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, 3 (69), 2017

SIMULATION OF BRCA1&2 CASE-CONTROL MUTATION SCREENING AS AN APPROACH TO CHARACTERIZE INTERMEDIATE-RISK SUSCEPTIBILITY GENES

D.T. BABIKYAN

Center of Medical Genetics and Primary Health Care,
Yerevan State Medical University after Mkhitar Heratsi
davidbio@yahoo.com

While high-risk susceptibility genes are traditionally found and/or analyzed by linkage analysis-related methods, the case-control association study-based methods are used to assess candidate-modest risk genes. At least a subset of these intermediate-risk genes can be assessed in a case-control mutation screening format using pooled evidence from a set of genetic variants that are intrinsically likely to alter gene function, comprised of a mix of truncating, splice junction, missense and regulatory variants selected via explicit analysis-based criteria. In the present study a systematic approach to assembling such a TSMR pool and use BRCA1 and BRCA2 mutation screening data in simulated tests of association to demonstrate the utility of the TSMR pool/case-control mutation screening strategy.

Case-control mutation screening – TSMR+ analysis

Եթե բարձր նախատրամադրվածության գեները սովորաբար բացահայտվում և ուսումնասիրվում են շղթայակցման վերլուծության մեթոդներով, ապա դեպք-ստուգիչ ասոցիացիոն հետազոտության վրա հիմնված մեթոդները կիրառվում են միջին ռիսկի գեների գնահատման նպատակով: Միջին ռիսկայնության գեների մի մասը կարող են հետազոտվել դեպք-ստուգիչ մուտացիոն սկրինինգի եղանակով, այդ նպատակով կիրառելով այդ գեներում բացահայտված և գենի ֆունկցիան իրենց բնույթով փոփոխող գենետիկական տարբերակների խումբը՝ բաղկացած կրճատող, սպլայս միացումների, միսսենս և կարգավորիչ տարբերակներից, որոնք ընտրվել են որոշակի վերլուծության վրա հիմնված չափանիշներով: Այս հետազոտության մեջ ներկայացվում է TSMR տարբերակների խմբի կանոնավոր հավաքման մոտեցումը և BRCA1 և BRCA2 գեների մուտացիոն սկրինինգի տվյալների կիրառումն ասոցացման մոդելավորված թեստերում՝ TSMR տարբերակների խմբի և դեպք-ստուգիչ մուտացիոն սկրինինգի ռազմավարությունը ցուցադրելու նպատակով:

Դեպք-ստուգիչ մուտացիոն սկրինինգ – TSMR+ վերլուծություն

Наследственная предрасположенность высокого риска определяется путем анализа сцеплений. Многие гены промежуточного риска можно оценивать методами мутационного скрининга системы случай-контроль. С этой целью исследуют группу генетических вариантов, способных изменить функцию гена, включая сокращенные, сплайсинговые, миссенс и регулирующие варианты, отобранные с помощью специальных критериев. В настоящем исследовании представлена целесообразность применения системного подхода выбора группы вариантов TSMR с использованием данных мутационного скрининга BRCA1 и BRCA2 в моделируемых тестах ассоциаций.

Случай-контроль мутационный скрининг – TSMR+ анализ

Evidence for a genetic component of risk for common cancers back at least to breast cancer (BC) pedigree studies complemented with linked genealogy, twin studies, and segregation analyses. However, only 25% of the genetic bases of BC can currently be attributed to specific genes. What genes, and what classes of sequence variants in those genes, are responsible for the as yet unexplained genetic risk of BC?

For the common cancers, any common high-risk variants would have been found long ago by linkage analysis and are not possible given constraints on evidence and observed familial risk. Uncommon high-risk variants, such as the Ashkenazi BRCA1 variant 185delAG, are sometimes found as founder mutations in specific populations. Linkage analyses followed by positional cloning led to the discovery of susceptibility genes BRCA1 and BRCA2 that harbor many rare, high-risk variants. Failure to identify any other equivalently informative susceptibility genes since 1996 has led some to argue that few genes harboring high-risk variants responsible for one or more of the common cancers remain to be identified. Analyses of risk attributable to high-risk genetic variants in the known BC susceptibility genes are not consistent with their being responsible for more that about 5% of the overall risk of BC.

