

Practice guidelines for the Interpretation and Reporting of Unclassified Variants (UVs) in Clinical Molecular Genetics.

Prepared and edited by Jennie Bell¹, Danielle Bodmer², Erik Sistermans³ and Simon C Ramsden⁴

1. West Midlands Regional Genetics Laboratory, Birmingham Women's Hospital NHS Trust, Metchley Park Road, Edgbaston, Birmingham, B15 2TG, United Kingdom.

2. Dept of human genetics, Radboud University Nijmegen Medical Centre, PO box 9101, 6500 HB Nijmegen, The Netherlands

3. Dept of clinical genetics, VU University Medical Center, van der Boechorststraat 7, 1081 BT Amsterdam, The Netherlands

4. National Genetics Reference Laboratory (Manchester), Dept of Medical Genetics, Saint Mary's Hospital, Hathersage Road, Manchester, M13 0JH, United Kingdom.

Guidelines ratified by the UK Clinical Molecular Genetics Society (11th January, 2008) and the Dutch Society of Clinical Genetic Laboratory Specialists (Vereniging Klinisch Genetische Laboratoriumspecialisten; VKGL) (22nd October, 2007).

1. INTRODUCTION

With the increased demand for molecular genetic testing over recent years there has been a marked change in the scale and sensitivity of molecular genetic analysis within the service environment. Inevitably this has resulted in a rapid increase in the detection of novel sequence variations of unknown pathogenicity. Whilst research laboratories may have large resources at their disposal to investigate individual variants, routine diagnostic service laboratories must undertake this analysis within a limited timescale and budget.

It is essential, therefore, that diagnostic laboratories have a set of agreed standards to assist in the determination of the clinical significance of variants identified in routine testing. In addition guidelines should be designed to educate referring clinicians as to possible testing outcomes so that they may inform their patients and families appropriately.

The standards outlined here have been drawn up as a guide to assess variants of unknown clinical significance for situations where there is likely to be a clinical benefit. It may not be appropriate to perform this analysis on all identified variants. The authors and the ratifying bodies (CMGS and VKGL) recognise that these guidelines are aspirational and the practicalities of implementation may lead to future revision.

2. SCOPE OF THE GUIDELINES.

These guidelines set agreed standards for the interpretation and reporting of sequence variants of uncertain pathogenicity in genes known to cause inherited Mendelian diseases in which molecular genetic testing has a proven clinical validity and utility. They do not consider the additive pathogenic effect of multiple low-penetrance alleles (Weedon *et al.*, 2006; Johnson *et al.*, 2007) although this needs to be kept under close review. The types of variants considered include missense and synonymous changes of unknown clinical significance as well as putative splice site variants outside of the invariant

boundaries. This document does not consider changes that alter the invariant AG/GT boundaries nor nonsense mutations however we do recognize that these changes cannot be exclusively regarded as pathogenic.

3. QUALITY STANDARDS.

3.1 Minimum quality standards for laboratories interpreting and reporting unclassified variants.

It is essential that the interpretation and reporting of variants of unknown pathogenicity is carried out by appropriately qualified and experienced staff working within certified laboratories that are working to recognized international quality standards (such as ISO 17025 and 15189).

3.2 Test Validation and External Quality Assessment/ Proficiency Testing

All technologies used to identify sequence variants must be appropriately validated to ensure that they meet acceptable performance standards and are fit for the purpose for which they will be used. Validation can be particularly difficult for genetic testing for rare disorders when it may be difficult to obtain suitable positive mutation controls. There is little guidance on the minimum requirements for validation. However the Clinical and Laboratory Standards Institute (www.clsi.org) has published guidance on the use of molecular diagnostic methods for genetic disease which includes a comprehensive section on test validation.

The use of reference materials and participation in external quality assessment (otherwise known as proficiency testing) schemes can contribute to this process. Laboratories should have regular independent assessment of the technical performance of their tests in order to ensure that analytical measurements made in one laboratory are consistent with those made in another laboratory.

