Power and limitations of RNA-Seq: findings from the SEQC (MAQC-III) Consortium

Paweł Piotr Łabaj

Chair of Bioinformatics Research Group, Boku University Vienna, Vienna, Austria

We present primary results from the Sequencing Quality Control (SEQC) project by US-FDA MAQC consortium. Here we present a multi-centre cross-platform study introducing a landmark RNA-Seq reference dataset comprising 30 billion reads. In addition to NGS also several microarray and qPCR platforms were examined. The study design supports large variety of complementary benchmark metrics by featuring known mixtures, high-dynamic range ERCC-spikes, as well as nested replication structure. With no independent 'gold standard' feasible, these built-in truths support an objective assessment of performance and are critical for the development and validation of novel or improved algorithms and data processing pipelines. We find that measurements of relative expression are accurate and reproducible across sites and platforms if specific filters are used. Comparisons with microarrays identified complementary strengths, with RNA-Seq at sufficient read-depth detecting differential expression more sensitively, and microarrays achieving higher rank-reproducibility. Measurement performance depends on the platform and data analysis pipeline, and variation is large for transcript-level profiling. On the other hand, even at read-depths >100 million, we find thousands of novel junctions, with good agreement between platforms, and with qPCR validation-rates >80%. We also have shown that the modelling approaches for inferring alternative transcripts expression-levels from read counts along a gene can be applied to probes along a gene in high-density next-generation microarrays. This has advantages in quantitative transcript-resolved expression profiling.