Breast Cancer Risk after Diagnostic Gene Sequencing (BRIDGES)

EU – H2020 project grant agreement no 634935

Website: bridges-research.eu

Deliverable Number	4.2
Deliverable Title	A catalogue of naturally occurring alternative splicing
Short Title	Alternative Splicing
Lead beneficiary	SERMAS
Del. Date (Annex 1)	01/09/2017
Achieved Date	15/08/2017
Nature	Demonstrator
Dissemination Level	PU
Document Filename	D42-AlternativeSplicing-PU-v1.0.pdf

Date	Authors/Reviewers	Remarks	Version
02/08/2017	De la Hoya/		0.1
31/08/2017	Vreeswijk		0.2
31/08/2017	Thomassen	Final Version submitted to EC	1.0

Contents

1. Executive Summary

One of the objectives of WP4 is to provide a comprehensive description of naturally occurring alternative splicing in BRIDGES genes that are (or are expected to be) clinically relevant, i.e. genes that are expected to be incorporated in routine diagnosis gene panels.

As already explained in the first periodic report, the selection of target genes for splicing analysis has evolve with the project and, based in part on expert opinion by member of WP3, we now have produced a catalogue of naturally occurring alternative splicing events in 11 BRIDGES genes (*ATM, BARD1, BRIP1, CHEK2, FAM175A, MRE11, PALB2, RAD51C, RAD51D*, *RECQL*, and *XRCC2*).

Deliverable 4.2 has been achieved in time.

This is a descriptive deliverable. Because of confidentiality issues, details have been omitted in this report. As soon as data are published this information will become publically available.

2. Introduction and Overview

The existence of alternative splicing in predisposition genes tends to be overlooked in the genetic diagnosis setting. However, our experience with the breast cancer predisposition genes *BRCA1* and *BRCA2* indicates that a comprehensive description of alternative splicing (including, if any, the characterization of tissue specific splicing) might be relevant for an accurate clinical classification of genetic variants in tumour suppressor genes. More specifically, alternative splicing might be particularly relevant in at least four related but independent aspects:

2.1 Clinical classification of intronic variants at dinucleotide splicing consensus sites: deleterious or uncertain significance? In absence of *in vitro* splicing data, research studies assume that genetic variants in cancer predisposition genes targeting splicing consensus sites (IVS±1,2) are *loss-of-function* variants equivalent to nonsense or frame-shift Indels (Easton et al, 2015). However, this is an oversimplification useful in research studies, but potentially harmful if translated directly to the clinic. Although in general, IVS±1,2 variants will affect splicing, the outcome of the splicing alteration might be in-frame, and thus of uncertain significance or likely non-pathogenic (if functional domains are not targeted). Thus, naturally occurring in-frame alternative splicing events, in particular small in-frame donor and acceptor shifts (but also cassette events in which the whole exon is deleted) might predict a likely non-pathogenic outcome for certain IVS±1,2 variants. For that reason, the ENIGMA BRCA1/2 Variant Classification Criteria (https://enigmaconsortium.org/library/general-documents/) recommend considering IVS±1,2 variants in certain donor/acceptor sites of *BRCA1* and BRCA2 (in absence of in vitro splicing studies) as variants of uncertain significance (for instance, variants at the

acceptor site of *BRCA1* exon 8, because the naturally occurring alternative splicing isoform *BRCA1* $\Delta 8p_3$ predicts a minor in-frame effect for these variants, a 3 nucleotides deletion).

2.2 Clinical classification of truncating variants at alternatively spliced exons: Loss-of-function variants? In general, truncating variants (frameshifts, Indels and nonsense variants) in cancer predisposition genes are assumed to be loss-of-function variants, and hence clinically relevant (Easton et al, 2015). Yet, it is conceivable that this is not always the case, in particular if truncating variants are located in non-constitutive exons, and the alternatively spliced isoform not containing the exon (and hence not containing the truncating variant) codifies for a fully functional protein. A recent report suggests that this might be the case for truncating variants in *BRCA1* exons 9 and 10, since the naturally occurring *BRCA1* splicing isoform $\Delta 9,10$ (representing approx. 20% of the overall *BRCA1* expression) is (partial) functional (de la Hoya et al, 2016).

