Clin Chem Lab Med 2009;47(11):1343–1350 © 2009 by Walter de Gruyter • Berlin • New York. DOI 10.1515/CCLM.2009.319

Opinion Paper

Considerations for the development of a reference method for sequencing of haploid DNA – an opinion paper on behalf of the IFCC Committee on Molecular Diagnostics[1)]

International Federation of Clinical Chemistry and Laboratory Medicine

François Rousseau[1,*], David Gancberg[2], Heinz Schimmel[3], Michael Neumaier[4], Alexandre Bureau[5], Cyril Mamotte[6], Ron van Schaik[7], Deborah Payne[8], Mario Pazzagli[9] and Ian Young[10]

[1] Department of Medical Biology, Faculty of Medicine, Université Laval and CHUQ, Québec, Canada

[2] Directorate-General Research, European Commission, Brussels, Belgium

[3] European Commission, Joint Research Centre, Institute for Reference Materials and Measurements, Geel, Belgium

[4] Institute for Clinical Chemistry, Mannheim, Germany

[5] Department of Social and Preventative Medicine, Faculty of Medicine, Université Laval, Québec, Canada

[6] Biomedical Sciences and Curtin Health Innovation Research Institute, Curtin University of Technology, Perth, Australia

[7] Department of Clinical Chemistry, Erasmus MC, Rotterdam, The Netherlands

[8] University of Texas Southwestern, Dallas, USA

[9] Department of Clinical Chemistry, University of Florence, Florence, Italy

[10] Wellcome Research Laboratories, Belfast, UK

Abstract

Following the completion of sequencing of the human genome, there has been a very rapid increase in the development of new molecular diagnostic tests. However, the numerous genetic tests and genetic testing technologies offered do not always satisfy essential quality criteria required to ensure confidence in the results that are produced. This is of particular importance for genetic tests since many patients may be tested for a particular genetic defect only once in their lifetime. Thus, there is a pressing need for comprehensive guidelines for the validation of molecular diagnostic tests and procedures, including DNA sequencing, the latter being a fundamental aspect of the development and validation of most genetic tests. To that end, the International Federation of Clinical Chemistry and Laboratory Medicine (IFCC) Committee for Molecular Diagnostics has prepared the following paper that describes a possible approach to the development of a reference method for sequencing of haploid DNA. We discuss various aspects which should be considered before, during and after applying the sequencing procedure, in order to achieve results with a known level of confidence, including robustness and assessments of quality.
Clin Chem Lab Med 2009;47:1343–50.

Keywords: DNA sequencing; nucleic acids; position paper; reference method.

Introduction

As a supra-national professional body, the International Federation of Clinical Chemistry and Laboratory Medicine (IFCC) has an important role in the harmonization of test procedures, and the promotion of quality management of testing services in Clinical Chemistry and Laboratory Medicine. This paper, prepared by the IFCC Scientific Division Committee on Molecular Diagnostics, is intended to address the important issue of whether a reference method for the sequencing of haploid DNA can be developed, and discusses a number of factors which will need to be considered if this is to be achieved. Unlike other areas of laboratory medicine, many emerging genetic technologies do not benefit from the existence of suitable reference materials or reference methods which might otherwise be used to validate methods and commutability of materials. While there are documents which detail the procedure required for sequencing per se, including those developed by the Clinical Laboratory and Standards Institute (CLSI) (1) [(previously The National Committee for Clinical Lab-

[1)] This paper was commissioned by the IFCC Scientific Division (SD) but does not carry any official endorsement by the IFCC SD.

*Corresponding author: Dr. François Rousseau,
Département de Biologie Médicale, Faculté de Médecine,
Centre Hospitalier Universitaire de Québec et Université
Laval, 10 rue de l'Espinay, Québec, Québec G1L 3L5,
Canada
Phone: +1 418 525 4470, Fax: +1 418 525 4429,
E-mail: francois.rousseau@mac.com

Received July 17, 2009; accepted July 28, 2009;
previously published online October 12, 2009

oratory Standards (NCCLS)], American College of Medical Genetics (ACMG) (2), The Clinical Molecular Genetics Society (CMGS) (3), the College of American Pathologists (4), and the popular ''molecular cloning: a laboratory manual'' (5), there are no guidelines that consider the entirety of the process, from DNA preparation to issuing a report. Particular aspects requiring attention include methods for estimating the level of confidence of a sequencing procedure, and acceptable thresholds for high order estimation of the true identity of a sequence. Some of the challenges are to raise the awareness of the importance of the quality of the result, and to recommend a potential high order (or reference) method for DNA sequencing.