3.3 Mutation Nomenclature

It is essential that diagnostic laboratories adopt a consistent approach to naming all variants. Nomenclature guidelines are available from the Human Genome Variation Society (HGVS:

<http://www.hgvs.org/>), the international body for defining gene variation nomenclature under the umbrella of the Human Genome Organization (HUGO) and the International Federation of Human Genetics Societies (IFHGS). We recommend that laboratories follow the HGVS guidance on mutation nomenclature in order to facilitate the unequivocal sharing of data via managed databases. An appropriate GenBank reference sequence must also be cited (e.g. NM_004004.3:c.3G>T). The Mutalyzer project is developing tools to support correct mutation nomenclature use according to HGVS guidelines – see www.LOVD.nl/mutalyzer.

3.4 Mutation Submission

To increase our knowledge of all gene variants identified and to ensure that existing databases are as up-to-date as possible it is recommended that any variants are submitted to an appropriate database (preferably a locus specific database, LSDB) at the earliest opportunity.

4. LINES OF EVIDENCE

There are numerous means of interpreting variants of unknown clinical significance. These may be relevant to all disease genes or disease specific. It is essential that a minimum set of standards is clearly set out in order to ensure that all patients are offered the same quality of care.

4.1 Mutation databases including LSDBs.

LSDBs are an essential means of recording all variation within a gene. The most successful databases contain accurate (curated), clearly referenced data naming variants at the DNA, RNA and protein level and include all relevant comments relating to the clinical interpretation of the variant. It is important to record the technology used for screening, when possible, whether a full gene screen has been completed, and when predicting an RNA defect it is essential to record if this has been confirmed by RNA studies and not just assumed. Whilst some databases will record each variant only once, from a diagnostic perspective those which allow for the repeated submission of the same variant must be regarded as the ideal. It is also helpful to record which combinations of variants were found and whether phase was established, particularly for recessive conditions.

Core databases such as HGMD, OMIM, dbSNP/Ensembl can also provide important information as can thorough search of the literature via PubMed, Google Scholar or the Web of Science.

Summary: It is *essential* that LSDBs are used where available and that staff carrying out searches should be appropriately trained in the use of databases. It is *highly desirable* that all laboratories working within the public health sector have a policy of submitting all variants (excluding known non-pathogenic polymorphisms) to LSDBs where this facility is available. One central, curated mutation database to which reference alignments are attached would be ideal, however this is not currently available.

4.2 Testing (matched) controls

The use of matched controls is a useful means of helping to exclude the possibility that a variant represents a pathogenic change, however there are a number of factors that must be taken into account. When deciding on the number of controls to screen it is helpful to consider the power of this approach. In general, the probability of drawing a given number of alleles (k) of one type from a total number of chromosomes tested (n) follows the binomial distribution:

$$P_{(k \text{ out of } n)} = \frac{n!}{k!(n-k)!} (p^k)(q^{n-k})$$

So for example in order to have a 95% chance of observing a variant with an allele frequency of 1 in 100 at least once we would have to screen 298 chromosomes. For certain diseases the screening of ‘normal’ alleles is achieved as part of the routine diagnostic service. However, caution must be exerted for autosomal recessive conditions where pathogenic variants may exist at a high carrier frequency in certain populations such as p.Phe508del in the *CFTR* gene.

If a laboratory has not screened 298 chromosomes (see above) as part of the routine service then the analysis of matched controls could be considered. Matched controls should be appropriately sourced however it is important to bear in mind that in the case of later onset disorders this may include a number of ‘at risk’ individuals.

Summary: Testing of matched controls is an *acceptable* means of determining whether a variant is to be found segregating in a population of normal chromosomes. However variants are rare and it is *essential* that laboratories are aware of the limitations of this approach. It is *essential* that ethnicity is considered and that all controls are completely anonymised.

4.3 Co-occurrence (*in trans*) with known deleterious mutations.

The identification of a variant in an individual in whom a clear pathogenic mutation has already been identified in the same gene may help further to classify the variant.

