Association for Clinical Genetic Science
Part of the British Society for Genetic Medicine

VKGL

Practice Guidelines for the Evaluation of Pathogenicity and the Reporting of Sequence Variants in Clinical Molecular Genetics.

Yvonne Wallis[1], Stewart Payne[2], Ciaron McAnulty[3], Danielle Bodmer[4], Erik Sistermans[5], Kathryn Robertson[6], David Moore[7], Stephen Abbs[8], Zandra Deans[9] and Andrew Devereau[6]

1. West Midlands Regional Genetics Laboratory, Birmingham Women's NHS Foundation Trust, Mindelsohn Way, Edgbaston, Birmingham, B15 2TG, United Kingdom.
2. Kennedy-Galton Centre (NW Thames) Regional Genetics Centre, Level 8V, Northwick Park & St Mark's Hospitals, Watford Road, Harrow, HA1 3UJ, United Kingdom.
3. Northern Genetics Service, Institute of Human Genetics, Central Parkway, Newcastle-upon-Tyne, NE1 3BZ, United Kingdom.
4. Dept of Human Genetics, Radboud University Nijmegen Medical Centre, PO box 9101, 6500 HB Nijmegen, The Netherlands
5. Dept of Clinical Genetics, VU University Medical Center, van der Boechorststraat 7, 1081 BT Amsterdam, The Netherlands
6. National Genetics Reference Laboratory (Manchester), Dept of Medical Genetics, Saint Mary's Hospital, Hathersage Road, Manchester, M13 OJH, United Kingdom.
7. South East Scotland Genetic Service, David Brock Building, Western General Hospital, Crewe Road, Edinburgh, EH4 2XU
8. East Anglian Medical Genetics Service, Genetics Laboratories, Addenbrooke's Hospital, Cambridge, CB2 0QQ
9. UK NEQAS for Molecular Genetics, UK NEQAS [Edinburgh], Department of Laboratory Medicine, Royal Infirmary of Edinburgh, Edinburgh, EH16 4SA

Original 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).

Guidelines updated by the Association for Clinical Genetic Science (formally Clinical Molecular Genetics Society and Association of Clinical Cytogenetics) and the Dutch Society of Clinical Genetic Laboratory Specialists (approved September 2013).

1. INTRODUCTION

With the increased demand for molecular genetic testing 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 rare or novel sequence variations. 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. It is important to liaise with clinical teams when possible. Discussion of detailed phenotype information may contribute significantly to the interpretation of some variants as well as to deciding the nature of follow up tests.

The standards outlined here have been drawn up as a guide to assess the clinical significance of DNA sequence variants 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 (ACGS and VGKL) recognise that these guidelines are aspirational and the practicalities of implementation may lead to further revision.

2. SCOPE OF THE GUIDELINES

These guidelines set agreed standards for the interpretation and reporting of sequence variants in genes known to cause inherited Mendelian and acquired 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. This document does not consider changes that are considered to be clearly pathogenic not requiring any further interpretation, such as frameshift mutations, changes that alter the consensus AG/GT boundaries and nonsense mutations. However, we do recognize that these changes cannot be exclusively regarded as pathogenic. There are examples of nonsense and frameshift changes occurring in the last exon of a gene (or within 50 bases of the second to last exon) that are not pathogenic. It is important to consider the impact of alternate transcripts if known.

3. QUALITY STANDARDS

3.1 Minimum quality standards for laboratories interpreting and reporting sequence variants

It is essential that the interpretation and reporting of sequence variants is carried out by appropriately qualified and experienced staff working within certified laboratories that are working to recognised 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 Variant 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 variant nomenclature in order to facilitate the unequivocal sharing of data via managed databases. Tools are available to support correct use of HGVS nomenclature and are listed in Appendix A. An appropriate reference sequence must be cited including the version number (e.g., NM 004004.3). It is recommended that if available an LRG (Locus Reference Genomic) sequence is used. LRG sequences are defined as stable genomic reference sequences designed to standardise the reporting of sequence variants (see Appendix A for website).

3.4 Variant 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. It is therefore essential for laboratories to have the capability of recording variants of uncertain clinical significance accurately. Easy access to previously recorded variants of uncertain clinical significance will also facilitate the review process, helping to identify quickly those patients in whom a sequence variant has been reclassified as either pathogenic or of no clinical significance.

4. LINES OF EVIDENCE

For those variants that are not clearly pathogenic, there are numerous means of interpreting 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. It is ***recommended*** that an appropriate literature search is performed as part of the standard evidence gathering process.