Objectives

The main objectives of the present paper are to: a) discuss the requirements for a reference DNA sequencing method in order to establish the ''true value'' of a haploid DNA sequence with a known level of confidence, a known applicable method range and description of quality parameters; and b) more generally, to increase awareness among Laboratory Medicine professionals of quality issues in genetic testing, and to encourage the practice of Laboratory Medicine to the highest standards, including the use of reference materials and their traceability to ''high order'' methods.

Scope and limitations

There are a number of published guidelines on the technical aspects of sequencing (1–4) that propose methods for performing DNA sequencing in routine molecular diagnostic laboratories. However, to our knowledge, no reference method has been proposed for the determination of the nucleotide sequence of a DNA fragment, be it haploid (for instance plasmid DNA) or diploid (for instance human genomic DNA). In addition, apart from the sequencing experiment per se, a reference method should cover the complete process from primer design, to reporting of the results. A reference method is not necessarily intended for routine use in a diagnostic setting because its objective is trueness of the sequence, and the highest possible degree of confidence in the results. The objective of a reference method is not the production of a result with a clinically acceptable confidence level, with an acceptable turn-around time and at reasonable cost (which are objectives of routine diagnostic testing).

This paper focuses on the required characteristics of a sequencing method for haploid genomic DNA that would fulfill the criteria for being a ''reference method'' (6), including matrix category, high-purity material (definition of purity), basic method (see below), description of quality parameters, and applicability of the estimated ''uncertainty range''. Other mandatory elements according to ISO 15193 (7) – which specifies requirements for the drafting of a reference measurement procedure – are also discussed. This includes other means of validation, credentialization or certification by professional organizations, measured definition, patent issues, multiple testing sites requirements, and method instructions.

Thus, we will discuss the present proposal in the same order and terms as the above mentioned requirements for a reference method.

A proposed framework of guiding principles

Matrix category

DNA sequencing requires high-quality DNA templates in order to insure high-quality sequencing readouts. The purity of extracted DNA differs from one extraction method to another, and needs to be verified before sequencing. Ideally, the addition of RNAse and a precipitation step in the purification process will enhance the quality of the recovered DNA. Although UV spectral analysis of a DNA solution can generate information about DNA purity, it appears from previous studies (8, 9) that the quality of the DNA extracted cannot be assessed using the 260 nm/280 nm OD ratio only. This ratio is expected to be between 1.8 and 2.0 for pure DNA. Note that the ratio is sensitive to pH and ionic strength, and a pH of 8.0–8.5 is recommended (10). The presence of particulate matter, which may influence absorbance readings, may be assessed by reading the absorbance at 320 nm (11). Also, a reading at 230 nm may also be useful to assess the presence of phenol if the latter is used in the extraction process. Analysis of the extracted DNA for validation of the size and integrity of the DNA by use of agarose gel electrophoresis and ethidium bromide should also be conducted. Finally, validation of the copy number of DNA by use of real-time polymerase chain reaction (PCR) should also be considered. This additional step would also serve as a check on possible contamination of the extracted DNA by inhibitors of the PCR reaction. The presence of PCR inhibition can be detected by analysis of several dilutions of the samples. The efficiency (ε) of PCR amplification can then be calculated from the resulting calibration curve according to the formula: (12)

$$\varepsilon = 10^{(-1/\text{slope})} \qquad [1]$$

All methods used for DNA extraction and further purification need to be validated (13). In our opinion, this approach should not be replaced by the use of sequence quality scores or of a ''formula'' to compute the expected error rate of a given sequence run afterwards.

Basic method

Primer design for amplification of genomic DNA/preparation of sequencing templates (PCR of genomic DNA) In addition to adequate DNA purity, a minimum DNA quantity (in number of copies) is required in order to achieve successful sequencing, with the exact number required dependant on the source of the DNA (e.g., plasmid DNA, genomic DNA, or PCR product). Until direct genomic sequencing becomes available, sequencing methods rely on a preliminary step of amplification, usually by PCR, of the genomic region of interest. Technical failure at this step seriously compromises the reliability of the results (14, 15). The design and optimization of the amplification process is therefore of paramount importance. We propose the use of appropriate strategies to minimize the probability of technical failure in this important part of the process (16–18).