With *BRCA1* for example there is now clear evidence for an absence of homozygotes and compound heterozygotes in humans and in addition homozygosity for *Brca1* mutations appears to be embryonically lethal in mice (Gowen *et al.*, 1996; Liu *et al.*, 1996; Hohenstein *et al.*, 2001). With *BRCA1* therefore the existence of an unclassified variant *in trans* with a known pathogenic mutation can provide a powerful means of excluding pathogenicity.

For this reason it is considered desirable to complete a screen on a patient wherever practicable regardless of whether or not a pathogenic change is identified early in the screening process. Whilst the data obtained may not add any information to the consultant, it will generate useful data for other families and add to the understanding of the gene. The decision as to whether a complete screen should include the entire gene should be based on published evidence. In genetically heterogeneous disorders it is not expected that all the relevant genes should be screened. All unclassified variants identified should be reported i.e. the pathogenic variant as well as others.

Summary: If the evidence is sufficiently strong for a given gene it is *acceptable* to use co-occurrence (*in trans*) of a variant with a known deleterious mutation in dominant disorders as

evidence of non-pathogenicity. It is *essential* to establish phase with the pathogenic mutation and, for this parental samples may be required.

4.4 Co-segregation with the disease in the family.

Segregation studies require that appropriate samples are available from family members and can be useful for establishing linkage to a particular disease locus. It is important to keep in mind the limitations of this approach and to consider the possibility of phenocopies and partial penetrance. Family structure is also important and the issue of non-paternity should be considered. This approach is most useful as a means of excluding pathogenicity in cases where a variant does not segregate with a given disorder. It should be borne in mind that an unclassified variant, possibly without known pathogenic consequences may appear to be pathogenic because it is *in cis* with an unidentified pathogenic mutation in a particular family or families. Caution must therefore be exercised in assigning pathogenicity to such variants, although they may act as acceptable linked markers in certain diseases. Such changes should be referred to as a 'variants without known phenotypic consequence or clinical significance'. In these circumstances, interpretation is made easier by complete and comprehensive screening of all appropriate genes and gene regions. It is noted that this strategy carries a high risk of misinterpretation. Laboratories can include a LOD score to indicate the value of such an association. However this type of analysis should be undertaken with extreme caution to avoid providing biased data.

Summary: Segregation analysis is an *acceptable* means of determining whether a disease segregates with a candidate gene. It is *acceptable* to recommend segregation analysis without referral to a genetic counselor in order to help determine pathogenicity, however this must be carried out ethically taking into account issues of non-penetrance. It is *unacceptable* to test unaffected individuals without referral to a genetic counselor if the information has a predictive value.

4.5 Occurrence of a new variant concurrent with the (sporadic) incidence of the disease.

If a variant is identified in a strong candidate disease gene concurrent with the sporadic incidence of the disease and is not present in parental samples this could be considered as strong evidence of pathogenicity. However it is not necessarily proof on its own, and should be considered in the clinical and amino acid context. It is important to consider that a deletion may appear *de novo* and yet be derived from a parent heterozygous for the deletion where the second chromosome carries a duplication (e.g. SMA). If non-paternity has not been excluded then a statement should be included in the report to highlight that the interpretation is based on family relationships being as stated.

Summary: It is *acceptable* to use occurrence of *de-novo* variant concurrent with the incidence of a sporadic disease as a strong indicator of pathogenicity.

4.6 Species conservation.

When performing conservation analysis the quality of the alignment used is key. If amino acid conservation is to be used as a predictor of pathogenicity then it is essential that reference sequence alignments for the gene of interest are available. As a minimum standard the alignments should include the full length sequence of eight orthologous genes and include at least five mammalian orthologues (drawn from different orders of mammals) plus chicken, frog and fish (either puffer or zebra) e.g. *Homo sapiens* (human), *Pan troglodytes* (chimpanzee), *Gorilla gorilla* (gorilla), *Pongo pygmaeus* (orangutan), *Macaca mulatta* (Rhesus Macaque), *Mus musculus* (mouse), *Canis familiaris* (dog), *Bos taurus* (cow), *Monodelphis domestica* (grey, short tailed opossum), *Gallus gallus* (chicken), *Xenopus laevis* (frog) and *Tetraodon nigroviridis* (pufferfish). If it is not possible to obtain clearly orthologous gene sequences from chicken, frog, and fish then additional mammalian sequences should be added to the alignment.