4.1 Variant 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 a 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 can also provide important information as can thorough search of the literature. See Appendix A for a list of the most widely used databases.

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. LSDBs that contain references to the published literature should be used in preference to those that do not. 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 the most appropriate LSDBs where this facility is available. One central, curated variant database to which reference alignments are attached would be ideal, however this is not currently available.

4.2 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), although with the possibility of inheritance with non-penetrance or age dependent effects, caution should be applied. SNP databases are a quick means of viewing 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 no assumption about the minimum allele frequency, therefore the data may contain both pathogenic and non-pathogenic variants.
On-going large-scale sequencing projects are beginning to provide more reliable population frequency information for variants. Datasets from the 1000 Genomes Project and the NHLBI GO Exome Sequencing Project are now being used by clinical scientists to obtain informative minor allele frequency information about variants of interest. A quick look-up tool is available from NGRL to search the 1000 Genomes dataset for frequency information on individual variants (https://ngrl.manchester.ac.uk/1kg_querytool/). The scope of the project and the populations sequenced (and potential disease phenotypes) should always be considered when using such datasets.

Summary: It is ***essential*** that SNP databases are reviewed on discovery of a novel sequence variant; however, they should be used with caution. It is ***essential*** to determine the source of the data since multiple reports of a given SNP may actually have arisen from the same original data source. It is ***essential*** that variants are only classed as SNPs if they are validated and are reported with convincing frequency information. It is ***unacceptable*** to use the presence of a SNP in such databases as evidence of non-pathogenicity in the absence of convincing frequency information.

4.3 Testing (matched) controls

The use of matched controls can be 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 exists 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.

4.4 Co-occurrence (*in trans*) with known deleterious variants

The identification of a variant in an individual in whom a clear pathogenic variant has already been identified in the same gene may help further to classify the variant. In the case of dominant conditions where a pathogenic variant has already been identified, the presence of a second sequence variant seen *in trans* may help exclude pathogenicity, e.g., where a second pathogenic variant would be lethal. However, it is important to recognise that this must be interpreted in the context of the detailed clinical information and will vary depending on the disorder. Careful consideration should be made as to whether, in a dominant condition, a second pathogenic variant seen *in trans* does or does not lead to a more severe clinical phenotype. In the case of recessive conditions, the presence of a second sequence variant *in trans* with a pathogenic variant does not exclude or confirm 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 consult and, 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. It is recommended that all variants of uncertain clinical significance identified are reported, i.e., the pathogenic variant as well as others.

Summary: If the evidence is sufficiently strong for a given gene it is ***recommended*** that co-occurrence (*in trans*) of a variant with a known deleterious variant in dominant disorders is used as evidence of non-pathogenicity. It is ***essential*** to establish phase with the pathogenic variant and for this the parental samples may be required.

4.5 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 when relevant. This approach can be definitive as a means of excluding pathogenicity in cases where a variant does not segregate with a given disorder.

Co-segregation of a variant with disease can provide useful information when statistically quantified as proposed by Møller *et al.*, (2011) who developed a simplified method for segregation analysis (SISA). This method considers a pedigree in which a number of affected members have been tested and found to have the variant in question. The number of informative meioses in the family is defined as the number of affected individuals with the variant minus 1 (to correct for ascertainment bias). For example, a family with an affected grandmother, three affected children and one affected grandchild (five affected members) will have four informative meioses. All affected members need to be shown to have the gene variant except obligate carriers. The basis of SISA is that if a variant is NOT disease causing, then the chance of co-segregation with the disease is 1 in 2 for each informative meiosis in a family, i.e., $(1/2)^n$ (where n = number of informative meiosis); in the example above $(1/2)^4 = 0.0625$. Using this method, the probability that the variant therefore causes disease is calculated as 1-0.0625 or 0.9375. As expected the value of $P = 1-(1/2)^n$ increases with each additional affected individual who has inherited the variant (informative meiosis) in the family and a virtue of this method is that multiple families with the same variant can be combined. If a second family were found to co-segregate the same variant as the family above with two informative meioses then n would be 6 and the likelihood of pathogenicity would be estimated to be $1-(1/2)^6 = 0.9844$. The authors, however, do not state the required threshold for pathogenicity.

It should be borne in mind that a sequence variant, possibly without known pathogenic consequences, may appear to be pathogenic because it is *in cis* with an unidentified pathogenic variant 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 may be referred to as '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 may also be appropriate to exclude linkage to alternative loci.