On the other hand, the disequilibrium structure of the human genome and gene pool is such that there tend to be few common SNPs at any given locus. This feature dramatically reduces the number of markers required to carry out genome wide SNP association studies as well as the degree of multiple testing inherent in such studies. Linkage analysis followed by mutation screening and segregation analysis has provided a powerful tool for finding high-risk susceptibility genes. BRCA1 and BRCA2 are currently the best characterized susceptibility genes and individually map roughly between 1% and 3% of population (genetic) attributable risk (PAR). Large scale case-control genotyping studies should provide a similarly powerful tool for finding common susceptibility genes. But what of uncommon to rare intermediate-risk susceptibility genes and deleterious variants which in them lay in the gap between the strengths of common genetic epidemiology and molecular epidemiology study designs and cannot be analyzed by neither of them?

In the present study, a case-control mutation screening was simulated that has adequate power to address the challenge of genes that harbor uncommon or rare intermediate risk variants. The key to case-control mutation screening is to generate a pool of sequence variants, in a single candidate gene, that are intrinsically likely to alter gene function. From basic molecular biology considerations, there are 4 classes of sequence variants that need to be pooled: Truncating mutations, splice junction mutations, the subset of missense substitutions (MS) that are intrinsically likely to alter protein function, and when eventually possible the subset of Regulatory sequence variants that are genuinely likely to alter gene expression. A combination across these classes of variants would constitute a TSMR+ pool, where the "+" denotes the need for sequence analysis required to distinguish between genetic variants in each class that are intrinsically likely to alter function and those that are not. In the following study, a subset of mutation screening results of 68,000 subjects underwent full re-sequencing of BRCA1 and BRCA2 to illustrate of a TSMR+ pool and tests of association based upon that pool.

Materials and methods. **BRCA1 and BRCA2 mutation set:** BRCA1 and BRCA mutation screening data were taken from the Breast Cancer Information Core (BIC) database which functions as an open repository of sequence alterations in BRCA1 and BRCA2.

TSMR+ pooling criteria. Truncating mutations. For BRCA1 and BRCA2, any truncating mutation located at or before the last conserved residue of the last well conserved amino acid sequence element (5 of 10 amino acids having GV≤61.3) in the protein would be retained in the TSMR pool. Splice junction consensus sequences lying within exons were considered. The last

two nucleotides of each exon are part of the splice donor consensus, and their canonical sequence in AG. With a simple rule, only the following substitution were retained in the TSMR+ pool with predicted interference with splicing, if the reference sequence at a) -2 is A, b) -2 is G substituted with C or T, c) -1 is G, and d) -1 is A substituted with C or T. MS. In addition to calculated GV, the Grantham Deviation (GD) of each MS was calculated as a measure of the fit between MSs and the range of variation observed at their position in a PMSA. GV and GD were used to divide MS into several groups from most likely to least likely to alter protein function: invariant site MSs (MI) (GV=0); MSs with non-conservative physical characteristics (MA) falling at variable positions (GV>0 and GD>61.3+GV). Any MS at a position with GV=0 will be outside of the cross-species range of variation. GV=61.3 is the outer limit of conservative substitution. For this analysis, all other MSs that do not fit those two criteria were excluded.

Results and Discussion. **Simulation of a case-control association analysis:** Under the hypothesis that the breast/ovarian cancer risk for a BRCA1:BRCA2 double carrier is not dramatically higher than the risk for a simple BRCA1 or BRCA2 carrier, the appearance of a double carrier in the BIC set is largely explained by either deleterious variant alone. Hence, subjects who are clear BRCA2 mutation carriers are used as pseudo-controls for the analysis of BRCA1 mutation screening data, and vice versa. This underlying reasonable biological hypothesis follows from the observation that the two genes function in the same biochemical pathway and loss of function of the wild copy of BRCA1 or BRCA2 is unlikely to be either the initiating or the rate limiting step of tumorigenesis in mutation carriers [5].