Existing guidelines for DNA sequencing propose, at a minimum, sequencing analysis of both strands of the target DNA, one in the forward and one in reverse direction (1). We agree with this approach, and this mandates the design and synthesis of two different sequencing primers (one for each strand).

Production of the sequencing templates In order to minimize the possibility of introducing de novo mutations in the amplicon that are not present in the genomic template, a proof-reading DNA-polymerase and long-range PCR reagents should be used to produce the sequencing template from genomic DNA. There should be no true heterozygote positions for haploid genomic DNA, and its presence in the sequence suggests either diploid or multiploid DNA or the introduction of de novo mutations during processing.

The quality of the sequencing templates (forward and reverse) should also be analyzed prior to undertaking the sequencing reactions. The same criteria used for the evaluation of the quality of genomic DNA should be used. The sequencing template must be homogeneous and be of the expected molecular weight, as analyzed by gel electrophoresis. Further purification of the sequencing template (1) may be performed if the quality of templates does not meet these criteria.

The sequencing method per se We do not propose a specific sequencing method. Instead, we will highlight the characteristics of a high-quality sequencing method and propose methods for the estimation of the confidence level for a sequence result.

The characteristics of a high-quality sequencing method are embedded in its validation (13), performed using the same platform as the one that will be used for experimental analysis and using the same reagents/kits. The validation includes:

- the determination of the working range for high-quality sequence data in bp (in general from 500 bp to 900 bp for Sanger-based methods)
- the number of runs recommended for each sequence
- signal-to-noise ratio (peaks and calibration/maintenance of the instrument)
- sensitivity to template quality
- specificity
- minimum sample intake
- robustness
- interpretation tools

Different sequencing methodologies and instruments are available at the present time (14, 15). Most rely on PCR dideoxy-terminator and primer extension sequencing, where products are size-separated on gel or capillary-based sequencing instruments. It is important that the method be well characterized and robust, and that the mean error rate per base be known with a high degree of certainty. The overall error rate should be no greater than 1 error for every 100 base within a single run (or 1%). Maintaining excellent signal-to-noise ratios is one way to achieve low error rates. The working range of the method that will be used must be known. However, there is no minimal or maximal range, provided that the error rate per single run remains 1% or less. The number of runs recommended for a given sequence depends on the target error rate of the laboratory.

In the context of determining the ''high order'' DNA sequence of a nucleic acid, the reference laboratory should aim for a very high standard. This calls for performing more replicate sequences than in a routine diagnostic sequencing context (see below). Also, the sequencing method should be as robust as possible to template lack of quality. Finally, at a minimum, the sequencing instrument should be able to provide a validated quality score, such as PHRED scores (19, 20; see below). Scores should be provided for each sequence position, for each sequencing run, and in an output format that can be exported for statistical analysis and computation of confidence levels. In addition, useful tools have been proposed where a database is constituted of accumulated sequence data from different samples. In this database, each position's peak height is used to compute an expected sequencing electrophoregram (or its equivalent) that is used to interpret sequencing data from new samples. Similarly, a database can also store all the variants observed for a given genomic region.

Quality parameters of the DNA sequence

Sequencing platforms are now equipped with dedicated software that allow analysis and evaluation of the quality of the results. There have been reports on the various platforms in use and their relative market shares (14, 15).

One of these platforms that is probably the most popular, is related to the PHRED scores (19, 20). A PHRED score is a quality value (QV) of a single base call based on peak mobility and shape. It provides a score that represents the likelihood of error at each position that is called (whether an A, G, C or T). PHRED scores are also based on trace features, such as peak spacing, uncalled/called peak ratio and peak resolution.

The PHRED score is a function of the probability of error (PE) at a specific position for which it is being computed, and is represented as $-10\times\log_{10}$ (PE) (Table 1). In many sequencing software packages, each base in the electrophoregram is highlighted with a particular color corresponding to its reliability.

