JMG Online First, published on July 13, 2005 as 10.1136/jmg.2005.033878

Comprehensive statistical study of 452 BRCA1 missense substitutions with classification of eight recurrent substitutions as neutral.

Sean V. Tavtigian[1a], Amie M. Deffenbaugh[2], Luo Yin[1], Thaddeus Judkins[2], Tom Scholl[2], Paul B. Samollow[3], Deepika de Silva[1], Andrey Zharkikh[2], Alun Thomas[4]

1. International Agency for Research on Cancer, Lyon, France
2. Myriad Genetic Laboratories, Inc., Salt Lake City, USA
3. Southwest Foundation for Biomedical Research, San Antonio, USA
4. Department of Medical Informatics, University of Utah, Salt Lake City, USA

[a] To whom correspondence should be addressed
Sean V. Tavtigian
International Agency for Research on Cancer
150 Cours Albert Thomas
69372 Lyon France

Email: tavtigian@iarc.fr
Tel: +33 (0)4 72 73 85 12
Fax: +33 (0)4 72 73 83 88

ABSTRACT

Introduction. Genetic testing for hereditary cancer syndromes contributes to the medical management of patients who may be at increased risk of one or more cancers. BRCA1 and BRCA2 testing for hereditary breast and ovarian cancer is one such widely used test. However, clinical testing methods with high sensitivity for deleterious mutations in these genes also detect many unclassified variants, primarily missense substitutions.

Methods. We developed an extension of the Grantham Difference, called A-GVGD, to score missense substitutions against the range of variation present at their position in a multiple sequence alignment. Combining two methods, co-occurrence of unclassified variants with clearly deleterious mutations and A-GVGD, we analyzed most of the missense substitutions observed in BRCA1.

Results. A-GVGD was able to resolve known neutral and deleterious missense substitutions into distinct sets. Additionally, eight previously unclassified BRCA1 missense substitutions observed in trans with one or more deleterious mutations, and within the cross-species range of variation observed at their position in the protein, are now classified as neutral.

Discussion. The methods combined here can classify as neutral about 50% of missense substitutions that have been observed with two or more clearly deleterious mutations. Furthermore, odds ratios estimated for sets of substitutions grouped by A-GVGD scores are consistent with the hypothesis that most unclassified substitutions that are within the cross-species range of variation at their position in BRCA1 are also neutral. For most of these, clinical reclassification will require integrated application of other methods such as pooled family histories, segregation analysis, or validated functional assay.

KEY WORDS

BRCA1, missense substitution, co-occurrence, Grantham difference, A-GVGD

INTRODUCTION

Genetic testing for hereditary cancer syndromes is increasingly contributing to the medical management of patients who may be at markedly increased risk of one or more cancers. Testing of a patient who has a strong family history will ideally result in the discovery of a clearly deleterious cancer predisposing mutation, leading to options such as aggressive screening, prophylactic surgery, or chemopreventive strategies. Unfortunately, in many cases where no clearly deleterious mutation is found, a sequence variant of uncertain clinical significance, most often a missense substitution, is found. In such cases, patients, and the healthcare providers who counsel them, are left with ambiguous test results that are of little help in determining appropriate cancer risk-reduction strategies.

Germline loss of function mutations in BRCA1 and BRCA2 confer high-risk of breast cancer and ovarian cancer and confer elevated risks of a number of other cancers [1-5]. Testing for mutations in these genes has become one of the most widely used hereditary cancer tests, with over 70,000 patients tested to date. Although 13.5% of patients tested through full sequence analyses of both BRCA1 and BRCA2 at Myriad Genetic Laboratories (MGL) are found to carry a deleterious mutation, 12% of patients who do not carry a clearly deleterious variant are found to carry an uncertain variant (database query updated from [6]).

Recently, Goldgar et al. [7] developed a method for analysis of unclassified missense substitutions in BRCA1 and BRCA2 that integrates four types of data: segregation of sequence variants of interest in pedigrees; pooled family histories of index cases who carry the variant versus all index cases tested; co-occurrence of the variant with clearly deleterious variants in the same gene; and cross-species protein multiple sequence alignment followed by comparison of the physico-chemical characteristics of the amino acids observed at the point of the mutation (Grantham analysis). Each of these data types has its strengths and weaknesses. Segregation analysis and pooled family history analysis are both pure human genetics measures; they provide fairly direct measures of disease susceptibility. However, both require accurate family history data, which may be difficult to obtain. Co-occurrence of variants of interest with clearly deleterious mutations takes advantage of the highly penetrant embryonic lethal phenotype conferred by most BRCA1-null genotypes [8-14] and is thus both a human genetics measure and a type of functional assay, but one in which genetics and function are measured indirectly. Sequence alignments and Grantham analyses are measures of evolutionary fitness that are only indirectly tied to disease susceptibility. In contrast to segregation analysis and summary family history, co-occurrence data can be obtained directly from systematically compiled mutation screening databases. Similarly, sequence alignment/ Grantham analysis can readily be applied to any observed missense substitution.

One strength of the integrated method is that each of the four types of data analysis that it has integrated was developed as an independent estimator of the likelihood that a sequence variant confers a high cancer risk versus being neutral or of little clinical significance (neutral/ LCS). Likelihood calculations from the four methods are multiplied to reach a final result. One does not need to use all four types of data to analyze any particular variant; analyses using 2 or 3 data types are also perfectly valid. In the following study we have used co-occurrence data and a modified approach to sequence alignment/ Grantham analysis to look at all 452 missense substitutions observed in BRCA1 in a series of 40,000 full-sequence BRCA1 and BRCA2 tests conducted at MGL (B1&2 40K set), with a goal of identifying missense substitutions that are neutral/ LCS.

METHODS

BRCA1 and BRCA2 testing (BRACAnalysis®). Full sequence analyses of BRCA1 and BRCA2 were performed by direct gene sequencing as previously described [15]. In order for a test to be performed, the test request form must be completed by the ordering health care provider and the form must be signed by an appropriate individual indicating that "informed consent has been signed and is on file". Patient samples were each assigned a unique bar-code for robotic specimen tracking. Most samples were received as 7 ml of anti coagulated blood, from which DNA was extracted and purified from leukocytes isolated from each sample. Aliquots of patient DNA were each subjected to polymerase chain reaction (PCR) amplification. The amplified products were each directly sequenced in the forward and reverse directions using fluorescent dye-labeled sequencing primers. Chromatographic tracings of each amplicon

were analyzed by MGL's sequence analysis software followed by visual inspection and confirmation, assisted by comparison of the proband sequence to a consensus wild-type sequence constructed for each amplicon. Each genetic variant (exclusive of non-reportable polymorphisms) was independently confirmed by repeated analysis including PCR amplification of the indicated gene region(s) and sequence determination.

All mutations and genetic variants were named according to the convention of Beaudet and Tsui [16], and all of them have been submitted to the BIC database <http://research.nhgri.nih.gov/bic/>. Nucleotide numbering starts at the first transcribed base of BRCA1 according to GenBank entry U14680. (Under this convention, the mutation commonly referred to as "185delAG" is named "187delAG". However, in this paper, we refer to this variant by its more common name.)

Co-occurrence analysis. The expression for using co-occurrence data to calculate the likelihood that a variant is deleterious versus neutral/ LCS was developed in Goldgar et al. [7]. Briefly, if the sequence variant of interest was observed *n* times, *k* of which were in individuals who also carry a clearly deleterious variant, the appropriate binomial likelihood ratio is:

$$\frac{(p_2)^k(1-p_2)^{n-k}}{p_1{}^k(1-p_1)^{n-k}} \quad \text{[eq 1].}$$

where p_1 is the probability that an individual in the test population who carries an unclassified neutral variant also carries (in trans) a deleterious mutation, and p_2 is the probability that an individual in the test population who carries an unclassified deleterious variant also carries (in trans) a deleterious mutation. The overall frequency of clearly deleterious BRCA1 mutations found by BRACAnalysis In the B1&2 40K set was 8.1%, and we take p_1 to be 50% of that frequency. BRCA1 homozygote and compound heterozygote genotypes are quite likely embryonic lethal and consequently extremely rare; accordingly, we have set p_2=0.0001 for these calculations [7].

In addition, our basic query of the BRACAnalysis database gave us the identity of every missense substitution observed, the number of times each was observed, the number of times each was observed with a clearly deleterious mutation in BRCA1, the number of different clearly deleterious mutations with which each was observed, and the number of times each was observed with a clearly deleterious mutation in BRCA2. We were also able to query the co-occurring deleterious mutations to see how often they were seen independently of the missense substitution of interest. In our analysis of unclassified missense substitutions, only co-occurrences with independently observed deleterious mutations were used in the likelihood calculation.

Haplotype analysis. Using 14 common polymorphisms in BRCA1 that are within the sequences covered by BRACAnalysis (exon 4 -49 C>T; IVS8 -58delT; Q356R; D693N; S694S; L771L; P871L; E1038G; S1040N; K1183R; R1347G; S1436S; S1613G; and M1652I), we have defined the 10 most common haplotypes in our test population ([17] and T. Scholl, manuscript in preparation). Due to the simple haplotype structure of BRCA1, no two genotypes that result from pairs of these haplotypes are identical. All of the sequence variants that we needed to analyze, both the unclassified substitutions and the deleterious variants with which they co-occurred, were examined for the haplotype contexts in which they were seen. For sequence variants that were observed more than ~5 times, this usually resulted in a single unambiguous haplotype assignment. Once variants had been assigned to specific haplotypes, we looked at the genotypes of the patients in whom co-occurrences were observed. In those cases where we were able to determine haplotypes for both the mutation and the unclassified substitution, it was sometimes clear that the test subject was a heterozygote for the two haplotypes and that the two variants were therefore in trans.

