Use of long sequencing reads to investigate RNA splicing

Vertebrate transcription generates a large number of RNA transcripts, which undergo further processing through splicing to generate an even greater number of unique molecules entities. RNA splicing is an essential cellular process carried out at the major and minor spliceosomes in collaboration with multiple splicing factors[1]. The control of splicing is of fundamental importance in the control of gene expression and abnormalities of splicing are increasingly recognised in diverse pathological states ranging from heritable single gene disorders to diseases in which splicing is widely disrupted, including several types of cancer[2,3]. Furthermore, in recent years splicing is increasingly seen as a potential node for therapeutic intervention in diverse contexts including neovascularisation and cancer[3,4]. As a result of such advances, there has been a significant interest in improving our understanding of splicing in both health and disease.

An improved understanding of splicing relies on a comprehensive annotation of splicing events/alternative splicing in health, in order to capture natural variation and physiological control of the process. With this in place, the study of splicing in pathological states is likely to be much more informative and better able to both define the role of splicing in disease and identify ways to manipulate splicing as a therapeutic intervention (REF). Advancement of these aims requires an improved way to investigating splicing both at the sequencing and at the analysis stage. To date splicing analyses have utilised datasets derived using short sequencing reads which are characteristic of the prevailing sequencing technologies. ***In this proposal, we propose that the application and analysis of long sequencing reads has the potential to give new insights into i) the annotation of splicing events in health and ii) the impact of splicing aberrations in disease and therapy*.**

The Brazma & Vassiliou groups

The *Brazma group* has a long track record in the study of the transcriptome and specific expertise in analysis of splice variants in human using RNA-seq data. We have developed an RNA-seq data analysis pipeline[5] that includes differential transcript analysis and used it to demonstrate that for protein coding genes in normal tissues most genes express one dominant transcript[6] and that in some cases switch of the dominant transcript lead to loss of function in cancer[7]. Our methods have also been used to annotate human genome[8]. The results of these analyses have had a direct impact on the development of the Expression Atlas at the EBI[9]. With single cell RNAseq the long transcript sequencing becoming possible, we are currently working to extend our analysis methods and pipeline for these types of data. Most recently, we are also extending our methods to non-coding transcript analysis to uncover their potential impact on human disease.

The *Vassiliou group* study haematological cancers with a particular focus on myeloid malignancies, including acute myeloid leukaemia (AML) and the myelodysplastic syndromes (MDS). Advances in cancer genomics have highlighted the importance of splicing gene mutations in the development of MDS and AML[2,10-13] and we recently described the first knock-in mouse model of *SF3B1* K700E, the most common mutation in human MDS, and described its global impact on alternative splicing, which recapitulated the abnormalities seen in human cancers with *SF3B1* mutations[14]. Our interest in splicing increased further in the last few months upon identifying, through a CRISPR-Cas9 genome-wide, the kinase SRPK1 as a novel therapeutic vulnerability of AML cells. SRPK1 phosphorylates the spliceosome protein SRSF1 at serine-arginine (SR) dinucleotides and to understand the basis of our findings, we analysed the impact of SRPK1 inhibition on genome-wide alternative splicing, using

industry standard tools DEXSeq[15] and MATS[16]. Our analysis gave us useful insights into the global splicing effects of SRPK1 inhibition, highlighting a number of transcripts whose splicing was altered by both genetic and pharmacological disruption of the kinase. Amongst these genes was BRD4, whose role as a therapeutic target in AML is well established. However, whilst the effect of SRPK1 inhibition on BRD4 splicing was statistically significant and affected many exons of the gene, it appeared very modest (Figure 1A). Nevertheless, because of the pedigree of this gene we decided to test the impact of the altered splicing on BRD4 protein. To our surprise this was very dramatic, showing an almost complete switch from the short to the long isoform (Figure 1B). As the short isoform is the one involved in transcriptional activation in AML, this change is likely to be an important mediator of the effects of SRPK1 inhibition. The fact that a number of other splicing changes are seen affecting the same transcript suggest that these may also have an effect at the protein level, potentially by coordinate effects at multiple exons within the same transcript. This can only be determined by long read sequencing and analysis

