Հայաստանի Գիտությունների Ազգային Ակադեմիա
Национальная Академия Наук Армении
National Academy of Sciences of Armenia

Հայաստանի Կենսաբանական Հանդես
Биологический Журнал Армении
Biological Journal of Armenia

• Փորձարարական և տեսական հոդվածներ •Экспериментальные и теоретические статьи• •Experimental and theoretical articles•

Biolog. Journal of Armenia, 4 (68), 2016

THE UTILITY OF SUFFICIENTLY INFORMATIVE PROTEIN MULTIPLE SEQUENCE ALIGNMENT FOR CANCER RISK ASSESSMENT OF BRCA2 MISSENSE SUBSTITUTIONS

D.T. BABIKYAN

Center of Medical Genetics and Primary Health Care, Yerevan, Armenia
davidbio@yahoo.com

Although the breast cancer risk, that is specific to BRCA1 and BRCA2 genes, is similar one to another, the analysis of missense substitutions has focused mainly on BRCA1. However, more than 40% of BRCA2 full-sequence tests in which a reportable DNA sequence variant is found present unclassified missense variants. The results of this study present amino acid sequence conservation across BRCA2 protein using protein multiple sequence alignment of 12 BRCA2 orthologs capturing enough evolutionary time to predict the role of unclassified missense variants of BRCA2 gene on BRCA2 protein function by means of bioinformatics analysis.

Missense substitutions – BRCA2 gene – protein multiple sequence alignment

Չնայած BRCA1 և BRCA2 գեներին բնորոշ է կրծքագեղձի քաղցկեղի զարգացման նույնանման գենետիկական ռիսկ, հիմնականում ուսումնասիրվել են BRCA1 գենի միսսենս տեղակայումները: Այնուամենայնիվ, գենետիկական թեստերի միջոցով հայտնաբերված BRCA2 գենի ամբողջ հաջորդականության փոփոխությունների 40% իրենցից ներկայացնում են չդասակարգված միսսենս տարբերակներ: Այս հետազոտության արդյունքները ցույց են տալիս BRCA2 սպիտակուցի ամինաթթվային հաջորդականության կոնսերվացիան՝ 12 BRCA2 օրթոլոգ սպիտակուցների բազմակի հաջորդականությունների համադրման կիրառմամբ: Վերջինը ընդգրկում է բավարար էվոլյուցիոն ժամանակահատված՝ BRCA2 սպիտակուցի ֆունկցիայի վրա BRCA2 գենում չդասակարգված միսսենս տարբերակների դերի կանխորոշման համար՝ կենսաինֆորմատիկ վերլուծության միջոցով:

Միսսենս տեղակալումներ – BRCA2 գեն – սպիտակուցների բազմակի հաջորդականությունների համադրում

Мутации генов BRCA1 и BRCA2 способствуют повышенному риску развития рака молочной железы. Поиск миссенс-мутаций был сосредоточен главным образом на гене BRCA1. Тем не менее, при секвенировании последовательностей гена BRCA2 выявлено более 40% неклассифицированных миссенс-вариантов. В данном исследовании использованы методы выравнивания последовательностей 12 ортологов белка BRCA2 с использованием биоинформатического метода. Полученные результаты свидетельствуют о стабильности аминокислотной последовательности белка BRCA2 на протяжении длительного эволюционного периода, достаточного для прогнозирования влияния неклассифицированных миссенс-вариантов гена BRCA2 на функции белка BRCA2.

Миссенс-замены – ген BRCA2 – множественные выравнивания белковых последовательностей

Since identification of BRCA1 and BRCA genes, extensive research has been carried out to elucidate their function in breast and ovarian cancer. BRCA1 and BRCA2 are both tumor suppressor genes involved in the signaling and repair of DNA. To date, more than 1000 and 500 deleterious mutations have been found in BRCA1 and BRCA2, respectively. Mutations in either gene lead to a compromise of DNA repair [14], resulting in DNA replication errors. In the current working model, this extra mutational load ultimately leads to the development of breast or ovarian cancer [2].

Classifying missense variants is difficult because genetic testing is performed on small families or isolated individuals in the contrary to large pedigrees. This makes it difficult to use segregation analyses to classify. In addition, most of the unclassified substitutions are too rare for standard genetic and/or epidemiological analyses. As opposed to the linkage analysis, the sequence analysis of BRCA2 from a diverse set of organisms allows identify positions within the alignments that are functionally constrained or not-constrained. Sequence analysis/bioinformatics is a frequency-independent way for the observed substitutions to be initially put in sets that are enriched for neutral substitutions or deleterious substitutions using quantitative measures of amino-acid physical-chemical similarity [1].