It is essential to attempt to capture the appropriate amount of evolutionary time and it is important to note the depth of the alignment. Examples of appropriate alignments for *BRCA1* and *BRCA2* can be viewed at the Align GVDG website (<http://agvgd.iarc.fr/index.php>). It is important to emphasise that a) this is not currently available in most cases and b) weighting GV/GD scores may vary according to the gene (or domain) type and especially depth of evolutionary comparison (e.g. the Walker motif in *MSH2* is conserved back to eubacteria but in *BRCA1* is not as highly conserved). Systematic evaluation is required and we recommend the building of suitable gene specific alignments which can be made freely available to the diagnostic community via locus specific databases. ClustalW can be used to build alignments (<http://www.ebi.ac.uk/clustalw/index.html>) and genome browsers may be a good starting point to find multiple sequence alignments.

Summary: It is *acceptable* that inter-species comparisons are carried out for all missense changes of unknown clinical significance and that the depth of conservation is recorded in any documentation relating to the analysis, although it doesn't need to be included in the report. The documentation should outline which species and/or orthologues were included in the comparison and what comparison method and parameters were used. The diagnostic community may develop suggested values and appropriate species in the future, although this may be better co-ordinated by locus-specific databases. The interpretation of species conservation should be cautiously applied, particularly if only one species is identified for which the variant is not conserved. Any overall conclusions on the clinical significance of the variant should be based on more than one line of evidence.

4.7 In silico prediction of pathogenic effect.

Once reference alignments are available (see section 4.6), it is possible to use them to predict if the substitution is tolerated. Align GVDG combines the Grantham matrix (Grantham 1974) (assessing the biochemical distances between amino-acids – hence, the severity of a change) with sequence alignments. SIFT (Sorting Intolerant From Tolerant) and Polyphen are both based on multiple sequence alignments although it should be noted that the latter program a) does not permit the use of the user's own alignments and b) takes into account 3D data when available

More details on the use and interpretation of these scores are given in the Additional file 1.

The aim of this analysis is to establish the degree of variation tolerated. The variation can be described by listing the amino acids observed at the relevant position in the alignment and quantified using the Grantham Variation (GV). Simple reference categories for GV are:

Invariable – conservative GV=0

Variable – conservative $0 < GV < 62$

Variable – non-conservative $GV > 62$

The substitution can then be scored using web-based prediction software such as SIFT or Align-GVGD (only the *BRCA1* and *BRCA2* alignments are currently available at the website, for other genes users must build their own alignments) or Polyphen.

A domain search using *Ensembl* can also be performed. It is recommended that laboratories use agreed reference sequence alignments and use them within each prediction site. Analysis of a particular variant should be performed using at least two different programs. Most of the time similar results will be obtained however dissimilar results can be generated which must be taken into account when deciding the likelihood of pathogenicity.

As well as Grantham scores, information on the chemical and functional properties of the amino-acids can be obtained from the Russell web-site (<http://www.russell.embl.de/aas/>).

It is recommended that users gain experience with the outputs of available informatic tools by observing the interpretations generated for mutations of well established pathogenicity, or neutrality, given various input parameter settings.

A domain search using *Ensembl*, *Uniprot* and *Prosite* can also be performed. It should be noted that *Prosite* provides clustal multiple alignments useful to determine the most conserved residues of a given domain.

Summary: It is *acceptable* to predict the severity of an amino acid change using *in-silico* methods. It is *unacceptable* to rely solely on these predictions to assign pathogenicity to a previously unclassified variant. Records of this work must specify the parameters and methods used to estimate the severity of the amino acid change.

