

Practice guidelines for Sanger Sequencing Analysis and Interpretation.

Prepared and edited by Sian Ellard¹, Ruth Charlton², Michael Yau³, David Gokhale⁴, Graham R Taylor⁵, Andrew Wallace⁶ and Simon C Ramsden⁶. Ratified by the CMGS Executive Committee on 7th August, 2009.

1. Department of Molecular Genetics, Royal Devon & Exeter NHS Foundation Trust, Barrack Road, Exeter, EX2 5AD, UK.
2. Yorkshire Regional DNA Laboratory, St James's University Hospital, Leeds LS9 7TF, UK.
3. Guy's & St Thomas's Hospital NHS Trust, Genetics Centre, 8th Floor, Guy's Tower, Guy's Hospital, London SE1 9RT, UK
4. Liverpool Women's NHS Foundation Trust, Crown Street, Liverpool, L8 7SS, UK.
5. Genomics Facility, Leeds Institute of Molecular Medicine & Leeds Teaching Hospitals, 6.2 Clinical Sciences Building, St James's University Hospital, Leeds, LS9 7TF, UK.
6. Regional Genetic Service, St Marys Hospital, Hathersage Road, Manchester, M13 0JH, UK.

1. INTRODUCTION

1.1 General Introduction

DNA sequencing is the most commonly used approach for both mutation scanning and mutation testing; it is widely regarded as the gold standard. Agreed practice guidelines for both the sequencing process and the interpretation of results are important to achieve a high quality approach with common quality standards across different laboratories. These guidelines do not constitute an experimental protocol or troubleshooting guide, rather they aim to establish consensus standards for identifying and reporting mutations.

Different standards will be required for clinical diagnostics than would be acceptable for a sequence-based research project. Since germline changes are most frequently being analysed, results will stand for the lifetime of the individual and may have implications for relatives of the proband.

This document considers quality aspects of the whole process of sequencing and makes the assumption that the analytical process takes place in an appropriate, accredited laboratory setting where routine aspects of good laboratory practice such as sample tracking and record keeping are in place.

Local sequencing practices may vary both in terms of reasons for investigation, chemistry, hardware, software and reporting of results. These guidelines have been updated from an earlier version by Ravine *et al* (no longer available) that followed a CMGS Sequencing Best Practice Meeting held in 2001. These guidelines identify common elements for each part of the process and specify quality criteria that should be met or exceeded.

Guidelines are described as either:

- **Essential** practice which must be implemented to ensure quality of service
- **Recommended** practice where more than one approach is satisfactory, however there is a clear advantage in following the advice given.

- **Not acceptable**, which highlights areas where the quality of service may be compromised.

1.2 Reasons for diagnostic sequencing

The reason for the sequencing investigation may influence the quality standards required. Two types of investigation can be considered:

- 1) Confirmation or exclusion of a known sequence variant (genotyping).
- 2) Full characterisation of a defined region of DNA (mutation scanning or re-sequencing).

From the perspective of the required quality standards these two processes are not equivalent; whereas genotyping is concerned only with base changes previously identified, fully characterising a region of DNA implies that each base has been determined with high confidence. So if the average probability of error for each base call of a 1 kilobase region was 1% (equivalent to a Phred score of 20 – see later), then one would expect 10 erroneous base calls which is clearly unacceptable for diagnostic purposes.

2. QUALITY ASPECTS OF THE LABORATORY PROCESS

2.1 Patient material

Suitable material for sequence analysis is essential. This requires correct identification of the proband, appropriate clinical diagnosis and the sample must be collected, identified, recorded and stored under quality controlled conditions appropriate for diagnostic testing. For example if a case has been identified as part of a research project it may be necessary to collect an additional sample. Genomic DNA from peripheral white blood cells is the typical starting material. Alternative sources such as fixed tissue or cDNA may raise quality control issues that are beyond the scope of these guidelines.

