It's been 10 years since the first draft of the human genome was published. The rough draft of our genetic make-up was published on 26 June 2000 in a joint announcement by the then US President, Bill Clinton, and UK Prime Minister, Tony Blair. The sequencing was deemed a major milestone and promised much to improve our understanding in areas of medicine and biotechnology.

In the decade since the draft was published, gene sequencing techniques have come a long way. The first complete human genetic code took 13 years to map, with the complete sequence being published in April 2003. Now, scientists at the Wellcome Trust's Sanger Institute in Cambridge can sequence an entire genome in 13 hours. And where it took millions of dollars for the Human Genome Project, it now costs around $10,000 to sequence a complete human genome. In fact, to celebrate the 10-year anniversary of the original draft code, the Wellcome Trust has launched the UK10K project to sequence 10,000 human genomes over three years. It will look at the genetic makeup of 4,000 healthy volunteers and 6,000 known sufferers from serious medical conditions including neuro-developmental disorders, congenital heart disease and obesity.

Current sequencing technology can produce terabytes of data, which puts it firmly in the realm of high-performance computing (HPC). The Sanger Institute can generate up to 120 terabytes of raw data per week from its sequencing machines, and has a total of 12 HPC clusters, eight of which run Platform LSF, an HPC management software solution.

‘Traditionally, genomics wasn’t a field in which the HPC industry was interested,’ comments Justin Johnson, director of bioinformatics at EdgeBio. ‘Now that sequencers can generate terabases of data in a week, institutes conducting genomics work are starting to think on the petabyte and exabyte scale for data generation. The introduction of next-generation sequencers has turned the traditional compute infrastructure on its head. It’s not just a software tools problem, but also a compute problem.’

EdgeBio is a research reagents company providing nucleic acid purification products. Within the last 18 months, it has expanded its services to include next-generation sequencing and bioinformatics consulting services. The company operates two SOLiD 4 and four SOLiD 3Plus sequencers from Applied Biosystems, which generate terabytes of data on a weekly basis. It is carrying out sequencing for a methylation project, between Life Technologies and Virginia Commonwealth University, and involving sequencing 1,575 human samples.

‘Traditionally, the cost was in the sequencing itself, because it was such a lengthy process,’ says Johnson. ‘Cost of sequencing has now lowered dramatically, but the informatics hurdles involved in finding value in the data remain and are expanding due to new applications possible with next-gen technologies.’

EdgeBio uses software from CLC Bio, a company providing a wide variety of next-generation sequence analysis tools, and has built its informatics infrastructure around CLC Bio’s genomics server. It has incorporated its own internally-developed algorithms and tools, as well as using open-source tools.

Linking genetics to disease

Human health research uses genomic technologies to gain a better understanding of the genetics behind certain conditions. Research centres will tap into clinical samples stored in biorepositories and conduct genotyping to search for any genetic variations that could be linked to the disease. The Respiratory, Genetics and Epidemiology division of the Channing Laboratory, itself a research division of Brigham and Women’s Hospital and Harvard Medical School, runs a high-throughput genotyping laboratory, which looks for novel mutations associated with disease risk.

The lab handles a biorepository for research
and clinical trials and LabVantage’s Sapphire Laboratory Information Management System (LIMS) with its BioBanking functionality is used for sample tracking and workflow management.

The laboratory conducts genotyping using Genome-Wide Association Studies (GWAS), which involves interrogating the entire genome for associations between specific SNPs (single-nucleotide polymorphisms) or other markers and the disease of interest. Fine mapping sequencing is also used to improve the resolution surrounding the SNP association found in GWAS analysis.

‘GWAS data sets are very large and informatics solutions are necessary to manage these,’ explains Jody Sylvia, senior bioinformatics project manager of the respiratory, genetics and epidemiology division. ‘Where it used to be single SNP genotyping on a sub-set of people, now a very large set of people are genotyped for a whole range of markers and then the researchers drill down into the data to find the particular SNPs that are associated with whatever disease or sub-set of phenotypes they are looking for.’

LabVantage’s Sapphire LIMS is used to manage the biorepository along with storage of subject data, some of the phenotype data, and the case control status, while a tier of other LIMS software is used to manage project-level work. Genetic markers and the sequencing platform used are linked to any other data within the project-level LIMS, so that the investigator has a good understanding of the project.

‘Managing the work in this way means investigators have all relevant information at their fingertips,’ says Sylvia. ‘It also makes collaboration much easier, because each project can be accessed by different collaborators. This is important for how we work; we support a team of investigators and most of them are working on projects in which they want to look at existing data and associate it with a subset of data a fellow researcher has completed.’

The high-throughput laboratory uses sequencing, not only for fine mapping regions of genes, but also for discovery. ‘We have samples from a lot of rare diseases,’ Sylvia says. ‘We’re looking for a specific genetic variant that might not be in the overall population, but for the subset of the disease group this variant means a lot – it changes the severity of the disease.’

Needle in a haystack
With genomics technologies, it’s not only the pure data size that’s a challenge, but also the data mining. ‘It’s like finding a needle in a haystack,’ remarks Dr Jens Hoekrens, head of Genedata Expressionist at software provider, Genedata. ‘There is this huge number of measurements and what you’re looking for is a handful of biomarkers. There’s now more and more hay while the number of needles has remained more or less the same.’ Data mining is where informatics solutions, like Genedata Expressionist, play an important role.

The Genomics Centre at King’s College London is using Qlucore’s Omics Explorer, along with other informatics packages, to interpret its microarray gene expression data. ‘Informatics software is the gateway to the result,’ states Dr Matthew Arno, manager of the Genomics Centre. ‘Anybody can apply samples to microarrays and generate the raw data – that’s extremely straightforward now, robots can do it. But to be able to interpret and analyse the huge amount of raw data and generate, at the very least, a list of significantly differentially expressed genes is a challenge. The informatics software we use makes that possible.’