Of the 68000 subjects, 4697 and 3561 were carriers of a clearly deleterious BRCA1 or BRCA2 mutation, respectively. And only 25 subjects carried clearly deleterious mutations in both genes. Thus, for BRCA1 we assume to have 64439 cases and 4867 pseudo-controls. Of these, 4842 and 25 carried clearly deleterious BRCA1 mutations, respectively. This data set is dominated by truncating mutations that require almost no sequence analysis and yields a pooled OD of 11.5, 80% power (alpha=0.05) with 170 cases& controls, and >99% power with 500 cases&controls. Overall, these and corresponding BRCA2 data are probably unrealistically powerful to serve as a model for any candidate susceptibility gene. Therefore, to focus the analysis on less powerful data that require sequence analysis to generate a useful TSMR+ pool, the analysis of BRCA1 was limited to all single nucleotide substitutions observed in the RING domain (aa 1-102) and the BRCT domain (aa 1641-1863). For BRCA2, all single nucleotide substitutions observed in the DNA binding domain (aa 2401-3110) were considered.

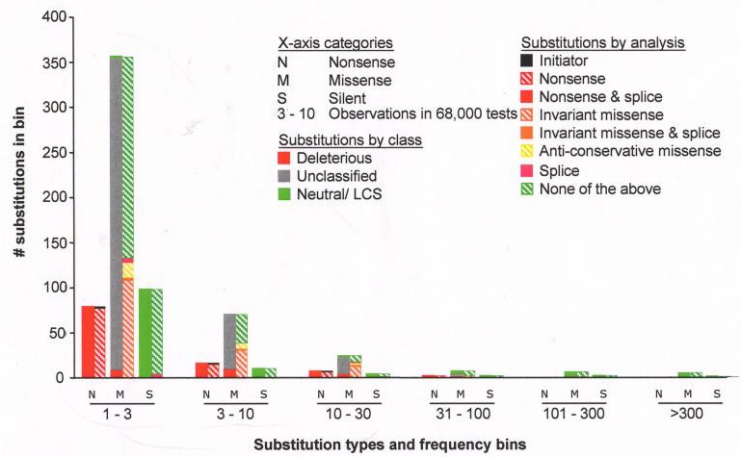

Fig. 1. Distribution of 698 distinct single nucleotide substitutions observed in BRCA1&2

By combining across these three domains of BRCA1 and BRCA2, the mutation screening revealed 698 distinct single nucleotide substitutions with different frequency distribution (fig. 1). Six of these, BRCA1 M1652I and BRCA2 A2466V, I2490T, V2728I, A2951T, and S2414S were observed more than 300 times each. Scaling the mutation screening data from 68000 subjects to 1000 cases and 1000 controls, these six substitutions are the only ones that would expect to use in standard tests of association. However, all of them are known neutral variants with OD no more than 1.2 in this case/pseudo-control format. The two most common deleterious BRCA1 variants in this set are the RING domain's MSs C61G and the nonsense mutation R1835X observed 180 and 56 times, respectively. The two most common deleterious BRCA2 variants are the nonsense mutation R2520X and the MS D2723H observed 52 and 42 times, respectively.

Hence, from 689 substitutions in the mutation screening set, there is only one recognized deleterious variant that one would expect to observe, on average, more than once in a set of 1000 cases. Furthermore, 664 of the substitutions were observed 30 or fewer times. Thus the vast majority of substitutions observed during the mutation screening are so rare, individually, that one would have less than 50% probability to observe any one of them even once during full mutation screening of a set of 1000 cases and 1000 controls. Nonetheless, cumulatively, such a mutation screening experiment would observe many of these sequence variants. The question is can a pool of such rare variants contribute useful data to a test of association? To address this question, TSMR+ pools were constructed from the observed sequence variants and then simulated case-control association studies by randomly sampling 100000 replicates of 500, 1000, and 1500 cases and pseudo-controls from this overall sample series.