Summary: Segregation analysis is an ***acceptable*** means of determining whether a disease segregates with a candidate gene. It is ***acceptable*** to recommend segregation analysis in affected family members without referral to a genetic counsellor in order to help determine pathogenicity, though it is important to take into account the possibility of non-penetrance. It is ***unacceptable*** to test unaffected individuals without referral to a genetic counsellor if the information has a predictive value.

4.6 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 inci-

dence 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 it is recommended that a statement is 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.

4.7 Species conservation

The aim of conservation analysis is to measure the degree to which amino acid properties at any given position are conserved across evolution. True invariant sites should be considered essential for protein function whereas variant sites can evidently accommodate at least a degree of substitution. The basis for this approach has been summarised as i) missense substitutions at evolutionarily constrained positions are often pathogenic, and those that are not often have neutral or minimal impact, and; ii) missense substitutions falling outside of the cross-species range of variation at a position in a multi-species alignment (MSA) are often pathogenic, and those that do not often have neutral or minimal impact (Tavtigian *et al.*, 2008).

When performing conservation analysis the quality of the MSA used is key. There are a number of studies that have sought to define the optimal MSA parameters (Mathe *et al.*, 2006; Greenblatt *et al.*, 2003; Abkevich *et al.*, 2004; Tavtigian *et al.*, 2008, 2009) and it is clear that the sequences to be included in an alignment are dependent on the protein being studied (Mathe *et al.*, 2006). The choice of sequences should provide sufficient evolutionary time for the invariant positions to have undergone substitution. Work by Greenblatt *et al.*, (2003) has shown that an alignment should contain an average of at least three substitutions per position to significantly reduce the likelihood ($P<0.05$) that an invariant position has occurred by chance. There is evidence that for some genes more substitutions per position are required to reach the required likelihood (Tavtigian *et al.*, 2008). Sequences that are not functionally equivalent must be excluded from the alignment, accordingly, paralogue sequences should not be included and the function of more distantly related sequences verified.

Uniprot (www.uniprot.org) provides comprehensive and freely accessible protein sequence information: orthologous protein sequences may be sourced from genome browsers such as NCBI and Ensembl or ortholog databases (http://www.treefam.org/, http://cegg.unige.ch/orthodb4).

The use of MSAs for *in-silico* prediction tools ideally requires systematic evaluation (see following section) and we recommend building of gene specific MSAs, ideally through manual selection of sequences and curation of the alignment, which can be made freely available to the diagnostic community

There are a number of web-based sequence alignment tools available to create MSAs (recently reviewed by Kemena and Notredame, 2009). Aniba *et al.*, (2010) describe benchmark alignments which may be used to compare alignment tools by individual laboratories. In some instances third party software tools provide or generate their own automatically generated or manually curated MSAs.

Similar to protein conservation, nucleotide conservation may be used to identify conserved and therefore functionally important nucleotide positions. The available tools generate conservation scores based on different resolution levels. Users should be aware at which resolution the programs are analysing the data as this will affect the interpretation of the data. Nucleotide conservation data should be considered as evidence for pathogenicity but should not be used in place of or exclude protein conservation analysis.

Summary: It is ***recommended*** that interspecies 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 does not need to be included in the report. The documentation should outline which species and/or orthologs were

included in the comparison and what comparison method and parameters were used. The diagnostic or research communities may develop suggested values and appropriate species in the future. 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.

4.8 *In silico* prediction of pathogenic effect

There are a number of web-based classification tools, employing different algorithms, which can be used to predict the impact of missense changes on protein function (recently reviewed by Karchin, 2009; Thusberg & Vihinen, 2009). The classification generated from the prediction tools must not be considered definitive and as with all analytical approaches, must form one aspect of a wider investigation.

A number of studies undertaken to validate and compare the predictive value of a number of *in silico* tools have highlighted the fact that no one tool can be considered superior nor achieve complete accuracy. Furthermore, the tool generating the optimal predictive score will depend upon the gene under investigation and parameters used (Chan *et al.*, 2007; Dorfman *et al.*, 2010).

A number of prediction tools require multiple sequence alignments (MSAs) as part of the input parameters. The MSA used for these tools is crucial to their effectiveness (Mathe *et al.*, 2006; Hicks *et al.*, 2010) and users should follow the guidelines outlined in section 4.7 for their creation. Users should be particularly aware of any gaps in the alignments at the position of the change as these can lead to poor predictions, in particular false negative results.

A number of the computational methods generate confidence scores to qualify the predictions made and recent work has shown that optimal predictive scores are achieved by setting gene specific thresholds.