A PHRED score of 30 for a base means that the likelihood of error at that base is 0.1%. The PHRED score does not depend on the length of the sequence. In

Table 1 Quality values calculated as QV=−10 $\log_{10}$ (PE), where PE represents the probability of error.

QV	PE, %	QV	PE, %	QV	PE, %
1	79.0	21	0.790	41	0.0079
2	63.0	22	0.630	42	0.0063
3	50.0	23	0.500	43	0.005
4	39.0	24	0.390	44	0.0039
5	31.0	25	0.310	45	0.0031
6	25.0	26	0.250	46	0.0025
7	20.0	27	0.200	47	0.002
8	15.0	28	0.150	48	0.0015
9	12.0	29	0.120	49	0.0012
10	10.0	30	0.100	50	0.0010
11	7.9	31	0.079	60	0.0001
12	6.3	32	0.063	70	0.00001
13	5.0	33	0.050	80	0.000001
14	4.0	34	0.040	90	0.0000001
15	3.2	35	0.032	99	0.000000012
16	2.5	36	0.025		
17	2.0	37	0.020		
18	1.6	38	0.016		
19	1.3	39	0.013		
20	1.0	40	0.010		

general, a PHRED score of 20 usually means that the sequence (the base call) is reliable for routine sequencing of small fragments. As proposed by CLSI (1), we recommend that a quality score be 40 or higher for every position in the sequence in the context of high order determination of the sequence of a DNA fragment (see level of confidence below).

In addition to a good score for each base pair, the DNA sequencing should be performed in duplicate, at a minimum, and on both strands (forward and reverse) to achieve a lower target error rate. Outside of the context of a reference method (i.e., high order determination of a DNA sequence), the need for a very high level of confidence in the sequence may be relaxed and a higher target error rate may be chosen.

When used for diagnostic purposes, the sequencing method must be fully validated and the platform regularly maintained and performance checked (peak intensity, baseline fluctuations, signal-to-noise ratio); similar to what is required by several International Organization for Standardization (ISO) systems, such as ISO 17025 (21). In addition, a control should be included in each run.

Limits of applicability of the method

The method we propose here for DNA sequencing of ''high order quality'' also has its limits. These include the presence of copy number variants (CNVs) or repeated sequences encompassing any part of the sequencing template. These will sometimes be detected through the presence of heterozygous positions within the expected homozygous haploid sequence data. Such observations should prompt further validation of the copy number in the genomic region of interest using complementary and quantitative methods that can detect the presence of multiple copies of a given sequence. In the context of CNVs, we do not believe that the sequencing method presented here will perform as well as expected. Another instance where the proposed method may not perform well is the presence of pseudogenes that overlap with primer sequences for the sequencing template. These will generate an amplicon of the expected length, but from another genomic region. Again, the presence of heterozygous positions should trigger complementary experiments and prevent reporting of the sequence produced. In presence of known pseudogenes, primer design becomes even more critical. It is important to make sure that primers target sequences that are absent from the pseudogenes (22).

Thus, it is important to have extensive knowledge of the genomic sequence being analyzed in order to correctly set up the initial amplification step and produce the sequencing template. Also, one must be able to correctly interpret the data and report only results that show a determined and low PE (see below).

We do not recommend the use of the present method for diagnostic purposes or for high order analysis of diploid genomic sequences. This is because further precautions need to be taken in preparing the sequencing templates (especially to prevent allelic drop-out), and because computation of the level of confidence is complicated by the presence of two superimposed sequence reads. We plan to propose a reference method for diploid genomic sequencing in a future manuscript.

Estimating the level of confidence of the reference method

Figure 1 shows an example of the frequency distribution of individual position's log (PE) of 10 typical sequencing runs, based on the PHRED scores. This distribution is clearly not Gaussian. Therefore, the distribution of PEs (individual position probability of error) would be even more skewed. Thus, it does not seem appropriate to use a parametric method based on the Gaussian distribution to estimate the level of confidence of a sequencing run.

