SUPPLEMENTARY APPENDIX

This supplementary appendix contains two sections: a supplementary methods section that expands on some of the items in the main manuscript Methods section and a supplementary figures section that contains two color figures that were referred to in the Discussion.

SUPPLEMENTARY METHODS.

In the following Supplementary Methods, the section on the Ascertainment Ratio (AR) is preceded by the definition of the reference set of neutral (or nearly so) missense substitutions required as its reference category. The section on the Enrichment Ratio for Single Nucleotide Substitutions (ERS) provide slightly more detail than did it counterpart in the main text document. The sections on Gaussian Smoothing, risk surface calculations, selection of contour curves, and bootstrap confidence interval estimations are much more detailed than their counterparts in the main text document.

If we think of missense substitutions as points on a graph with GV as the X-axis and GD as the Y-axis, the logic underlying Align-GVGD dictates that deleterious substitutions should either map near the Y-axis (substitutions at positions with low GV), into the upper portions of the graph (substitutions with high values of GD), or both (low GV and high GD). In contrast, neutral substitutions should either map to the right edge of the graph (substitutions at positions with high GV), into the lower portions of the graph (substitutions with low values of GD), or both (high GV and low GD). However, we have not previously made a systematic exploration of the distribution of evidence of genetic risk in the GV-GD plane.

3. RISK ESTIMATIONS

Missense substitutions

Starting with all of the missense substitutions observed during full-sequence mutation screening of *BRCA1* and *BRCA2* from a series of 68,000 subjects, we focused on those observed in the *BRCA1* RING, *BRCA1* BRCT repeat, and *BRCA2* DNA binding domains because these are the only domains of these proteins where some missense substitutions have already been proven to confer relatively high risk of breast cancer due to missense-induced effects *per se* [Goldgar et al., 2004; Easton et al., 2007]. We excluded the substitutions observed at M1 (because they are likely deleterious due to interference with translation initiation independent of any missense effect *per se*) and those observed at the 9 canonical C3HC4 RING residues (because substitutions at cysteine have very high Grantham Differences, inclusion of which could bias our results towards very high values of GD in a way that might only be characteristic of the minority of proteins that are functionally dependent on multi-cysteine motifs). After these exclusions, our data set contained a total of 453 missense substitutions. Only one of these, the known neutral variant *BRCA1* M1652I, had a carrier frequency above 1% (2.8%). 225 of the substitutions were observed exactly once each.

As will become clear below, calculation of the AR requires a reference set of missense

substitutions that are neutral or very nearly so. The reference set of missense substitutions that we used were extracted from the B1&2 68K set via the following 4 criteria:
i) located outside of the RING, BRCT, and DBD domains;
ii) underlying nucleotide substitution does not fall on the last two nucleotides of an exon (to exclude likely splice variants);
iii) frequency in the B1&2 68K set <1% (i.e., rare variants);
iv) GD=0 based on the *BRCA1* or *BRCA2* alignment from human to frog (i.e., variants are within the observed cross-species range of variation).
These are more stringent criteria for likely neutral substitutions than those which we have used previously to identify substitutions that are neutral or of little clinical significance [Tavtigian et al., 2006], and, on the basis of Easton et al's FamHx-LR calculations, the fraction of these that are expected to be high-risk variants is 0.00 [Easton et al., 2007]. 189 BRCA1 missense substitutions and 365 BRCA2 missense substitutions from the B1&2 68K set met these criteria.

Note that the data available to us report only the number of each variant seen in the *BRCA1/2*/+/- groups. Thus, an individual carrying two variants in M, or a variant in both M and the neutral reference set, will be double counted. However, as we are analyzing only rare variants, such individuals will be rare; any bias that such double counting introduces will be diluted across the unconfounded genotypes of many other individuals in the B1&2 68K set and should thus be negligible.

Ascertainment ratio

Consider those individuals in the B1&2 68K test series who carry no reportable high-risk variants in *BRCA1* but who do carry *BRCA1* missense variants in some potentially deleterious set M. Let the number of these who have a clearly high-risk mutation in *BRCA2* be a_1 and those who do not be b_1. Under the rationale that the high-risk *BRCA2* variants carried by the a_1 subjects largely explains their presence in the sample series, the allele frequencies for deleterious *BRCA1* variants in these subjects should be closer to population allele frequencies than will be the allele frequencies for deleterious *BRCA1* variants in the b_1 subjects. Thus the a_1 subjects can be thought of as pseudo-controls, the b_1 subjects can be thought of as pseudo-cases, and the ratio b_1/a_1 is an (non-normalized) estimate of the odds for breast cancer for a carrier of a missense substitution in M. Comparing this ratio with the analogous quantity d_1/c_1 for a clearly benign set of sequence variants, the reference set, gives us the AR for *BRCA1*, which is an estimate of the odds ratio for carriers of *BRCA1* missense substitutions in M [Tavtigian et al., 2006]. For calculation of the AR for *BRCA2* define similarly a_2, b_2, c_2, and d_2.