Constructing the TSMR+ pools: Analysis of BRCA1 RING domain and BRCT domain substitutions revealed 35 nonsense mutations; 10 substitutions that fell on the last 2 nucleotides of an exon, 9 of which interfered with the AG splice donor consensus and were retained in the TSMR+ pool; and 4 mutations at the translation initiation codon which one could consider either as truncating, missense, or regulatory mutations. For this analysis, these substitutions comprise the TS component of the BRCA1 TSMR+ pool. MSs were analyzed using Align-GVGD (http://agvgd.iarc.fr/alignments.php) software. Of the 167 BRCA1 MSs observed in these domains, 72 fell at invariant position in the alignment and were retained in the TSMR+ pooled as "invariant substitutions, M(I)". 12 additional MSs falling at slightly variable positions met the criterion GD>61.3+GV and were retained as anti-conservative substitutions, M(A). Analysis of BRCA2 DNA binding domain substitutions revealed 65 nonsense mutations and 10 substitutions that fell on the last 2 nucleotides of an exon, 9 of which interfered with the AG splice donor consensus; these were retained as the TS component of the BRCA2 TSMR+ pool. MSs were analyzed using the same Align-GVGD program. Of the 305 BRCA2 MS observed I this domain, 88 were retained in the TSMR+ pool as M(I) invariant substitutions and 13 were retained as M(A) anti-conservative substitutions. The frequency distribution of the overall set of substitutions (Fig. 1) showed that the vast majority of MSs observed 30 or fewer times are unclassified, but approximately 43% of these were retained in the TSMR+ pool. All of the known deleterious substitutions were in fact retained in the TSMR+ pool.

Tests of association: To stimulate realistic mutation screening scenarios, equal case and control sets of 500, 100, and 1500 individuals were sub sampled from the complete cases and control sets of BRCa1 and BRCA2. For each pair of samples we then compared the frequency of various categories of variant in the case and control sets using a t-test. The samples are large enough that the normal approximation to the binomial distribution involved in this test should be valid. Nonetheless we checked p-values under the null hypothesis and found them to be correct. For each gene, size of sub

sample, and TSM<R+ formulation, 100000 replicate case and control sets were sampled, and used these to evaluate the power to detect a difference using a 5% significance test.

The variant classes considered to explore utility of the TSMR+ pools, were: (1) nonsense only; (2) nonsense and splice site substitutions; (3) nonsense, splice sites, and invariant missense substitutions M(I); (4) nonsense, splice sites, invariant missense substitutions M(I), and anti-conservative missense substitutions M(A). Further, to explore the consequence of MS, the following were also considered: (5) nonsense, splice sites, and all MSs with frequency <1%; (6) all MSs with frequency <1% but not splice, M(I), or M(A); and (7) M(I) invariant and M(A) anti-conservative missense substitutions alone.

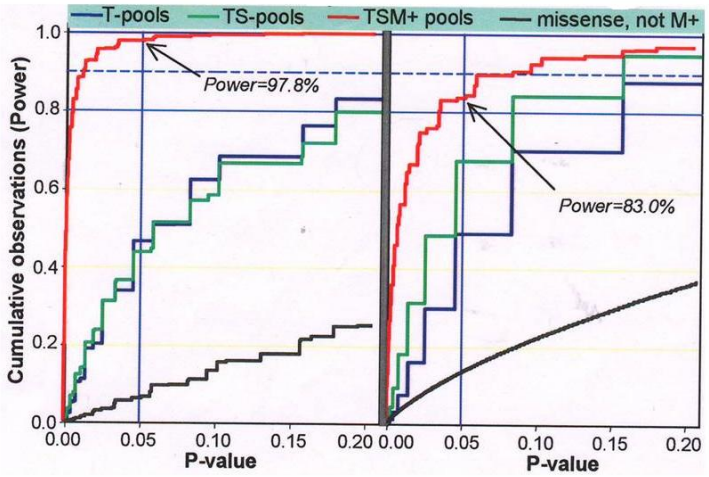

Fig. 2. BRCA1 TSMR+ analysis (left) and BRCA2 TSMR+ analysis (right)

