Practice guidelines for Targeted Next Generation Sequencing Analysis and Interpretation.

Prepared and edited by Sian Ellard[1,2], Helen Lindsay[3], Nick Camm[3], Chris Watson[3], Steve Abbs[4], Yvonne Wallis[5], Chris Mattocks[6], Graham R Taylor[7] and Ruth Charlton[3].

1. Department of Molecular Genetics, Royal Devon & Exeter Hospital, Exeter, EX2 5AD, UK.
2. University of Exeter Medical School, Barrack Road, Exeter, EX2 5DW, UK.
3. Yorkshire Regional DNA Laboratory, St James's University Hospital, Leeds LS9 7TF, UK.
4. East Anglian Medical Genetics Service, Addenbrooke's Hospital, Cambridge CB2 0QQ, UK.
5. West Midlands Regional Genetics Laboratory, Birmingham Women's NHS Foundation Trust, Birmingham, B15 2TG, UK
6. Wessex Regional Genetics Laboratory, Salisbury District Hospital, Salisbury, SP2 8BJ, UK
7. Department of Pathology, University of Melbourne, Melbourne, VIC 3010, Australia.

Original guidelines ratified by the Clinical Genetics Molecular Genetics Society (December 2012).

Guidelines updated by the Association for Clinical Genetic Science (formally Clinical Molecular Genetics Society and Association of Clinical Cytogenetics) approved May 2014.

1. INTRODUCTION

DNA sequencing is the most commonly used approach for mutation scanning and is widely regarded as the gold standard. Agreed practice guidelines for both the sequencing process and data analysis 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.

Next generation sequencing, also described as "second generation" or "massively parallel" has replaced Sanger sequencing as the primary methodology employed by researchers to identify novel disease genes. The ability to simultaneously analyse multiple or very large genes at a cheaper cost per base makes next generation sequencing an attractive solution for clinical diagnostic testing to identify the disease-causing mutation (or mutations) in patients with genetically heterogeneous disorders. Targeted next generation sequencing describes a strategy where a specific set of genes related to the patient's phenotype are analysed within the context of a genetic "test". For practical purposes a broader spectrum of genes may actually be sequenced in order to achieve a streamlined laboratory pipeline, described as the "assay". This approach minimizes the possibility of identifying highly penetrant mutations in genes unrelated to the patient's phenotype.

Targeted next generation sequencing services were introduced into the NHS in 2010 and whilst the laboratory and analytical protocols are both varied and rapidly evolving, these guidelines aim to describe the general principles that underlie the quality requirements of this technology. This document considers quality aspects of the whole process of targeted next generation sequencing. It is essential that this process 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).
Local sequencing practices may vary both in terms of the targeting strategy, sequencing chemistry/hardware, analysis software and reporting of results. These guidelines are based on the principles established for Sanger sequencing (CMGS 2009 http://www.cmgs.org/BPGs/Best_Practice_Guidelines.htm). They 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.

2. VALIDATION OF NEXT GENERATION SEQUENCING TESTS

It is essential to validate any new laboratory test. This includes validation of all aspects of the test process, i.e. targeting method, sequencing process and data analysis. The validation requirements will vary according to the methodology employed and the context of the clinical diagnostic application (*Mattocks et al 2010*). It is important to understand the technical weaknesses of the methodology in order to ensure that these are assessed during the validation (for example homopolymer tract errors that vary according to platform and the lower sensitivity for detecting base substitutions compared to insertions and deletions.

2.1 Reproducibility

It is essential to derive information on reproducibility and robustness (particularly in terms of horizontal coverage) during the validation phase. It is recommended that validation samples are analysed from at least three independent sequence runs. Run-to-run comparisons will determine the level of multiplexing possible to ensure minimum diagnostic vertical coverage.
If the same test protocol is applied to multiple genes then it is acceptable to perform the validation at the level of the process rather than the gene. Verification of gene panels on a smaller scale can subsequently be performed to check that equivalent results are obtained on a given panel compared to validation data. This should include analysis of read depth (vertical coverage) across the targeted region, and sequencing of positive control samples when available. It is essential that the positive controls are representative of the disease mutation spectrum.