Ideally, therefore, it is recommended that the choice of tool to investigate a variant for a particular gene is based on the comparative validation of a number of tools (and if applicable MSA) against sequence variants of known effect.

It is acknowledged that such validation may not be feasible and therefore it is acceptable for users to use tools that have not been validated, however users should be aware of the limitations of any such prediction.

Analysis of a particular variant should be performed using at least three different programmes (to reduce the prevalence of either a false negative or false positive result). Users should ideally use tools based on different algorithms: a list of tools grouped by algorithm is provided at http://www.ngrl.org.uk/Manchester/page/missense-prediction-tools. Similar results will usually be obtained, however dissimilar results can be generated which must be taken into account when deciding the likelihood of pathogenicity.

Grantham scores can be used to assess the biochemical distances between amino-acids and hence the severity of a change. Grantham scores range from 5 to 215 with the higher scores representing more radical changes in amino acid properties. Grantham scores are designated as conservative (0-50), moderately conservative (51-100), moderately radical (101-150), or radical (≥151) according to the classification proposed by Li *et al*. (1984). 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).

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.

4.9 *In silico* splice site prediction

It is ***acceptable*** to assign nucleotide changes that disrupt the consensus dinucleotide splice sites (+/-1 and +/-2) as clearly pathogenic requiring no further investigation. It is important however to consider the impact of alternative transcripts if known. Disruption of a consensus dinucleotide may cause an in-frame single exon deletion known to be missing in alternate transcripts present in the normal population (Li *et al*, 2009). It may also be appropriate to consider the impact of cryptic splice sites in the

vicinity of disrupted wild type consensus splice sites (Houdayer *et al.*, 2012).
All other potential splice site variants (including intronic changes, missense and synonymous changes at or near the consensus splice sites) should be investigated using appropriate prediction tools.
Splice prediction tools commonly used by diagnostic laboratories are listed in Appendix A. These are generally valid when used correctly and within the scope of their applicability (note also the scope of these guidelines as stated in section 2). An NGRL study (http://www.ngrl.org.uk/Manchester/sites/default/files/publications/Informatics/NGRL_Splice_Site_Tools_Analysis_2009.pdf) showed that the better performing tools were capable of a good degree of accuracy, and that users can therefore be confident of the safe interpretation of results as part of the assessment of a variant. However, they must be used with caution and should not be relied upon alone. This study also showed that these tools are likely to be useful beyond the invariant acceptor and donor sites at intron/exon boundaries, but not beyond positions +7 and -10 of exons. Laboratories should however be aware that any sequence changes (not necessarily just those adjacent to intron/exon boundaries) may create new splice sites and as such are amenable to analysis using this software. Scientific knowledge should be applied to deeply embedded intronic variants as detailed prediction analysis may not be appropriate. In the case of silent variants and missense variants, testing them for an effect on splicing should be considered, especially when AG or GT dinucleotide sequences are formed.
Splice tools are usually made available via specific websites maintained by their developers. They may also be downloaded by users and installed locally, or be provided through third-party applications or websites where they may be implemented by incorporating the software as part of the application or as a web link to the original developer's website. Users should ensure that they know how a tool is being implemented and therefore which version is being used, and should satisfy themselves that results obtained agree with those from the developer's implementation. Users should also access documentation to ensure that they are using the tool correctly and within its intended range of applicability, if necessary accessing documentation or publications provided on the original developer's website. The user may be able to adjust the settings on the prediction tools: it is considered appropriate to use the default settings unless otherwise stated.
Tools for prediction of exonic/intronic splice enhancers and inhibitors (ESEs, ISEs, ESIs and ISIs) are also available but their use in a diagnostic setting has not been validated to the same extent as splice-site prediction tools. Therefore their use is not currently recommended.
It is considered good practice to use more than one splice-site prediction tool. The NGRL study did not show that overall accuracy improved significantly when result from tools were combined, however use of three tools guards against errors due to incorrect or invalid use of a tool for a specific variant and does not add significantly to the analysis time, particularly where a third-party interface is used.
It is not currently possible to set criteria for the change in prediction tool scores which should be considered significant (e.g., a 10% deviation from the wild-type score). This remains a matter for local judgement and agreement.

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. It is, however, acceptable to suggest further investigations based on the outcome. If this method of prediction is used it is ***recommended*** to arrive at an interpretation based upon a consensus of at least three splice site prediction programmes.