Eq 1
$$AR(M_{B1}) = \frac{b_1}{a_1} \times \frac{c_1}{d_1} \qquad AR(M_{B2}) = \frac{b_2}{a_2} \times \frac{c_2}{d_2}$$

If the AR is used to estimate the odds ratio for N mutually exclusive sets of missense substitutions, then the data can be arrayed and analyzed in a standard Nx2 contingency table. The AR should be useful for estimating the odds ratios for pools of rare missense substitutions observed in pairs or sets of genes that meet the following 3 criteria: (1) detrimental variants in both genes confer similar phenotypes, (2) the odds ratio for carrying deleterious variants in both genes is sub-multiplicative, and (3) both genes are completely mutation screened as a routine part of the testing process.

Enrichment Ratio for Single nucleotide substitutions

For each nucleotide in a canonical DNA sequence, there are three possible single nucleotide substitutions. However, these substitutions are not equally likely to occur because of differences in the underlying substitution rate constants. Using the dinucleotide substitution rate constants given by Lunter and Hein [2004], averaging sense and anti-sense orientations, we can calculate a relative substitution rate for every possible single nucleotide substitution to a DNA sequence, r_i. In a protein coding sequence, each of the possible single nucleotide substitutions can also be classified as silent, nonsense, or missense. The probability that a new mutation (i.e., a new germline sequence variant at the moment that it comes into existence) will fall into a particular class C is given by:

Eq 2 $$p_c = \sum_{i \in c} r_i \Big/ \sum_{all\_i} r_i$$

Hence, under the null hypothesis of no selection, we can obtain from the total number of variants observed in a mutation screening study, o_T, the number expected in any class, $e_C = p_C o_T$, and compare this to the actual number observed, o_C. A missense substitution classifier will stratify the missense variants into subsets that may be enriched for deleterious substitutions or enriched for neutral substitutions, depending on the definition of the classifier. We can thus obtain the following diagram as a basis for evaluating the classifier:

Class	Observed	Expected
Silent	O_S	$e_S=p_So_T$
Nonsense	O_N	$e_N=p_No_T$
Missense 1	O_{M1}	$e_{M1}=p_{M1}o_T$
Missense 2	O_{M2}	$e_{M2}=p_{M2}o_T$
…..	…..	…..
Missense n	O_{Mn}	$e_{Mn}=p_{Mn}o_T$

From this diagram we define the ERS for the missense substitutions in a set M as the ratio of observed to expected counts for M, divided by the same ratio for silent substitutions, to normalize for potential variation in overall substitution rates within a given gene. Note that O_T cancels, leaving the ratio of observed substitutions in M to the probability of M, normalized by same ratio for silent substitutions.

Eq 3 $$ERS(M) = \frac{o_M}{e_M} \Big/ \frac{o_S}{e_S} = \frac{o_M}{p_M} \Big/ \frac{o_S}{p_S}$$

Note that the ERS can be calculated for nonsense substitutions, which can provide a clear loss-of-function standard; indeed, the ERS for the combined pool of all nonsense substitutions observed in the B1&2 68K set is 6.25. The ERS can also be calculated for other definable sets of substitutions such as specific subsets of splice junction variants as long as these are predicted by an algorithm that can select the required set from a list of all possible substitutions in the gene of interest. Thus the ERS could be used to explore the efficacy of other missense substitution analysis programs, splice junction mutation prediction algorithms, exonic splice enhancer mutation prediction algorithms, etc.

4. DISTRIBUTION OF GENETIC RISK WITHIN THE GV-GD PLANE

While there are only a finite number of points in the GV-GD plane that can be occupied depending on the specific amino acids seen in the alignment and the set of observed and possible substitutions, it is useful to envision the AR and ERS as smooth two dimensional surfaces. We constructed such surfaces using Gaussian smoothing of the observed AR and ERS data and combined them to give a single joint risk estimate for any point in the quadrant GV, GD ≥0. These surfaces are shown as heat maps in the figures detailed below. In order to define ordered categories (grades) of risk, we drew boundary lines of the form $GD = GD_0 + \tan(a)\ GV^b$, $b>0$ and $0< a < \pi/2$, which approximately follow the contours of the surface while still being a simple, convenient functional form. We used bootstrapping to evaluate the robustness of our surface estimates to variation in the data. Full details of these procedures are given in the following section, Risk Surface Calculations.