Sequencing of *Monodelphis* BRCA1. Peptide sequences of the individual coding exons of human BRCA1 were searched against genomic sequence reads from the *M. domestica* genome sequencing project by tBLASTn [18]. Most exons of the *M. domestica* BRCA1 ortholog were identified with little ambiguity.

PCR and sequencing primers were designed based on the predicted exon sequences; primers were selected so that predicted PCR product lengths would be between 1 kb and 2kb, missing exons in the assembly would be spanned, and products would overlap to allow complete sequencing. cDNA was prepared from colon and testis samples of two individual opossums. After PCR, products were gel purified and sequenced with Big-Dye dye terminator chemistry. Our sequence, which has 47% amino acid sequence identity to human BRCA1, has been submitted to Genbank under accession # AY994160 and also used in the alignment described below.

Creation and analysis of the multiple sequence alignment. The BRCA1 protein multiple sequence alignment used for this analysis contained 12 full-length BRCA1 sequences. The evolutionary relationships and % sequence identities between most of these sequences were described previously [19]. The multiple sequence alignment was made with the alignment program 3DCoffee, which also incorporates alignment to X-ray and NMR structures [20]. 3DCoffee was run using Mlalign_id_pair, Mslow_pair, and Mclustalw_aln to generate amino acid alignments and Mfugue_pair to generate structure sequence alignments. GenBank accession numbers for BRCA1 protein sequences used in the alignment were as follows: human, NP_009225; chimpanzee, AAG43492; gorilla, AAT44835; orangutan, AAT44834; rhesus macaque, AAT44833; mouse, AAD00168; dog, AAC48663; cow NP_848668; opossum, AAX92675; chicken, NP_989500; *Xenopus*, AAL13037; *Tetraodon*, AAR89523. For the structure component of the alignment, we used BRCA1 RING NMR sturucture 1JM7.pdb and the BRCA1 BRCT repeat crystal structures 1JNX.pdb and 1T29.pdb [21-23]. A parsimony based method was used to calculate the minimum number of missense substitutions required to create the observed alignment, taking into account the underlying phylogenetic tree [19, 24]. Because we are interested in human disease genetics, our subsequent analyses only considered sequence variation at positions in the alignment where the human sequence has a residue. On the other hand, absence of an amino acid in the alignment of a non human BRCA1 at a position where the human sequence does have a residue was considered a sequence variation.

The number of slowly substituting positions (Class 2 or SS), fast substituting positions (Class 3 or FS), and the relative odds that a position in BRCA1 where 0,1,2,...n substitutions are observed is either SS or FS, were calculated from the protein multiple sequence alignment using the modified Fitch covarion model "Model 3" of Abkevich et al. [19, 24]. The relative odds that a position in BRCA1 is either SS or FS is identical to the sequence conservation likelihood ratio used in Goldgar et al. [7].

Grantham analysis. For two amino acids i and j with sidechain compositions *Ci* and *Cj*, polarities *Pi* and *Pj*, and volumes *Vi* and *Vj*, the standard Grantham Difference formula is:

$$50.723\times\sqrt{\left[1.833(C_i-C_j)^2\right]+\left[0.1018(P_i-P_j)^2\right]+\left[0.000399(V_i-V_j)^2\right]} \quad \text{[eq 2].}$$

C-, *P*-, and *V*-values for the 20 common amino acids are given with Grantham's definition of the measure [25].

The set of amino acids observed at a particular position in a protein multiple sequence alignment will have minimum and maximum values of *C*, *P*, and *V*. For calculation of the Grantham Variation (GV) of a position, *Cmax* replaces *Ci*, *Cmin* replaces *Cj*, and so on, in eq 1. A sample calculation of GV is given in **Figure 1c**. Difficulties arise at gaps in the sequence alignment. We arrange our alignments so that the first sequence is the human sequence and successive sequences are from species that are successively more distantly related to humans. GVs are calculated sequentially from closely related species to distantly related species. At the first appearance of a gap at a particular position in the alignment, we set *Cmin* and *Cmax* to 0 and 3, respectively (the highest value of *C* for an amino acid is 2.75); we use *Pmin* and *Pmax* from the observed amino acids; and we set *Vmin* to 0 but set *Vmax* to the largest *V* from the observed amino acids. If the position is again gapped in more distantly related species, and if the positions immediately before and immediately after the position of interest are also gapped, then we set all *Cmin*, *Pmin*, and *Vmin* =0 and *Cmax*, *Pmax*, and *Vmax* to 3, 14, and 175 (which are all slightly above the highest values for normal amino acids).

Each missense substitution is characterized by values *Cm*, *Pm*, and *Vm*. These are either below, within, or above the range of variation captured in the GV of the position of interest. For calculation of the Grantham Deviation (GD), if *Cm* < *Cmin*, then the *C* component of eq 1 is replaced by (*Cmin* - *Cm*); if *Cmin*≤*C*≤*Cmax*, then *C* component of eq 1 =0; if *Cm* > *Cmax*, then the *C* component of eq 1 is replaced by (*Cm* - *Cmax*). Corresponding substitutions are made for the *P*- and *V*-components of eq 1. A sample calculation of GD is given in **Figure 1d**. The overall method combining sequence alignment with calculation of GV and GD is called A-GVGD.

Statistical methods. Calculations of GV and GD were implemented in APL (APLX Version 2.0.9, MicroAPL Ltd). The BRCA1 with BRCA2 ascertainment for BRACAnalysis odds ratios were calculated in a series of contingency tables. For the disease status axis of these tables, "controls" are individuals who carry a clearly deleterious mutation in BRCA2, and "cases" are individuals who were not found to carry a deleterious BRCA2 mutation. Categories on the genotype axis are pooled sets of sequence variants that meet specific selection criteria, as described in the legend to **Figure 3**. Confidence intervals were estimated using Miettinen's test-based approximation.

RESULTS

Co-occurrence analysis. The idea of using co-occurrence of clearly deleterious sequence variants with unclassified missense substitutions to classify missense substitutions in BRCA1 arises from two independent sources. One is a series of mouse and *Xenopus* studies which show that complete loss of BRCA1 function confers a highly penetrant embryonic lethal phenotype [8-14]. The second is observation of a clear deficit of Ashkenazi BRCA1 homozygotes/ compound heterozygotes in MGL's BRACAnalysis database [6, 19]. Of the 452 missense substitutions in the B1&2 40K set, 72 have been observed in an individual who also carried a clearly deleterious BRCA1 mutation. Fifteen of the 16 missense substitutions that currently are classified as neutral/ LCS fall into this group, as do 57 unclassified missense substitutions (**Table 1**). None of the 28 missense substitutions that are currently classified by MGL as Deleterious or Favor Deleterious were observed in a patient who carries another clearly deleterious mutation in BRCA1.

Table 1. Observations of missense substitutions in BRCA1 in individuals who also carry a clearly deleterious BRCA1 mutation.

Count[a]	Polymorphism[o]	Unclassified	Deleterious[o]	Total
0	1(1)	351(351)	28(28)	380(380)
1	0(0)	39(46)	0(0)	39(46)
2	1(1)	4(3)	0(0)	5(4)
3	1(2)	6(7)	0(0)	7(9)
4-5	2(1)	2(0)	0(0)	4(1)
6-10	1(1)	2(1)	0(0)	3(2)
11-100	3(4)	4(0)	0(0)	7(4)
>100	7(6)	0(0)	0(0)	7(6)
	16(16)	408(408)	28(28)	452(452)

[a] Each row is a co-occurrence count bin. Thus the first row gives the number of missense substitutuions that were not observed to co-occur with a clearly deleterious mutation. The third row gives the number of missense substitutions that were observed to co-occur with exactly 2 clearly deleterius mutations, or (in parantheses) exactly 2 different deleterious mutations.

[o] Includes missense substitutions recently classified in either Goldgar et al. [7] or Phelan et al. [26].

At a superficial level of analysis, the main pitfall of co-occurrence data is that, for purposes of classification of missense substitutions, an observation of co-occurrence is only meaningful if the unclassified missense substitution and the clearly deleterious mutation are in trans. Fortunately, as a

consequence of full sequence testing, we have additional data that can help to discriminate cases of cis from trans co-occurrence. First, in addition to the number of times that each missense substitution has been observed with a clearly deleterious mutation, we also know the number of different deleterious mutations with which it was observed to co-occur and their individual identities. Second, for each deleterious variant that has been observed in a patient who also carried an unclassified substitution of interest, we know the number of times that the deleterious variant was observed with and without the substitution of interest. Combined, these data allow us to identify substitutions that actually are in cis with a clearly deleterious mutation. Third, for recurrent sequence variants we can usually determine, using single nucleotide polymorphisms (SNPs), independent of any disease consideration, the common haplotype on which that variant is found.