2.2 Sensitivity

The required sensitivity of the test will depend upon the clinical application. For example, replacing a Sanger sequencing test with a high pick-up rate may demand a higher sensitivity per gene than a large panel (10s to 100s of genes) in which the prior likelihood of finding a mutation is low and testing has not previously been available.
It is essential that laboratories are able to demonstrate 95% confidence that the error rate for heterozygote/homozygote mutation detection is <5%. This requires concordant results for a minimum of 60 unique variants tested by the new method in an independent, blinded analysis and compared with the gold standard (*Mattocks et al 2010*). To achieve 95% confidence that the error rate for is <1% requires concordant results for 300 unique variants. It is recognized that this is unlikely to be feasible in the diagnostic setting. For laboratories using common methodologies (i.e.

targeting methodology and sequencing platform) evidence might be obtained by pooling data as was previously done during the validation of unidirectional semi-automated Sanger sequence analysis by the CMGS (*Ellard et al 2009)*, but a multi-centre validation will be subject to variation in analysis pipelines between laboratories.

3. QUALITY ASPECTS OF THE LABORATORY PROCESS

3.1 Patient material

It is essential that suitable material for sequence analysis is available, that the proband has been correctly identified and that the appropriate clinical diagnosis and/or phenotype information is provided. 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 may raise quality control issues that are beyond the scope of these guidelines.

3.2 Targeting methodology

Targeting methodologies include PCR amplification (long range or standard amplicon), hybridization (liquid capture using RNA or DNA "baits") or methods developed specifically for next generation sequencing (eg Haloplex). Issues relating to specific methodologies are beyond the scope of these guidelines, but for PCR-based methods it is essential to check for primer binding site SNPs that could cause allelic drop out (CMGS 2009 http://www.cmgs.org/BPGs/Best_Practice_Guidelines.htm).

It is essential to consider that some regions of the human genome are very difficult or impossible to sequence. These include:

a) genes where the presence of pseudogenes makes unique targeting difficult
b) highly GC rich regions
c) repetitive elements

3.3 Template library preparation

Sample multiplexing is an integral part of next generation sequencing protocols and individual sample identification is obtained by "index tagging" or using "molecular barcodes". This refers to the process of adding a unique DNA sequence to each patient sample during the library preparation process in order that the sequences obtained from that patient sample can be extracted from the total sequence data. It is essential that there is a robust system (witness or barcode check) to ensure that the index tag recorded for the patient matches the index tag added during library preparation and when index sorting at the analysis stage. Where custom index tags are used it is essential to have more than 1bp difference (or edit distance) between tags to minimize the risk of errors during synthesis or sequencing that could generate two identical tags. An edit distance of at least 3bp is recommended. Laboratories might also consider typing a set of SNPs by an alternative method to check that genotypes generated from the sequencing are concordant and confirm sample identity.

Any PCR steps within the template preparation require inclusion of a negative control to check for sample contamination.

3.4 Sequencing

A choice of sequencing chemistries/platforms is available but issues relating to specific methodologies are beyond the scope of these guidelines.

3.5 Outsourcing

It is recommended that outsourced work only be performed by certified laboratories working to recognised international quality standards (such as ISO 17025 and 15189).

The outsourcing laboratory may choose to prepare the sequencing libraries in-house but outsource the sequencing to an accredited external provider. In this situation the sample identity is defined by the index tag or molecular barcode incorporated at the library preparation stage and any sample mix-up in the outsourcing laboratory should not result in an incorrect test result for the patient provided sufficient unique identifiers are available for use.

4. DATA ANALYSIS

4.1 Data quality and depth of coverage

Quality metrics are generated at multiple stages of the analytical process, for example quality scores associated with the base, the read, alignment, variant call or strand bias. The acceptable thresholds should be determined during the validation process.

It is recommended that the following data quality markers are logged as part of the sequencing audit trail. These technical details are not routinely required on a clinical report, but in some circumstances it may be useful to include some of this data to help explain the results.

