S
oftware developed by Dr Gerton Lunter\(^1\) and colleagues from the Wellcome Trust Centre for Human Genetics at the University of Oxford is enabling the NHS and Genomics England, coordinators of the UK’s ‘100,000 Genomes’ project, to more effectively utilise high-throughput sequencing data. BBSRC funding allowed the researchers to develop software to carry out the vital first steps required to process DNA sequencing data ten times faster and more accurately than existing software tools.\(^2\) The project is also developing algorithms to ensure the software can be used with the new ‘third generation’ sequencing technologies currently being developed by companies such as Oxford Nanopore.

The software is now being used by researchers at the Institute of Cancer Research and The Royal Marsden NHS Foundation Trust to identify gene variants that could predispose patients to cancer.\(^3\) As the process is now much quicker and cheaper, the researchers can examine nearly 100 genes in all of the patients they see, rather than just two genes in a handful of patients.

Researchers from NHS Blood and Transplant are also working with Lunter to use the software to speed up the selection and matching of haematopoietic stem cell donors and recipients and to improve the outcome of these transplants.

Lunter was one of four founders of spin-out company Genomics plc.\(^4\) The company, which, among other activities, has developed variant caller software building on the researchers’ experiences during the BBSRC-funded project, was created to support Genomics England\(^5\) with the analysis of sequencing data for the national ‘100,000 Genomes’ project.

**Addressing the challenges of high-throughput sequencing**

High-throughput sequencing is widely used in research to study genes and entire genomes, and increasingly in the clinic to identify specific gene variants that could underpin medical conditions. Modern ‘second generation’ sequencing technologies can generate DNA sequences rapidly; it is now possible to sequence the entire human genome (3,000,000,000 bases, or 3 GB, in size) in one day. However, analysing the large quantities of data produced by high-throughput sequencing is a significant challenge.

Lunter was initially interested in the first step of processing required for high throughput sequencing data, called ‘read mapping’. Instead of producing a single long sequence of DNA, high throughput sequencing generates a series of ‘reads’ – shorter sequences that must then be stitched

**IMPACT SUMMARY**

BBSRC funding enabled researchers at the Wellcome Trust Centre for Human Genetics at the University of Oxford to develop innovative software for analysing the data from high throughput DNA sequencing.

Researchers at the Institute of Cancer Research and The Royal Marsden NHS Foundation Trust are using the software to help identify gene variants that can predispose people to cancer. By reducing the cost of testing, the software allows the researchers to look at almost 100 genes in all of their patients, rather than just two in a fraction of patients.

NHS Blood and Transplant are investigating how the software can improve the process of matching stem cell donors and recipients.

Lunter and colleagues also established spinout company Genomics plc to support data analysis for the 100,000 Genomes project. The company has received more than $20M investment, employs 40 people and is now working with leading pharmaceutical companies.

The research was supported by a £626K responsive mode grant from BBSRC.
DNA sequence analysis software aids clinical practice

Lunter developed read mapper software, and then shifted his attention to the next step in the process. Once the sequence has been assembled, the next step is to compare it to a reference sequence to identify any variations, a process known as ‘variant calling’. Existing software was often slow and struggled with a specific type of variant called an ‘indel’, where bases have been added to or deleted from the DNA sequence. Lunter’s BBSRC project enabled him to develop and refine software that could tackle both variant calling and read mapping, and which could be applied to data produced by future high throughput sequencing technologies.6

Screening for cancer gene variants

Creation of the software led to a collaboration between Lunter and Professor Nazneen Rahman7 at the Institute of Cancer Research. Rahman was leading a Wellcome Trust-funded project to develop a ‘panel’ of genes that could be screened to identify variants that predispose patients to cancer. Guidelines from NICE (the National Institute for Health and Care Excellence) limit the costs of such screening, meaning that researchers were only able to screen two genes (BRCA1 and BRCA2; both linked to breast cancer) in a fraction of patients. Rahman was interested in using high throughput sequencing to bring down costs and enable researchers to screen more genes in more people.