Risk Surface Calculations

To visualize the distribution of evidence of genetic risk among *BRCA1* and *BRCA2* missense substitutions, we calculated GV and GD for each substitution at each of three depths of alignment: frog, pufferfish, and sea urchin. We then created a grid of points in the GV-GD plane and calculated two measurements of genetic risk, the AR and the ERS, at each point on the grid. These risk estimates used a log-distance Gaussian weighting approach to load the required observational data (number of observations in pseudocases, number of observations in pseudocontrols, underlying dinucleotide substitution rate constants) onto each grid point before the AR and ERS for that grid point were calculated. Point risk estimates were then displayed as heat maps.

We constructed a square grid with, starting at (0,0), points at intervals of 5 GV or GD units (which are on the same scale because of the underlying definitions [Tavtigian et al., 2006]) from 0-100 and at intervals of 10 units from 110-280 (GV) or 110-220 (GD). The upper-right portion of this grid is very sparsely populated. All but 0.2% of coordinates at which substitutions can occur were within the subjectively fitted segment of an ellipse, described by $0.55GV^2 + (GD+10)^2 = 44{,}000$. The 6 most extreme observed substitutions lie very nearly along the straight line GD= -0.568(GV) + 212 (r^2=0.99). To avoid unwarranted extrapolation and to improve computation efficiency, we bounded the grid to include all observed substitutions and allow some limited extrapolation to possible substitutions. This was done by setting the boundary 1/3 of the way from the linear limit of observed substitutions toward the elliptical limit of possible substitutions, measured along the GD direction.

To achieve an approximately uniform distribution of possible missense substitutions in the GV-GD grid, we log-transformed the GV and GD distances. These were first offset from zero by 10, which was chosen subjectively to give a satisfactory dispersion of data points. Thus, given a point i with coordinates GV_i,GD_i and a missense substitution j with GV and GD values (at some depth of P-MSA) giving it coordinates GV_j and GD_j, we calculate the distance D_{ij} between their log transformed positions as

Eq 4 $$D_{ij} = \sqrt{\left(\ln(10+GV_i) - \ln(10+GV_j)\right)^2 + \left(\ln(10+GD_i) - \ln(10+GD_j)\right)^2}$$

These distances were then used to calculate a two-dimensional Gaussian weighting factor G_{ij}

Eq 5 $$G_{ij} = \frac{e^{-D_{ij}^2 / 2\sigma^2}}{2\pi\sigma^2}$$

The Gaussian weighting factor requires standard deviations. After exploring some preliminary heat maps, we settled on a standard deviation for the ERS of 0.3 and for the AR of 0.6. These were determined empirically as the lowest standard deviations that prevented individual grid values from markedly exceeding the directly calculated values for the pool of presumably high-risk variants near the GV=0 axis. Note that in practice the denominator need not be included in the calculation because the AR and ERS are both ratios of ratios, so the denominator will cancel out.

The data for all of the substitutions in the data set were then loaded onto every grid point i as

Eq 6 $$DATA_i = \sum_{all.j} G_{ij} \times DATA_j$$

The *BRCA1* AR, *BRCA2* AR, and ERS for every grid point were then calculated according to equations 1, 2, and 3. Thus observational data from all 453 BRCA1 RING/ BRCT and BRCA2 DBD missense substitutions recorded in the B1&2 68K set (after excluding BRCA1 M1 and the 8 canonical C3HC4 residues of the BRCA1 RING motif) contribute to the AR calculations, and these plus all 9,234 possible missense substitutions that can result from single nucleotide substitutions to the same gene segments (with the same exclusion as above) contribute to the ERS calculations at every grid point.

We then took the weighted geometric mean of the AR and the ERS to create a joint risk estimate. The AR estimates based on *BRCA2* data were assigned twice the weight of those based on the *BRCA1* data because the number of *BRCA2* sequence variants is 2x the number of B1 sequence variants. ERS estimates were given twice the weight of the combined AR estimates, again reflecting the approximate relative amount of data. Hence the final weighted joint risk estimates were ($AR_{BRCA1}^{0.11}$)x($AR_{BRCA2}^{0.22}$)x($ERS^{0.67}$). Joint risk estimates for each grid point were calculated at three depths of P-MSA (through frog, pufferfish, and sea urchin), yielding 3 heat maps.

