Non-coding mutation analysis reveals previously unrecognized pathways in lymph node-invasive breast cancer.

Short title: Splicing mutations in breast cancer

Dorman, S.N.[1], Viner, C.[2], Shirley, B.C.[3], Rogan, P.K[1,2,3].
[1]Department of Biochemistry and [2]Department of Computer Science, University of Western Ontario, London, ON, Canada, [3]Cytognomix Inc., London, ON, Canada.

Large-scale DNA sequencing studies have elucidated genomic landscapes of breast cancer (BC) tumours to identify driver mutations and dysregulated pathways contributing to tumourigenesis. Reported somatic mutations largely consist of protein coding mutations, whereas analyses of mRNA splicing mutations have been limited, even though these mutations are prevalent in many genetic disorders. We conducted an independent study of 445 matched normal and BC tumour exomes from The Cancer Genome Atlas Consortium (TCGA) using software for large-scale prediction and validation of novel splicing mutations. A variant caller, Strelka, was used to detect somatic mutations by comparing matched tumour-normal genotypes. The Shannon Human Splicing Pipeline predicted 5 206 splicing mutations, of which 1 130 are classified as cryptic splice site variants, 1 355 and 2 721 as inactivating or weakening natural sites, respectively. We predicted 385 (90%) of 429 splicing variants reported by the TCGA. Software was developed for high throughput validation of these variants using matched RNA sequencing data and controls to confirm expected consequences due to the variant's effect on the strength of the cognate splice site: cryptic splicing (extension or truncation of an exon), exon skipping or intron inclusion attributable to missplicing. Statistical significance was computed based on counts of the expected isoforms relative to their occurrence in non-variant containing samples. The splicing mutations were found to be in genes previously implicated in BC (*TP53, MLL3, CDH1, MAP3K1, PTEN, PIK3CA, GATA3 and RB1*) as well as novel significantly mutated genes identified by the TCGA (*CBFB, PIK3R1* and *NF1*). Pathway analysis of mutated genes based on splicing variants alone revealed overrepresentation of 29 pathways including 14 collagen, 4 extracellular matrix (ECM), and other pathways previously associated with BC such as the cell cycle or *ERBB2* signaling. Many, but not all, of these pathways were shown to be overrepresented by the TCGA. Inclusion of splicing mutations revealed enrichment in 8 *NCAM1* related pathways in samples with evidence of lymph node involvement, which was not observed in the lymph node-negative subset ($p < 0.05$). *NCAM* pathway-related mutations explain variability within significant components of the data, and are correlated with tumour stage and receptor status. Based on our findings, we hypothesize that NCAM1 and associated mutations contribute to tumour metastasis, whereas, overall, tumours are enriched for collagen/ECM mutations regardless of lymph node status. A non-negligible fraction of splicing variants are also predicted to overlap codons. We propose that comprehensive reporting of DNA sequencing data should consider both protein coding and non-trivial splicing analyses to avoid missing clinically-significant deleterious splicing mutations, which may contribute to novel metastasis-associated pathways.