with age,” he adds.

What Teschendorff’s group observed in the study is that the DNA methylation landscape of aged normal cells gradually deviates from that of normal younger cells, independently of tissue or cell type – a phenomenon known as ‘epigenetic drift’. Some of the drift observed between random individuals could be due to genetic differences, but most of it is probably caused by environmental influences. Epigenetic drift measured by Teschendorff’s group has since been extensively validated by many other researchers.

Because age is the strongest demographic risk factor for many cancers, scientists hypothesise that molecular changes accumulating in normal cells as a function of age may ultimately predispose these cells to becoming cancerous. “So far, DNA methylation changes represent the most convincing demonstration of the validity of this hypothesis,” Teschendorff asserts.

**HAND2**

In another project focusing on endometrial cancer development, Teschendorff’s group developed a novel statistical algorithm based on network theoretical concepts, to integrate DNA methylation and gene expression data. This systems approach revealed a whole molecular pathway of epigenetic deregulation, identifying HAND2 – a gene encoding a transcription factor expressed in the endometrial stroma – as one of the most commonly hypermethylated and silenced genes in endometrial cancer and the target of this altered pathway. In collaboration with Widschwendter’s group, these results were validated using candidate gene methylation analysis in multiple clinical sample sets of tissues from a total of 272 additional women.

Because Teschendorff’s analysis clearly demonstrated that HAND2 methylation is a common and crucial molecular alteration in endometrial cancer, it could potentially lead to the use of the HAND2 gene as a biomarker. “We can easily measure HAND2 DNA methylation levels by collecting DNA from vaginal swabs. This non-invasive test is of very high accuracy and could serve for the early detection of endometrial cancer and as a predictor of treatment response,” he explains. However, further validation of the true clinical utility of HAND2 DNA methylation is required in prospective studies.

**CLINICAL APPLICATIONS**

Teschendorff is passionate about the impact that statistical methods can have on the outcome of a study: “It is always very gratifying when a new discovery is made thanks to the application of a novel statistical algorithm or concept,” he reflects. When analysing genomic data, the details of which statistical method to use are often not of that much importance, as slightly different methods will yield very similar answers. “Hence, when a novel method leads to a dramatic difference in interpretation or end result, it is extremely satisfying, especially if it then leads to an important biological insight or eventually to a clinical test,” he adds.

For instance, in 2008 he used a novel method to discover that the clinical outcome of hormone receptor-negative breast cancers could be explained by a relatively simple immune response gene expression signature. This work was followed up, generalised and validated by his collaborators and many other groups, and very likely will be used as a means of identifying breast cancer patients who can be spared aggressive and potentially harmful chemotherapy. “Likewise, with HAND2 or EVORA, I hope that these tools will eventually make it into the clinic in one form or the other,” Teschendorff enthuses.