Thus, creation of "a sufficiently informative" protein multiple sequence alignment (PMSA) capturing a combined evolutionary time (cladogram path length) of at least 2,000 million years is a prerequisite for such an analysis.

Materials and methods. **Creation of BRCA2 PMSA:** The BRCA2 PMSA was constructed from existing full length sequences of 6 mammals (human, U43746; chimpanzee XP_509619; monkey XP_001118184; rat, AAB71378; dog, NP_001006654; cow, XP_583622), chicken (AAL89470), green spotted pufferfish (EF564374) and fugu pufferfish. Alignment was prepared using the M-Coffee server (http://www.tcoffee.org/Projects/mcoffee/index.html), which rather than computing a multiple sequence alignment on its own, it uses other packages to compute the alignments. Using Mlalign_id_pair, Mslow_pair, Mmuscle_msa, Mclustalw_msa and Mt_coffeee_msa methods, the server then used T-Coffee method to combine all these alignments into one unique final alignment. The initial PMSA was then re-aligned using the Genetic Database Environment program (version MacGDE 2.4), resulting in the alignment used for the analyses presented here.

Assembling BRCA2 Sequence from the Genomic Sequences: Three new sequences - opossum (Monodelphis domestica), frog (Xenopus tropicalis) and sea urchin (Strongylocentrotus purpuratus) BRCA2 were also targeted for addition to the PMSA. For these, a combination of tBLASTn (The National Center for Biotechnology Information) and splice junction prediction was used to build initial gene models from the available genomic sequences. First, the ortholog protein of more diverged organism was subjected to BLAST and gene orthologs were predicted. Genes were referred to as unique if they do not satisfy the criteria of bi-directional best hits. Such a unique gene prediction of a given organism was then searched against the genome of the other model organism to exclude any missing prediction because of the limitation of the tBLASTn program. Only when a unique gene and no other match were detected then the gene ortholog was considered as unique and was taken into the PMSA.

Molecular genetic confirmation of low confidence segments: The parts of three gene models with low rate of conservation were PCR-amplified from initially synthesized cDNAs and sequenced by using the same primer sets and Dye-terminator sequencing reactions (BigDye Terminator, version3.1, LifeTechnologies). Sequencing reaction products were then run on a 96-capillary Spectrumedix Sequencer (Transgenomics) in accordance with the manufacturer's recommendations.

Analysis of sequence invariance in the alignment: Three statistical models were considered for calculating, from the PMSA, the number of positions in the alignment where human BRCA2 has a residue that remain invariant due to common ancestry and strong sequence conservation as against those that are invariant due to common ancestry even though the position is probably under weak functional constraint. In the Model 1, probability of missense substitutions

does not depend on position in the sequence. In the Model 2, the alignment is treated as if it consisted of two groups positions – invariant and variable positions. In the Model 3, the alignment is treated as if it consisted of two groups positions – slowly and rapidly substituting positions. All three models follow and the expected number of invariant positions for all three models was calculated as explained in the Fitch's development of the covarion hypothesis [3, 5, 6].

Results and Discussion. The contribution and significance of germline BRCA1&2 mutations to breast cancer, both in high-risk patients and in the general population, is still incomplete and remains unclear. In cases of clearly deleterious mutations (frameshift, nonsense and splice junction mutations, large gene rearrangements), the test results are of precious importance for subsequent clinical implication. However, missense substitutions observed in these genes are a difficult challenge as their classification as neutral variants or deleterious mutations presents a serious obstacle for medical genetics. The most successful approaches used to date combine cross-species sequence comparison of the gene of interest with analysis of the physical characteristics of wild-type and mutant amino acids. The amino acid sequence of BRCA2 is poorly conserved across mammalian species. For example, the sequence identity of human BRCA2 with mouse Brca2 is 59% [13]. This is far below the average (>80% identity) for human/mouse gene comparisons. Only the C-teminal DNA binding domain (DBD) and BRC repeats in the central portion of BRCA2 are also somewhat well-conserved [4].

Therefore, the accurate assessment of unclassified and rare missense variations identified across the BRCA2 gene would require creation of a sufficiently informative PMSA of BRCA2 which is the central aim of this study.

