Measuring the activity of multiple signal transduction pathways in cancer using RNA sequencing on cells and FFPE tissue

***F. van Hemert*, B. de Regt, A. van Brussel, G.C. Dam-de Veen, S. Kuijpers, W. Verhaegh, A. van de Stolpe, S. van den Bosch**

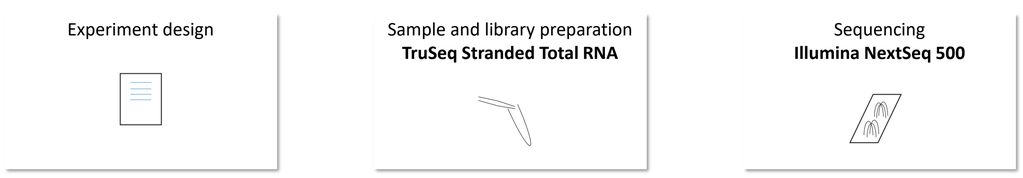
Philips Research, Eindhoven, The Netherlands, contact: freek.van.hemert@philips.com

Summary

With the introduction of high throughput genomics technologies, extensive molecular cancer information has become available, but translation into clinically actionable pathophysiological characteristics for improved treatment selection remains a challenge [1]. The past decade we have developed, and biologically validated, assays to quantitatively measure activity of clinically actionable signal transduction pathways (estrogen receptor, androgen receptor, PI3K-FOXO, Wnt, Hedgehog, TGFβ, Notch, NFκB, JAK-STAT1/2 , JAK-STAT3, MAPK pathways) in cancer cell/tissue samples, based on interpretation of expression levels of target genes of pathway-specific transcription factors via a Bayesian computational model [2],[3],[4],[5],[6]. While originally developed for Affymetrix microarrays, our knowledge-based approach enables assay translation to other mRNA measurement platforms, which was successfully done previously for RT-qPCR. We now successfully translated pathway activity models (estrogen receptor, androgen receptor, Wnt, Hedgehog and TGFβ) to RNAseq, enabling use of RNAseq data to quantify signaling pathway activity. Pathway target gene expression values measured with RNAseq correlate well to microarray and qPCR gene expression measurements, from the same samples. We will present results showing RNASeq-based signaling pathway model calibration and validation on cell line experiments with known signaling pathway activity. Using this approach we can reliably determine the activation status of multiple signal transduction pathways in cell lines as well as in clinical FFPE material.

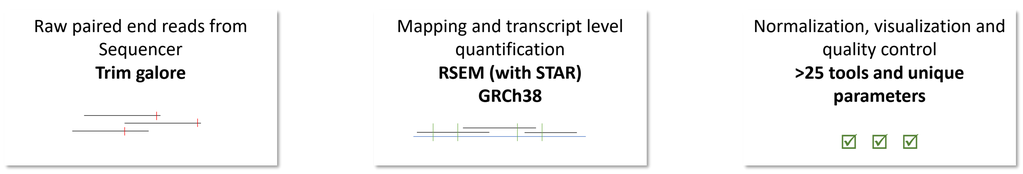
RNA sequencing workflow

Sample preparation

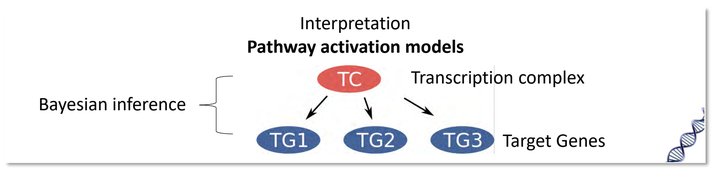


We use an Illumina NextSeq 500 (capacity 400 million reads). For FFPE samples, we typically sequence 100 million reads per sample prepared using the TruSeq Stranded Total RNA kit (Illumina).

The data analysis pipeline



Our Snakemake based data analysis pipeline uses bcl2fastq to generate de-multiplexed paired-end fastq files, performs adapter and quality trimming using trim galore, before proceeding with RSEM (with STAR) to perform transcriptome mapping and expression estimation based on Ensembl's GRCh38 reference and annotation. After each step we perform extensive Quality Control and generate reports (using MultiQC and a custom made QC module).



The renormalized gene level quantification values serve as input for our pathway activation models which uses Bayesian inference to compute the probability of a specific transcription complex and upstream signaling cascade being active [2].

Re-normalization to improve reproducibility

In our data analysis pipeline, we renormalize the RSEM generated gene level Transcripts Per Million (TPM) values (derived from the weighted transcript based quantification) on a selected set of genes. These adapted TPM (aTPM) values (where the sum of the expression values of this gene set is normalized to 1 million) are more stable across technical replicates with a large difference in quality (figure 1) and when comparing Fresh Frozen (FF) to Formalin Fixed Paraffin Embedded (FFPE) samples (figure 2).