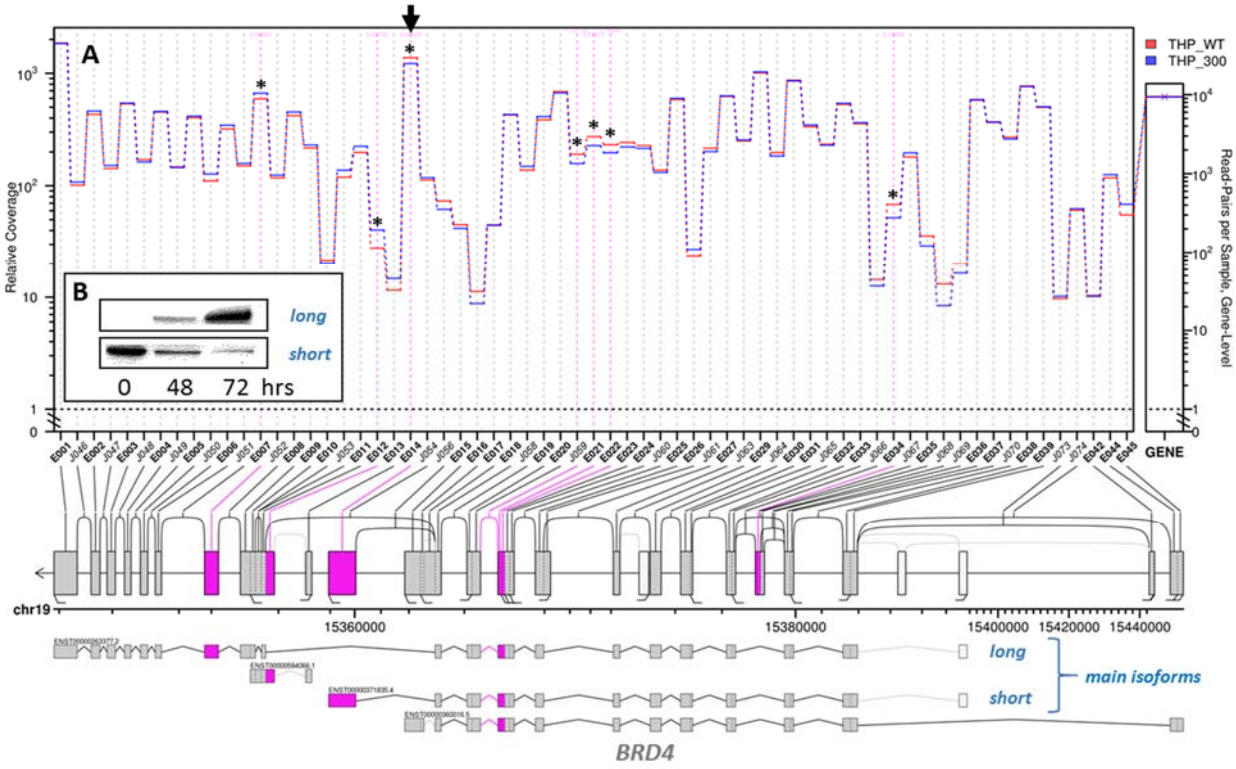

Figure 1. The impact of SRPK1 inhibition on *BRD4* mRNA splicing

A. Splicing changes at the *BRD4* locus from whole transcriptome RNAseq analysis, showing normalised read counts per exon in the absence (red) and presence (blue, 300nM, 48 hours) of the SPRK1 inhibitor SPHINX31. Exons whose splicing is significantly altered (FDR 0.2, genome-wide) are highlighted (*). The splicing event at the last exon of the short BRD4 isoform is also indicated (arrow). B (inset). Western blot with an antibody that detects both BRD4 protein isoforms (long and short). In contrast to the apparently modest effects of SRPK1 inhibition on mRNA splicing, this shows a striking protein level switch from the short to the long isoform.

Proposed work

We propose to use long sequencing reads to understand the co-regulation of distant exons of the same transcript in order to improve annotation in normal cells and then to study the impact of splicing changes in disease and I response to genetic or chemical/therapeutic perturbations. We plan to generate RNA-seq data using long sequencing reads including 300bp paired-end reads using Illumina MiSeq and 5-10kb reads using PacBio Iso-Seq (http://www.pacb.com/blog/intro-to-iso-seq-

method-full-leng/) technology. We will then adapt existing analytical tools developed by us and others to understand the co-regulation/co-dependence of distant splicing events genome-wide. The current RNA-seq pipelines at EBI are aimed at short read analysis, however reconstruction of full-length transcripts from short reads is challenging. To adapt this pipeline for long reads, we will investigate the most appropriate available sequence mapping algorithms, which take into account the long transcript sequencing error models. For transcript quantification we will explore using Bayesian methods to combine long and short reads, for improving the coverage. Finally, we will look into adapting existing differential expression methods to distinguish between splice variants in two conditions.