Arno explains that there’s a lot of redundancy at the level of the probes on the array, as there are multiple probes for each gene. Therefore, multiple measurements are summarised to calculate a gene-level expression value. In addition, factors like the annotation of the gene come into play. Arno also notes that being able to visualise the data is important as it allows scientists to identify trends in the data. ‘The software [Qlucore’s Omics Explorer] doesn’t just produce an anonymous table; researchers can see what’s happening in the data throughout the course of a gene expression study occurring in real-time.’ Visualisation is important for analysts working with genomics data and is functionality available in most software packages used for genomics, including Genedata Expressionist and CLC Bio’s software.

Standardisation
A problem that Sylvia, of the Channing Laboratory, identifies is tracking what algorithm or what analysis tool is used in a study. ‘Everybody uses different tools,’ she says. Not having standard file types complicates working with and exchanging data, although there are groups, such as the Genomic Standards Consortium (GSC), that are trying to implement greater levels of standardisation in the area of genomics.

The problem of non-standardisation is probably an inevitable one. William Mounts, of Pfizer, comments: ‘Data management and data analysis are not the first things the vendor has to address – the first thing is the technology platform itself.’

Predicting genetics
A University of Toledo doctoral candidate in biomedical sciences has used supercomputers at the Ohio Supercomputer Center (OSC) to compute a novel algorithm for the prediction of exon and intron genomic sequences.

Samuel Shepard based the predictions on mid-range sequence patterns of 20 to 50 nucleotides in length, which are said to display a non-random clustering of bases referred to as mid-range inhomogeneity (MRI). Shepard and his team hypothesised that the MRI patterns were different for exons and introns and would serve as a reliable predictor.

Shepard developed a technique known as binary-abstracted Markov modelling (BAMM), which involves creating rules that reduce mountains of nucleotide information into a much smaller binary code, based upon the DNA ‘word’ length and the nucleotide bases found within those words.

To test rules for long word lengths, Shepard utilised the Ohio Supercomputer Center and its 9,500-node IBM Cluster 1350. Shepard comments: ‘During the project, our algorithm read 12 million nucleotides of exons and introns each, and three million each were used to test the predictions.’

www.scientific-computing.com
Another informatics bottleneck that Mounts identifies is linking results data to other sample attributes and providing the information to a broad array of scientists in an intuitive way. Pfizer is using Genedata Expressionist and other informatics platforms to provide an interface for researchers to access and analyse disparate genomic data types.

Intuitive presentation of data is something that Ruth Burton, product manager clinical solutions at Oxford Gene Technology (OGT), also considers vital. OGT provides microarray products and services and Burton comments: ‘The challenge is going from a tiff image of a microarray to having data that people can work with, especially in a cytogenetics lab where researchers may be less familiar with microarray analysis. Considering the number of data points involved, it is not a trivial step to go from the image to the results data.’

OGT’s CytoSure Interpret software is used to analyse the data generated from CytoSure microarrays, providing an overview of the genomic locations where signal intensity differs between sample and control DNA.

**Only a piece of the puzzle**

Of course, genomics is not the be all and end all when it comes to investigating human physiology and a lot of institutes are trying to link the data generated by genomics with that of other omics technologies as well as clinical data. Trish Meek of Thermo Fisher Scientific comments: ‘When genomics is paired with proteomics, we can gain a real understanding of the human body.’

Thermo Fisher Scientific’s Nautilus LIMS supports configurable sample and aliquot hierarchies to allow laboratories to tailor the sample structure to reflect the way they work. Information can then be referenced at each level in the hierarchy, enabling laboratories to track and maintain all of their sample data.

One of the goals of the Molecular Genomic Core of the University of Texas Medical Branch, for implementing GenoLogics’ Genexus platform to manage gene expression data, was to integrate the data generated from PCR analysis into a broader view that includes traditional microarray analysis. ‘We have a limited view when we just look at the gene expression data. Now you can take a larger view,’ says Trish Meek of OGT.

**Where next for next-gen?**

‘An important point about his whole area is that it is rapidly changing,’ states Sheldon. ‘It’s evolving at such a pace that there has to be software that can cope.’

Sheldon also argues that, with regards to next-generation sequencing, the beauty of it is that it could give some uniformity to bioinformatics. In addition to sequencing genomes, it can be used to carry out SNP detection and whole transcriptome profiling – sequencing can give an idea of expression profiles instead of being reliant on microarray technology. ‘Gene sequences become the currency, if you like, of all these different genomic technologies,’ he says. ‘Sequence data as a uniform currency across all these techniques will make the modelling of the data much more straightforward.’

However, Johnson of EdgeBio warns that as the cost of next-generation sequencing continues to fall, the danger is that everyone starts doing it their own way and the bioinformatics becomes even more non-standardised. Companies that don’t have the capabilities of large genome centres tend to do whatever they can to get an answer. It’s a big educational issue within the community to help people to try to use standardised tools.

Next-gen sequencing is still too expensive to be a widely-used genomics tool – the $1,000 genome is generally considered the holy grail, in terms of a price for which next-gen sequencing could begin to have clinical applications.

‘One of the things I hear, almost on a daily basis, is that sequencing is almost free now,’ says Johnson. ‘That’s hard to see when there is a lot of capital expenditure, such as the sequencing machines, the staff, the project management, the bioinformatics, the compute infrastructure, etc. The cost of the reagents for sequencing is dropping significantly – the cost of reagents for Applied Biosystems’ SOLID system to sequence an entire human genome is $6,000. What comes before and after is where people need to understand how to invest their time and money and those are significant challenges.’