2.3 Designing the right *in vitro* splicing assays: We anticipate that with the incorporation of BRIDGES genes to multi-gene panels used in the clinic, the identification of likely spliceogenic variants, and thus the demand for *in vitro* splicing assays performed with diagnosis purposes (RT-PCR assays in carriers and/or minigene assays) will increase enormously. To be reliable, the design of these assays should take into consideration the possible existence of alternative splicing in the vicinity of the variant under scrutiny, in particular multi-cassette events, as demonstrated by the *in vitro* splicing analysis of *BRCA1* c.591C>T (a variant located at the end of exon 9), which only renders meaningful results if the alternative splicing isoform *BRCA1* $\Delta 9,10$ is incorporated into the analysis (Dosil et al, 2010).

2.4 RT-PCR splicing assays in blood samples? In routine diagnosis, RT-PCR assays performed in carriers use blood as a source of RNA. Yet, if alternative splicing is relevant for the design and clinical interpretation of these assays, it is obvious that the existence of tissue specific splicing will question blood-based analyses.

In short, by providing a comprehensive catalogue of high confident alternative splicing events in 11 BRIDGES genes, we expect to provide information useful for the clinical classification of genetic variants within the BRIDGS project, but also in the diagnosis setting.

3. Science and Technology

To produce a catalogue of naturally occurring alternative splicing events in clinically relevant BRIDGES genes, we have analyzed RNAs from LCLs, PBMCs, healthy breast, ovary and fimbria tissues, by combining RNAseq with capillary electrophoreses (CEP) analysis of RT-PCR products (RT-PCR/CE). In the BRIDGES proposal, the original plan was to characterize alternative splicing exclusively by RT-PCR/CE, not with RNAseq. Yet, thanks to collaboration with other partners in the BRIDGES and ENIGMA consortiums, we have been able to use both approaches, discovering that those techniques complement each other (see 3.4).

3.1 RNA-seq (LCLs): The raw data used in BRIDGES (STAR and TopHat output files for all 11 BRIDGES genes) has been kindly provided by Dr. Logan Walker, an ENIGMA collaborator from Otago University, Christchurch, New Zealand. In brief, whole transcriptome sequencing analysis was undertaken on a lymphoblastoid cell line derived from a healthy female, cultured with and without cycloheximide treatment. cDNA libraries were constructed from total RNA using a TruSeq Stranded total RNA sample preparation kit (Illumina, San Diego, CA) following manufacturer's instructions, and sequenced on a HiSeq2000 (Illumina). Reads were mapped to the Homo_sapiens.GRCh37.72 reference genome, downloaded from Ensembl, using the two pass approach of the STAR (Spliced Transcripts Alignment to a Reference) aligner (Dobin et al, 2013) using the default settings, except maximum intron length was set to 100,000. Reads were mapped also using TopHat2 (Kim et al, 2013).

3.2 RNAseq (Normal breast an fimbria tissue): The raw data used in BRIDGES (STAR and TopHat output files for *ATM, BARD1, BRIP1, CHEK2, PALB2, RAD51*C and *RAD51D*) has been kindly provided by Dr. Anders Kvist, a BRIDGES collaborator from the University of Lund, Sweden (P4). In brief, whole transcriptome sequencing analysis was undertaken in nine normal breast tissue samples from nine women with breast cancer, and two normal fimbriae tissue samples from prophylactic oophorectomies. Methods for tissue sampling, preservation, RNA extraction, sequencing library preparation and sequencing of the breast tissue samples are described in Saal et al 2015. Libraries were paired-end sequenced on an Illumina HiSeq 2000(2x50bp, 2 samples) or NextSeq 500 sequencer (2x75bp, 9 samples).

Reads were aligned using STAR (version 2.4.1d) to the human genome reference GRCh37.The genome index was created by using reference genome and annotation from Ensemble, distributed by iGenomes. The only adjusted parameter for index construction was sjdbOverhang 91 (Dobin et al., 2013). Other explicitly adjusted parameters used with STAR include --outFilterMultimapNmax 2, --outFilterMismatchNmax 20 and --chimSegmentMin 0.