Using a p-value of 0.05 as the criterion for significance, the combination of initiator, nonsense, and splice junction mutations in BRCA1 had 50% power with 1000 cases and pseudo-controls. Addition of the M(I) invariant missense substitution increased that power to 97.8%, and addition of the M(A) anti-conservative substitutions further increased the power to 98% (fig. 2). On the other hand, inclusion of all MS with a frequency of <1% led to a small decrease in the power, back to 97%. Although the BRCA1 TSMR+ pool that contained all of the MSs with frequency <1% retained good power, the rare MS do contain an identifiable subset that is enriched for neutral substitutions. Thus the set of RING and BRCT MSs with frequency <1% but neither invariant, nor anti-conservative, nor located near a splice donor had essentially no power to detect evidence of risk. Taking the BRCA1 TSMR+ pool consisting of initiator, nonsense, splice junction, M(I) invariant missense, and M(A) anti-conservative MSs as near optimal, the power of this pool was explored as a function of the number of subjects. Overall, 59%, 98%, and 99% power was observed with 500, 1000, and 1500 cases and pseudo-controls, respectively.

For BRCA2, a combination of nonsense and splice junction mutations had 67% power with 1000 cases and pseudo-controls. Addition of the DNA binding domain M(I) invariant substitutions increased that power to 83% (Fig. 2). Further addition of the M(A) anti-conservative substitutions led to a small reduction in power (to 67%). In contrast to the BRCA1 results, inclusion of all BRCA2 MSs with a frequency of <1% completely destroyed power to detect evidence risk. As with BRCA1, the set of MS with frequency <1% but neither invariant, nor anti-conservative, nor located near a splice

donor had essentially no power to detect evidence of risk. Power of the BRCA2 TSMR+ pool including M(I) invariant missense, and M(A) anti-conservative missense substitutions as a function of sample size observed 48%, 83%, and 88% power with 500, 1000, and 1500 cases and pseudo-controls, respectively.

The particular set of BRCA1 and BRCA2 sequence variant analyzed here was selected to illustrate a strategy of genetic case-control tests of association that could be applicable to known and candidate susceptibility genes. The genes, BRCA1 and BRCA2 were chosen since a large series of controls within the 68,000 subjects were defined with complete mutation screening data. Focusing the research on single nucleotide substitutions found in the RING and BRCT domains of the BRCA1 , and DBD domain of BRCA2 provided opportunity to demonstrate that alignment based sequence analysis such as Align-GVGD or SIFT, can specifically add subsets of missense substitutions that are intrinsically likely to alter protein function into the tests of association. The algorithmic approach to creating a TSMR+ pool used here demonstrates a systematic method of pooling rare sequence variants that should be applicable to analyses of candidate intermediate risk susceptibility genes, irrespective of whether the candidate gene's mutation profile is dominated by truncating, splice, or missense variants.

The contribution of individual components of the TSMR+ pool to the power of the simulated association analysis reveals two important points. (1) Addition of the pool of M(I) invariant and M(A) anti-conservative missense substitutions markedly increases the power of both the BRCA1 and BRCA2 analyses at all sample sizes. (2) The MSs have considerable power on their own. For BRCA1, the combined M(I+A) missense substitutions were considerably more powerful than the combination of nonsense and splice junction mutations. For BRCA2, the combined M(I+A) missense substitutions had approximately the same power as did the nonsense mutation alone.