4.8 *In silico* splice site prediction

There are several splice prediction programs commonly used by diagnostic laboratories (see additional file 1). It is important to note however that these programs have not been fully validated for use in a diagnostic setting and consequently they must be used with caution. The user is able to adjust the settings on these sites and no information is available at this time to advise how best to alter these settings. It is considered appropriate to use the default settings unless otherwise stated. Laboratories should be aware that any sequence changes and not necessarily just those adjacent to intron/exon boundaries may constitute splice site mutations and as such are amenable to analysis using this software (<http://www.umd.be/SSF>). Scientific knowledge should be applied to deeply embedded intronic variants as detailed prediction analysis may not be appropriate. In the case of silent mutations and missense mutations, testing them for an effect on splicing should be

considered, especially when AG or GT dinucleotide sequences are formed.

Summary: It is *acceptable* to use *in-silico* splice site prediction, however it is *unacceptable* to base an unequivocal clinical interpretation solely on this line of evidence. If this method of prediction is used it is *essential* to arrive at an interpretation based upon a consensus of at least three splice site prediction programs.

4.9 RNA Studies

Where possible, RNA studies are the best means of interpreting the consequences of a splicing mutation. Therefore it is recommended that RNA studies be performed, in the following context:

- If RNA from an appropriate and validated tissue or cell type is available
- If an unclassified variant is found that may affect splicing where two prediction programs give an indication for a splice site alteration
- Where other lines of evidence support an effect on splicing
- If the entire gene has been tested and no other mutations have been found

It should also be noted that sequence confirmation on cDNA is essential, though not possible if the variant is located within the intron. In addition we recommend:

- In cases where no splice errors are apparent on mRNA analysis, biallelic expression by molecular analysis of a variant should be proven to conclude that the unclassified variant has no effect such as nonsense mediated decay which may prevent *in vitro* observation of aberrant splice product.
- Naturally occurring alternate splice variants should be excluded by the testing of matched controls and the use of RNA from appropriate and validated tissue or cell types.

Summary: Given the high predictive value of RNA studies they must be regarded as *essential* for the definitive interpretation of putative splicing mutations. However it is recognized that not all laboratories have the facilities to perform these analyses and that limited expression patterns may mean that the required tissue is not available for analysis.

4.10 Functional Studies

A reliable functional assay is generally regarded as one of the best means of confirming pathogenicity, however this is rarely available as part of a routine diagnostic service. Occasionally an affiliated research group will undertake functional studies. It is important to note however, that this is very gene/disease specific and the assay must be in an appropriate system. When relying on reports in the literature laboratories should be careful to scrutinise the context in which the assay is performed as the diagnostic evaluation of a functional assay can be difficult. In certain circumstances, analysis of cell and tissue-specific expression of proteins by means of immunohistochemistry (e.g. dystrophin and DNA mismatch repair proteins) can be considered a form of *in vivo* functional study, and as such can be an extremely useful aid in interpreting DNA/RNA findings.

Summary: If a reliable functional assay is available it must be regarded as *essential* for the definitive interpretation of a variant of unclear clinical significance. However it is recognised that the sensitivity and specificity of assays vary and where less

reliable assays are all that is available their use in interpretation is only *desirable*.

4.11 Loss of Heterozygosity (LOH)

LOH can indicate the presence of a tumor suppressor gene in the deleted region. Often, the remaining copy of the tumor suppressor gene has been inactivated by a point mutation and consequently LOH may increase the likelihood of pathogenicity of a variant on the remaining chromosome. The classical example is hereditary retinoblastoma, in which one parent's contribution of the tumor suppressor *RBI* carries a pathogenic change. Although most cells will have a functional second copy, chance loss of heterozygosity events in individual cells almost invariably lead to the development of this retinal cancer in the young child. There are published descriptions in the literature regarding the objective inclusion of LOH data in linkage analysis (Lustbader *et al.*, 1995; Rohde *et al.*, 1997).

Summary: It is *acceptable* to use LOH to assist in the prediction of pathogenicity of variants in tumour suppressor genes, however this evidence is unlikely to be convincing in the absence of other lines of evidence.