3.3 RT-PCR/CE (LCLs, PBMCs, Normal Breast and Ovary tissues) Multiple combinations of forward and reverse primers located at exonic regions (as defined by the reference transcripts, see Table 1) were used to amplify cDNAs. In total, we have designed and validated 94 ATM, 16 BARD1, 32 BRIP1, 16 CHEK2, 4 FAM175A, 19 MRE11, 22PALB2, 15 RAD51C, 12 RAD51D, 16 RECQL, and 5 XRCC2 RT-PCR primers. Different combinations of forward and reverse primers were used to produce overlapping RT-PCRs scanning the whole reference coding sequence of the 11 target genes. The RT-PCR products were analyzed by capillary EP in an ABI 3130 genetic analyzer (Applied Biosystems), using LIZ-500/600 as internal size markers. Size calling was performed with GeneMapper v4.0 software (applied biosystems). Capillary EP peaks were annotated as described previously (Colombo et al, 2014). An example is shown in figure 1.

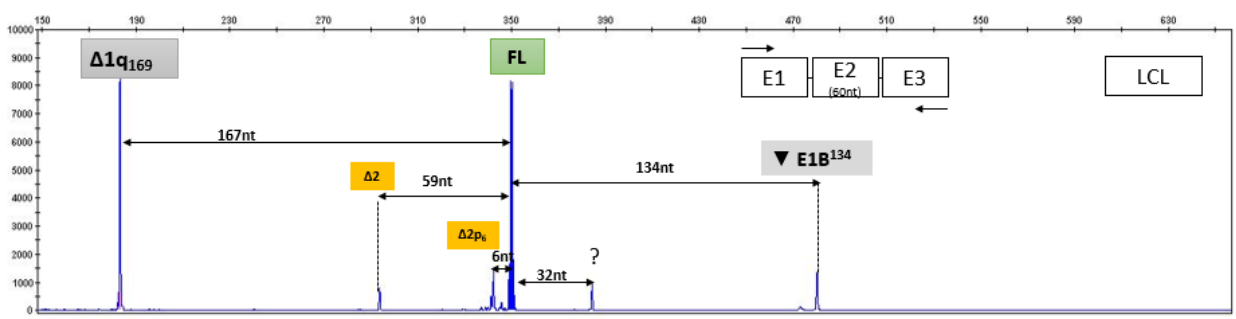

Figure 1. Example of a Gene X RT-PCR analyzed by capillary EP. Up to 6 different RT-PCR products are detected. The inferred annotation of alternative splicing events is based on size differences with the full-length (FL) product compatibles with the use of alternative GT/AG splicing sites. Splicing event supported (or not supported) by RNAseq are highlighted in gray (orange).

3.4 Merging RNAseq and RT-PCR/capillary EP data: Splicing events annotated from STAR and TopHat output files were merged with annotations inferred from capillary EP experiments. Splicing events supported by both approaches are considered *high confident events*. We have identified up to 372 high-confident splicing events in 11 BRIDGES genes (Table 1). Of these, 290 (78%) are novel events not described in GENCODE. Of particular relevance are 115 high confident in-frame events, as they may have a relevant role in the clinical classification of genetic variants.

Splicing events supported only by RNAseq or capillary EP are considered *low confident splicing events*. In total, we have identified 282 low confident events (238 of them supported only by RNAseq data, the remaining 44 supported only by capillary EP data). Most likely, some of these represent true alternative splicing events expressed at very low levels, while others certainly represent artefacts introduced by aligners or RT-PCR experiments.

We have observed that in some cases, STAR and TopHat aligners showed discrepancies. In these cases, capillary EP data is particularly useful to produce a list of high confident alternative splicing events. Apart from fixing discrepancies between STAR and TopHat aligners, the combination of RNAseq and capillary EP data is particularly useful to characterize novel exons (data not shown).

4. Summary

For the purpose of this study, we define alternative splicing event as those incorporating splice junctions not present in the corresponding reference transcript (Table 1). We have identified up to 372 high-confident splicing events in 11 BRIDGES genes (Table 1). Of these, 290 (78%) are novel events not described in GENCODE. Of particular relevance are 115 high confident in-frame events, as they may have a relevant role in the clinical classification of genetic variants.

