Genome Biology:

Techniques to study *de novo* mutations.

Dr Alex Moorhouse, Centre for Health Effects of Radiological and Chemical Agents, Institute of Environment, Health and Societies, Brunel University London.

Members of the nuclear test community identified some years ago a need for more research into the potential for genetic damage (changes to DNA) amongst nuclear test veterans, and the possibility of veterans having transmitted genetic alterations to their children. To address these questions, researchers at the Centre for Health Effects of Radiological and Chemical Agents (CHRC) are undertaking a cytogenetic (chromosomal) and genetic (DNA mutation) study to check if any genetic changes are found in a sample population of nuclear test veteran families and to compare this with veteran families with no association to nuclear test sites.

To support the community to better understand the potential outcomes of this work, we have created a series called 'Genome Biology' to provide background scientific knowledge relevant to different aspects of this project. The first and second instalments in this series (published in *Campaign*, summer 2016 and spring 2017) examined how chromosomes may become structurally altered and also, the techniques used to detect such chromosomal changes.

In the third of our 'Genome Biology' series we will look at the key techniques used to detect DNA mutations in human cells, and how we will employ this advanced technology in the 'Genetics' study.

The rapid development of new DNA sequencing technologies during the last 15 years has fundamentally changed how we approach many scientific questions. The ability to deliver fast, accurate and affordable DNA sequence data enables us to answer genetic questions in much greater detail than previously thought possible *and* to ask many new ones. Challenges and limitations remain, and a continued improvement of the different technologies involved is required if we are to shed light on the most complex and difficult scientific questions. Genetics and genome biology have come a long way over the last 50 years, and advancements within these areas continue at a rapid pace.

On the shoulders of giants

In 1953 James Watson and Francis Crick working at Cambridge University made a remarkable observation from an X-ray photograph produced by Rosalind Franklin and Maurice Wilkins at King's College London. They described two stands of a molecule running anti-parallel to one another forming a distinctive double helix with pairing of complementary bases (nucleotides) in the centre. Already known to be the hereditary material, the study of deoxyribonucleic acid (DNA) subsequently benefited from several further key developments. In the mid-1970s Frederick Sanger in Cambridge developed a method to determine the sequence (code) of nucleotides along a strand of DNA (Figure 1). During the 1980s a method for the targeted amplification (to make multiple copies) of specifiable regions of DNA, termed as the polymerase chain reaction (PCR), was developed from pioneering work by Kary Mullis in California. Both Scientists were rewarded with Nobel prizes for developing these elegant and powerful new techniques.

These innovative advances enabled work on several challenging projects to commence in earnest, including DNA sequencing of the entire human genome, investigating genetic diversity among populations worldwide, sequencing ancient DNA from fossil remains, and the introduction of DNA fingerprinting in forensics. Indeed, genetic research on heritable human diseases, infection and immunity, evolution and biodiversity have all benefitted from these technological advances.

What is Next Generation Sequencing?

Following the turn of this century, new technologies for DNA sequencing have improved significantly. Several competing approaches capable of 'massively parallel' DNA sequencing increased data (information) output and reduced the cost of experiments substantially (Figure 1). Next Generation Sequencing (NGS) brings together developments in engineering, robotics, imaging, microfluidics and biochemistry. The more advanced of these instruments are today yielding over 6000 billion nucleotides (bases in the DNA code), equivalent to sequencing 50 human genomes within 2 days. These advances led to a series of ground breaking discoveries and ushered the field of genomics into a new era.

Next Generation Sequencing, also known as 'short read', or 'shotgun' sequencing, requires genomic DNA to be fragmented into many millions of smaller pieces before it can be sequenced. Because fragmentation causes DNA pieces to be jumbled up, data from NGS are like pieces of a jigsaw that need to be fitted together. To do this a reference DNA sequence is used, like looking at the picture of a jigsaw box. DNA sequence 'reads' overlap one another and can be paired which helps them fit together correctly. Sequencing an entire genome using NGS produces tens of millions of sequence reads and piecing these together is a computationally intensive process requiring

specially developed software. Once completed, it is then possible to compare genomes from different populations to detect any differences or 'variants' between them. These differences might be single nucleotide variants (SNVs), variants that are short insertions or deletions (indels), or larger structural variants (SVs), such as gene copy number variants (CNVs), or more complex rearrangements. The two populations to be compared in the 'Genetic' study underway at CHRC are nuclear test veteran families and veteran families with no known participation at test sites.