- Average base call quality scores for each position as a phred-like value for data to be used in analysis. Filtering or cutoff criteria for the exclusion of reads or bases from downstream analysis.
- Mapping quality scores if genome wide alignment is performed.
- Number of reads mapped and percentage of target covered at the minimum coverage required.
- The alignment algorithm and alignment settings (seed length, mismatch tolerance, mismatch penalties, gap penalties and gap extension penalties) should be recorded.

Minimum depth of coverage will depend upon the required sensitivity of the assay, the targeting/sequencing method and the type of mutation detected. The minimum read depth should be evidence based and will be established during the test validation process. Regions of sequence not meeting the required read depth might be tested using other methods, e.g., Sanger sequencing (especially important when replacing an existing Sanger test), described within the report as low coverage or might not be required for the clinical report if a definite pathogenic mutation (or mutations) has already been identified.

4.2 Defining the region of interest (ROI)

It is essential that the extent of the analysis is defined. For example, a minimal region of horizontal 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 or local laboratory policy. The ROI may extend to the minimal promoter, known enhancers, branch sites, 3' untranslated region (eg the polyadenylation site) and/or additional intronic sequence depending on the distribution of known mutations.

4.3 Analysis pipeline to identify variants

Software for data analysis may be supplied commercially or be open source. Most commercial programs cover the analytical process from read alignment to variant annotation but in-house pipelines utilize different programs for read alignment, removal of duplicate reads, indel realignment, quality calibration, variant calling and annotation. Accurate versioning is essential and each software upgrade requires revalidation. There are multiple settings options and these again must be determined during the test validation. Validation of software updates may be achieved using an existing data set and does not necessitate additional laboratory work.

4.4 Copy number analysis

For some targeting strategies (eg hybridization capture) it is possible to detect large deletions or duplications that would not be detected by Sanger sequencing because they span one or both primer binding sites. This is achieved through comparative depth of coverage analysis between normal controls and patient samples. However, at the time of writing copy number analysis is not available within commercial software programs and open source programs require validation within the clinical diagnostic setting. The size thresholds for detection of indels by variant calling within reads or by copy number analysis have not yet been determined.

4.5 Annotation of variants

It is essential that variants are described in accordance with the Human Genome Variation Society (HGVS) recommendations (http://www.hgvs.org/mutnomen/). The reference sequence (with a version number if appropriate) should be included on the report. It is recommended that the genomic coordinates with hg build number are also recorded. The HGVS guidelines recommend use of a LRG (Locus Reference Genomic sequence) if available for the gene of interest (see http://www.lrg-sequence.org/). LRGs aim to remain fixed to ensure standardization of nomenclature. In the absence of an LRG a coding DNA reference sequence may be used. The coding sequence should preferably be derived from the RefSeq database (http://www.ncbi.nlm.nih.gov/RefSeq/). The report should 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. The transcript version used to annotate the variant must be specified

4.6 Filtering of variants

The strategy for filtering variants to remove polymorphisms of no clinical significance will depend upon the likely mode of inheritance. For example a heterozygous variant reported in an unaffected adult is unlikely to be the cause of a dominant, congenital disorder, but other heterozygous variants found in unaffected adults may be recessive mutations. It may be appropriate to filter out polymorphisms according to minor allele frequency, but threshold settings remain to be established and some recessive mutations are relatively common in certain populations. There are multiple variant databases (eg. DMuDB, Exome Variants Server, 1000 genomes, dbSNP, HGMD, DECIPHER and in-house) which vary according to the population (including age, disease status, ethnicity), variant type, data quality and annotation (including pathogenicity status). An understanding of the structure and content of these databases is essential in order to utilize them effectively. In house variant databases can easily be established by export of variant files (eg .vcf) to LOVD3 (http://www.lovd.nl/3.0/).

5. DATA STORAGE

Guidelines from the Royal College of Pathologists and the Institute of Biomedical Science (2005) recommend that data and records pertaining to pathology tests are retained for a minimum of 25 years.

Storage of next generation sequence data challenges this guideline since (1) it is not feasible to retain the raw image files due to their size (2) Analysis methods are continually updated and hence repeating the historical data analysis is unlikely to be possible and (3) DNA samples are stored indefinitely so the raw material will be available for re-testing.

