Improved Footprinting of DNase-seq data

DNase-seq is a state-of-the-art approach to analyse the global pattern of protein-DNA binding in regulatory regions of DNA. In this assay, DNA is digested by DNaseI and high-throughput sequencing is applied to reveal the genome-wide pattern of DNA digestion in hypersensitive regions. Regions of DNA bound by regulatory proteins are protected from digestion and leave behind “footprints”, regions of reduced cut frequencies.
A single DNaseI-seq data set provides a “snapshot” of the global occupancy pattern of regulatory proteins on DNA, revealing binding of a large number of different proteins at once which is highly informative. For example, our collaborators at Birmingham, Constanze Bonifer and Peter Cockerill, are using this type of data to study aberrant regulatory events in leukaemic cells.

In order to derive meaningful results from this type of data it is of paramount importance to understand the main factors that influence the read distribution as it is not merely determined by transcription factor binding events alone. One major factor is the distribution of nucleosomes on the DNA as was clearly shown in a recent Nature Methods paper (see below). In this project we are going to make use of publically available pairs of data sets where both DNase-seq and MNase-seq data are available. MNase-seq data reveals nucleosome positions as the MNase enzyme cuts predominantly in regions of linker DNA (i.e. DNA not occupied by nucleosomes). We will use the MNase-seq data to infer nucleosome positions and based on this inference correct the DNase-seq data for the effect of well-positioned nucleosomes. Time permitting, we will use existing software to evaluate if the detection of footprints can be improved in this way.

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 Ott).
We will develop programming skills further, but also train data analysis and interpretation, statistics, and presentation of results.

The following paper is describing a new method for the analysis of single DNase-seq data sets that we published recently. It is becoming very popular and we know that it is being used by a number of leading labs in the field. We are keen to build on the current momentum to establish this tool and related tools we are developing in the near future as a widely used means for DNase-seq data analysis.

http://nar.oxfordjournals.org/content/41/21/e201.full?keytype=ref&%2520ijkey=C5ERIRQ7eNX7whC

This paper makes a big recent improvement to our understanding of what influences (or biases) DNase-seq data:

“Refined DNase-seq protocol and data analysis reveals intrinsic bias in transcription factor footprint identification”, Nature Methods 11, 73–78

(Warwick do not subscribe to Nature Methods (!), but I can send you the PDF-files)

In a PhD-project in this area we would expand current analysis techniques to the case of comparing two datasets, for example data from cancerous cells versus normal cells. We would develop statistics that distil out the differences between the data sets, in particular differences that are likely to be attributable to changes in binding of regulatory proteins. We can further explore the use of paired-end sequencing data and the use of other enzymes.