4.12 Presence or absence on Single Nucleotide Polymorphisms (SNP) Databases.

The presence of a variant in an unaffected individual at population risk can be used as evidence of non-pathogenicity (see section 4.1). SNP databases are a quick means of viewing normal variation in a gene and can help in this process. However the most commonly used SNP database, dbSNP, contains a large, but uneven sample, of genome diversity and includes SNPs, deletion/insertion polymorphisms (DIPs), short tandem repeats (STRs), multiple nucleotide polymorphisms (MNPs) and 'No Variation' data, pathogenic variants are also recorded. There is often no assumption about the minimum allele frequency, therefore the data may contain both pathogenic and non-pathogenic variants.

Summary: It *essential* that SNP databases are reviewed on discovery of an unclassified variant. It is *unacceptable* to use the presence on a SNP database as evidence of non-pathogenicity in the absence of convincing frequency information.

4.13 Value of a Bayesian approach.

A Bayesian approach combining different strands of evidence can be used to produce a final probability of pathogenicity (Goldgar *et al.*, 2004). However there are currently no recognised statistical tools available to diagnostic laboratories for this purpose.

4.14 General Points

It is *essential* that laboratories keep accurate records of all databases and prediction programs used in the interpretation of a variant and the dates of all searches. It is *essential* that laboratories record the dates, settings and scores of all web-based analyses. It is *recommended* to use a checklist to standardize laboratory investigations of variants of unknown clinical significance in order to ensure that all procedures are documented. Examples of two checklists currently used are included in the additional files

2 and 3. Laboratories may choose to adapt one of these for their own requirements.

5. REPORTING STANDARDS

5.1 Reporting UVs.

It is *essential* that laboratories have mechanisms in place to submit results to existing databases (especially LSDBs) and it is *essential* that laboratories issue an updated clinical report as new information becomes available to them (reports should be re-issued when a UV becomes clearly pathogenic or is not pathogenic anymore). However a caveat can be included in any clinical report to point out that the information represents the best interpretation of the data at the time of reporting and that the most appropriate interpretation of UVs may change with time.

It is *essential* that the unqualified terms 'polymorphism' and 'mutation' are not used in reports. However it is *acceptable* to use terms such as 'variant of no known phenotype' and 'pathogenic mutation'. An *acceptable* alternative term for UV is 'sequence variant of unknown clinical significance'.

When reporting UVs it is *essential* to include the extent of the screening performed, if a complete laboratory analysis has not been performed then this must be detailed in the report. If no other variants have been identified this should also be highlighted.

It is *not essential* to document all the lines of evidence obtained in the report, however complete records must be stored in the laboratory. On occasions where reference to specific reports in the literature will support the re-classification of a variant it is *acceptable* to include more details of the evidence.

5.2 To whom is it acceptable to report UVs?

It is *essential* that reports of UVs should be issued to appropriately trained clinicians. In the vast majority of cases this is likely to be clinical geneticists or genetic counsellors. However we recognize that other healthcare professionals may be equally competent in the interpretation of UVs. Extreme caution should be taken when issuing a report of a UV to any professional who is not conversant with the complexities of such information. In these cases it is *essential* that careful unambiguous wording is used and it is *essential* to suggest discussion with a clinical geneticist. It is acceptable to request further samples from a clinician in order to facilitate the interpretation of a UV without referral to a genetics clinic (see section 4.4 above).

5.3 Which UVs should be reported

It is *not essential* to report all non-pathogenic polymorphisms, and in fact this could lead to misinterpretation outside of the laboratory, but it is *acceptable* to include a disclaimer that they are not being reported. It is, however *essential* to report all UVs where the clinical significance is uncertain and it is *essential* that all UVs are recorded within the laboratory.

Some laboratories have adopted an internal classification system

Class 1 – Certainly not pathogenic

Class 2 – Unlikely to be pathogenic but cannot be formally proven

Class 3 – Likely to be pathogenic but cannot be formally proven

Class 4 - Certainly pathogenic

This system is for laboratory use only and should not be incorporated into the patient report, it is not considered essential to use a classification system such as this.