4.10 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 (i.e., one known to express the transcript of interest) is available.
- If a variant of uncertain significance is found and prediction programs give an indication for a splice site alteration.
- Where other lines of evidence support an effect on splicing.

- Only if the entire gene has been tested and no other variants have been found.

It should also be noted that sequence confirmation on cDNA is essential when possible (i.e., if the variant of interest is located within exonic sequence). 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 recognised 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.

4.11 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 ***recommended*** for the interpretation of a variant of uncertain 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***.

4.12 Loss of Heterozygosity (LOH)

LOH can indicate the presence of a tumour suppressor gene in the deleted region. Often, the remaining copy of the tumour 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. 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.

4.13 An integrated evaluation of sequence variants

The outcomes of the analyses and observations described here may be integrated using Bayesian statistical inference in a method called the 'integrated evaluation' or 'multifactorial method' (Vallée *et al.,* 2011). This is finding application to diseases including breast cancer and Lynch syndrome (Spurdle, 2010). However, application of this method is not routinely accepted in diagnostic practice, and is suited to projects or communities with centralised resources and a committee approach to variant classification. It is therefore ***recommended*** that laboratories seek such classifications from projects or groups where they exist and are validated, so that laboratories are satisfied of the reliability of their evaluations, and contribute data as recommended in section 4.1.

4.14 General Points

It is ***essential*** that laboratories standardise the process of variant interpretation and a checklist is ***recommended*** in order to ensure that all procedures are documented. For each search performed it is ***essential*** to record the following information: the date of the search; the tool

version; any changes made to default settings; alignments used; database build; and the scores of all analyses. It is not always possible to determine the specific differences between versions and/or settings of tools and therefore recording the date and version is sufficient rather than the actual differences. It is ***desirable*** that regional genetic centres have procedures in place to review the status of previously reported variants of uncertain clinical significance.

5. REPORTING STANDARDS

5.1 Reporting variants.

It is ***essential*** that laboratories develop mechanisms 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 variant of uncertain clinical significance becomes clearly pathogenic, or a variant 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 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'. The term 'variant of uncertain clinical significance' (VUS) is the preferred terminology for variants which cannot be classified as either *clearly pathogenic* (class 5), *likely to be pathogenic* (class 4), *unlikely to be pathogenic* (class 2), or *clearly not pathogenic* (class 1), see section 5.2 below.

When reporting a VUS 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.

Guidelines for reporting molecular genetics results can be found at: www.cmgs.org/BPGs/Reporting%20guidelines%20Sept%202011%20APPROVED.pdf.

5.2 Classification of variants

A number of different classification systems exist, using 3, 4 or 5 different variant classes. It is ***recommended*** that a 5 class system is adopted as described in the table below for internal assessment purposes, with interpretation based on local policies. The use of a 5 class system is considered ***essential*** for the standardisation of report wording and follow up studies. The numbering itself however should not be incorporated into the patient report as this is considered likely to cause confusion with clinical users.

Class	Description
1	Clearly not pathogenic
2	Unlikely to be pathogenic
3	Unknown significance (VUS)
4	Likely to be pathogenic
5	Clearly pathogenic

5.3 Report wording for each variant class

It is ***recommended*** that laboratories adopt where possible, taking into account local policies, standard wording within reports to describe each of the 5 variant classes as suggested in the table below. If local policies preclude the use of the recommended wording, it is ***essential*** that consistent wording is used for each class of variant, and the wording is at least comparable to the recommendations below, so that readers of reports will interpret the wording consistently regardless of their origin.

Class	Wording to include within reports
1	**Not** pathogenic "Common" polymorphism and therefore not reported
2	**Unlikely** to be pathogenic Diagnosis not confirmed molecularly
3	**Uncertain** pathogenicity **Does not** confirm or exclude diagnosis
4	**Likely** to be pathogenic **Consistent** with the diagnosis
5	Predicted to be **pathogenic** This result **confirms** the diagnosis

5.4 Which variants to report

It is important that reports meet the needs of users and therefore local policies may determine which of the 5 classes of variant are reported. Even though not all variants will be reported it is ***essential*** that all variants are recorded within the laboratory. It is ***not essential*** to report class 1 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. Local policy will determine whether class 2 variants are reported and again it is acceptable to include a disclaimer as above. It is ***essential*** to report classes 3, 4 and 5. Classes 3 and 4 may warrant follow up studies to clarify the significance of a variant, which should be stated clearly within the report. The nature of follow up tests will be condition and scenario specific and may require appropriate discussion with the clinical team. It is important to offer follow up studies if there is clinical benefit. However, the wording used on reports should be chosen carefully, as certain studies may not be possible or appropriate for a given patient or family. Therefore, use of words such as "recommended" should be avoided. Instead, the report should use wording such as "further possible studies…", and/or offer testing of relatives "if available" or "if appropriate".