Taking advantage of these additional data, we can add four layers of caution to our calculation of the likelihood ratio using co-occurrence data (cooc-LR). (i) When the deleterious variant involved in a co-occurrence has been observed only once, we do not know whether the unclassified substitution and the deleterious variant are independent. For application of cooc-LR to the unclassified variants, these are ignored. (ii) When the unclassified variant and deleterious variant are not usually observed independently, co-occurrences of this pair of variants are ignored. (iii) If these precautionary subtractions leave two or more distinct co-occurrences, we can assume that at least n-1 of them are in trans; this point is explained further in the next paragraph. (iv) If these precautionary subtractions leave only one distinct co-occurrence, we use our knowledge of SNP-based haplotypes to determine whether the individual who carries the two variants of interest has a SNP genotype that is compatible with being a heterozygote for the SNP haplotypes on which the unclassified missense substitution and the deleterious mutation are usually observed and incompatible with being a homozygote for either of those haplotypes. Meeting this condition confirms a trans co-occurrence; the confirmed trans co-occurrence(s) is then used in the calculation.

When an individual is observed to carry two rare sequence variants, in the same gene, *that usually segregate independently*, the most likely explanation by far is that the two sequence variants have been inherited in trans. However, there can be exceptions. One possibility is that the subject could have inherited a very rare recombinant chromosome that carries the two sequence variants in cis. In this scenario, for BRCA1, it is extremely unlikely that both sequence variants are deleterious. This is because the recombination that brought the two sequence variants into cis would have to have taken place in an ancestor who inherited the two sequence variants in trans. But if both variants are deleterious, that ancestor's genotype would have been a highly penetrant embryonic lethal [8-13], and, if not lethal, would have interfered directly with the process of recombination [14]. An alternative possibility is that the subject could have inherited a chromosome on which the second of the two sequence variants is a new mutation that happens to be identical to the rare sequence variant that normally segregates independently of the first variant (or vice versa). However, excepting substitutions at CpG dinucleotides and length variations in repeated sequence elelements, multiple independent origins of the same human sequence variant are quite rare. Thus the probability that we will have observed a BRCA1 allele bearing one of the relatively rare unclassified missense substitutions of interest that has been hit by the 2nd origin of a specific deleterious BRCA1 mutation is low, and the probability the we will observe an unclassified missense substitution-bearing allele that has been hit twice by 2nd origins of deleterious mutations is infinitesimal. This logic also holds for the occurrence of a 2nd independent origin of one of the unclassified missense substitutions of interest on a deleterious mutation-bearing BRCA1 allele. Thus, for rare sequence variants, we can assume that at most one observation of co-occurrence with an independently observed deleterious mutation is due to occult co-occurrence in cis.

We applied this analysis to the 33 missense substitutions that were observed to co-occur with a deleterious variant 2 or more times. Of these, 15 are classified as neutral/ LCS while 18 are currently unclassified. cooc-LRs for all 15 known neutral/ LCS variants were below $1x10^{-2}$ (**Table 2A**). The neutral substitution with the fewest co-occurrences was Y856H (observed once with each of two different deleterious mutations). Even though we ignored on of its two co-occurrences in the calculation, its cooc-LR was below 0.01.

Table 2. Missense substitutions in BRCA1 that have been observed with 2 or more known deleterious mutations.

A. Known neutral/ LCS missense substitutions

Misssense substitution		Observations					Haplotype test Req'd?	result	cooc-LR
aa name	nt name	het	homo	With B1*	Diff B1•	Indep∞			
Q356R	1186A>G	4,650	(198)	364	111	NA	no	--	< 1.0E-10
D693N	2196G>A	5,307	(219)	305	131	NA	no	--	< 1.0E-10
K820E	2577A>G	95	(4)	8	6	NA	no	--	< 1.0E-10
R841W[o]	2640C>T	117	(0)	4	4	NA	no	--	3.9E-09
Y856H	2685T>C	31	(1)	2	2	NA	no	--	8.8E-03
P871L	2731C>T	17,722	(5399)	1,696	440	NA	no	--	< 1.0E-10
E1038G	3232A>G	17,356	(4234)	1,557	407	NA	no	--	< 1.0E-10
S1040N	3238G>A	1,611	(22)	114	67	NA	no	--	< 1.0E-10
S1140G	3537A>G	97	(1)	4	3	NA	no	--	2.9E-04
K1183R	3667A>G	17,440	(4278)	1,562	406	NA	no	--	< 1.0E-10
R1347G	4158A>G	433	(2)	27	17	NA	no	--	< 1.0E-10
S1512I	4654G>T	278	(1)	15	12	NA	no	--	< 1.0E-10
S1613G	4956A>G	17,454	(4330)	1,565	409	NA	no	--	< 1.0E-10
M1628T[a]	5002T>C	82	(0)	3	3	NA	no	--	1.6E-04
M1652I	5075G>A	1,049	(8)	70	38	NA	no	--	< 1.0E-10

B. Unclassified missense substitutions

Misssense substitution		Observations					Haplotype test Req'd?	result	cooc-LR
aa name	nt name	het	homo	With B1*	Diff B1•	Indep∞			
Y105C	433A>G	15	(0)	2	2	2	no	trans	1.1E-05
Y179C	655A>G	40	(1)	3	3	1	yes	trans	1.2E-02
S186Y	676C>A	28	(1)	3	1	0	no	cis	2.9E+00
L246V	855T>G	67	(0)	32	3	2	no	trans	2.6E-05
F486L	1575T>C	41	(1)	3	3	1	yes	trans	1.2E-02
R496H	1606G>A	73	(0)	7	3	2	no	trans	9.2E-05
R504H	1630G>A	11	(0)	5	1	0	no	cis	1.3E+00
N550H	1767A>C	40	(1)	3	3	1	yes	trans	1.2E-02
L668F	2121C>T	22	(0)	16	1	0	no	cis	1.3E+00
V772A	2434T>C	44	(0)	34	3	2	no	trans	9.4E-06
M1008V	3141A>G	12	(0)	7	6	4	no	trans	< 1.0E-10
M1008I~	3143G>A	103	(1)	3	1	1	yes	trans (3)	9.6E-07
E1060A	3298A>C	2	(0)	2	1	0	no	cis	1.0E+00
E1250K	3867G>A	14	(0)	2	2	2	no	trans	1.0E-05
D1546N	4755G>A	22	(0)	2	2	2	no	unclear	5.6E-03
L1564P	4810T>C	11	(0)	4	3	1	yes	trans	3.3E-03
P1637L	5029C>T	54	(0)	49	1	0	no	cis	1.2E+00
D1733G	5317A>G	3	(0)	3	1	0	no	cis	1.0E+00

* Co-occurrence with known deleterious mutations in BRCA1

• Co-occurrence with different know deleterious mutations in BRCA1

∞ Confirmed independent occurrence of the unclassified and deleterious variants

[o] Classified Neutral/LCS in Goldgar et al. [7].

[a] Classification in Phelan et al. [26] meets the criteria for Neutral/ LCS of Goldgar et al. [7].

~The deleterious mutation with which Ile1008 was observed is 185delAG; all 3 co-occurrences are trans.

Following the cautious approach outlined above, cooc-LRs were also calculated for the 18 unclassified missense substitutions that were nominally observed to co-occur 2 or more times (**Table 2B**). Although not strictly necessary, a haplotype based cis-trans test was made for at least one double carrier of each of these substitutions. Seven of the substitutions were observed with 2 or more different, independently segregating deleterious mutations; cooc-LRs for these were all below $1x10^{-2}$. One of these substitutions, D1546N, was observed once each with two different deleterious mutations. Asn1546 and the two deleterious mutations were all on the same SNP haplotype, rendering the cis-trans test uninformative. However, because two independently segregating deleterious mutations were involved in these co-occurrences, we conclude that at least one of these, and probably both, are bona fide trans co-occurrences. Five of the substitutions were observed with only one independently segregating deleterious mutation. In each case, BRCA1 SNP genotypes for the double carriers were as expected for a trans-carrier of the two variants of interest, resulting in cooc-LRs for these 5 substitutions of $1.2x10^{-2}$ or less. The result with M1008I was particularly interesting. Ile1008 was observed 103 times; 3 of these observations were with a clearly deleterious mutation, but always the same mutation. Superficially, one might expect from this pattern that a rare mutation had occurred on an Ile1008 chromosome and that the co-occurrences would be in cis. However, Ile1008 is most often observed in individuals of Ashkenazi ancestry. The mutation with which it has been observed 3 times is 185delAG. Among Ashkenazim, the frequency of 185delAG is higher than the summed frequency of all other clearly deleterious mutations in the gene; consequently, it is not surprising to have observed three individuals who carry this pair of variants in trans. Finally, 6 of the unclassified missense substitutions were observed with a deleterious mutation more than once, but always with the same deleterious mutation and in observation patterns best explained by co-occurrences in cis.

Measures of evolutionary variation and observed deviation. The basic logic behind the use of protein multiple sequence alignments to identify missense substitutions that are likely to be either neutral/LCS or deleterious breaks down into two components: (1) missense substitutions at positions that are highly functionally constrained tend to alter protein function while substitutions at positions that are not so constrained are less likely to alter function, and (2) missense substitutions that are outside the range of variation that is evolutionarily tolerated at their position in the protein tend to alter protein function whereas those that are within the range of variation tend to have little effect on protein function. Previously, we have used two different approaches to this problem [7, 19]. The approach taken in Abkevich et al. [19] made only qualitative use of evidence that a position in the BRCA1 was functionally constrained or not and then made quantitative use of the fit between an observed missense substitution and the range of variation at its corresponding position in a BRCA1 protein multiple sequence alignment. Neither component of that analysis was formatted as a likelihood ratio; consequently, that approach shared with SIFT and PolyPhen [27, 28] the flaw that it is not easily integrated into a proper multi-model likelihood calculation. The approach taken in Goldgar et al. [7] decomposed alignment/ Grantham analysis into two likelihood expressions: the first was based on the ratio of probabilities that the position at which a missense substitution is observed is functionally constrained or not, and the second was based on the Grantham Difference [25] for the missense substitution vs. the canonical human residue at its position in BRCA1. The first component, which we shall refer to as the constrained position likelihood ratio (con-LR), in agreement with Abkevich et al. [19], appeared to have good predictive power. However, the way that the Grantham Differences were used in the second component of the analysis did not correlate well with the other measures used in Goldgar et al. This was probably because Grantham Differences, used by themselves, do not make use of the fit between an observed missense substitution and the range of variation at its corresponding position in a BRCA1 protein multiple sequence alignment.