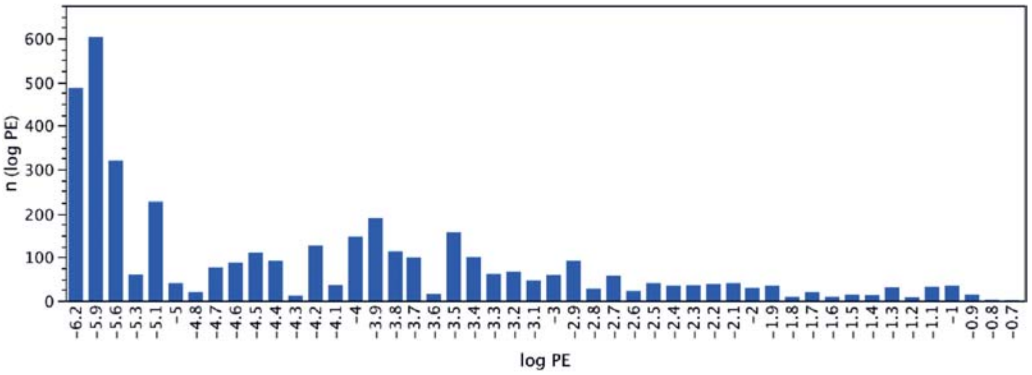

Figure 1 Frequency distribution of individual position's log (PE) of 10 typical sequencing runs.

However, as discussed above, whereas the error rate (PE) is independent of sequence length and is directed solely by the likelihood of error at each position (for instance PE = 1%, PHRED = 20), in the context of sequencing a given DNA fragment with high order quality, it is also relevant to compute the level of confidence that the whole sequence reported corresponds to the true sequence in the sample analyzed.

To compute the expected total likelihood of error of a sequencing experiment, i.e., the probability that a sequence result is erroneous, even in the presence of concordant replicate (or opposite strand) sequencing runs, we propose using a method which draws from the individual probability of error (PE) of each base for each run.

The total expected number of errors of a sequencing experiment is as follows:

Single run:

$$TE_s = PE_1 + PE_2 + PE_3 + \ldots + PE_i \qquad [2]$$

where TE_s is the total number of expected positions with miscalls per sequencing run (single run). The $PE_{1 \ldots i}$ are the PE of each base position in the sequence from position 1 to i.

Thus, in a 500-bp sequence performed in singlicate with a mean PE of 1%, it is expected that there will be 5 sequencing errors. This level of confidence is obviously unacceptable in a clinical setting. Thus, clinical sequencing laboratories have adopted various sets of methods to improve the level of confidence of sequencing results. These methods include performing multiple runs and sequencing the upper and lower strands, expecting concordant replicates and sequences. For instance, CLSI recommends one duplicate sequence or one sequence of the opposite strand (1) for a total of two sequencing runs per DNA segment. New sequencing methods are reported to be more reliable than previous ones (23–25). The expected (a priori) error rate (see below) calculations shown in Table 2 cover sequencing error rates down to 0.01%.

When multiple runs are performed and are concordant, there still exists a calculable level of confidence related to PE for an individual base, the sequence length and to the total number of duplicate or opposite strand (all concordant) runs of the same sequence. Also, given that both errors at any given position need to match with respect to each other (concordant erroneous sequence duplicates), only one of the three possible erroneous nucleotide bases of the second sequencing run will be concordant with the erroneous base introduced at a given position in the first run. This results in a factor of 1/3 over each error probability at each position.

Duplicate runs (1 and 2):

$$TE_{dup} = 1/3 \times [(PE_{11} \times PE_{12}) + (PE_{21} \times PE_{22}) + \ldots + (PE_{i1} \times PE_{i2})] \qquad [3]$$

where TE_{dup} is the total number of expected positions with miscalls per concordant duplicate sequencing runs. The $PE_{11 \ldots i1}$ are the PE of each base position in the sequence run "1" from position 1 to i. The $PE_{12 \ldots i2}$ are the PE of each base position in the sequence run "2" from position 1 to i.

For more than 2 sequence replicates, this factor would be $(1/3)^{(d-1)}$, where d is the number of repli-

Table 2 Concordant-runs probability of error (cPE) at each position in a haploid DNA sequence relative to replicates.

