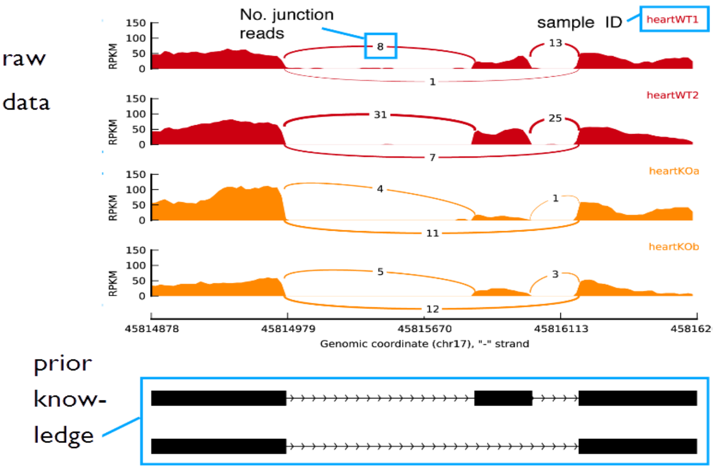
Project: *Biomedical Big Data Analysis (II): Probe optimization for alternative transcript discrimination*

Supervision & Resources: We can arrange access to a modern CSC four-screen analysis workstation at Systems Biology in Warwick and a state-of-the-art high-performance compute cluster in Vienna. We can meet in person for project kick-off and towards the end of the work period, complemented by weekly meetings by video-conference. Alternatively, if you want to spend some time with the group, we can also fly you to Vienna for a shorter or longer visit. If you have any questions, please feel free to contact me at any time (D.Kreil@warwick.ac.uk).

Why: In the molecular life sciences, we study complex living systems by large-scale parallel measurements that collect very high dimensional data sets characterizing the constituent molecules and their activities. The profiling of gene expression, in particular, gives functional snapshots of which parts of the genome are actively used. Typically, the number of parameters *p* which we want to investigate (such as inferring which genes interact) is much larger than *n*, the number of samples (patients, say). This is exacerbated by the ever improving resolution of genome-scale assays. Recent developments now allow the discrimination of alternative gene transcripts, which increases the measured variables more than 5-fold to well over 250,000. This is exciting because alternative transcripts are known to have major roles in determining cell types and tissues, have been associated with neuronal and immune-system specific functions, and are implicated with a number of diseases. Measurements based on next-generation sequencing, however, often depend on signals from very specific gene regions to discriminate alternative transcripts. In untargeted assays, there can be very few sequencing reads falling into these regions, giving extremely noisy estimates of transcript abundances. Latest developments indicate that next-generation high-density microarrays with hybridization probes targeting these key regions could provide a highly efficient alternative [2] using modern probabilistic models [3,4].



What: Models developed for next-generation sequencing assume a uniform signal along the gene, although it is known that coverage can vary ~100-fold, just as for microarrays. For microarrays, these variations can be modelled empirically [4] or by thermodynamic simulation [5]. The challenge now is to select sequence regions to probe along the genes that maximize sensitivity and specificity, choosing both probe numbers and locations, and taking gene model complexity into account to minimize prediction uncertainty. You will implement and evaluate an algorithm for a highly parallel cluster computing environment. If time permits, alternative algorithms can be tested and compared.

PhD option: An exciting extension of this work would be the application to over 5,000 microarrays from blunt trauma patients that have been collected by collaborators at Harvard. Improved high-level analysis methods will then need to be developed and implemented that can exploit the uncertainties of transcript expression level estimates in a probabilistic framework [4], to allow the efficient testing of functional groups or the identification of latent factors.

References

1. Thierry-Mieg D and Thierry-Mieg J. (2006) AceView: a comprehensive cDNA-supported gene and transcripts annotation. *Genome Biol* **7**, S12.1-14.
2. The SEQC Consortium (2013) Power and Limitations of RNA-Seq. *Nature Biotechnology*, in press – we hold joint first authorship and joint senior/corresponding authorship.
3. Katz Y, Wang ET, Airoldi EM, and Burge CB. (2010) Analysis and design of RNA sequencing experiments for identifying isoform regulation. *Nature Methods* **7**, 1009-15.
4. Liu X, Gao Z, Zhang L, and Rattray M. (2013) puma 3.0: improved uncertainty propagation methods for gene and transcript expression analysis. *BMC Bioinformatics* **14**, 39.
5. Leparc GG, Tüchler T, Striedner G, Bayer K, Sykacek P, Hofacker IL, and Kreil DP. (2009) Model-based probe set optimization for high-performance microarrays. *Nucleic Acids Res* **3**, e18.