The fundamental problem is that the standard Grantham Difference is a pairwise comparison pressed into service for an application that would benefit from a genuine multiple comparison. As formulated in Goldgar et al., Grantham Differences are being used in a simple pairwise format. As formulated in Abkevich et al. and also in the recently proposed Grantham Ratio [29], the Grantham approach is extended to a multiple pairwise comparison; this is better but not entirely satisfactory. In order to achieve a true simultaneous multiple comparison, we introduce the Grantham Variation (GV) and the Grantham Deviation (GD) scores as follows.

In a 3-space where the axes are measures of amino acid sidechain composition (*C*), polarity (*P*), and

volume (*V*), the original Grantham Difference for a pair of amino acids is the Euclidean distance between the *C*-, *P*-, and *V*-values for those two amino acids (with scaling constants applied to squared *C*-, *P*-, and *V*-differences) [25]. For the amino acids observed at a specific position in a protein multiple sequence alignment, we consider the smallest rectangular box that contains all the corresponding points in Grantham space. The GV is then the length of the longest diagonal of the box. Given the point in Grantham space corresponding to the amino acid caused by a missense substitution, we define the GD as the shortest distance from that point to the bounding box. Any point inside the box has GD = 0 (see Methods for exact formulas). Consequently, using the same scale as Grantham Differences, these two measures provide a numerically precise yet quite natural description of the magnitude of sequence variation at any position in a protein multiple sequence alignment, and the fit between any given human missense substitution and the sequence variation present at its position in a multiple sequence alignment.

Characterization of the multiple sequence alignment and the Grantham Deviation. Calculations of GV, GD, and the con-LR for the 452 missense substitutions observed in the B1&2 40K set are based on a 12-sequence alignment that contains full length BRCA1 sequences from 8 placental Mammals plus *Monodelphis domestica* (Gray, short-tailed opossum), *Gallus gallus* (Chicken), *Xenopus laevis* (African clawed frog), and *Tetraodon nigroviridis* (Green-spotted pufferfish). The sequence alignment averages an absolute minimum of 3.25 amino acid substitutions per position, thus meeting the criterion of 3 substitutions per position derived by Greenblatt et al. for an alignment that is sufficiently informative to make predictions based on sequence conservation [30]. There are 130 invariant positions in the alignment. The con-LR at these invariant positions is 27, indicating odds of 27:1 that the individual invariant positions are functionally constrained. The con-LR drops to 2.6, 0.25, 0.025, and 0.003 at positions where a minimum of 1, 2, 3, and 4 substitutions are required in order to account for the observed alignment, respectively. Thus the likelihood that a position in BRCA1 is under strong functional constraint drops steeply as the number of evolutionarily tolerated substitutions at that position increases.

Grantham Deviations of known deleterious, unclassified, and known neutral missense substitutions are displayed in **Figure 2**. As GDs are calculated for successively more informative iterations of the multiple sequence alignment, a dramatic trend in the data becomes apparent: GDs for known deleterious mutations are essentially the same as standard Grantham Differences and change very little as further diverged sequences are included in the analysis. In contrast, GDs for known neutral substitutions are lower than standard Grantham Differences and drop precipitously as further diverged sequences are included in the analysis. This is explained by visualising the bounding boxes containing positions in the alignment at which deleterious and neutral substitutions have been observed. Because there is very little cross-species sequence variation at the positions at which most deleterious BRCA1 substitutions have been observed [19], the corresponding bounding boxes remain small. In contrast, there tends to be considerable sequence variation at the positions of known neutral substitutions. At these positions, as we consider greater evolutionary breadth, the corresponding bounding boxes expand and in so doing include the space into which many possible missense substitutions fall.

We could, in principle, model the distribution of GD data, estimate the probability density functions for the distributions of scores for known neutral/ LCS and deleterious variants, and then format a likelihood expression. However, such a likelihood expression would not be independent of the con-LR because both depend on sequence variation in the alignment. Thus the con-LR and a likelihood expression based on GDs cannot be used as independent factors in an integrated analysis. Simultaneous use requires that they be combined in a way that accounts for this co-dependence, which is beyond the scope of this analysis. We note, however, that evidence in favour of neutrality will be greatest when GD = 0. Hence, if we use the con-LR by itself in these cases it will be a very conservative approximation to the likelihood ratio that the combined analysis would provide. Until a complete likelihood expression has been developed, we will regard the con-LR as inappropriate for the application of identifying neutral variants if the GD > 0.

The appropriate evolutionary tree. The last question is whether data from all species in the BRCA1 sequence alignment should be used to calculate GV and GD. Beyond the alignment of 8 placental mammal BRCA1s, the average GD for known neutral/ LCS variants drops with the addition of each further diverged sequence. The sequential decrease with addition of the opossum, chicken, and frog sequences is substantial and the standard deviations of the GD also decrease. In contrast, the decrease in

both GD and its standard deviation with the addition of the pufferfish sequence is quite small (**Figure 2**).

One of the more difficult issues in BRCA1/2 genetics has been estimation of odds ratios for deleterious mutations. The basic problem is that, with a summed allele frequency for all of the high-risk missense substitutions of <<1%, testing the required number of controls is prohibitively expensive. Because BRACAnalysis is a full sequence test of both BRCA1 and BRCA2, with results tracked in a single database, we know which carriers of interesting sequence variants in BRCA1 also carried a clearly deleterious mutation in BRCA2. 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 BRCA2 carrier, the appearance of a double carrier in the B1&2 40K set is largely explained by the deleterious BRCA2 variant. Hence, we can use the BRCA1-chromosomes of the BRCA2 mutation carriers as a kind of control and thereby calculate a BRCA1 with BRCA2 ascertainment-for-BRACAnalysis odds ratio (B1:2 A-OR). That the underlying biological hypothesis is reasonable follows immediately from the observation that the two genes function in the same biochemical pathway and loss of function of the good copy of BRCA1 or BRCA2 is unlikely to be either the initiating or the rate limiting step of tumorigenesis in mutation carriers (for discussion see [31]).

Twelve of the known neutral missense substitutions in BRCA1 have allele frequencies of < 10%. Many of the carriers of one or another of these neutral BRCA1 substitutions also carry a clearly deleterious mutation in BRCA2. If we use this stratum as the reference category, then the B1:2 A-OR for the 4 common neutral missense substitutions in BRCA1 is 0.98 (95% CI 0.9 - 1.1). In contrast, the B1:2 A-OR for all truncating mutations in BRCA1 is 9.2 (95% CI 6.4 - 13.2). The B1:2 A-OR is emphatically not the odds ratio for carriage of a variant of interest in BRCA1, but it may well track with the true odds ratio. **Figure 3** gives B1:2 A-ORs for missense substitutions at positions in the alignment that are invariant; missense substitutions that are outside the range of variation observed at variable positions in the alignment (i.e., GV > 0 and GD > 0); and missense substitutions that are inside the range of variation observed at variable positions in the alignment (i.e., GV > 0 and GD = 0). Two points emerge. First, many of the missense substitutions falling at invariant positions in the alignment must be deleterious and the longer the period over which the position has been invariant, the stronger the evidence that this is so. Second, the pooled evidence is in accord with a hypothesis that missense substitutions that fall at variable positions in the alignment of vertebrate BRCA1s *and are within the range of variation observed at those positions* are neutral. Perhaps surprisingly, even the 36 additional missense substitutions that are brought within the range of variation by the step from frog to pufferfish have a B1:2 A-OR, as a group, of 0.89 (95% CI 0.48 - 1.65). Thus, for the narrow application of validating the con-LR, our results including pufferfish are not substantially different than those that include *Xenopus* but exclude pufferfish. Nevertheless, we note that the B1:2 A-OR for variable position substitutions that are outside of the range of variation at those positions takes a sharp (and possibly important) upswing in the step from *Xenopus* to pufferfish, suggesting that the validity of the analysis might be reaching some sort of limit at this evolutionary distance. If so, then future data from even more distant (e.g., non-vertebrate) genomes might clearly reveal this limit by adding variants that abruptly change the "curves" for the variable sites both within and outside of the range of variation. In view of the objective of providing patients and their health-care advisors with unambiguous guidelines for interpreting the implications of rare BRCA1 variants (as neutral vs. deleterious), we believe that truncating the analysis of GD between *Xenopus* and pufferfish makes the best sense, at least until sequence data from more distant species are analyzed and utilized in these calculations.

Grantham analysis. Of the 15 neutral/ LCS missense substitutions that we analyzed for co-occurrence, 14 fall at positions in the protein that have substantial cross species sequence variability. Twelve of these have a con-LR of ≤0.003 and the other 2 have con-LR=0.025. The remaining neutral variant, P871L, falls at a position that is leucine in all of the other species in the alignment. Eleven of these 15 neutral/ LCS missense substitutions, including P871L, have GDs of 0 in the alignment from human to *Xenopus*. For these 11 substitutions, we conclude that the con-LR is appropriate for inclusion in an integrated analysis (**Table 3**).