In order to address the above mentioned shortcomings of the present state of understanding the splicing landscape we intend to employ longer RNA-seq reads from MiSeq and Iso-Seq sequencers, design suitable analytical and statistical methodology to accurately quantify splice junctions/variants and comprehensively annotate the transcriptome. Using these tools we intend to investigate i) the role of altered splicing in the pathogenesis of myeloid leukaemias and ii) the impact of genetic or pharmacological perturbations on the transcriptome of normal and leukaemic cells, in order to improve our understanding of the pathogenesis of myeloid leukaemias and aid the development of rational therapeutic strategies against these diseases.

References

1 Scotti, M. M. & Swanson, M. S. RNA mis-splicing in disease. *Nat Rev Genet* **17**, 19-32, doi:10.1038/nrg.2015.3 (2016).

2 Papaemmanuil, E. *et al.* Somatic SF3B1 mutation in myelodysplasia with ring sideroblasts. *N Engl J Med* **365**, 1384-1395, doi:10.1056/NEJMoa1103283 (2011).

3 Lee, S. C. & Abdel-Wahab, O. Therapeutic targeting of splicing in cancer. *Nat Med* **22**, 976-986, doi:10.1038/nm.4165 (2016).

4 Gammons, M. V. *et al.* Topical antiangiogenic SRPK1 inhibitors reduce choroidal neovascularization in rodent models of exudative AMD. *Invest Ophthalmol Vis Sci* **54**, 6052-6062, doi:10.1167/iovs.13-12422 (2013).

5 Fonseca, N. A., Rung, J., Brazma, A. & Marioni, J. C. Tools for mapping high-throughput sequencing data. *Bioinformatics* **28**, 3169-3177, doi:10.1093/bioinformatics/bts605 (2012).

6 Gonzalez-Porta, M., Frankish, A., Rung, J., Harrow, J. & Brazma, A. Transcriptome analysis of human tissues and cell lines reveals one dominant transcript per gene. *Genome Biol* **14**, R70, doi:10.1186/gb-2013-14-7-r70 (2013).

7 Scelo, G. *et al.* Variation in genomic landscape of clear cell renal cell carcinoma across Europe. *Nat Commun* **5**, 5135, doi:10.1038/ncomms6135 (2014).

8 Wright, J. C. *et al.* Improving GENCODE reference gene annotation using a high-stringency proteogenomics workflow. *Nat Commun* **7**, 11778, doi:10.1038/ncomms11778 (2016).

9 Petryszak, R. *et al.* Expression Atlas update--an integrated database of gene and protein expression in humans, animals and plants. *Nucleic Acids Res* **44**, D746-752, doi:10.1093/nar/gkv1045 (2016).

10 Haferlach, T. *et al.* Landscape of genetic lesions in 944 patients with myelodysplastic syndromes. *Leukemia* **28**, 241-247, doi:10.1038/leu.2013.336 (2014).

11 Papaemmanuil, E. *et al.* Clinical and biological implications of driver mutations in myelodysplastic syndromes. *Blood* **122**, 3616-3627; quiz 3699, doi:10.1182/blood-2013-08-518886 (2013).

12 Sperling, A. S., Gibson, C. J. & Ebert, B. L. The genetics of myelodysplastic syndrome: from clonal haematopoiesis to secondary leukaemia. *Nat Rev Cancer* **17**, 5-19, doi:10.1038/nrc.2016.112 (2017).

13 Cancer Genome Atlas Research, N. Genomic and epigenomic landscapes of adult de novo acute myeloid leukemia. *N Engl J Med* **368**, 2059-2074, doi:10.1056/NEJMoa1301689 (2013).

14 Mupo, A. *et al.* Hemopoietic-specific Sf3b1-K700E knock-in mice display the splicing defect seen in human MDS but develop anemia without ring sideroblasts. *Leukemia* **31**, 720-727, doi:10.1038/leu.2016.251 (2017).

15 Anders, S., Reyes, A. & Huber, W. Detecting differential usage of exons from RNA-seq data. *Genome Res* **22**, 2008-2017, doi:10.1101/gr.133744.111 (2012).

16 Shen, S. *et al.* MATS: a Bayesian framework for flexible detection of differential alternative splicing from RNA-Seq data. *Nucleic Acids Res* **40**, e61, doi:10.1093/nar/gkr1291 (2012).