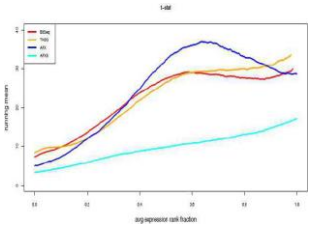
Project: *Big Data Analysis (I): Characterization & compensation of detection bias in differential gene expression profiling*

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). Classical approaches thus include feature selection, variable screening, and latent factor analyses, in the hope that the selected genes or the identified factors correspond to a relevant biological process for further investigation (*e.g.*, by experimental perturbation or drug treatment). To overcome the limitations of individual experiments, modern analyses combine complementary data sets. We distinguish 'lateral' data integration of conceptually equivalent data sources (*e.g.*, gene expression profiles from multiple laboratories, different technologies) and 'vertical' integration of different data types (*e.g.*, gene expression and genome variation). The development and demonstration of methodological improvements are challenging and at the centre of current research efforts. Advances have an immediate impact on thousands of biomedical studies each year.

What: One of the most common analyses is the search for biologically relevant differences between two conditions (*e.g.*, diseased *vs* healthy, therapy responder *vs* non-responder). We will first focus on 2-group comparisons of cancer patient gene expression profiles. While the non-linear dependency of the measurement variance on the signal level is well known, and transforms and adjusted statistical tests have been developed to account for that heteroscedasticity [1,2], the non-linear distortion of the measured signal as a function of the true signal level has received little attention. We can show that both are specific to the methods chosen for an experiment (technology type, data processing), creating detection bias (see figure) and impeding an efficient integration of data sets. In collaboration with the US FDA [3], we have collected a large multi-laboratory reference data set that now lets us characterize the combined impact of both effects and evaluate the effectiveness of different approaches to compensating for it. Depending on your methodological background and interests, you will adapt classical frequentist statistics [4] and/or approaches adjusting priors in a Bayesian framework, and assess them at gene and function levels (testing the implication of GeneOntology classes, and the specificity and stability of unsupervised methods).



PhD option: A logical extension of this project is the 'vertical' integration of quantitative and discrete data sets. In collaboration with colleagues in Boston, we work on extending an algorithm for latent pathway analysis [5] that adapts the Google page rank algorithm to join multi-track evidence along biological networks. Work will focus on selected subsets of the The Cancer Genome Atlas repository, providing clinical data, expression profiles, and genome variation tracks including single nucleotide polymorphisms (SNP), copy number variation, and DNA methylation.

References

1. Huber W, von Heydebreck A, Sültmann H, Poustka A, and Vingron M. (2002) Variance stabilization applied to microarray data calibration and to the quantification of differential expression. *Bioinformatics* **18**, S96–S104.
2. Oshlack A and Wakefield MJ. (2009) Transcript length bias in RNA-seq data confounds systems biology. *Biol Direct* **4**, 14.
3. The SEQC Consortium (2013) Power and Limitations of RNA-Seq. *Nature Biotechnology*, in press – we hold joint first authorship and joint senior/corresponding authorship.
4. Young MD, Wakefield MJ, Smyth GK, and Oshlack A. (2010) Gene ontology analysis for RNA-Seq: accounting for selection bias. *Genome Biol* **11**, R14.
5. Pham L, Christadore L, Schaus S, and Kolaczyk ED. (2011) Network-based prediction for sources of transcriptional dysregulation using latent pathway identification analysis. *Proc Natl Acad Sci U.S.A.* **108**, 13347-52.