2.2 Template preparation, sequencing reactions and post-sequencing clean-up

Issues relating to specific methodologies are beyond the scope of these guidelines. Appropriate negative controls (ie water blanks) should be included at the PCR and sequencing stages. It is not necessary to check successful PCR amplification prior to sequencing unless difficulties arise during the sequencing process. Suitable templates should contain a single PCR fragment, or in the case of multiplex sequencing, defined PCR multiplex fragments. A variety of template purification, sequencing chemistries and methods of dye terminator removal have been found to be satisfactory for diagnostic testing.

2.3 Checking for SNPs

For predictive tests where the familial mutation is not detected, it is *essential* that the primer sequences are checked for the presence of SNPs. A useful tool for this purpose, Diagnostic SNPcheck, can be found at <http://ngri.man.ac.uk/>. When using this tool it should be verified that the target primer binding sites are correct and that they match the primer fully. When a SNP is found within a primer sequence it is *recommended* that the primer is re-designed and re-checked.

The re-designing of primers encompassing published SNPs is a vital component to the quality assurance of this procedure, however as more SNPs are described with smaller minor allele frequencies it is recognised that it may not always be possible to identify primer binding sites entirely free from published SNPs. Presence in the dbSNP database alone does not indicate SNP frequency and it is outside the scope of this document to recommend a minor allele frequency cut off for the requirement to re-design primers.

2.4 Data storage

Any system that enables unambiguous sample tracking through the testing, results and reporting procedure is acceptable. The raw data in the form of the sequence output files should be stored together with details of the test and analyses. The results should be indexed with the reporter and result checker identified as well as the date of the analysis and/or report. The reference sequence should be an EMBL or GenBank reference sequence with the accession number and version identified.

2.5 Reference sequences

The complete reference sequence identifier should be included in any description of variants according to HGVS nomenclature guidelines. The latest version of the sequence should be obtained from NCBI.

2.5.1 RefSeq sequences

RefSeq sequences are preferred as they aim to be more comprehensive and better-annotated than other GenBank reference sequences. In addition RefSeq sequences are curated and non-redundant, i.e. there is only one RefSeq sequence per gene or isoform. They are identified by the prefix NM_ for mRNA sequences, NP_ for peptides and NG_ for genomic sequences. The latter are a product of the RefSeqGene project which aims to provide gene-centered genomic sequences annotated with flanking and intronic

coordinates to act as stable references for variation data and exon numbering.

2.5.2 LRG Sequences

Locus Region Genomic (LRG) sequences are currently being developed to provide a stable genomic reference sequence for each gene region annotated with transcripts and other associated data. LRGs aim to remain fixed to ensure standardization of nomenclature, so new versions will not be produced.

RefSeqGenes and LRGs are new standards that have not yet become fully established in practice however their use should be considered where possible.

2.6 Outsourcing

Outsourced work should meet the same criteria and labs must be able to demonstrate good practice e.g. by ISO 15189 (clinical diagnostic laboratories) and/or ISO 17025 (testing and calibration laboratories). Sequence output files should be supplied with the result for independent quality assessment.

3. ASSESSMENT OF SEQUENCE QUALITY

High quality sequence data is *essential* but this is difficult to measure because there is no standardized parameter for assessment of either raw or processed sequence quality, nor are any of the hardware platforms supplied as fit for diagnostic use. An international quality assurance scheme reported large variation in sequence quality across 64 participating laboratories from 21 countries (*Patton et al 2006*). The quality parameters of one example data set have been described in more detail in the context of a CMGS study (*Ellard et al 2009*) but there is a need for further empirical studies to clarify performance criteria across different platforms and using different software. At present the following options exist:

1. *Visual inspection of base calling quality of a region of DNA from a single read.* In this case, suitable quality means that base spacing is regular, although band compression might be seen in G rich regions. Peak heights should be relatively even and minor peaks due to background noise should on average be less than 5% of the intensity of the major peak. Any part of a sequence that does not meet these criteria should be excluded from the reported result.

2. *Quality estimates based on signal:noise ratios for the entire sequence or the sequence across a user defined region of interest (ROI).* For example, the Mutation Surveyor software (Softgenetics, Pennsylvania, US) generates quality scores where 100 is equivalent to $\leq 1\%$ noise, a score of 50 denotes 2% noise and 20 corresponds to 5% noise. (Please note that in Mutation Surveyor versions 3.10 and 3.20, there is a maximum ROI quality score of 50 for a sequence with $\leq 1\%$ noise and a score of 20 denotes 2.5% noise although the scores for the entire sequence are as stated above). It is *recommended* that diagnostic quality sequence should meet or exceed a mean ROI quality score equivalent to $\leq 2.5\%$ noise at least in one orientation.