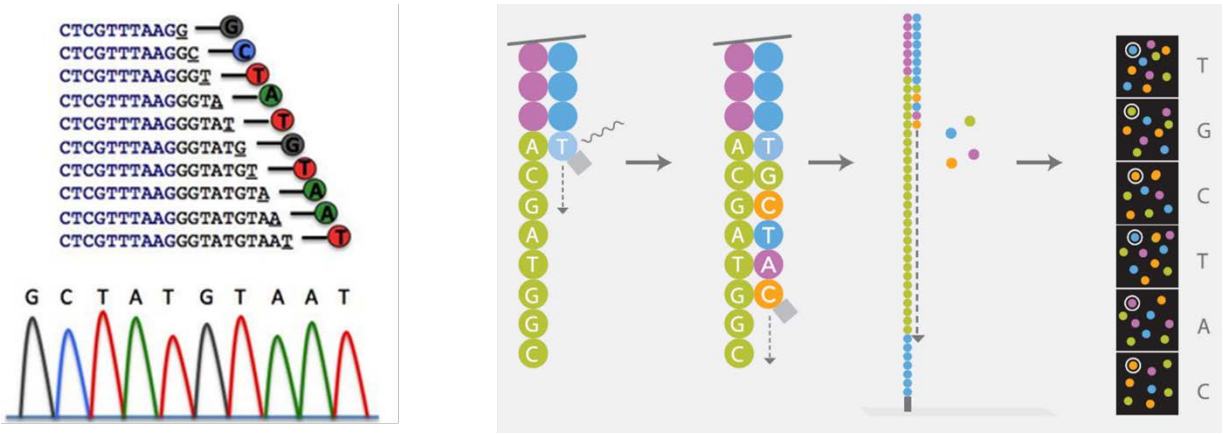

Figure 1) **The Sanger sequencing method** (left) developed in the 1970s determines the sequence of nucleotides (or bases comprising the DNA code) along a DNA molecule. The DNA sequence can be determined because there are many copies of each possible length of the DNA molecule and because each copy terminates in one of the four fluorescently (different coloured) labelled bases. Detection is performed by a camera that measures the fluorescent colour signal as the molecules pass through a tube. Sequence lengths using this method are usually 300-1000 nucleotides. **Next-generation sequencing (NGS)** (right) speeds up this process using new chemistry that allows the removal of the fluorescent coloured label after detection, thereby permitting successive cycles of base incorporation and detection. The term 'massively parallel' derives from the sequencing of tightly packed clusters of molecules, where each cluster contains many single stranded copies of a different DNA molecule. Many millions of clusters can be sequenced simultaneously.

De novo (newly arising) DNA mutations

New DNA mutations contribute to the natural process of DNA variation and arise as a result of DNA damage. DNA damage occurs naturally throughout life and is usually corrected very efficiently by DNA repair processes. However, when damaged DNA is not correctly repaired then errors called mutations accumulate within the genome. An excess of DNA mutations can occur when a cell's response to DNA damage is compromised or overworked, and when a cell is exposed to chemical

mutagens or ionising radiation. A new DNA mutation that arises in a germ cell (egg or sperm) of one of the parents and which is then transmitted to the child or children, or in a fertilized egg cell itself, is called a *de novo* or newly arising mutation (DNM) (Figure 2).