Using just 13% of the clear BRCA1 mutation data or 7% of the clear BRCA2 mutation data, plus data from many unclassified MSs, excellent power was generated in a case-control format to show that these are susceptibility genes. Given the high-throughput mutation scanning techniques such as high resolution melt curve analysis [1], it should be possible to detect evidence of risk for candidate susceptibility genes with attributable risks as low as 10% of that of BRCa1 or BRCA2. The output of tests of association based on TSMR+ pool will be global Ors and frequencies for each gene tested. For genes actually found to contribute to disease susceptibility, it is unlikely that every variant in the pool will confer the same OR. Yet, for genes where evidence of association is detected, many of the variations in the TSMR+ pool will play a causal role in disease susceptibility.

Building sufficiently informative multiple sequence alignments to empower classification of MS is no trivial. Classification of MSs by this method is both dependent on the classification algorithm and the depth of the underlying alignment. Finally, the TSMR+ method is better suited for analysis of genes where loss-of-function confers increased risk than for genes where gain-of-function increases risk. Therefore, the TSMR+ method could be applied for many candidate genes with intermediate risk.

Neither of two large multicenter genome scans for new prostate cancer and BC susceptibility loci [6,4] found particularly strong evidence of linkage (LOD>3.5) at any locus. Another gene, CHEK2 was selected as a biochemically plausible candidate gene located under a weak (LOD=1.2) linkage peak [2], and the interpretation is that an important subset of not statistically significant peaks identified by a linkage analysis are actually attributable to modest risk susceptibility loci. The TSMR+ pool and case-control mutation screening strategy described here provides powerful approaches to analyze candidate intermediate risk susceptibility genes that are likely to emerge from positional cloning within linkage peaks, biochemical pathway-based re-sequencing as well as Next

Generation sequencing projects. The single most important point from this analysis is that the case-control mutation screening method is able to test candidate intermediate-risk susceptibility genes and provide the data required to calculate genetic population attributable fraction to those genes that actually contribute substantially to BC susceptibility. The secondary point is that the analysis ends up with more likely deleterious sequence variants in the TSMR+ pool. Analysis of ATM gene [3] demonstrated in molecular epidemiology terms that ATM is indeed a BC susceptibility gene but did not include the missense substitutions. The value of applying a systematic analysis of missense substitutions is a substantial improvement in the statistical strength of the result. Thus, TSMR+ analysis should yield a more accurate measurement of risk attributable to intermediate-susceptibility genes than a less complete analysis.

REFERENCES

1. *Chou L.S., Lyon E, Wittwer C.T.* A comparison of high-resolution melting analysis with denaturing high-performance liquid chromatography for mutation scanning: cystic fibrosis transmembrane conductance regulator gene as a model. Am J Clin Pathol. 124(3):330-8, 2005.
2. *Meijers-Heijboer H., van den Ouweland A., Klijn J., et al*. Low-penetrance susceptibility to breast cancer due to CHEK2(*)1100delC in noncarriers of BRCA1 or BRCA2 mutations. Nat Genet. *31*, 1, 55-9, 2002.
3. *Renwick A, Thompson D, Seal S, Kelly P, Chagtai T, Ahmed M, North B, Jayatilake H, Barfoot R, Spanova K, McGuffog L, Evans DG, Eccles D; Breast Cancer Susceptibility Collaboration (UK), Easton DF, Stratton MR, Rahman N.* ATM mutations that cause ataxia-telangiectasia are breast cancer susceptibility alleles. Nat Genet. *38*, 8, 873-5, 2006.
4. *Smith P, McGuffog L, Easton DF, Mann GJ, Pupo GM, et al.* A genome wide linkage search for breast cancer susceptibility genes. Genes Chromosomes Cancer. *45*, 7, 646-55, 2006.
5. *Venkitaraman A.R.* Cancer susceptibility and the functions of BRCA1 and BRCA2. Cell. 25;108(2):171-82, 2002.
6. *Xu J., Dimitrov L., Chang B.L., Adams T.S., et al.* A combined genomewide linkage scan of 1,233 families for prostate cancer-susceptibility genes conducted by the international consortium for prostate cancer genetics. Am J Hum Genet. 77(2):219-29, 2005.

Received on 08.09.2016