5.5 To whom is it acceptable to report VUS

Variants of uncertain significance (class 3) have the potential to cause confusion. It is ***recommended*** that the reports describing the identification of a VUS are issued to appropriately trained clinicians. In the vast majority of cases this is likely to be clinical geneticists or genetic counsellors. However we recognise that other healthcare professionals may be equally competent in the interpretation of VUSs. Extreme caution should be taken when issuing a report of a VUS 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 ***appropriate*** 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 VUS without referral to a genetics clinic (see section 5.4 above).

5.6 Predictive and prenatal testing

Testing for a VUS in a predictive and prenatal 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 counselling.

Acknowledgements:

This document updates guidelines originally prepared and edited by Jennie Bell, Danielle Bodmer, Erik Sistermanns and Simon Ramsden (January 2008) and represents the output of the workshop to review the unclassified variant practice guidelines held on 29th March 2011 at Birmingham Women's Hospital NHS Foundation Trust. Changes were agreed by attendees representing the UK Clinical Molecular Genetics Society, with further changes made by representatives of the Dutch Society of Clinical Genetic Laboratory Specialists. Final adjustments were made to the revised edition following a UK NEQAS for Molecular Genetics participants meeting on Unclassified Variants held in Edinburgh, March 2013.

REFERENCES

Abkevich V, Zharkikh A, Deffenbaugh AM, Frank D, Chen Y, Shattuck D, Skolnick MH, Gutin A, Tavtigian SV (2004) Analysis of missense variation in human BRCA1 in the context of interspecific sequence variation. J Med Genet. **41**(7):492-507.

Aniba MR, Poch O, Thompson JD (2010) Issues in bioinformatics benchmarking: the case study of multiple sequence alignment. Nucleic Acids Res. **38**(21):7353-63.

Chan PA, Duraisamy S, Miller PJ, Newell JA, McBride C, Bond JP, Raevaara T, Ollila S, Nystrom M, Grimm AJ, Christodoulou J, Oetting WS, Greenblatt MS (2007) Interpreting missense variants: comparing computational methods in human disease genes CDKN2A, MLH1, MSH2, MECP2, and tyrosinase (TYR). Hum Mutat. **28**:683–693.

Dorfman R, Nalpathamkalam T, Taylor C, Gonska T, Keenan K, Yuan XW, Corey M, Tsui LC, Zielenski J, Durie P (2010) Do common *in silico* tools predict the clinical consequences of amino-acid substitutions in the CFTR gene? Clin Genet. **77**(5):464-73.

Greenblatt MS, Beaudet JG, Gump JR, Godin KS, Trombley L, Koh J, Bond JP (2003) Detailed computational study of p53 and p16: using evolutionary sequence analysis and disease-associated mutations to predict the functional consequences of allelic variants. Oncogene. **22**(8):1150-63.

Hicks, S., Wheeler, D. A., Plon, S. E. and Kimmel, M. (2011), Prediction of missense mutation functionality depends on both the algorithm and sequence alignment employed. Human Mutation, **32**: 661–668. doi: 10.1002/humu.21490.

Houdayer, C., Caux-Moncoutier, V., Krieger, S., Barrois, M., Bonnet, F., Bourdon, V., Bronner, M., Buisson, M., Coulet, F., Gaildrat, P., Lefol, C., Léone, M., Mazoyer, S., Muller, D., Remenieras, A., Révillion, F., Rouleau, E., Sokolowska, J., Vert, J.-P., Lidereau, R., Soubrier, F., Sobol, H., Sevenet, N., Bressac-de Paillerets, B., Hardouin, A., Tosi, M., Sinilnikova, O. M. and Stoppa-Lyonnet, D. (2012), Guidelines for splicing analysis in molecular diagnosis derived from a set of 327 combined *in silico*/*in vitro* studies on *BRCA1* and *BRCA2* variants. Hum. Mutat., **33**: 1228–1238

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.

Karchin R (2009) Next generation tools for the annotation of human SNPs. Brief Bioinform **10**(1):35-52.

Kemena C, Notredame C (2009) Upcoming challenges for multiple sequence alignment methods in the high-throughput era. Bioinformatics. **25**(19):2455-65.