Table 3. Alignment/ Grantham analysis of missense substitutions in BRCA1 that have been observed with 2 or more different known deleterious mutations.

A. Known neutral/ LCS missense substitutions

Missense substitution	con-LR	Appropriate yes/ no	Standard Grantham Difference	Grantham Deviations 8 placental Mammals	+Mdom	+Ggal	+Xlae	+Tnig	Hsap	Ptro	Ggor	Ppyg	Mmul	Mmus	Cfam	Btau	Mdom	Ggal	Xlae	Tnig
Q356R	0.025	no	42.8	39.5	39.5	5.1	5.1	0.0	Q	Q	Q	Q	Q	P	Q	Q	Q	K	-	-
D693N	0.003	yes	23.0	2.0	0.0	0.0	0.0	0.0	D	D	D	D	A	A	A	A	V	N	N	H
K820E	< 3E-3	yes	56.9	0.0	0.0	0.0	0.0	0.0	K	K	K	K	E	N	K	K	S	Q	-	-
R841W[o]	0.003	no	101.3	101.3	85.0	85.0	85.0	78.3	R	R	R	Q	Q	Q	Q	Q	S	S	Q	T
Y856H	0.025	yes	83.3	68.0	68.0	68.0	0.0	0.0	Y	Y	Y	Y	Y	Y	C	Y	Y	Y	H	Q
P871L	2.568	yes	97.8	0.0	0.0	0.0	0.0	0.0	P	L	L	L	L	L	L	L	L	L	L	L
E1038G	0.003	yes	97.9	97.9	81.1	81.1	0.0	0.0	E	E	E	E	E	E	E	E	V	R	-	S
S1040N	< 3E-3	no	46.2	45.8	2.0	2.0	2.0	0.0	S	S	S	S	S	G	G	S	D	G	-	K
S1140G	< 3E-3	yes	55.3	0.0	0.0	0.0	0.0	0.0	S	S	S	S	S	G	S	R	T	H	-	-
K1183R	< 3E-3	yes	26.0	0.0	0.0	0.0	0.0	0.0	K	R	R	R	R	R	S	K	Q	K	-	G
R1347G	0.003	yes	125.1	103.5	81.1	0.0	0.0	0.0	R	R	R	R	R	M	R	R	E	-	R	-
S1512I	0.003	no	141.8	113.0	113.0	95.4	50.7	50.7	S	S	S	S	S	G	S	S	S	P	K	E
S1613G	< 3E-3	yes	55.3	28.4	28.4	0.0	0.0	0.0	S	S	S	S	S	A	S	N	T	-	F	-
M1628T[a]	< 3E-3	yes	81.0	0.0	0.0	0.0	0.0	0.0	M	M	M	M	M	V	S	R	R	-	K	-
M1652I	0.003	yes	10.1	0.0	0.0	0.0	0.0	0.0	M	M	M	M	L	M	M	M	L	I	F	L
B. Unclassified missense substitutions																				
Y105C	0.003	no	193.7	191.7	131.3	131.3	131.3	131.3	Y	Y	Y	Y	Y	F	Y	Y	Q	R	Q	L
Y179C	0.025	no	193.7	193.7	193.7	184.1	184.1	0.0	Y	Y	Y	Y	Y	Y	Y	Y	Y	L	F	-
S186Y	2.568	no	143.1	143.1	143.1	143.1	143.1	116.0	S	S	S	S	S	S	S	S	S	S	S	-
L246V	0.025	yes	30.9	30.9	0.0	0.0	0.0	0.0	L	L	L	L	L	L	L	L	V	L	-	A
F486L	< 3E-3	yes	21.8	0.0	0.0	0.0	0.0	0.0	F	F	F	F	L	F	C	F	V	T	L	F
R496H	< 3E-3	yes	28.8	0.0	0.0	0.0	0.0	0.0	R	R	R	R	R	Q	H	Q	H	G	G	N
R504H	0.003	yes	28.8	0.0	0.0	0.0	0.0	0.0	R	R	R	R	R	R	H	C	R	R	S	-
N550H	< 3E-3	no	68.4	65.5	65.5	40.5	13.2	0.0	N	N	N	N	N	S	N	S	N	-	E	R
L668F	0.251	no	21.8	21.8	21.8	21.8	21.3	21.3	L	L	L	L	L	L	L	L	L	L	M	G
V772A	0.251	no	65.3	65.3	65.3	1.0	1.0	1.0	V	V	V	V	V	V	V	V	V	S	D	V
M1008V	< 3E-3	yes	20.5	0.0	0.0	0.0	0.0	0.0	M	M	M	M	V	S	M	T	I	L	N	K
M1008I	< 3E-3	yes	10.1	10.1	0.0	0.0	0.0	0.0	M	M	M	M	V	S	M	T	I	L	N	K
E1060A	0.251	no	106.7	106.7	77.3	77.3	51.8	0.0	E	E	E	E	E	E	E	E	K	E	-	-
E1250K	0.025	yes	56.9	51.6	51.6	34.4	0.0	0.0	E	E	E	E	E	Q	E	E	E	-	K	-
D1546N	0.025	yes	23.0	0.0	0.0	0.0	0.0	0.0	D	D	D	D	D	N	D	D	-	-	D	-
L1564P	0.025	yes	97.8	97.8	0.0	0.0	0.0	0.0	L	L	L	L	L	L	L	L	P	-	L	N
P1637L	0.025	no	97.8	97.8	97.8	63.6	63.6	63.6	P	P	P	P	P	P	P	P	S	E	P	E
D1733G	0.025	no	93.8	75.3	75.3	56.4	56.4	51.7	D	D	D	D	D	E	D	D	D	H	D	L

[o] Classified Neutral/ LCS in Goldgar et al [7].
[a] Classification in Phelan et al. [26] meets the criteria for Neutral/ LCS proposed in Goldgar et al [7].

Species abbreviations are: Hsap, *Homo sapiens*; Ptro, *Pan troglodytes*; Ggor, *Gorilla gorilla*, Ppyg, *Pongo pygmaeus*; Mmul, *Macaca mulatta*; Mmus, *Mus musculus*; Cfam, *Canis familiaris*; Btau, *Bos taurus*; Mdom, *Monodelphis domestica*; Ggal, *Gallus gallus*; Xlae, *Xenopus laevis*; Tnig, *Tetraodon nigroviridis*.

Of the 18 unclassified variants that have co-occurred at least twice with a clearly deleterious BRCA1 mutation, 14 fall at positions in the protein that have substantial cross species sequence variability. Seven have a con-LR of ≤0.003 and 7 have con-LR=0.025. Nine of the 14 missense substitutions with relatively low con-LRs also have GDs of 0 in the alignment from human to *Xenopus*. For these 9 substitutions, we again conclude that the con-LR is appropriate for inclusion in an integrated analysis (**Table 3**).

Integrated analysis. Integration of the co-occurrence data with the alignment/ Grantham data is achieved by simply multiplying the two likelihood ratios, taking into account whether the con-LR was appropriate. Goldgar et al. [7] discussed and then set thresholds for declaring an unclassified variant either deleterious or neutral/ LCS on the basis of the analysis. If the integrated likelihood is >1,000 or <0.01, then the variant is considered deleterious or neutral/ LCS, respectively. If the score is between those two thresholds, the variant remains unclassified. In the present two method analysis, when the con-LR is inappropriate, we consider that only one valid analysis has been done. Similarly, if the observed co-occurrences did not meet our precautionary criteria, we consider that only one valid analysis has been done. As neither of these two cases are an *integrated analysis*, the variant would remain unclassified.

Of the 15 known neutral/ LCS variants that we analyzed, 11 would have been classified as neutral/ LCS by the approach taken here (**Table 4**). The 4 substitutions that would not have been classified neutral/ LCS fail because the human missense substitution is outside the range of variation observed from human to frog.

For 8 of the 18 unclassified variants that we analyzed, we calculated valid likelihood ratios below 0.01. We conclude that these missense substitutions are neutral/ LCS (**Table 4**). The B1:2 A-OR for these 8 substitutions is 0.89 (95% CI 0.55-1.44) also indicating that, in aggregate, this group of variants confers no greater risk of familial breast/ ovarian cancer than do the other missense substitutions that have already been classified neutral/ LCS. The human missense substitutions defining nine of the other unclassified variants fall outside the range of variation observed from human to frog; these remain unclassified. One additional substitution, R504H, met the criterion of GD=0 but remains unclassified because it is almost certainly in cis with the deleterious variant with which it has repeatedly co-occurred. The B1:2 A-OR for this group of 10 substitutions is 1.15 (95% CI 0.54-2.48).

The 8 substitutions that we have classified as neutral/ LCS were observed a total of 346 times in 40,000 tests, 268 times with no clearly deleterious mutation in BRCA1 or BRCA2. However, fewer than 268 unclassified variant reports are resolved because some patients carried more than one unclassified variant. For example, Y179C, F486L, and N550H are often seen together and probably constitute a rare haplotype. One of these three substitutions, F486L met the criteria for neutral/ LCS, but the other two did not. Thus the unclassified variant reports on the 38 individuals who carried all 3 of these substitutions, but no clearly deleterious variant, are not resolved. In the end, we find 202 patient reports that would move from containing a reportable unclassified variant to no reportable variants as a consequence of this analysis.