3. *Confidence estimates using base calling QA programs such as Phred that are calculated for individual bases or can be expressed as an average across a sequence or ROI.* These methods have the benefit of providing an objective measure of base calling quality, usually expressed as 10x the log of

likelihood of error, although the reliability of these estimates is uncertain. Phred is the prototype base caller and quality assessor (*Ewing and Green, 1998*). Using peak mobility and shape analysis, a likelihood of an error in base calling is as follows: Phred 20 means a 1/100 likelihood of error (ie 99% confidence that the base is called correctly), Phred 30 means 1/1000 (99.9% confidence), Phred 40 means 1/10,000 and so on. Phred scores were originally developed to maximise the information gained from genome sequencing projects, and they indicate the likelihood of a base being mis-called for de-novo sequencing. However in the context of re-sequencing we have a prior knowledge of the sequence and are also making a comparison to a reference trace. Phred scores are therefore probably over-estimating the probability of an error at the level of a re-sequenced base. Where Phred scores are used to measure quality of sequencing data of the ROI a mean Phred of 20 is **recommended** for bi-directional data and 30 is **recommended** for unidirectional data.

The above options relate to quality assessment of processed data. The possibility of obtaining an erroneous result from a failed PCR reaction due to low level cross-contamination by another patient's PCR product highlights the need to check the raw data. This may be achieved either by gel electrophoresis (to check that PCR failures correlate with sequencing failures), by using software tools such as Sequence Scanner 1.1 (freely available from Applied Biosystems) to examine parameters such as raw signal intensity, or by extracting these data directly from the .ab1 files.

4. IDENTIFICATION OF VARIANTS

4.1 Sequence-based assessment for a known mutation

Accurate base calling should extend at least 10 bases either side of the mutation in question, or 10 bases upstream of an insertion/deletion mutation.

Bi-directional sequencing is **not essential**, but it is **essential** that a visual inspection of the base in question is undertaken to investigate the possibility of a mosaic mutation present in an unaffected parent.

It is **recommended** that a positive control is included in the same experiment. In order to minimise the chance of a false negative result caused by allele dropout, the positive control should, where possible, be from the same pedigree as the proband. If a positive control is not available, this should be noted on the report. A high quality normal control sequence must be available for comparison with the patient sample but it is not necessary to include a normal control in each batch of samples in which case a stored sequence from a previous run is **recommended**.

4.2 Sequence-based scanning for an unknown mutation

The extent of the analysis must be defined. For example, a minimal region of coverage would include the coding regions of the gene and the conserved splice sites. There is currently no consensus within splice site prediction software regarding the minimum region of interest and consequently the extent of intronic sequence screened must be defined according to the distribution of known mutations in a particular gene.

This ROI will be gene specific and may extend to the minimal promoter, branch sites, 3' untranslated region (eg the polyadenylation site) and/or additional intronic sequence depending on the distribution of known mutations.

Bi-directional sequencing is **not essential** for the detection of heterozygous (or homozygous) mutations provided that high quality sequence data is obtained for the complete ROI in one direction. There is no known biological reason for a heterozygous base substitution being visible in only one direction and bi-directional analysis is not always possible due to the presence of intronic insertion/deletion polymorphisms that generate mixed sequences in heterozygotes. However, bi-directional sequencing is **recommended** for the detection of mosaic mutations.

Sole reliance on unassisted visual inspection of the sequence data is **not acceptable** due to the possibility of operator error, particularly for homozygous mutations (unpublished data, CMGS sequencing study 2003). A multitude of software analysis programs are available for the identification of variants in patient samples by comparison with the reference (Genbank) sequence, but published data regarding their sensitivity and specificity in the clinical diagnostic setting is often lacking. One exception is Mutation Surveyor where a CMGS-led study across three laboratories demonstrated >99.59% sensitivity (with 95% confidence) for unidirectional analysis of heterozygous base substitutions in a data set comprising 701 variants (*Ellard et al 2009*).