It is essential to store the output file from the variant annotation step (eg. vcf file) and some laboratories may choose to also retain the FastQ, SAM or BAM files in order to re-analyse the read data in the future. These data should be stored together with a log of the informatics processing that was applied to the raw data in order to make the sequence and/or mapping files.

6. REPORTING

6.1 General principles

Reports should follow the general principles described in the ACGS reporting best practice guidelines (http://www.acgs.uk.com/quality-committee/best-practice-guidelines/). Where possible reports should integrate sequence data with the clinical information that has been provided and variants should be annotated as described in 4.5 above.

Because of the established use of previous nomenclature, it is helpful for the common names to be referenced alongside the HGVS version.

6.2 Positive results

In the absence of a proven, robust tube transfer checking system (eg. barcode scanning, witnessed transfers) or a secondary test (eg SNP assay) to assure sample identity at each stage, it is essential to confirm mutations or variants included in a clinical report by an independent test from a new DNA dilution. This will provide additional data regarding test specificity (false positive rate) to support discontinuation of confirmatory testing once a tube transfer checking process or secondary identity test has been validated.
It is recommended to include references for previously reported missense mutations and splicing mutations outside the conserved splice donor/acceptor sites.

6.3 Negative results

For negative results it is essential to include the following information:

(a) The expected diagnostic yield if known (i.e. the proportion of cases with the phenotype in question in which a mutation is detected by testing strategy that has been employed). Where the diagnostic yield is not known, it is recommended that patient results are stored in order to establish evidence for the diagnostic yield in the future.
(b) The extent of the test defined as the genes or regions analysed (horizontal coverage);
(c) The analytical sensitivity of the test defined by the read depth (vertical coverage);
(d) The spectrum of mutations that are detectable;
(e) Limitations of the assay that may result in false negative results. Examples include:

- Incomplete coverage of targeted regions
- PCR primer binding site polymorphisms that cause allelic dropout (amplicon sequencing methods only)
- Tissue mosaicism
- Large deletion/duplication in the absence of copy number analysis
- Translocation or inversion

If horizontal coverage is incomplete and/or read depth (vertical coverage) does not meet the defined minimum it may be appropriate to recommend further testing depending on the clinical diagnosis.

If insufficient space is available within the clinical report, this information may be provided via a website link (and documented within the SOP).

6.4 Variants of unknown clinical significance

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 ACGS website (www.acgs.uk.com).
Next generation sequencing tests for large gene panels may identify multiple variants of uncertain clinical significance. The prior probability of finding a pathogenic mutation in most genes within large panels will be lower than if just testing for the most common genetic causes. Consequently it may be appropriate to set higher thresholds for follow up tests required to provide further evidence of pathogenicity. In light of this it is acceptable to report variants of uncertain significance in a separate "technical report" as appropriate without confirmation by a second method as long as this is clearly stated in the clinical report.

6.5 Submission of variants to databases

It is recommended that as a minimum, and assuming appropriate patient consent is in place, all mutations included in the clinical report are submitted to DMuDB (https://secure.dmudb.net/ngrl-rep/).

Submitting these mutations to additional appropriate databases is also desirable in order that the variant data is publicly accessible. Complete upload of all variants (including polymorphisms) and associated phenotype information from every patient is the ultimate goal and software enhancements to facilitate automated data export from laboratory databases are under development.

Acknowledgements
This document updates guidelines originally produced from the next generation sequencing good practice meeting held on July 19th 2012 in Leeds attended by representatives of the UK Clinical Molecular Genetics Society. Revisions to the original document were made following the "Practice Guidelines for Targeted Next Generation Sequencing Analysis and Interpretation" workshop held in London, 13th November 2013.

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. 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*, **13**, 381-6.
3. Mattocks CJ, Morris MA, Matthijs G, Swinnen E, Corveleyn A, Dequeker E, Müller CR, Pratt V and Wallace A (2010) A standardised framework for the validation and verification of clinical molecular genetic tests. *Eur J Hum Genet* **18**, 1276-88.
4. The Royal College of Pathologists and the Institute of Biomedical Science (2005). The retention and storage of pathological records and archives (3rd edition).