Table 4. Integrated analysis.

Misssense substitution aa name	nt name	cooc-LR	con-LR	Appropriate yes/ no	Integrated likelihood ratio	Conclusion
A. Known neutral/ LCS missense substitutions						
A. Previously classified missense substitutions						
D693N	2196G>A	< 1.0E-10	0.003	yes	< 1.0E-10	Known neutral/ LCS
K820E	2577A>G	< 1.0E-10	< 3.0E-3	yes	< 1.0E-10	Known neutral/ LCS
Y856H	2685T>C	8.8E-03	0.025	yes	2.2E-04	Known neutral/ LCS
P871L	2731C>T	< 1.0E-10	2.568	yes	< 1.0E-10	Known neutral/ LCS
E1038G	3232A>G	< 1.0E-10	0.003	yes	< 1.0E-10	Known neutral/ LCS
S1140G	3537A>G	2.9E-04	< 3.0E-3	yes	< 8.7E-7	Known neutral/ LCS
K1183R	3667A>G	< 1.0E-10	< 3.0E-3	yes	< 1.0E-10	Known neutral/ LCS
R1347G	4158A>G	< 1.0E-10	0.003	yes	< 1.0E-10	Known neutral/ LCS
S1613G	4956A>G	< 1.0E-10	< 3.0E-3	yes	< 1.0E-10	Known neutral/ LCS
M1628T	5002T>C	1.6E-04	< 3.0E-3	yes	< 4.8E-7	Known neutral/ LCS
M1652I	5075G>A	< 1.0E-10	0.003	yes	< 1.0E-10	Known neutral/ LCS
Q356R	1186A>G	< 1.0E-10	0.025	no	---	Known neutral/ LCS
R841W	2640C>T	1.6E-06	0.003	no	---	Known neutral/ LCS
S1040N	3238G>A	< 1.0E-10	< 3.0E-3	no	---	Known neutral/ LCS
S1512I	4654G>T	< 1.0E-10	0.003	no	---	Known neutral/ LCS
B. Unclassified missense substitutions						
L246V	855T>G	2.6E-05	0.025	yes	6.5E-07	Neutral/ LCS
F486L	1575T>C	1.2E-02	< 3.0E-3	yes	< 3.6E-05	Neutral/ LCS
R496H	1606G>A	9.2E-05	< 3.0E-3	yes	< 2.7E-07	Neutral/ LCS
M1008V	3141A>G	< 1.0E-10	< 3.0E-3	yes	< 1.0E-10	Neutral/ LCS
M1008I	3143G>A	9.6E-07	< 3.0E-3	yes	< 2.9E-09	Neutral/ LCS
E1250K	3867G>A	1.0E-05	0.025	yes	2.6E-07	Neutral/ LCS
D1546N	4755G>A	5.6E-03	0.025	yes	1.4E-04	Neutral/ LCS
L1564P	4810T>C	3.3E-03	0.025	yes	8.3E-05	Neutral/ LCS
Y105C	433A>G	1.1E-05	0.003	no	---	Unclassified
Y179C	655A>G	1.2E-02	0.025	no	---	Unclassified
S186Y	676C>A	2.88	2.568	no	---	Unclassified
R504H	1630G>A	1.28	0.003	yes	3.8E-03	Unclassified
N550H	1767A>C	1.2E-02	< 3.0E-3	no	---	Unclassified
L668F	2121C>T	1.28	0.251	no	---	Unclassified
V772A	2434T>C	9.4E-06	0.251	no	---	Unclassified
E1060A	3298A>C	1.00	0.251	no	---	Unclassified
P1637L	5029C>T	1.23	0.025	no	---	Unclassified
D1733G	5317A>G	1.00	0.025	no	---	Unclassified

DISCUSSION.

We conclude that 8 currently unclassified missense substitutions in BRCA1, L246V, F486L, R496H, M1008V, M1008I, E1250K, D1546N, and L1564P, should be considered neutral/ LCS. Our approach to classifying these missense substitutions might seem unusual to some because we used neither segregation analysis nor association study nor functional assay. Instead, we used two methods, co-occurrence with clearly deleterious mutations in BRCA1 and A-GVGD. The analysis was carried out under the assumption that neither method is inerrant.

Co-occurrence is really a test for embryonic lethality due to inheritance of a compound heterozygous null genotype. The underlying assumption is that inheritance of a genuine high-risk missense substitution in BRCA1, along with a clearly deleterious mutation, will either lead to death during embryogenesis or a severe phenotype such as Fanconi's anemia. Internal statistics of the BRACAnalysis database provide support for this hypothesis. But it is difficult to exclude the possibility that some BRCA1 genotype combinations of interest could result in a high cancer risk with no other obvious phenotype. The method also has the pitfall that the clearly deleterious mutation and the sequence variant of interest must be in trans for the analysis to be valid. On this latter point we have taken the cautious approach of only analyzing missense substitutions for which at least two co-occurrences have been observed and, when the number of distinct co-occurrences was very small, confirming by SNP haplotype analysis that at least one was in trans.

In order to measure the range of sequence variation that has occurred at specific residues in BRCA1 during vertebrate evolution and the fit between observed missense substitutions and that range of variation, we extended the concept of Grantham scores to multiple sequence alignments. The method, A-GVGD, can be used to identify sets of missense substitutions that are either enriched for deleterious variants or enriched for neutral variants. However, A-GVGD does not account for the possibility that sequence variation that has been permissible during the evolution of BRCA1 in one group of non-human vertebrates is not permissible in human BRCA1. It also does not take into account that the nucleotide substitution underlying a missense variant may interfere with mRNA splicing or have some other deleterious effect at the level of DNA or RNA.

In integrating these methods, we required that the cooc-LR is <1 and that GD=0. Thus we only make the classification in those 8 cases where combined evidence from co-occurrence data, variability in the sequence alignment, and GD, support the conclusion neutral/ LCS. Essentially, we have used these methods in such a way that each complements the potential for error in the other. The integrated likelihoods for each of these 8 missense substitutions are two or more orders of magnitude below the threshold of 0.01 set by Goldgar et al. [7], thus our confidence in the classification, in each of these cases, is very strong. There were 10 unclassified substitutions that we analyzed closely and did not classify as neutral/ LCS. The B1:2 A-OR for this set of substitutions (1.15) was not much higher than that for the group that we did classify (0.89). There is no substantial evidence that there is more risk associated with the 10 substitutions that we did not classify neutral/ LCS than the ones that we did; rather, that these 10 variants remain unclassified may be viewed as an example of how applying very strict criteria in order to minimize type I error can reduces our power to classify substitutions that may well be neutral.

In the 40,000 test data set there remain 39 unclassified missense substitutions that have been observed to co-occur one time with a clearly deleterious variant. Nineteen of these have GV=0 in the alignment from human to *Xenopus*; consequently, with the pair of methods used here, we would expect to be able to reclassify many of these as neutral/LCS. However, there also remain 164 unclassified missense substitutions with GD=0. As the B1:2 A-OR for this pool of substitutions is 0.91 (95% CI 0.60 - 1.39), the vast majority of these must be neutral/ LCS. That only 19 of these co-occurred with a clearly deleterious mutation underscores the point that further clinical reclassification of superficially innocuous BRCA1 missense variants will require integrated application of other methods such as pooled family histories, segregation analysis, or validated functional assays. At this point, however, BRCA2 is creating a noticeably larger burden of unclassified missense substitutions than BRCA1. Thanks to model organism genome sequencing, sufficient genome sequences are available to allow the straightforward gene model creation, testing, and correction required to compile a BRCA2 protein multiple sequence alignment comparable to the BRCA1 alignment used here. As the same analysis model appears to apply to BRCA2, this should lead to reclassification of a considerable number of missense substitutions followed by a substantial reduction in the burden of tests that include a reportable unclassified variant.

ACKNOWLEDGEMENTS AND AFFILIATIONS.

This research was supported in part by the INHERIT BRCAs program from the Canadian Institute for Health Research, grant 7792 (to SVT) from the Association pour la Recherche sur le Cancer, and grant RR14214 (to PBS) from the US National Institutes of Health. None of the funding agencies had any influence on the direction of this work.

Competing interests. AMD, TJ, TS, and AZ are employees of Myriad Genetics, Inc., or Myriad Genetic Laboratories, Inc. SVT and AT own stock in Myriad Genetics, Inc. Results from this work bear on Myriad Genetic Laboratories' commercial test for mutations in BRCA1 and BRCA2.

FIGURE LEGENDS.

Figure 1. Sample calculations of GV and GD using M1652I as an example. (**a**) Thirty amino acid sequence alignment from the beginning of the BRCA1 BRCT domain. Position 1652 is marked in grey. (**b**) For each amino acid aligned to human position 1652, we give the amino acid sidechain composition (*C*), polarity (*P*), and volume (*V*). Starting with human and working sequentially to the *T. nigroviridis* sequence, we determine the minimum and maximum observed values of *C*, *P*, and *V*; these are used to calculate GV according to eq 2. Starting with human and working sequentially to the *T. nigroviridis* sequence, we then determine the difference between *C*, *P*, and *V* for the missense variant (Ile in this case; *C*, *P*, and *V* values for Ile are marked in grey) and the range of variation observed to that point in the alignment. The differences, given in the columns headed ΔC, ΔP, and ΔV, are used to calculate GD according to eq 2. (**c**) Sample calculation of GV. (**d**) Sample calculation of GD.