PE (individual base)	Number of concordant runs					
Mean PE	1 run	2 runs	3 runs	4 runs	5 runs	6 runs
0.02	0.02	0.00014	0.00000094	6.4E-9	4.4E-11	3E-13
0.01	0.01	0.000034	0.00000012	3.9E-10	1.3E-12	4.4E-15
0.005	0.005	0.0000084	0.000000014	2.4E-11	4E-14	6.7E-17
0.001	0.001	0.00000033	1.1E-10	3.7E-14	1.4E-17	nc
0.0005	0.0005	8.3E-8	1.4E-11	2.3E-15	nc	nc
0.0001	0.0001	3.3E-9	1.1E-13	nc	nc	nc

nc, not computed.

cates. Thus, for multiple replicate runs (of number d), the estimated a priori upper limit of the error rate for a concordant series of d runs is:

$$\begin{aligned}TE_d=(1/3)^{(d-1)}\times[&(PE_{11}\times PE_{12}\times \dots \times PE_{1d})\\&+(PE_{21}\times PE_{22}\times \dots \times PE_{2d})+\dots\\&+(PE_{i1}\times PE_{i2}\times \dots \times PE_{id})]\end{aligned}\qquad [4]$$

Multiplication of PEs at a single position for duplicate runs relies on the assumption of independence. Thus, in the context of high order sequencing we recommend performing each duplicate sequencing run starting with the raw sample (i.e., reextracting DNA from the sample for each duplicate run) to achieve independence between runs. It has been proposed that different sequencing primers for each sequencing read be used in order to insure even more independence between multiple reads on the same strand (26).

Table 2 shows the predicted concordant-runs PE (cPE) at each position in a sequence of a haploid genome relative to the mean PE of a sequencing method, and the number of concordant replicates of this sequence (either of the same strand or the opposite strand).

Table 3 shows the expected number of positions with miscalls, for a haploid DNA fragment of 500 bp. This number is the upper bound for the probability of erroneous concordant sequences (Pe) by application of Markov inequality (27). When the expected number of positions with miscalls is >1 (such as in the 1-run column), this means that there will be at least one sequencing error in each sequencing experiment, with near certainty. (nc=not computed; $<$1E-17).

With a mean PE of 1% and two runs, the expected number of errors is 34:1,000,000 base positions, or one miscall (concordant on both replicate sequence runs) every 58 runs of a 500 bp sequence. In our opinion, this may not be acceptable in a clinical diagnostic setting, unless one is simply confirming a mutation or sequence at one or a few positions. In the context of determining the ''true value'' of a sequence fragment of 500 bp, we believe that the confidence level must be much higher. Three concordant runs with a mean PE of 1% would provide a Pe of 1:10,000,000 bp and four concordant runs a Pe of 1:2,500,000,000 bp. For sequencing a 500 bp DNA region, this translates into one erroneous concordant sequence every 20,000 concordant triplicates of the target sequence, and one every 5,000,000 concordant quadruplicates of the same length (e.g., two runs for both the upper and lower strands).

Of course, for sequences to be determined for a length different than the 500 bp used here as an example, calculations must be made accordingly.

Reporting of results

This position paper does not aim for proposing guidelines for reporting sequence data in routine diagnostic laboratories. These types of guidelines already exist (1–4). Rather, in the context of ''higher order'' sequence determination it is important to report the details of the design of the sequencing method, including the quality measures of the genomic DNA; the position and primers used to generate the sequencing template, the quality measures of the sequencing template, the position and primers used for sequencing, the number (and strand) of replicate sequences for each segment of the sequencing template, the minimum PHRED score of each sequence run, the final sequence and the level of confidence per base and for the total sequence.

Traceability of a sequence determination derived from the proposed method to a higher-order one

The present paper being as an effort to propose, for the first time, a framework comparable to a reference method for the sequencing of haploid genomic DNA. There is no higher order method (or any high order method) to which a sequence determination derived from the present method can be traceable.

Discussion

We aim to fill a gap in the field of molecular diagnosis in the context of the production of reference materials for DNA testing and for certification of these materials according to ISO standards. Currently, there is neither a proposed nor accepted reference method for determination of the nucleotide sequence of a genomic DNA fragment. Although there are recommended guidelines for nucleic acid sequencing methods in diagnostic laboratory medicine as proposed by CLSI (1), these do not describe a validated method that would provide ''high order'' quality DNA sequences, such as those expected from a reference method, with an accurate estimation of the level of confidence in the results.

Table 3 Upper bound on the probability of erroneous concordant replicates (Pe) of a 500-bp haploid DNA fragment. Probability of erroneous concordant sequences for a 500-bp haploid sequence.