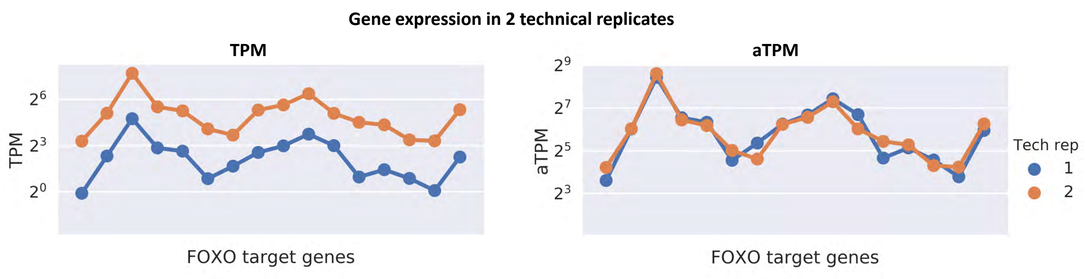


Figure 1, FOXO target gene expression values for 2 technical replicates of an Endometrioid adenocarcinoma of the uterus FFPE sample. Left; gene expression values in TPM. Right; the same gene expression data but normalized using the adapted TPM. The first technical replicate (1, blue) contained a lower fraction of uniquely aligned reads (21 vs 62%, 8.5 vs 19.4 million reads), fewer reads mapping to exons (2.4 vs 11%) and much more overrepresented sequences (probably due to inefficient rRNA removal). Although these quality control parameters result in lower TPM values, aTPM corrects for these effects and gives values similar across the 2 replicates, for all genes in this subset.

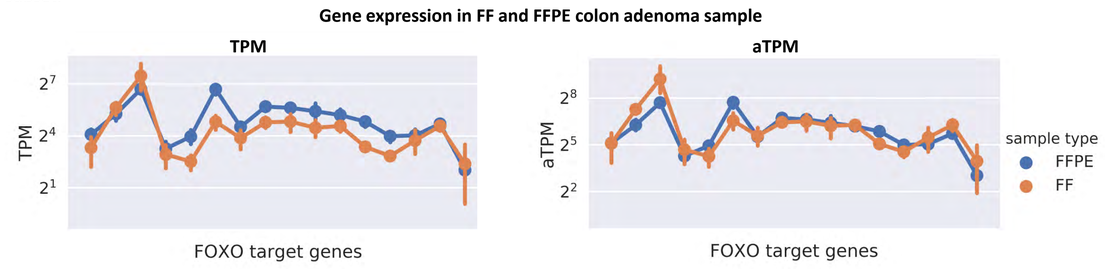


Figure 2, Average and standard deviation of FOXO target gene expression values for 9 FFPE (blue) and 5 Fresh Frozen (orange) colon adenoma samples. aTPM normalization makes it possible to directly compare gene expression values between FFPE and FF samples.

RNAseq results correlate well to Affymetrix and PCR based transcript quantification

In the past, pathway activation models developed on Affymetrix microarray data have translated well to qPCR based transcript quantification requiring only recalibration. In figure 3 we show that RNAseq provides similar gene expression patterns as qPCR and Affymetrix arrays. This means that our established pathway activation models will translate well to the RNAseq platform.