Rahman worked with Lunter to use his software to analyse the sequencing data for the project. As a result, “Rahman managed to set up an entire workflow that looked at almost 100 genes, not two, much quicker, much cheaper, and more accurately, with the result that under the same NICE guidelines she could analyse all of the patients she was seeing,” says Lunter. “One of the things she liked about our tool was that it was both accurate and easy to use.”

Genomics plc

In 2013 the UK Government established Genomics England as a company within the NHS to deliver the 100,000 Genomes Project. The project aims to sequence 100,000 genomes from NHS patients with cancer and rare diseases by 2017 to establish a ‘genomics medicine’ service for the NH, enable medical research, and to build a UK genomics industry.

The project, the largest such national sequencing project of its kind in the world, needs to accurately and rapidly process large quantities of sequencing data. Lunter, together with Professor Peter Donnelly, Professor Gil McVean and Dr Chris Spencer, established a spinout company, Genomics plc, to work with Genomics England on data analysis.

“We are all walking around with thousands of variants, or DNA mutations, and the vast majority are irrelevant to our health, but a tiny minority are relevant in some people.” Lunter explains. “The trick is to identify those and to use those variants to help with diagnosis and occasionally also with treatment. That’s a much bigger and longer-term project, and it’s one of the key focus areas for the company.”

To date, Genomics plc has raised more than $20M from...
investors, employs around 40 people, and is working with leading pharmaceutical companies.

**Matching unrelated stem cell donors and recipients**
The researchers are also working with Drs Cristina Navarrete, Lydia Quaye and their colleagues from NHS Blood and Transplant who are using the software to analyse next generation sequencing data and determine the high resolution HLA types of unrelated adult donors and recipients of ‘haematopoietic’ stem cells (i.e. stem cells that develop into all other blood cell types) and cord blood. The HLA region of the genome determines protein markers on human cells which play a role in immune responses. To improve the outcome of the transplants, stem cells (and organs) should only be transplanted between people with very similar or identical HLA types. Over 1,500 people received stem cell donations in 2012 in the UK.\(^8\)

NHSBT had a system for determining HLA types using existing commercial software but they were also interested in using a second application for cross-checking the HLA types.

“One of the drawbacks of the existing typing methods is that you can type the known types but you cannot so easily identify if there is something new. There are literally thousands of HLA types, most of which are known but some of which are not in the database,” Lunter explains.

When Lunter and colleagues first published details of their variant caller software, they included a demonstration of HLA typing. Expanding on that, and using a database of HLA types from NHS Blood and Transplant to validate their results, Lunter and his team were able to create an improved HLA typer algorithm. The next step is to design a graphical user interface for the software to improve accessibility.

“The HLA typer is an impressive piece of software which has accurately assigned the correct HLA alleles to approximately 1300 samples at HLA-A, -B, -C, -DRB1 and DQB1 loci,” says Quaye.\(^3\)[Post-doctoral researcher] Dr Hang Phan has been instrumental in taking our comments and suggestions on board, analysing the data and explaining how the software works. “We are very grateful for her time and dedication in developing the software, implementing our suggested changes, as well as analysing and providing us with the results.”

“We are happy with the accuracy of the HLA typer and we are considering using it in future as part of our routine work in typing donor samples”, Quaye continues.

**REFERENCES**
1. Dr Gerton Lunter: http://www.well.ox.ac.uk/dr-gerton-lunter
3. Through the Welcome Trust’s Mainstreaming Cancer Genetics programme: https://mcgprogramme.com/
5. Genomics England
6. Project BB/102593X/1 ‘An integrated variant calling pipeline for third-generation sequencing technologies. £626K http://gtr.rcuk.ac.uk/projects?ref=BB/I02593X/1
7. Professor Nazneen Rahman