Exploratory analysis of next generation sequencing data

Ali M. Mosamam[1*], Constanze Bonifer[2] and David R. Westhead[1]

[1] Institute of Molecular and Cellular Biology, University of Leeds
[2] Institute of Molecular Medicine, University of Leeds

New massively parallel sequencing methods are set to revolutionise molecular biology through the generation of data of unprecedented volume and precision. Our preliminary work with a technique known as ChIP-Seq has shown that the method can delineate specific marking of chromatin and binding of the cellular transcriptional machinery to precisely defined regions of chromosomes, improving significantly in resolution and precision by comparison with older techniques involving DNA microarrays. We are now continuing this work to study the key process of cellular differentiation in cells of the immune system. This is a dynamic process during which cells develop into more specialised cell types by changing their programme of gene transcription. When it goes wrong can lead to cancer (leukaemia). To fully understand this process we will use sequencing to examine changes in chromosomal marking, the binding of key transcription factors and the expression of genes, dynamically as cells differentiate, both normally and in tumour development. This brings significant statistical challenges, in particular the need to quantify not only the genomic loci with specific markings and bound transcription factors but how these change in both position and amplitude with time. We present preliminary studies of these effects, employing visualization techniques based on the Hilbert curve for high length linear data (chromosomes), and an assessment of the likely performance of the method in assessing the necessary dynamic changes.