We have identified up to 30 high-confident in-frame donor/acceptor shifts that might be highly relevant to classify variants in canonical splice donor/acceptor sites (IVS±1,2).

We have identified a number of in-frame splicing isoforms that deserve further attention. If these isoforms codify for functional proteins (analysis will be performed by P1 in this WP), it is possible that truncating variants located in exons that are absent in functional isoforms are not equivalent to truncating variants in other exons in terms of risk. Furthermore, in-frame isoforms representing ≈10% of the overall gene expression level have been detected in other genes.

We have identified up to 57 high-confident multi-cassette events that are highly relevant for the design of *in vitro* splicing assays as well as minigene analysis.

We have no evidence of tissue specific isoforms for any of the 11 BRIDGES genes investigated (i.e. all isoforms present in breast and/or ovarian tissue was also observed in LCLs and/or blood). We cannot discard the presence of subtle quantitative differences but overall, the data supports the clinical relevance of *in vitro* splicing assays in BRIDGES genes performed in blood samples.

Table 1. Naturally occurring alternative splicing events detected in 11 BRIDGES genes

Gene	Ensembl Ref transcript	Biotype[1]						Functional Annotation
		All (novel)[2]	Donor Shift	Acceptor Shift	Cassette	Multi-cassette	Other	In-frame (novel)[2]
ATM	ENST00000278616.8	75(66)	12	24	25	11	3	20 (19)
BARD1	ENST00000260947	28 (20)	5	3	12	7	1	6 (5)
BRIP1	ENST00000259008.6	17 (17)	2	3	8	4	0	7 (7)
CHEK2	ENST00000328354.10	48 (29)	4	4	24	11	5	13 (2)
FAM175A	ENST00000321945.11	13 (10)	0	1	10	2	0	3(3)
MRE11	ENST00000323929.7	40 (35)	2	5	28	1	4	17(15)
PALB2	ENST00000261584	17 (11)	2	4	6	4	1	7(5)
RAD51C	ENST00000337432	65 (54)	7	8	28	9	13	26 (23)
RAD51D	ENST00000345365	33 (20)	1	5	9	8	10	10 (5)
RECQL	ENST00000444129	30 (23)	3	6	12	0	9	4 (4)
XRCC2	ENST00000359321.1	6(5)	0	1	5	0	0	2(2)
Total		372 (290)	38	64	167	57	46	115 (90)

[1]According to Mudge et al, 2013. [2]Not described in GENCODE (last checked 31/07/2017) .

5. Conclusion

The Deliverable 4.2 has been achieved in time.

6. Literature references

de la Hoya M, Soukarieh O, López-Perolio et al. Combined genetic and splicing analysis of BRCA1 c.[594-2A>C; 641A>G] highlights the relevance of naturally occurring in-frame transcripts for developing disease gene variant classification algorithms. Hum Mol Genet. 2016; 25(11):2256-2268.

Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, Batut P, Chaisson M, Gingeras TR. STAR: ultrafast universal RNA-seq aligner. Bioinformatics. 2013; 29(1):15-21.

Dosil V, Tosar A, Cañadas C, Pérez-Segura P, Díaz-Rubio E, Caldés T, de la Hoya M. Alternative splicing and molecular characterization of splice site variants: BRCA1 c.591C>T as a case study. Clin Chem. 2010; 56(1):53-61.

Easton DF, Pharoah PD, Antoniou AC et al. Gene-panel sequencing and the prediction of breast-cancer risk. N Engl J Med. 2015; 372(23):2243-57.

Kim D, Pertea G, Trapnell C et al. TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. Genome Biol. 2013; 14(4):R36.

Mudge JM, Frankish A, Harrow J. Functional transcriptomics in the post-ENCODE era. Genome Res. 2013; 23(12):1961-73.

Saal, L. H. et al. The Sweden Cancerome Analysis Network - Breast (SCAN-B) Initiative: a large-scale multicenter infrastructure towards implementation of breast cancer genomic analyses in the clinical routine. Genome Medicine. 2015; 7, 20,