Contour curves

To place portable gradations of genetic risk on the heat maps, we abstracted visible contour of the maps into simple equations. From inspection of the heat maps, we concluded that fractional monomials of the form

$GD = GD_0 + \tan(a)\ GV^b$, $b>0$ and $0< a < \pi/2$.

should fit the observed heat map contours reasonably well.

Specific contours were selected in 5 steps as follows:

(i) external reasoning was used to specify fixed values of GD_0, specifically GD_0=65, 55, 45, 35, 25, and 15;
(ii) values of the parameters *a* and *b* were varied by systematic grid search;
(iii) using the Gaussian weighting approach described above, joint risk estimates were sampled at 20 points spaced along each candidate curve at three depths of P-MSA (through frog, pufferfish, and sea urchin);
(iv) the variance from the joint risk estimate at GD_0 was calculated for each curve at each of the three depths of alignment, and the results averaged;
(v) preferred curves were those with the lowest mean variance. However if more than one candidate curve was within a tolerance of 2.5×10^{-5}, the curve with lowest variance for the sea urchin alignment was selected.

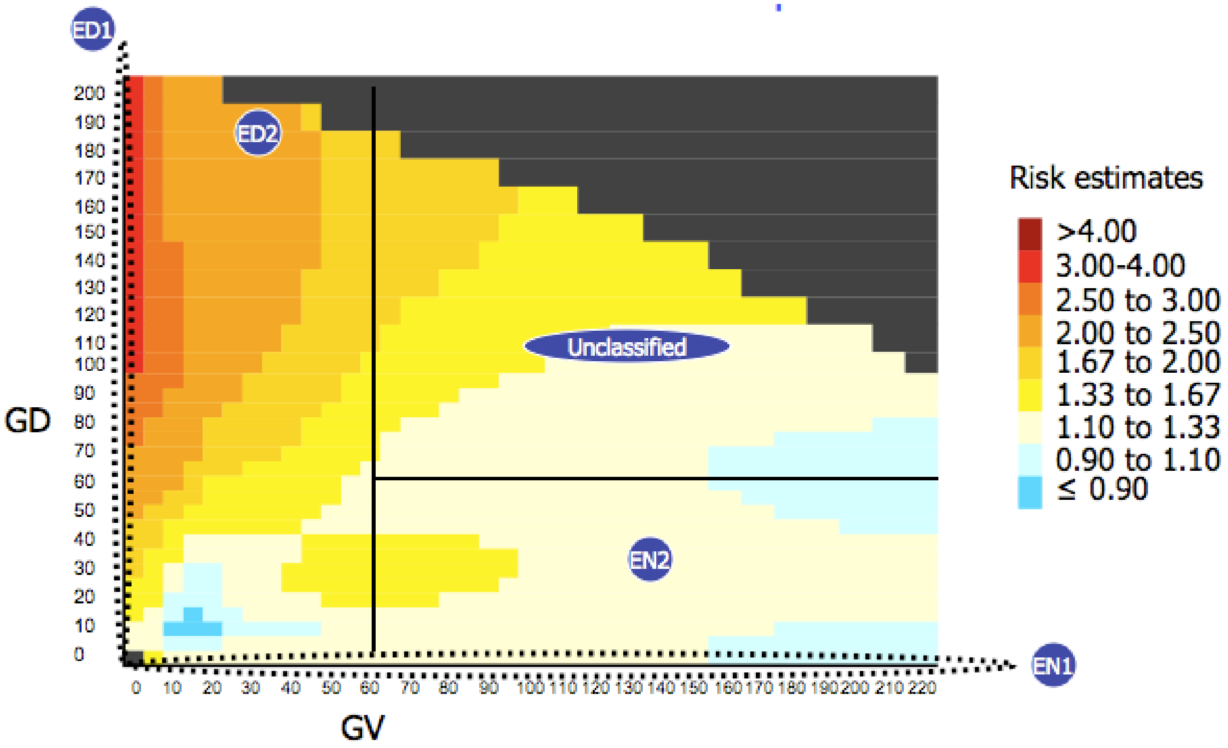
For the grid search, *a* varied from 10° to 75° in steps of 5° and *b* varied from 0.5 to 3.0 in steps of 0.1: there was little to be gained by further refinement given the available data.