Analysis of sequences using the Trace_diff module of the Staden sequence analysis package (*Bonfield et al 1998*) relies upon a comparison with a normal control sequence. This normal control must be of high quality control (see section 3 above) but it is not necessary to include a normal control in each batch of samples if a stored sequence is available that has been generated using the same chemistry and electrophoresis conditions. The "synthesis" control sequence generated by Mutation Surveyor can be used, but greater sensitivity may be achieved through the use of a normal control sequence (Patch, Guegan and Ellard, unpublished data).

4.3 Semi-automated mutation scanning

Semi-automated analysis relies upon the automatic analysis of sequence data with visual inspection confined to checking of variants and sequences, or bases within a sequence, that do not meet defined quality parameters. The choice of quality parameter(s) will depend upon the software package used, but might include a threshold signal:noise ratio for the entire region of interest and a minimum Phred score for individual bases within a sequence read. It is acceptable to use semi-automated analysis, but the thresholds for visual inspection should be determined using an evidence-based approach within individual laboratories to minimise the risk that a mutation is not detected due to poor quality sequence.

5. REPORTING

5.1 Positive results

In the absence of a robust tube transfer checking system (eg. barcode scanning, witnessed transfers) to assure sample identity at each stage, mutations identified for the first time in a given pedigree should be confirmed by a second test, using a fresh template.

5.2 Negative results

Negative results should be reported for the range of nucleotides where high quality sequence data was analysed. The extent of the analysis should be defined within the report.

There is a risk of false negative results from :

- a) PCR or sequencing primer binding site polymorphisms that cause allelic dropout
- b) Tissue mosaicism
- c) Preferential amplification of the smaller allele in a large insertion.
- d) Deletion of an entire exon/gene

For predictive tests where the familial mutation is not detected, it is *essential* that the primer sequences are checked for the presence of SNPs. A useful tool, Diagnostic SNP check, can be found at <http://ngri.man.ac.uk/>.

5.3 Reports

Where possible reports should integrate sequence data with the clinical information that has been provided.

The “test sensitivity” in the context of the clinical setting (i.e. the proportion of cases with the phenotype in question in which a mutation is detected by testing strategy that has been employed) should be reported for negative results.

It is *essential* that mutations are described in accordance with the Human Genome Variation Society (HGVS) recommendations <http://www.hgvs.org/mutnomen/>. These guidelines recommend use of a coding reference sequence wherever possible and specify that the A of the translation initiation codon ATG is base 1. Mutalyzer software (<http://www.humgen.nl/mutalyzer>) can be used to check sequence variant nomenclature. An mRNA reference sequence (NM_ref seq) with a version number should be included on the report. Because of the established use of previous nomenclature, it is helpful for the common names to be referenced alongside the HGVS version.

It is useful to include references if the mutation has previously been reported. Reasonable efforts should be made to check if any current locus-specific database has documented the mutation.

The pathogenic significance of missense or non-coding mutations is not always clear. Best practice guidelines for the interpretation of novel variants are available on the CMGS website (www.cmgs.org).

6. REFERENCES

1. Dunnen JT and Antonarakis SE (2000) Mutation nomenclature extensions and suggestions to describe complex mutations: a discussion. *Hum.Mutat.* **15**,7-12.
2. Bonfield, JK, Rada, C & Staden R (1998) *Nucleic Acids Res* **26**,3404-9
3. Ellard S, Shields B, Tysoe C, Treacy R, Yau S, Mattocks C & Wallace A (2009) Semi-automated unidirectional sequence analysis for mutation detection in a clinical diagnostic setting. *Genetic Testing and Molecular Biomarkers*, In Press
4. Ewing B and Green P (1998) Base-calling of automated sequencer traces using Phred. *Genome Res.* **8**,186-94.
5. Patton SJ, Wallace AJ & Elles R (2006) Benchmark for evaluating the quality of DNA sequencing: proposal from an international external quality assessment scheme. *Clin Chem* **52**,728-36.