Figure 2. A-GVGD analysis of missense substitutions in BRCA1. Excluding missense substitutions within 2 bp of a splice junction, there were a total of 438 unique BRCA1 missense substitutions in the B1&2 40K set. Using the BRCA1 protein multiple sequence alignment and Grantham Deviations (GDs), we have divided these into four data series. The four data series are: filled squares - missense substitutions coded as deleterious, but excluding those that fall within 2 bp of a splice junction; open squares - missense substitutions coded as deleterious, but excluding those at the initiator methionine, those that alter a canonical C3HC4 cysteine residue, and those that fall within 2 bp of a splice junction; closed circles - unclassified missense substitutions, but excluding those that fall within 2 bp of a splice junction; closed triangles - substitutions that are coded as neutral/ LCS. In this figure, the X-axis represents the depth of alignment at which each analysis was made. None = no alignment (i.e., standard Grantham Differences); 8 P M=alignment of 8 placental Mammals; + Mdom = addition of *Monodelphis domestica* (opossum); +Ggal = addition of *Gallus gallus* (chicken); +Xlae = addition of *Xenopus laevis* (frog); +Tnig = addition of *Tetraodon nigroviridis* (pufferfish). Error bars give +/- 1 standard deviation of GD except that some lower error bars are cut off at 0 because GDs cannot take a negative value. Error bars for the unclassified variant data series (closed circles) were omitted because they would tend to obscure evidence that the GDs of known neutral variants are resolved from those of known deleterious variants.

Figure 3. Ascertainment risk associated with groups of missense substitutions in BRCA1. Excluding missense substitutions at the initiator methionine, within 2 bp of a splice junction, and those already classified as neutral/ LCS, there were a total of 418 unique BRCA1 missense substitutions in the B1&2 40K set. Using the BRCA1 protein multiple sequence alignment and A-GVGD, we have divided these into three data series. The three data series are: closed triangles - substitutions that fall at variable positions in the alignment and are within the observed range of variation to that layer in the alignment (GV>0, GD=0); closed circles - substitutions that fall at variable positions in the alignment and are outside of the observed range of variation to that layer in the alignment (GV>0, GD>0); closed squares - substitutions that fall at positions that are invariant to that layer in the alignment (GV=0, GD>0). In this figure, the X-axis represents the depth of alignment at which each analysis was made. 8 P M = alignment of 8 placental Mammals; + Mdom = addition of *M. domestica*; +Ggal = addition of *G. gallus*; +Xlae = addition of *X. laevis*; +Tnig = addition of *T. nigroviridis*. The reference set (B1:2 A-OR = 1.0) was the group of 12 known neutral missense substitutions in BRCA1 that have allele frequencies of < 10%. Error bars on the closed circle and closed triangle data series are the 95% confidence interval adjusted for a three degree of freedom 2-sided test; error bars for the unclassified variant data series (closed circles) were omitted because they would tend to obscure evidence that the B1:2 A-ORs of known neutral variants are resolved from those of known deleterious variants.

REFERENCES.

1. Miki Y, Swensen J, Shattuck-Eidens D, Futreal PA, Harshman K, Tavtigian S, Liu Q, Cochran C, Bennett LM, Ding W, Bell R, Rosenthal J, Hussey C, Tran T, McClure M, Frye C, Hattier T, Phelps R, Haugen-Strano A, Katcher H, Yakumo K, Gholami Z, Shaffer D, Stone S, Bayer S, Wray C, Bogden R, Dayananth P, Ward J, Tonin P, Narod S, Bristow PK, Norris FH, Helvering L, Morrison P, Rosteck P, Lai M, Barrett JC, Lewis C, Neuhausen S, Cannon-Albright L, Goldgar D, Wiseman R, Kamb A, Skolnick MH. A strong candidate for the breast and ovarian cancer susceptibility gene BRCA1. *Science* 1994;**266**:66-71.
2. Wooster R, Bignell G, Lancaster J, Swift S, Seal S, Mangion J, Collins N, Gregory S, Gumbs C, Micklem G. Identification of the breast cancer susceptibility gene BRCA2. *Nature* 1995;**378**:789-792.
3. Tavtigian SV, Simard J, Rommens J, Couch F, Shattuck-Eidens D, Neuhausen S, Merajver S, Thorlacius S, Offit K, Stoppa-Lyonnet D, Belanger C, Bell R, Berry S, Bogden R, Chen Q, Davis T, Dumont M, Frye C, Hattier T, Jammulapati S, Janecki T, Jiang P, Kehrer R, Leblanc JF, Goldgar DE, et a. The complete BRCA2 gene and mutations in chromosome 13q-linked kindreds. *Nat Genet* 1996;**12**:333-337.
4. Antoniou A, Pharoah PD, Narod S, Risch HA, Eyfjord JE, Hopper JL, Loman N, Olsson H, Johannsson O, Borg A, Pasini B, Radice P, Manoukian S, Eccles DM, Tang N, Olah E, Anton-Culver H, Warner E, Lubinski J, Gronwald J, Gorski B, Tulinius H, Thorlacius S, Eerola H, Nevanlinna H, Syrjakoski K, Kallioniemi OP, Thompson D, Evans C, Peto J, Lalloo F, Evans DG, Easton DF. 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* 2003;**72**:1117-1130.
5. Thompson D, Easton D. The genetic epidemiology of breast cancer genes. *J Mammary Gland Biol Neoplasia* 2004;9:221-236.
6. Frank TS, Deffenbaugh AM, Reid JE, Hulick M, Ward BE, Lingenfelter B, Gumpper KL, Scholl T, Tavtigian SV, Pruss DR, Critchfield GC. Clinical characteristics of individuals with germline mutations in BRCA1 and BRCA2: analysis of 10,000 individuals. *J Clin Oncol* 2002;**20**:1480-1490.
7. Goldgar DE, Easton DF, Deffenbaugh AM, Monteiro ANA, Tavtigian SV, Couch FJ, Committee TBCICBICS. Integrated evaluation of DNA sequence variants of unknown clinical significance: Application to BCA1 and BRCA2. *Am J Hum Genet* 2004;**75**:535-544.
8. Gowen LC, Johnson BL, Latour AM, Sulik KK, Koller BH. Brca1 deficiency results in early

embryonic lethality characterized by neuroepithelial abnormalities. *Nat Genet* 1996;**12**:191-194.

9. Liu CY, Flesken-Nikitin A, Li S, Zeng Y, Lee WH. Inactivation of the mouse Brca1 gene leads to failure in the morphogenesis of the egg cylinder in early postimplantation development. *Genes Dev* 1996;**10**:1835-1843.
10. Hakem R, de la Pompa JL, Sirard C, Mo R, Woo M, Hakem A, Wakeham A, Potter J, Reitmair A, Billia F, Firpo E, Hui CC, Roberts J, Rossant J, Mak TW. The tumor suppressor gene Brca1 is required for embryonic cellular proliferation in the mouse. *Cell* 1996;**85**:1009-1023.
11. Ludwig T, Chapman DL, Papaioannou VE, Efstratiadis A. Targeted mutations of breast cancer susceptibility gene homologs in mice: lethal phenotypes of Brca1, Brca2, Brca1/Brca2, Brca1/p53, and Brca2/p53 nullizygous embryos. *Genes Dev* 1997;**11**:1226-1241.
12. Hohenstein P, Kielman MF, Breukel C, Bennett LM, Wiseman R, Krimpenfort P, Cornelisse C, van Ommen GJ, Devilee P, Fodde R. A targeted mouse Brca1 mutation removing the last BRCT repeat results in apoptosis and embryonic lethality at the headfold stage. *Oncogene* 2001;**20**:2544-2550.
13. Joukov V, Chen J, Fox EA, Green JB, Livingston DM. Functional communication between endogenous BRCA1 and its partner, BARD1, during Xenopus laevis development. *Proc Natl Acad Sci U S A* 2001;**98**:12078-12083.
14. Xu X, Aprelikova O, Moens P, Deng CX, Furth PA. Impaired meiotic DNA-damage repair and lack of crossing-over during spermatogenesis in BRCA1 full-length isoform deficient mice. *Development* 2003;**130**:2001-2012.
15. Frank TS, Manley SA, Olopade OI, Cummings S, Garber JE, Bernhardt B, Antman K, Russo D, Wood ME, Mullineau L, Isaacs C, Peshkin B, Buys S, Venne V, Rowley PT, Loader S, Offit K, Robson M, Hampel H, Brener D, Winer EP, Clark S, Weber B, Strong LC, Thomas A, et a. Sequence analysis of BRCA1 and BRCA2: correlation of mutations with family history and ovarian cancer risk. *J Clin Oncol* 1998;**16**:2417-2425.
16. Beaudet AL, Tsui LC. A suggested nomenclature for designating mutations. *Hum Mutat* 1993;**2**:245-248.
17. Hendrickson BC, Pruss D, Lyon E, Scholl T. Application of haplotype pair analysis for the identification of hemizygous loci. *J Med Genet* 2003;**40**:346-347.
18. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res* 1997;**25**:3389-3402.
19. Abkevich V, Zharkikh A, Deffenbaugh A, Frank D, Chen Y, Shattuck D, Skolnick MH, Gutin A, Tavtigian SV. Analysis of missense variation in human BRCA1 in the context of interspecific sequence variation. *J Med Gen* 2004;**41**:492-507.
20. Poirot O, Suhre K, Abergel C, O'Toole E, Notredame C. 3DCoffee@igs: a web server for combining sequences and structures into a multiple sequence alignment. *Nucleic Acids Res* 2004;**32**:W37-40.
21. Brzovic PS, Rajagopal P, Hoyt DW, King MC, Klevit RE. Structure of a BRCA1-BARD1 heterodimeric RING-RING complex. *Nat Struct Biol* 2001;**8**:833-837.
22. Williams RS, Green R, Glover JN. Crystal structure of the BRCT repeat region from the breast cancer-associated protein BRCA1. *Nat Struct Biol* 2001;**8**:838-842.
23. Shiozaki EN, Gu L, Yan N, Shi Y. Structure of the BRCT repeats of BRCA1 bound to a BACH1 phosphopeptide: implications for signaling. *Mol Cell* 2004;**14**:405-412.
24. Fitch WM, 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* 1970;**4**:579-593.
25. Grantham R. Amino acid difference formula to help explain protein evolution. *Science* 1974;**185**:862-864.
26. Phelan CM, Vesna A, Tice B, Favis R, Kwan E, Barany F, Manoukian S, Radice P, van der Luijt RB, van Nesselrooij BPM, Chenevix-Trench G, kConFab Caldes T, de la Hoya M, Lindquist S, Tavtigian

SV, Goldgar D, Borg A, Narod SA, Monteiro ANA. Classification of BRCA1 missense variants of unknown clinical significance. *Journal of Medical Genetics* 2005;**42**:138-146.

