A bioinformatics pipeline for analysing germline mutations in human breast cancer by exome sequencing

I.V. Bizin
Department of Bioinformatics, Peter the Great St. Petersburg Polytechnic University, St. Petersburg, Russia,
bizin@yandex.ru

A.P. Sokolenko, E.Sh. Kuligina, E.N. Imyanitov
Department of Tumor Biology, N.N. Petrov Institute of Oncology, St. Petersburg, Russia,
annasokolenko@mail.ru

D. Frishman
Department of Bioinformatics, Technische Universität München, Wissenschaftszentrum Weihenstephan, Freising, Germany,
Institute for Bioinformatics and Systems Biology, HMGU German Research Center for Environmental Health, Neuherberg, Germany,
Department of Bioinformatics, Peter the Great St. Petersburg Polytechnic University, St. Petersburg, Russia,
d.frishman@wzw.tum.de

The genetic basis of familial breast cancer remains poorly understood. Less then a half of cases are linked to germline mutations in known major hereditary breast cancer genes. Recent research demonstrated the need for different therapy approaches for familial and sporadic tumors. It is therefore critically important to identify a possibly complete set of genes causally related to familial breast cancers.

We have developed a comprehensive pipeline for detecting germline mutations in variants identified by exome sequencing. The main challenge in using exome sequencing to find novel disease genes is to distinguish disease-related alleles from the background of non-pathogenic polymorphisms and sequencing errors [1]. To address this problem our pipeline utilizes a three-step procedure: i) a search for recurrent variants found in several familial breast cancer samples, while absent in control samples, ii) collection of minor allele frequencies (MAF) from publicly available variation databases, such as dbSNP and NHLBI Exome Sequencing Project (ESP), and iii) scoring the deleteriousness of single nucleotide variants by the Combined Annotation Dependent Depletion (CADD) method [3]. We defined deleterious mutations as those introducing frame-shifts or in-frame stop codons, as well as those missense variants with a CADD-score of greater or equal 20, thus selecting the top 0.1% of the most damaging variants. We primarily sought to identify novel deleterious mutations with

MAF lower than one percent in those genes that do not harbor deleterious variants with MAF greater than one percent (i.e. common polymorphisms). For each gene we determined the number of those common variants present in variation databases that we defined as deleterious based on our scoring procedure. Newly found deleterious mutations in the genes that are generally less susceptible to disease variation are presumed to be more interesting targets for experimental verification and are thus given higher priority by our pipeline for additional experimentally verification.

We applied the computational pipeline described above to process exome sequencing data on familial breast cancer and control samples obtained in the Department of Tumor Biology at Petrov Institute of Oncology. Variants from 42 familial breast cancer samples and 10 control samples were analyzed. Out of 762980 variants found in all samples the pipeline highlighted 100 mutations, with six of them occurring in genes associated with breast cancer [4]. Some of these novel variants are currently being experimentally verified by Sanger sequencing.

1. M.J. Bamshad, S.B. Ng, A.W. Bigham, H.K. Tabor, M.J. Emond, D.A. Nickerson, J. Shendure (2011) Exome sequencing as a tool for mendelian disease gene discovery, *Nature Reviews Genetics*, **12:**745–755.

2. G.M. Cooper, D.L. Goode, S.B. Ng, A.B. Sidow, M.J. Bamshad, J. Shendure, D.A. Nickerson, A. Deborah (2010) Single-nucleotide evolutionary constraint scores highlight disease-causing mutations, *Nature methods*, **7:**250–251.

3. M. Kircher, D.M. Witten, P. Jain, B.J. O’Roak, G.M. Cooper, J. Shendure (2014) A general framework for estimating the relative pathogenicity of human genetic variants, *Nature genetics*, **46:**310–315.

4. A.P. Sokolenko, E.V. Preobrazhenskaya, S.N. Aleksakhina, A.G. Iyevleva, N.V. Mitiushkina, O.A. Zaitseva, O.S. Yatsuk, V.I. Tiurin, T.N. Strelkova, A.V. Togo (2015) Candidate gene analysis of BRCA1/2 mutation-negative high-risk Russian breast cancer patients, Cancer Letters, **42:**259–261.