Li WH, Wu CI, Luo CC (1984) Nonrandomness of point mutation as reflected in nucleotide substitutions in pseudogenes and its evolutionary implications. J Mol Evol.**21**:58–71.

Li, L., Biswas, K., Habib, L. A., Kuznetsov, S. G., Hamel, N., Kirchhoff, T., Wong, N., Armel, S., Chong, G., Narod, S. A., Claes, K., Offit, K., Robson, M. E., Stauffer, S., Sharan, S. K. and Foulkes, W. D. (2009), Functional redundancy of exon 12 of *BRCA2* revealed by a comprehensive analysis of the c.6853A>G (p.I2285V) variant. Hum. Mutat., **30**: 1543–1550.

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

Mathe E, Olivier M, Kato S, Ishioka C, Hainaut P, Tavtigian SV (2006) Computational approaches for predicting the biological effect of p53 missense mutations: a comparison of three sequence analysis based methods. Nucleic Acids Res. **34**(5):1317–1325.

Møller P, Clark N, Mæhle L. (2011) A simplified method for segregation analysis (SISA) to determine penetrance and expression of a genetic variant in a family. Hum Mutat **32**;1-4

Plon SE, Eccles DM, Easton D, Foulkes WD, Genuardi M, Greenblatt MS, Hogervorst FB, Hoogerbrugge N, Spurdle AB and Tavtigian SV (2008). Sequence variant classification and reporting: recommendations for improving the interpretation of cancer susceptibility genetic test results. Human Mutation, **29**: 1282–1291.

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

Spurdle AB (2010) Clinical relevance of rare germline sequence variants in cancer genes: evolution and application of classification models. Current Opinion in Genetics & Development, **20** (3): 315-323.

Tavtigian SV, Greenblatt MS, Lesueur F and Byrnes GB (2008) In silico analysis of missense substitutions using sequence-alignment

based methods. Human Mutation, **29**: 1327–1336.

Tavtigian SV, Oefner PJ, Babikyan D, Hartmann A, Healey S, Le Calvez-Kelm F, Lesueur F, Byrnes GB, Chuang S-C, Forey N, Feuchtinger C, Gioia L, Hall J, Hashibe M, Herte B, McKay-Chopin S, Thomas A, Vallée MP, Voegele C, Webb PM, Whiteman DC, Australian Cancer Study, Breast Cancer Family Registries (BCFR), Kathleen Cuningham Foundation Consortium for Research into Familial Aspects of Breast Cancer (kConFab), Sangrajrang S,. Hopper JL, Southey MC, Andrulis IL, John EM, Chenevix-Trench G. (2009) Rare, Evolutionarily Unlikely Missense Substitutions in ATM Confer Increased Risk of Breast Cancer. The American Journal of Human Genetics. **85(**4):427-446

Thusberg, J. and Vihinen, M. (2009), Pathogenic or not? And if so, then how? Studying the effects of missense mutations using bioinformatics methods. Human Mutation, **30**: 703–714. doi: 10.1002/humu.20938

Vallée MP, Francy TC, Judkins MK, Babikyan D, Lesueur F, Gammon A, Goldgar DE, Couch FJ and Tavtigian SV (2011) Classification of missense substitutions in the BRCA genes: A database dedicated to Ex-UVs. Human Mutation. doi: 10.1002/humu.21629

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.

Appendix A

Missense prediction tools

A catalogue of missense variant evaluation tools is available from http://www.ngrl.org.uk/Manchester/page/missense-prediction-tools. Tools are loosely grouped based on their algorithm into three categories:

1. Sequence and evolutionary conservation-based methods
2. Protein sequence and structure-based methods
3. Supervised learning methods

Splice site prediction tools

There are a number of tools available for splice site prediction. These include:

GeneSplicer (http://www.cbcb.umd.edu/software/GeneSplicer/gene_spl.shtml)
A method based on a combination of predictive approaches including Markov models.

Human Splice Finder (http://www.umd.be/HSF/)
A prediction method based on position weight matrices.

MaxEntScan (http://genes.mit.edu/burgelab/maxent/Xmaxentscan_scoreseq.html)
A prediction method based on the maximum entropy particle.

NetGene2 (http://www.cbs.dtu.dk/services/NetGene2/)
A prediction method based on neural networks.

NNSplice (http://www.fruitfly.org/seq_tools/splice.html)
A prediction method based on neural networks.

SpliceSiteFinder-like
A method based on position weight matrices and implemented through the Alamut software package.