Orthologs are genes in different species that evolved from a common ancestral gene by speciation. Normally, orthologs retain the same function in the course of evolution. Given the availability of complete genome sequences from related organisms, the BRCA2 PMSA was constructed from 6 mammalian (human, chimpanzee, monkey, rat, dog, and cow), chicken, green spotted and fugu pufferfish full length sequences. However, identification of new orthologs is critical for the construction of the PMSA of sufficient size for classification of missense variations. Therefore, the PMSA has to sample enough sequences at sufficient evolutionary remove from each other ("alignment depth") to contribute missense substitutions analysis with reasonable sensitivity and specificity.

For this reason, three model organisms – Monodelphis domestica, Xenopus tropicalis, and Sea urchin with recently available genome sequences were included in the developing BRCA2 PMSA. Opossums are the closest phylogenetic relatives of placental mammals. Marsupial genomes are approximately the same size as other mammalian genomes. The Monodelphis domestica is becoming an increasingly important model organism for comparative genomic research. X. tropicalis is an important model amphibian organism for the study of developmental genetics and vertebrate biology which genome contains some gene segments well conserved from human to frog and beyond. Sea urchin is a member of the Echinoids, invertebrate marine animals which are useful in evolutionary studies as non-chordate deuterostomes. Regions of synteny between sea urchins and mammals have been identified, and they are more closely related to humans and other chordates than flies or nematodes.

An approach of BLAST analysis with sequence conservation providing important clues for predicting gene structure was applied to achieve high quality gene prediction for these three newly-sequenced species. The predicted sequences are exon by exon predictions of BRCA2 gene which span only one sequence contig and are consequently predicted by a single gene model. Because of the high variability of BRCA2 orthologs, some predicted exons had wrong splice junctions or were missing from the predictions. Furthermore, it was not observed one-to-one correspondence of human BRCA2 with the

predicted orthologs. These problems were overcome by careful molecular-genetic analysis of cDNA of three model organisms and sequence confirmation of low confidence segments. The final mRNA and protein sequences of three BRCA2 orphologs are reported in the GenBank [7-12]. The PMSA was created by M-Coffee server and then was manually re-aligned by GDE software (fig. 1).

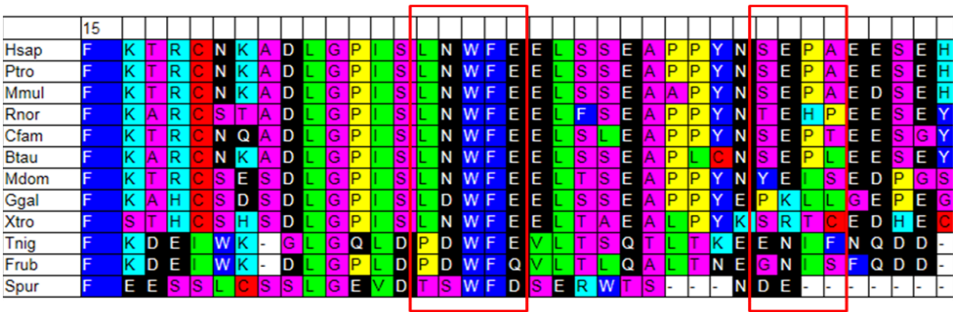

Fig. 1. A section of BRCA2 alignment from human residues 15-52 representing one well conserved (in the left) and one variable (in the right) segments (both in the red frames).

The BLAST and experimental analysis of three new orthologs approved that the genes actually encode all of the BRCA2 motifs documented in vertebrate BRCA2. The Sea urchin's BRCA2 is the most evolutionary distant BRCA2 gene currently known to be a good structured analog to its human counterpart.

The statistical likelihood that an invariant amino acid position is truly functionally constrained depends on the size of the evolutionary database used. The minimum number of substitutions was calculated by the Fitch-type 3 rate constant model of Fitch and Markowitz expressing as the number of substitutions per amino acid position, averaged across all positions in the gene [3,5]. Comparison of the observed and expected (based on the null hypothesis that variants should be uniformly distributed) numbers of invariant amino acids reports on the alignment depth to be informative. In general, if a PMSA contains >3.0 substitutions/position, then the probability that any amino acid will be invariant is <5%.