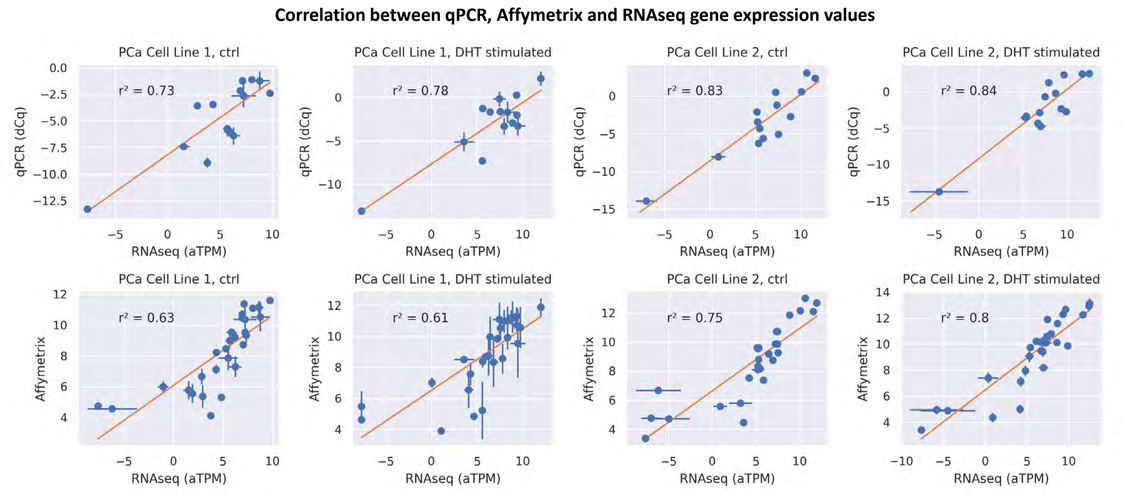


Figure 3. Scatter plots of RNAseq (log2 aTPM) versus PCR (reference gene based normalized Ct values) or Affymetrix microarray U133 P2.0 gene expression values for 2 prostate cancer cell lines, either vehicle stimulated or stimulated with Dihydrotestosterone.

Calibration and validation of AR and TGFβ pathway activation models

We used the RNAseq aTPM (log2) gene expression values to calibrate AR and TGFβ pathway activation models. For the AR pathway we used 2 prostate cancer cell lines (the same data as presented in figure 3), either vehicle stimulated or stimulated with DHT. The results below show that our models are able to discriminate between cells with a stimulated or an unstimulated AR pathway using both models (figure 4, left). For calibration of the TGFβ pathway model we used an adenocarcinoma and breast cancer cell line model, either vehicle stimulated or stimulated with hTGFB1. Again we can discriminate the activation status of the TGFβ pathway on both cell lines using models calibrated on any of the two cell lines. Different pathway activity values between models may be explained by differing baseline pathway activation. We have started testing and evaluating these models on relevant public data.

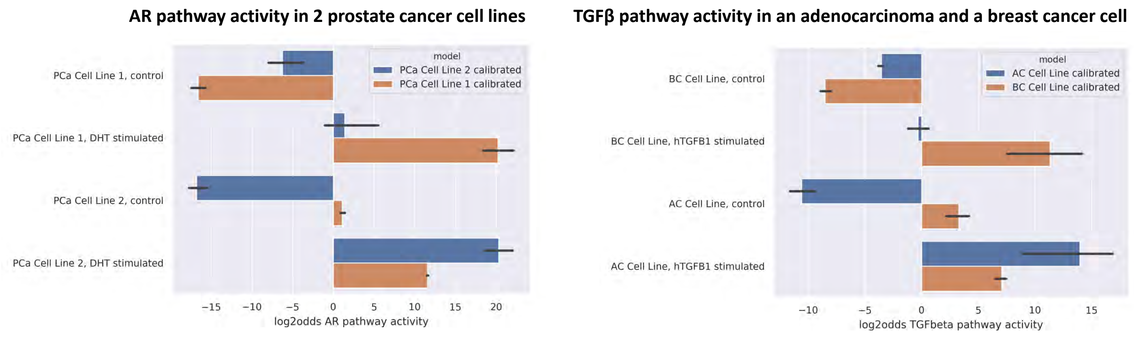


Figure 4. Left: Barplots of the log2 of the odds of the AR pathway being active vs. inactive (log2odds) according to 2 differently calibrated models (blue vs orange). Right: Barplots of the log2odds of the TGFβ pathway being active according to either a Breast Cancer cell line calibrated model (blue) or an adenocarcinoma cell line calibrated model (orange). In both plots the error bars represent the standard deviation over 3 replicates.

Conclusion

Using the Illumina Nextseq 500 sequencer and an adapted data analysis pipeline, we have been able to generate robust gene expression data which can serve as a stable input for our pathway activation models [2-6]. We have shown that by using our RNAseq pipeline and proper normalization we obtain similar gene expression values for FF and FFPE samples and that we can salvage poor quality samples. Our results correspond well to the results of gene expression platforms on which we have developed pathway activation models that provide clinically actionable results.

We are currently establishing RNA sequencing as a high quality method for generating gene expression data to use as input for our Bayesian signal transduction pathway activation models. Using these models we map aberrant signal transduction and provide means for improved therapy selection.

References

[1] C. Massard et al., Cancer Discov, vol. 7, no. 6, pp. 586–595, 2017.
[2] W. Verhaegh et al., Cancer Res., vol. 74, no. 11, pp. 2936–2945, Jun. 2014.
[3] W. Verhaegh and A. Van de Stolpe, Oncotarget, vol. 5, no. 14, pp. 5196–5197, Jul. 2014.
[4] H. van Ooijen et al., Am. J. Pathol., vol. 188, no. 9, pp. 1956–1972, Sep. 2018.
[5] A. van de Stolpe, et al., Sci Rep, vol. 9, no. 1, p. 1603, Feb. 2019.
[6] A. van de Stolpe, Cancers, vol. 11, no. 3, p. 293, Mar. 2019.