Mean PE	1 run	2 runs	3 runs	4 runs	5 runs	6 runs
0.02	10	0.069	0.000003	2.2E-8	1.5E-10	1E-12
0.01	5	0.017	1.9E-7	6.5E-10	2.2E-12	7.3E-15
0.005	2.5	0.0042	1.2E08	2E-11	3.3E-14	nc
0.001	0.50	0.00017	1.9E-11	6.9E-15	nc	nc
0.0005	0.25	0.000042	1.2E-12	nc	nc	nc
0.0001	0.050	0.0000017	nc	nc	nc	nc

nc, not computed.

The present method is a proposition, published with the objective of highlighting the lack of a consensus high order (or reference) method for DNA sequencing of haploid DNA. It also pursues the goal of generating discussion on various proposed approaches to determine with the highest accuracy the nucleic acid sequence of haploid genomic DNA.

In this first attempt to fill in the gap concerning methodology and high order, we needed to make some basic assumptions to cover the most frequent case of the locus to be sequenced having no CNV or pseudogene. However, we believe that our proposal provides an initial framework for many instances where high order determination of the true nucleic acid sequence needs to be performed. This includes sequencing haploid (e.g., plasmid-based) reference materials, quality control materials, internal controls for genotyping or sequencing commercial kits, etc.

We have attempted to cover the major requirements of a reference method according to the Joint Committee for Traceability in Laboratory Medicine (JCTLM) (7; http://www.bipm.org/en/committees/jc/jctlm/jctlm-wg1/). However, DNA sequencing is a qualitative method and, thus, we have defined the major requirements in this respect. We have proposed a matrix category and a definition of purity of the sample material. We have described the basic method in a general framework. This framework includes determination of the size of the sequencing template according to documented nucleotide diversity heterogeneity in the human genome (to confirm that there is no allelic exclusion), production of the sequencing template, assessment of the quality of the sequencing template, the sequencing experiments per se with specific replicate runs (on each strand), and quality parameters of the DNA sequence. We have stated the limits of applicability of the proposed method, namely with respect to CNVs and pseudogenes. We have proposed a means for estimating, a priori, the level of confidence of sequence results depending on the error rate of the sequencing method, the number of replicate sequences and the length of the sequenced fragment. We further propose a method for estimating the level of confidence of a sequence using the quality scores of the sequencing experiments.

There are other mandatory elements according to ISO 15193 for a reference method (7). In the present case, there could not be another means for validation of the proposed method. This is because there is no high order method of DNA sequencing at the present time. With respect to the measurand being clearly defined, the constituents of a DNA sequence are well known and are of qualitative nature (A, C, G, T). There are no patent issues relative to this proposed method as it is generic and relies on the performance of specific DNA sequencing reagents and apparatus that will be used. Indeed, the expected roll-out of a new generation of DNA sequencers with much lower rates of errors may pave the way to a simplified reference method, at least with respect to the recommended number of replicates. The tables we have provided allow the computation of error rates, as well as the total error rates of these new methods. We believe that if a reference laboratory used the method that is recommended here for determining the nucleic acid sequence of a DNA fragment, there is no need for multiple testing sites to obtain the true value of the sequence.

Finally, different levels of confidence are to be expected for routine sequencing, as compared to ‘‘high order’’ determination of a DNA sequence. High order determination should occur only in a reference laboratory that is certified/accredited (ISO certification/accreditation for reference laboratories). We do not propose using a potential reference method in routine diagnostic activities as it would be labor intensive as well as expensive. In the context of routine sequencing activities, however, the present paper proposes a method to compute the likelihood of observing a concordant duplicate sequence that is erroneous.