(The NNSplice, MaxEntSplice, GeneSplicer and SpliceSiteFinder-like methods have been assessed and found to perform well - http://www.ngrl.org.uk/Manchester/sites/default/files/publications/Informatics/NGRL_Splice_Site_Tools_Analysis_2009.pdf)

Tools for prediction of exonic/intronic splice enhancers and inhibitors (ESEs, ISEs, ESIs and ISIs) are also available:
ESEFinder (http://rulai.cshl.edu/cgi-bin/tools/ESE3/esefinder.cgi?process=home)

Resources for mutation analysis

Locus Reference Genomic website
LRG sequences provide a stable genomic DNA framework for reporting mutations with a permanent ID and core content that never changes
http://www.lrg-sequence.org/home

1000 Genomes Project (http://www.1000genomes.org/data)
The 1000 genome project is an international collaboration that aims to understand common variation in the human genome. When completed, the 1KG project aims to have defined a well-characterised dataset of common and rare variants with a frequency of at least 1%. Sequenced individuals have been grouped into populations based on their shared ancestry and include African, Asian, American and European. Variant MAFs can vary between populations and therefore ancestry can be an important consideration. More details on the project and populations used can be found here: http://www.1000genomes.org/about.

Alamut (http://www.interactive-biosoftware.com/alamut.html)
Alamut is a licensed software package available from Interactive Biosoftware and is commonly used in molecular diagnostic labs for evaluating sequence variants.

Café Variome (http://www.cafevariome.org/)
Café Variome acts as a portal for genetic variation data produced by diagnostic labs. It allows users to announce, discover and acquire a comprehensive listing of observed neutral and disease-causing gene variants in patients and unaffected individuals.

dbSNP (http://www.ncbi.nlm.nih.gov/projects/SNP/)
The single nucleotide polymorphism database is an archive of genetic variation within and across species. Variants held include Single Nucleotide Polymorphisms (SNPs), Deletion and Insertion Polymorphisms (DIPs), Short Tandem Repeats (STRs) and Multiple Nucleotide Polymorphisms (MNPs).

DMuDB (https://secure.dmudb.net/ngrl-rep/Home.do)
The diagnostic mutation database provides a resource for accessing and sharing human mutation data within and between diagnostic laboratories. Initially established to support UK genetic testing labs, access has now been extended to non-UK labs through a subscription based service.

HGMD (http://www.hgmd.org/)
The human gene mutation database collates genetic information associated with inherited human disease from the literature. Registration is required but mutation data is freely available 3 years after initial inclusion in the database. Up-to-date mutation data is available via a subscription.

LSDBs
Locus-specific databases store genetic variants of specific genes and diseases. These are often curated and maintained by experts. Lists of LSDBs are available from http://www.lovd.nl/2.0/index_list.php, http://www.hgmd.org/, http://www.gen2phen.org/data/lsdbs and http://www.hgvs.org/dblist/glsdb.html

OMIM (http://www.ncbi.nlm.nih.gov/omim)
The Online Mendelian Inheritance in Man database is a continuously updated catalogue of human genes and genetic disorders, with particular focus on the molecular relationship between genetic variation and phenotypic expression.

NHLBI GO Exome Sequencing Project(http://evs.gs.washington.edu/EVS/)
The Exome Sequencing Project (ESP) was established to discover novel genes contributing to heart, lung and blood disorders. Whereas individuals sequenced in other normal variation datasets are 'assumed healthy', ESP individuals, taken from European American and African American populations, are known to have certain heart, lung and blood diseases. When assessing ESP variants, it is important to consider

the potential phenotype of the patient as well as the populations in which these variants are found.

Mutalyzer (https://mutalyzer.nl/)
Mutalyzer allows checking of sequence variant nomenclature to HGVS guidelines.

Genome Browsers

Genome browsers offer visualisation and browsing of an entire genome, accompanied by annotation relating to specific attributes.

Ensembl (http://www.ensembl.org/index.html)
UCSC Genome Browser (http://genome.ucsc.edu/)
NCBI MapViewer (http://www.ncbi.nlm.nih.gov/mapview/)

Publication databases

PubMED (http://www.ncbi.nlm.nih.gov/pubmed/)
Google Scholar (http://scholar.google.co.uk/)

Other Resources

EMBL-EBI (http://www.ebi.ac.uk/)
The European Molecular Biology Laboratory - European Bioinformatics Institute provides a number of computational tools for the analysis for the analysis of genetic and protein data.

NCBI (http://www.ncbi.nlm.nih.gov/)
The National Center For Biotechnology Information provides databases, tools and resources for the analysis of genetic and protein data.