Analysis of TCR sequencing data

The key to adaptive immunity lies in recognition molecules called receptors. These are found on the surface of immune cells called T-cells. The T-cell receptors (TCRs) can recognise individual germs. A person is thought to have as many as a hundred million different TCRs. This is called the 'repertoire'. When we get an infection the best-fitting TCRs are selected to fight this pathogen. The cells with these TCRs then remain on 'red alert' for further infections and ensure that a secondary immune response is both faster and stronger. Different individuals, even identical twins, generally possess different TCRs although they may share some or even many. It is believed that TCRs are the root cause of autoimmune diseases such as type 1 diabetes, rheumatoid arthritis, multiple sclerosis and so on.
In theory, human V(D)J rearrangements could produce over 10^{18} different TCRs. Only an infinitesimally small fraction of these possibilities are ever used. The recent advances of sequencing techniques have made it possible to sequence many millions of TCR sequences at once. In a collaboration with Andrew Sewell at Cardiff we have recently obtained a large data set of this type. This allows us to analyse the distribution of V(D)J recombination events and to estimate frequencies of individual TCR clones.

We would like to explore the data testing some fundamental assumptions about V(D)J rearrangements and the generation of diversity in the TCR repertoire. These analyses will allow us to be among the first few groups to learn from such data and will lay the basis for future projects where we would like to compare TCR repertoires, for example between patients and controls, or at different points in the life time of an individual. The scope for future work in this area is vast.

Examples of specific questions we would like to try and answer within the short mini-project are: Can we find new, previously uncharacterised V, D, or J segments in the genome? Is recombination happening at clearly defined boundaries, or is it “messy”, i.e. joining partial segments together? What nucleotide changes do we find apart from recombination events (is there some form of editing at the linking regions, or even elsewhere)? Why does the fragment length distribution of our data show clear peaks at certain fragment lengths?

Programming skill is important for pursuing this project. If you are interested in this project, but unsure about your programming skills then please have a chat with me (Sascha).
We will develop programming skills further, but also train data analysis and interpretation, statistics, and presentation of results.

http://www.frontiersin.org/Journal/10.3389/fimmu.2013.00463/abstract

This paper is an example of a paper on computational methodology:
http://www.ncbi.nlm.nih.gov/pubmed/22806588

We have used some of the following software for our preliminary analyses:
http://www.pnas.org/content/early/2009/10/28/0909775106.full.pdf+html