First, we calculated the well-conserved DNA binding domain (DBD domain) (human residues 2401–3110) segment of the alignment after the addition of the experimentally determined sequences from the opossum, frog, and sea urchin BRCA2 orthologs. Based on a Fitch-type 3 rate constant model, approximately 45% of positions in the concatenated DBD domain alignment were under functional constraint, and the alignment had a maximum likelihood estimate of 3.3 substitutions per position. Given the much greater cross-species sequence variability of the BRCA2 gene, the proportion of truly conserved positions were smaller, requiring more substitutions per position. On the other hand, if the BRCA2 PMSA contained only 9 vertebrate sequences, it had 124 observed invariant positions (about 3.6% of BRCA2's 3418 amino acids), compared with the expected number of 67.1. Thus, up to 50% of the invariant positions may still have been invariant by chance. To reach that criterion required addition of three more sequences (from opossum, frog and sea urchin), and pushed the total number of substitutions per position past 5.

Consequently, the carefully constructed and curated BRCA2 PMSA containing 11 vertebrate and 1 invertebrate BRCA2 sequences captures enough evolutionary time (about 2,500 million years), has about 5 substitutions per position, and >95% of invariant positions are because of functional constraint. Therefore, the BRCA2 PMSA fully satisfying the three criteria is considered sufficiently informative for further classification of missense substitutions by bioinformatics alignment-based tools.

REFERENCES

1. *Abkevich V., Zharkikh A., Deffenbaugh A.M., Frank D., Chen Y., Shattuck D., Skolnick M.H., Gutin A., Tavtigian S.V.* Analysis of missense variation in human BRCA1 in the context of interspecific sequence variation. J. Med. Genet., *41*, 7, 492-507, 2004.
2. *Antoniou A., Pharoah P.D., Narod S., Risch H.A., Eyfjord J.E., Hopper J.L., Loman N., Olsson H., Johannsson O., Borg A., Pasini B., Radice P., Manoukian S., Eccles D.M., Tang N., Olah E., Anton-Culver H., Warner E., Lubinski J., Gronwald J., Gorski B., Tulinius H., Thorlacius S., Eerola H., Nevanlinna H., Syrjäkoski K., Kallioniemi O.P., Thompson D., Evans C., Peto J., Lalloo F., Evans D.G., Easton D.F.* Average risks of breast and ovarian cancer associated with BRCA1 or BRCA2 mutations detected in case Series unselected for family history: a combined analysis of 22 studies. Am J Hum Genet., *72*, 5, 1117-30, 2003.
3. *Fitch W.M., Markowitz E.* An improved method for determining codon variability in a gene and its application to the rate of fixation of mutations in evolution. Biochem Genet. *4*, 579-93, 1970.
4. *Jasin M.* Homologous repair of DNA damage and tumorigenesis: the BRCA connection. Oncogene, *58*, 16, 21, 8981-93, 2002.
5. *Markowitz E.* Estimation and testing goodness-of-fit for some models of codon fixation variability. Biochem. Genet., *4*, 595-601, 1970.
6. *Miyamoto M.M., Fitch W.M.* Testing the covarion hypothesis of molecular evolution. Mol. Biol. Evol., *12*, 503-13, 1995.
7. *Tavtigian S.V., Babikyan D., Monnier S., Voegele C.* BRCA2 (Strongylocentrotus purpuratus), GenBank: ABP57025.2, 2007.
8. *Tavtigian S.V., Babikyan D., Monnier S., Voegele C.* Strongylocentrotus purpuratus BRCA2 (brca2) mRNA, partial coding sequence, GenBank: EF523433.2, 2007.
9. *Tavtigian S.V., Babikyan D., Voegele C.* BRCA2 (Xenopus (Silurana) tropicalis), GenBank: ABP48763.1, 2007.
10. *Tavtigian S.V., Babikyan D., Voegele C.* Xenopus tropicalis BRCA2 (brca2) mRNA, partial coding sequence, GenBank: EF508681.1, 2007.
11. *Tavtigian S.V., Babikyan D., Voegele C., Samollow P.B.* BRCA2 (Monodelphis domestica), GenBank: ABP48762.1, 2007.
12. *Tavtigian S.V., Babikyan D., Voegele C., Samollow P.B.* Monodelphis domestica BRCA2 (brca2) mRNA, partial coding sequence, GenBank: EF508680.1, 2007.
13. *Warren M., Smith A., Partridge N., Masabanda J., Griffin D., Ashworth A.* Structural analysis of the chicken BRCA2 gene facilitates identification of functional domains and disease causing mutations. Hum Mol Genet., *7*1, 11, 841-51, 2002.
14. Welcsh P.L., King M.C. BRCA1 and BRCA2 and the genetics of breast and ovarian cancer. Hum Mol Genet., *7*, 10, 705-13, 2001.

Received on 08.09.2016