5.4 Predictive testing.

Testing for UVs in a predictive context is not recommended. However, there may be some circumstances where it may be considered appropriate to do so. This must only be offered within the context of appropriate genetic counseling.

Acknowledgements: This document represents the combined efforts of all attendees of the unclassified variants good practice meeting held on the 12th & 13th April 2007 at The Nowgen Centre, Manchester, UK. This meeting was attended largely by representatives of the UK Clinical Molecular Genetics Society (CMGS, <http://www.cmgs.org/>) and Dutch Society of Clinical Genetic Laboratory Specialists (Vereniging Klinisch Genetische Laboratoriumspecialisten; VKGL, <http://www.nav-vkgn.nl>). Special thanks are given to the speakers Prof Sean Tavtigian, Dr. Johan T. den Dunnen, Andrew Devereau and Dr Fiona Laloo. We gratefully acknowledge the financial support of the National Genetics Reference Laboratory (Manchester) and additional files 2 and 3 are included with thanks to Danielle Bodmer and Stewart Payne.

REFERENCES

Goldgar DE, Easton DF, Deffenbaugh AM, Monteiro AN, Tavtigian SV, Couch FJ; Breast Cancer Information Core (BIC) Steering Committee. (2004) Integrated evaluation of DNA sequence variants of unknown clinical significance: application to BRCA1 and BRCA2. *Am J Hum Genet.* **75**:535-44.

Gowen LC, Johnson BL, Latour AM., Sulik KK, Koller BH (1996) *BRCA1* deficiency results in early embryonic lethality characterized by neuroepithelial abnormalities. *Nat Genet* **12**:191-194

Grantham R. (1974) Amino acid difference formula to help explain protein evolution. *Science* **185**:862-4.

Hohenstein P, Kielman MF, Breukel C, Bennet LM, Wiseman R, Krimpenfort P, Cornelisse C, van Ommen GJ, Devilee P, Fodde R (2001) A targeted mouse *BRCA1* mutation removing the last BRCT repeat results in apoptosis and embryonic lethality at the headfold stage. *Oncogene* **20**:2544-2550.

Johnson N, Fletcher O, Palles C, Rudd M, Webb E, Sellick G, Dos Santos Silva I, McCormack V, Gibson L, Fraser A, Leonard A, Gilham C, Tavtigian SV, Ashworth A, Houlston R, Peto J. (2007) Counting potentially functional variants in BRCA1, BRCA2 and ATM predicts breast cancer susceptibility. *Hum Mol Genet.* **16**:1051-7.

Liu CY, Flesken-Nikitin A, Li S, Zeng Y, Lee WH (1996) Inactivation of the mouse *BRCA1* gene leads to failure in

the morphogenesis of the egg cylinder in early postimplantation development. *Genes Dev* **10**:1835-1843.

Lustbader E, Rebbeck TR, Buetow KH. (1995) Using loss of heterozygosity data in affected pedigree member linkage tests. *Genet Epidemiol.* **12**(4):339-50.

Ng PC, Henikoff S. (2001) Predicting deleterious amino acid substitutions. *Genome Res* **11**:863-74.

Rohde K, Teare MD, Santibáñez Koref M. (1997) Analysis of genetic linkage and somatic loss of heterozygosity in affected pairs of first-degree relatives. *Am J Hum Genet.* **61**:418-22

Rudd MF, Williams RD, Webb EL, Schmidt S, Sellick GS, Houlston RS. (2005) The predicted impact of coding single nucleotide polymorphisms database. *Cancer Epidemiol Biomarkers Prev.* **14**:2598-2604.

Weedon MN, McCarthy MI, Hitman G, Walker M, Groves CJ, Zeggini E, Rayner NW, Shields B, Owen KR, Hattersley AT, Frayling TM. (2006) Combining information from common type 2 diabetes risk polymorphisms improves disease prediction. *PLoS Med.* **3**:e374.

Xi T, Jones IM, Mohrenweiser HW. (2004) Many amino acid substitution variants identified in DNA repair genes during human population screenings are predicted to impact protein function. *Genomics* **83**:970-9.