References

1. NCCLS. Nucleic Acid Sequencing Methods in Diagnostic Laboratory Medicine; Approved Guideline. NCCLS document MM9-A [ISBN 1-56238-558-5]. NCCLS, 940 West Valley Road, Suite 1400, Wayne, Pennsylvania 19087-1898 USA, 2004.
2. American College of Medical Genetics; Standards and Guidelines for Clinical Genetics Laboratories; revised 02/2007 (http://www.acmg.net/AM/Template.cfm?Section=Laboratory_Standards_and_Guidelines&Template=/CM/HTMLDisplay.cfm&ContentID=3735).
3. CMGS guidelines, 2003 retired 2007.
4. College of American Pathologists. Checklist for Molecular Pathology. Chicago, IL.
5. Sambrook J, Fritsch EF, Maniatis T. Molecular cloning: a laboratory manual, vol. 1,2,3. New York: Cold Spring Harbor Laboratory Press, 1989.
6. ISO/IEC Guide 99:2007. International Vocabulary of Metrology – Basic and general concepts and associated terms (VIM). Geneva: International Organization for Standardization; 2007.
7. ISO/IEC 15193: 2002. In vitro diagnostic medical devices – measurement of quantities in samples of biological origin – presentation of reference measurement procedures. Int Org Standardization 2002.
8. Broothaerts W, Corbisier P, Emons H, Emteborg H, Linsinger TP, Trapmann S. Development of a certified reference material for genetically modified potato with altered starch composition. J Agric Food Chem 2007;55:4728–34.
9. Orlando C, Verderio P, Maatman R, Danneberg J, Ramsden S, Neumaier M, et al. EQUAL-qual: a European program for external quality assessment of genomic DNA extraction and PCR amplification. Clin Chem 2007;53:1349–57.
10. Wilfinger WW, Mackey K, Chomczynski P. Effect of pH and ionic strength on the spectrophotometric assessment of nucleic acid purity. Biotechniques 1997;22:474–6, 478–81.
11. Sahota A, Brooks AI, Tischfield JA. Preparing DNA from saliva for genotyping. Cold Spring Harb. Protoc., 2007. doi:10.1101/pdb.prot4831.
12. Kontanis EJ, Reed FA. Evaluation of real-time PCR amplification efficiencies to detect PCR inhibitors. J Forensic Sci 2006;51:795–804.

13. Eurachem Guide 1998: The fitness for purpose of analytical methods: a laboratory guide to method validation and related analysis. http://www.eurachem.org/guides/valid.pdf.
14. Ahmad-Nejad P, Dorn-Beineke A, Pfeiffer U, Brade J, Geilenkeuser WJ, Ramsden S, et al. Methodologic European external quality assurance for DNA sequencing: the EQUALseq program. Clin Chem 2006;52:716–27.
15. Patton SJ, Wallace AJ, Elles R. Benchmark for evaluating the quality of DNA sequencing: proposal from an international external quality assessment scheme. Clin Chem 2006;52:728–36.
16. Rådström P, Knutsson R, Wolffs P, Lövenklev M, Löfström C. Pre-PCR processing: strategies to generate PCR-compatible samples. Mol Biotechnol 2004;26:133–46.
17. Robertson JM, Walsh-Weller J. An introduction to PCR primer design and optimization of amplification reactions. Methods Mol Biol 1998;98:121–54.
18. Grunenwald H. Optimization of polymerase chain reactions. Methods Mol Biol 2003;226:89–100.
19. Ewing B, Hillier L, Wendl M, Green P. Base-calling of automated sequencer traces using Phred. I. Accuracy assessment. Genome Res 1998;8:175–85.
20. Ewing B, Green P. Base-calling of automated sequencer traces using phred. II. Error probabilities. Genome Res 1998;8:186–94.
21. ISO/IEC 17025:2005 General requirements for the competence of testing and calibration laboratories. Int Org Standardization, 2005.
22. Neumaier M, Gerhard M, Wagener C. Diagnosis of micrometastases by the amplification of tissue-specific genes. Gene 1995;159:43–7.
23. Li H, Ruan J, Durbin R. Mapping short DNA sequencing reads and calling variants using mapping quality scores. Genome Res 2008;18:1851–8.
24. Ten Bosch JR, Grody WW. Keeping up with the next generation: massively parallel sequencing in clinical diagnostics. J Mol Diag 2008;6:484–92.
25. Schuster SC. Next-generation sequencing transforms today's biology. Nature Methods 2008;5:16–8.
26. Stephens M, Sloan JS, Robertson PD, Scheet P, Nickerson DA. Automating sequence-based detection and genotyping of SNPs from diploid samples. Nat Genet 2006;38:375–81.
27. Ross SM. A first course in probability, 8th ed. Upper Saddle River. NJ: Pearson Prentice Hall, 2008.