BOOTSTRAPPING

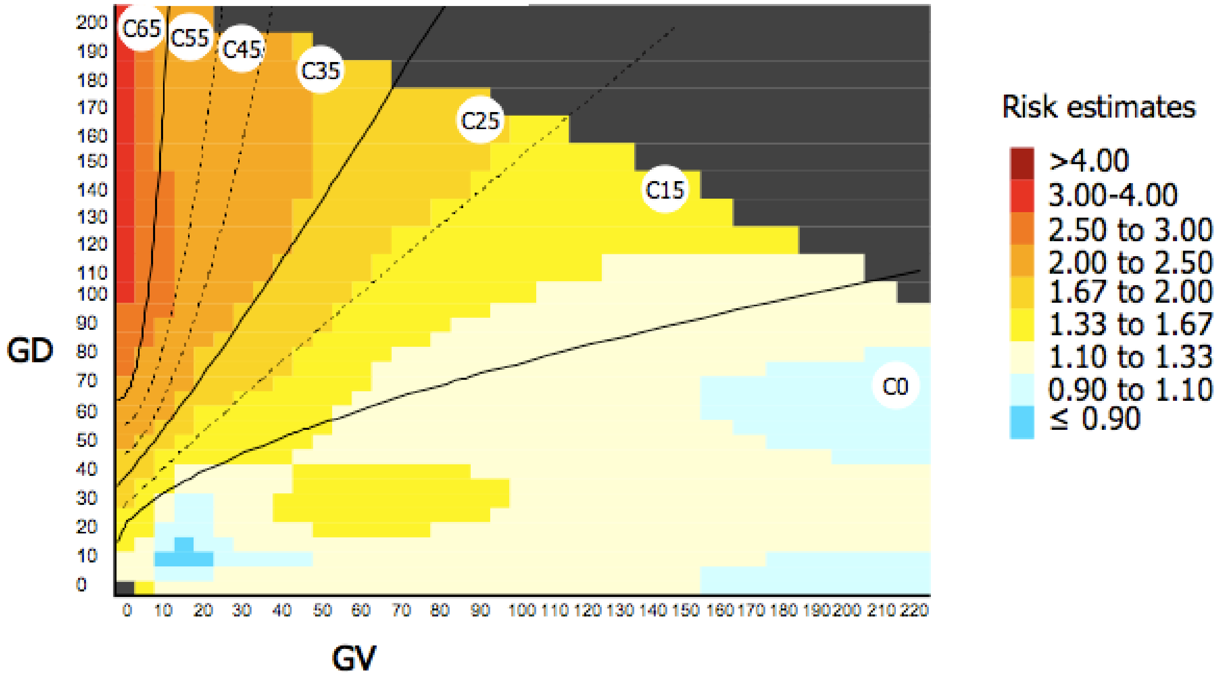
We used a bootstrapping approach to estimate confidence intervals and to provide an initial indication of their robustness of our results to sampling variation. Because they are fundamentally different from each other, the AR and ERS had to be bootstrapped independently. To bootstrap the AR, we constructed a table with one entry for each observation of a missense substitution in a pseudocase, one entry for each observation in a pseudocontrol, and including the GV-GD data for each underlying missense substitution. This table had a total of 2295 *BRCA1* entries and 3963 *BRCA2* entries. In each bootstrap cycle, 2295 entries were randomly selected, with replacement, from the *BRCA1* data and 3963 entries were randomly selected, with replacement, from the *BRCA2* data. Gaussian weighting was then used to load these data onto each grid point (heat map or contour curve or grade) as described above.

To bootstrap the ERS, we constructed a table with one entry for each possible single nucleotide substitution in the *BRCA1* and *BRCA2* data sets, including the GV-GD data for each of the variants and whether that variant had been observed in a subject; this table had a total of 2844 *BRCA1* and 6390 *BRCA2* entries. In each bootstrap cycle, 2844 entries were randomly selected, with replacement, from the *BRCA1* data and 6390 entries were randomly selected, with replacement, from the *BRCA2* data. Gaussian weighting was then used to load these data onto each grid point (heat map or contour curve or grade) as described above.

The resulting simulated data were then used to calculate ARs, ERSs and joint risk estimates for confidence interval estimations.



Supplementary Figure S1. The original Align-GVGD classifiers [Tavtigian et al., 1996] superposed on the joint risk estimate heat map for the PMSA through sea urchin.



Supplementary Figure S2. The original new Align-GVGD grades superposed on the joint risk estimate heat map for the PMSA through sea urchin.