In the study being undertaken at CHRC, we are looking for DNMs that may have arisen in the germline of British nuclear test veterans as a consequence of veterans participating at nuclear bomb test sites. DNMS are DNA variants present in the child but not in either of the parents' own genomes. To do this we are using NGS technology to sequence the genomes of test veteran families in father-mother-child trios, and comparing what we observe with a sample population whose veteran fathers did not attend nuclear test sites. Although DNMs induced by ionising radiation may themselves be largely indistinguishable from naturally occurring DNMs, we can look for deviations in the expected number and pattern of DNMs because we are comparing our data to a control group. Previous work has shown that DNMs arising from exposure to radiation may include more clustered SNVs and more CNVs, and that more DNA mutations may also occur at highly mutable repeat regions compared to the general population.

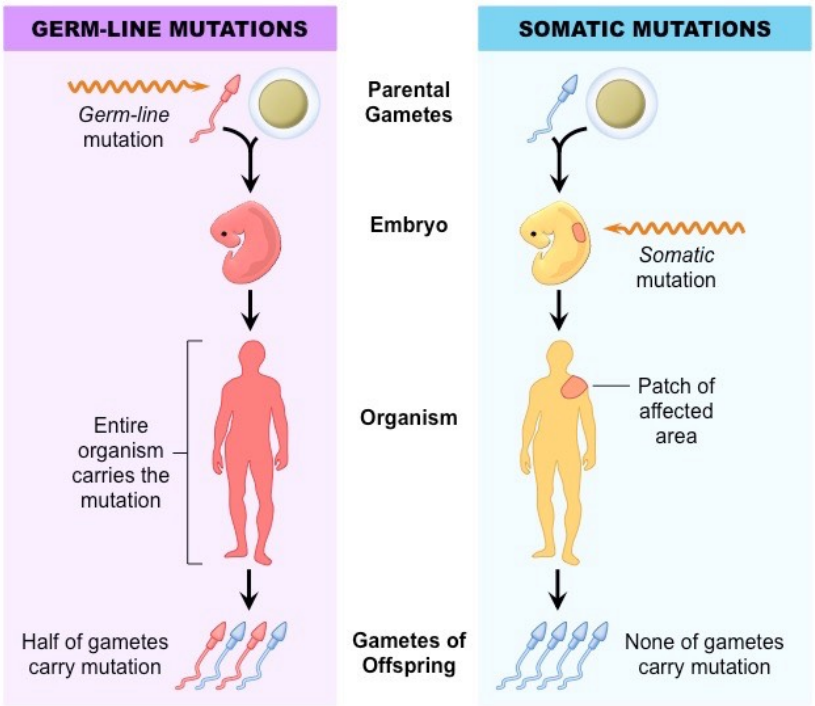

Figure 2) **A new mutation** that arises in a germ cell (egg or sperm) of one of the parents, or in the fertilized egg cell itself, is called a germ-line *de novo* mutation (DNM) and is present in the child's genome but not in either of the parents. Mutations also arise in non-germline cells, including those of developing embryos, these are called somatic mutations and are not transmitted to offspring.

Challenges and solutions

Remarkably good tools and techniques for sequencing and analysing DNA mutations have been developed over the last 15 years, although it is not yet possible to detect all mutations in the human genome with total accuracy. To mitigate against this we will sequence the entire genome of all participants in our study to a 'depth of coverage' of 30-35 times (where most areas of the genome have been read an average of 30-35 times), called 'high-depth' or 'deep' sequencing. DNA variant detection software also has its limitations and can report false positive DNMs (where a mutation is reported, but turns out not to exist). We will therefore again follow standard practice by using multiple detection tools which learn to recognise sequencing errors, and we will apply additional stringent filters to help sift out false positives. To ensure we are successfully detecting true DNA variants, we validate (double-check) DNMs by manually inspecting all candidate DNA mutations, and by sequencing regions containing DNA mutations a second time to confirm that DNMs are present.

The techniques and processes described above are currently being employed for the detection and comparison of DNA mutation variants in nuclear test and control family trios, and will take approximately two years to complete. We look forward to being able to report our findings to the nuclear test community once we have published our findings in the peer-reviewed scientific press.