27. Ng PC, Henikoff S. SIFT: Predicting amino acid changes that affect protein function. *Nucleic Acids Res* 2003;**31**:3812-3814.
28. Ramensky V, Bork P, Sunyaev S. Human non-synonymous SNPs: server and survey. *Nucleic Acids Res* 2002;**30**:3894-3900.
29. Vitkup D, Sander C, Church GM. The amino-acid mutational spectrum of human genetic disease. *Genome Biol* 2003;**4**:R72.
30. Greenblatt MS, Beaudet JG, Gump JR, Godin KS, Trombley L, Koh J, Bond JP. Detailed computational study of p53 and p16: using evolutionary sequence analysis and disease-associated mutations to predict the functional consequences of allelic variants. *Oncogene* 2003;**22**:1150-1163.
31. Venkitaraman AR. Cancer susceptibility and the functions of BRCA1 and BRCA2. *Cell* 2002;**108**:171-182.

Figure 1

a

Hsap	1646	V	N	K	R	M	S	M	V	V	S	G	L	T	P	E	E	F	M	L	V	Y	K	F	A	R	K	H	H	I	T
Ptro	1646	V	N	K	R	M	S	M	V	V	S	G	L	T	P	E	E	F	M	L	V	Y	K	F	A	R	K	H	H	I	T
Mmul	1646	V	N	K	R	M	S	L	V	V	S	G	L	T	P	E	E	F	M	L	V	Y	K	F	A	R	R	Y	H	I	A
Mmus	1589	A	D	R	D	I	S	M	V	V	S	G	L	T	P	K	E	V	M	T	V	Q	K	F	A	E	K	Y	R	L	T
Cfam	1649	V	N	K	R	I	S	M	V	A	S	G	L	T	P	K	E	F	M	L	V	H	K	F	A	R	K	H	H	I	S
Btau	1639	S	K	K	R	L	S	M	V	A	S	G	L	T	P	K	E	L	M	L	V	Q	K	F	A	R	K	H	H	V	T
Mdom	1623	G	N	R	K	I	S	L	V	S	S	G	L	T	P	K	E	N	M	L	V	Q	K	F	A	R	K	T	H	S	T
Ggal	1537	C	R	T	E	M	S	I	V	A	S	G	L	N	Q	S	E	H	L	M	V	Q	K	F	A	R	K	T	Q	S	T
Xlav	1370	S	R	R	N	L	S	F	V	A	S	G	L	N	Q	C	E	M	A	L	V	Q	R	F	S	K	T	T	Q	S	I
Tnig	1058	S	L	A	R	M	L	L	V	T	S	G	L	G	P	S	Q	Q	I	T	V	K	K	F	A	K	R	I	G	A	T

b

		C	P	V	Cmax	Cmin	Pmax	Pmin	Vmax	Vmin	**GV**	ΔC	ΔP	ΔV	**GD**
Hsap	M	0.0	5.7	105	0.0	0.0	5.7	5.7	105	105	**NA**	0.0	-0.5	6.0	**10.1**
Ptro	M	0.0	5.7	105	0.0	0.0	5.7	5.7	105	105	**0.0**	0.0	-0.5	6.0	**10.1**
Mmul	L	0.0	4.9	111	0.0	0.0	5.7	4.9	111	105	**14.3**	0.0	0.0	0.0	**0.0**
Mmus	M	0.0	5.7	105	0.0	0.0	5.7	4.9	111	105	**14.3**	0.0	0.0	0.0	**0.0**
Cfam	M	0.0	5.7	105	0.0	0.0	5.7	4.9	111	105	**14.3**	0.0	0.0	0.0	**0.0**
Btau	M	0.0	5.7	105	0.0	0.0	5.7	4.9	111	105	**14.3**	0.0	0.0	0.0	**0.0**
Mdom	L	0.0	4.9	111	0.0	0.0	5.7	4.9	111	105	**14.3**	0.0	0.0	0.0	**0.0**
Ggal	I	0.0	5.2	111	0.0	0.0	5.7	4.9	111	105	**14.3**	0.0	0.0	0.0	**0.0**
Xlav	F	0.0	5.2	132	0.0	0.0	5.7	4.9	132	105	**30.3**	0.0	0.0	0.0	**0.0**
Tnig	L	0.0	4.9	111	0.0	0.0	5.7	4.9	132	105	**30.3**	0.0	0.0	0.0	**0.0**

c

GV for the set {M,L,I,F} = $50.723 \times \sqrt{\left[1.833(0-0)^2\right] + \left[0.1018(5.7-4.9)^2\right] + \left[0.000399(132-105)^2\right]} = 30.3$

d

GD for I vs M = $50.723 \times \sqrt{\left[1.833(0-0)^2\right] + \left[0.1018(5.7-5.2)^2\right] + \left[0.000399(111-105)^2\right]} = 10.1$

Figure 2

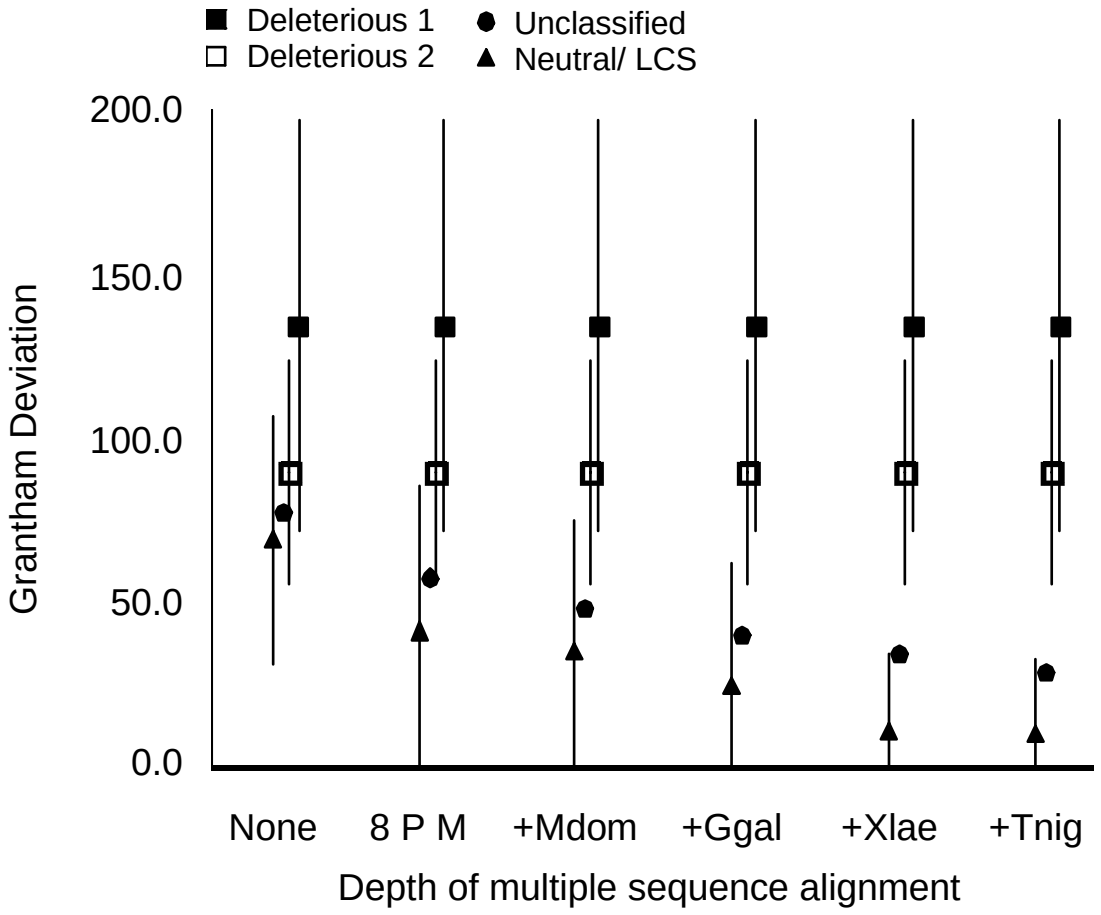

Figure 3