Focusing on the most relevant gene variants in inherited breast and ovarian cancer

Modern sequencing technology (which has begun to become adopted in clinical labs) can provide invaluable genetic information of individuals, such as those with a family history of early onset breast cancer. However, this information also results in an overwhelming number of new and uncommon gene variants with an unknown impact on the disease. Known loss-of-function variants, which are most likely deleterious, represent only a minor proportion of identified variants. The majority of the variants will change the protein code (missense), silent, or found in regions outside of the protein coding regions of the gene. Most will have little or no significant impact on protein function. This can result in inconclusive or uncertain genetic testing results.

To reduce the number of variants found in patients to a manageable amount for functional analysis, we propose a framework to prioritize those most likely to be disease causing. While mutations which lead to a change in protein coding are commonly studied, variants in non-coding regions which can affect transcriptional activation of a gene (turning on the expression of an mRNA transcript - which is a copy of the gene sequence), mRNA splicing (a process of converting the transcribed copy of RNA to its mature form), and mRNA stability (ensuring the longevity of the transcript in the cell) are often underreported. Recognition of DNA or RNA sequences by specific proteins can be affected by gene variants. An altered binding site needed for proper mRNA splicing can alter the encoding of the protein and its function. Abolition of a DNA sequence recognized by transcription factors will reduce the level of expression of the mRNA.

Our study used information theory, a mathematical theory for communication that can be applied to study shared genetic patterns, to evaluate the strengths of binding sites in DNA or RNA that are recognized by proteins. We previously found that this approach is robust and accurate at predicting variant-directed changes in mRNA splicing. Here, we sequenced and evaluate coding and non-coding variants for 7 genes known to harbour mutations that increase breast cancer risk (*ATM*, *BRCA1*, *BRCA2*, *CDH1*, *CHEK2*, *PALB2* and *TP53*) using an information theory-based framework to prioritize potentially disease-causing variants as candidates for further study.

Using microarray technology, we captured all non-repetitive regions of the 7 breast cancer-related genes for DNA sequencing of 102 patients with a high-risk for breast cancer. The variants identified from sequencing were further evaluated using information theory-based methods. Variants found in the untranslated regions of the gene were further evaluated for potential changes in RNA structure using SNPfold, and variants flagged by this method were evaluated further by measuring changes in RNA structure. Variants found in coding regions were evaluated with tools commonly used to predict the effect of a variant on protein function. Furthermore, a new approach was developed to identify potential large deletions by tracking these variants within gene sequences.

Among these patients, we found 15,311 unique variants, however only a small minority (245)

occurred in translated regions. With the unified information framework, we prioritized only 87 variants that were potentially functionally significant, a manageable number for further analysis. The relevance of several of these variants to breast cancer was supported by previous independent studies. We also identified 7 changes which would have major effects on protein translation, as well as one large interval which could be a potential deletion. Structural analysis of the mRNAs flagged 5 previously identified variants, including a *TP53* variant which caused a change in RNA structure that we confirmed experimentally.

Our sequencing approach highlights the importance of sequencing non-coding regions when potentially disease causing mutations are not evident by conventional sequencing of coding regions. Through this information theory-based framework (alongside coding sequence analysis), variants of unknown significance can be prioritized by predicted function and tested accordingly, and can be used as an intermediate bridge between sequencing and variant classification.

Eliseos J. Mucaki, Peter K. Rogan
Department of Biochemistry, Schulich School of Medicine and Dentistry, University of Western Ontario, Canada

Publication

A unified analytic framework for prioritization of non-coding variants of uncertain significance in heritable breast and ovarian cancer.
Mucaki EJ, Caminsky NG, Perri AM, Lu R, Laederach A, Halvorsen M, Knoll JH, Rogan PK
BMC Med Genomics. 2016 Apr 11