Perspective

Minor variant detection in amplicons using 454 massive parallel pyrosequencing: experiences and considerations for successful applications

Ina Vandenbroucke, Herwig Van Marck, Peter Verhasselt, Kim Thys, Wendy Mostmans, Stéphanie Dumont, Veerle Van Eygen, Katrien Coen, Marianne Tuefferd, and Jeroen Aerssens
Tibotec-Virco Virology, Janssen Pharmaceutical Companies of Johnson & Johnson, Beerse, Belgium

BioTechniques 53:167-177 (September 2011) doi 10.2144/000113733
Keywords: ultra-deep sequencing; 454 pyrosequencing; minor variant detection; DNA polymerase; sequencing error

Supplementary material for this article is available at www.BioTechniques.com/article/113733.

Ultra-deep sequencing (UDS) of amplicons is a major application for next-generation sequencing technologies, even more so for the 454 Genome Sequencer FLX. Especially for this application, errors that might be introduced during any of the sample processing or data analysis steps should be avoided or at least recognized, as they might lead to aberrant sequence variant calling. Since 454 pyrosequencing relies on PCR-driven target amplification, it is key to differentiate errors introduced during the amplification step from genuine minority variants. Thereto, optimal primer design is imperative because primer selection, primer dimer formation, and nonspecific binding may all affect the quality and outcome of amplicon-based deep sequencing. Also, other intrinsic PCR characteristics including amplification drift and the formation of secondary structures may influence sequencing data quality. We illustrate these phenomena using real-life case studies and propose experimental and analytical evidence-based solutions for effective practice. Furthermore, because accuracy of the DNA polymerase is vital for reliable UDS results, a comparative analysis of error profiles from seven different DNA polymerases was performed and experimentally assessed in parallel by 454 sequencing. Finally, intra- and interrun variability evaluation of the 454 sequencing protocol revealed highly reproducible results in amplicon-based UDS.

An important factor in any minor variant detection approach is the sensitivity for detecting DNA sequence variants or mutations in an excess of nonmutated genomes. This challenge is encountered across many applications, including the detection of neoplastic cells in a majority of normal cells, the detection of somatic mutations in tumor biopsies, and the detection of a minor viral variant in a background of a large viral population. Sanger sequencing (1) has long been regarded as the gold standard for mutation detection, because prior knowledge of mutations is not required and assay development is limited only by sequencing primer design and read length. This approach, however, is unreliable for detecting variants that constitute <20% of the total population of genomes in a sample (2,3), it generates only an average sequence of the PCR product, it does not allow determining linkage of mutations, and the analysis of samples with heterogeneous insertion-deletion mutations remains challenging. Minor variant detection methods that rely on subcloning of PCR products, in conjunction with conventional sequencing, are expensive, time-consuming, and suffer from the drawback that PCR-based errors are propagated into the cloned DNA and cannot be discriminated from bona fide mutations. On the other hand, nonsequencing-based assays for minor variant detection [e.g., allele-specific PCR or probe-based methods (4)] generally offer high sensitivity, but prior knowledge of the mutation of interest is required, and no information on the sequence context can be generated, nor is linkage of the identified mutations possible.

The power of the new sequencing technologies and their utility for variant detection derive from the ability to sequence single molecules in massive amounts. In this process, each of the single molecules of an amplicon is clonally amplified and is sequenced individually, allowing for the identification of rare variants and haplotype information over the whole read length. At present, the 454 pyrosequencing technology achieves the longest read lengths (400–500 bp reads), using the titanium chemistry on the Genome Sequencer FLX (GS FLX; 454 Life Sciences, Roche Applied Science, Branford, CT, USA). This technology enables the clonal sequencing of hundreds of thousands of molecules, which allows the ultra-deep sequencing (UDS) of amplicons at high coverage (5). The combination of high coverage and long read length has made the GS FLX a promising tool for sensitive and quantitative detection of

variants, with potential applications in various clinically relevant research areas, including virology, oncology and human genetics.

For example, monitoring HIV-1 drug resistance has become increasingly important for guiding treatment, especially for patients failing antiretroviral therapy (6,7). The conventional method uses bulk population genotyping of the viral quasispecies in an infected patient to predict HIV-1 drug resistance profiles. Some studies suggest that low-frequency (<20%) resistance mutations may have an impact on therapy outcome as a result of transmitted drug resistance or remnants of earlier drug selection in patients previously exposed to antiretroviral therapy, while other studies did not observe such correlation (8–15). With UDS becoming more widely available, the relevance of minor mutations in the context of different antiretroviral therapy regimens might help define the clinical benefit of low-frequency resistance testing. In oncology, UDS has been applied to identify rare somatic mutations in complex tumor samples, which might impact diagnostics and therapeutics. For example, Thomas et al. (16) reported the presence of low-abundance oncogene mutations in complex samples with low tumor content for which conventional Sanger sequencing was not informative. Another study screened 623 cancer-related genes in 188 human lung adenocarcinoma, revealing more than 1000 somatic mutations across the samples (17). UDS has also been applied to search for rare mutations in samples from patients suffering from tuberous sclerosis complex, an autosomal dominant neurocutaneous syndrome (18), and in samples from B cell chronic lymphocytic leukemia patients (19).

Over the past years, we have built experience in the design and optimization of UDS assays on amplicons using 454 massive parallel pyrosequencing technology, primarily with applications in virology and oncology. Here, we discuss and illustrate by means of case studies from our laboratory different sources of errors that may occur during UDS. Emphasis is on the experimentally controllable variables affecting fidelity, quality, and outcome of amplicon-based deep sequencing.

PCR primer design

The design and selection of a primer set that specifically targets the region of interest,

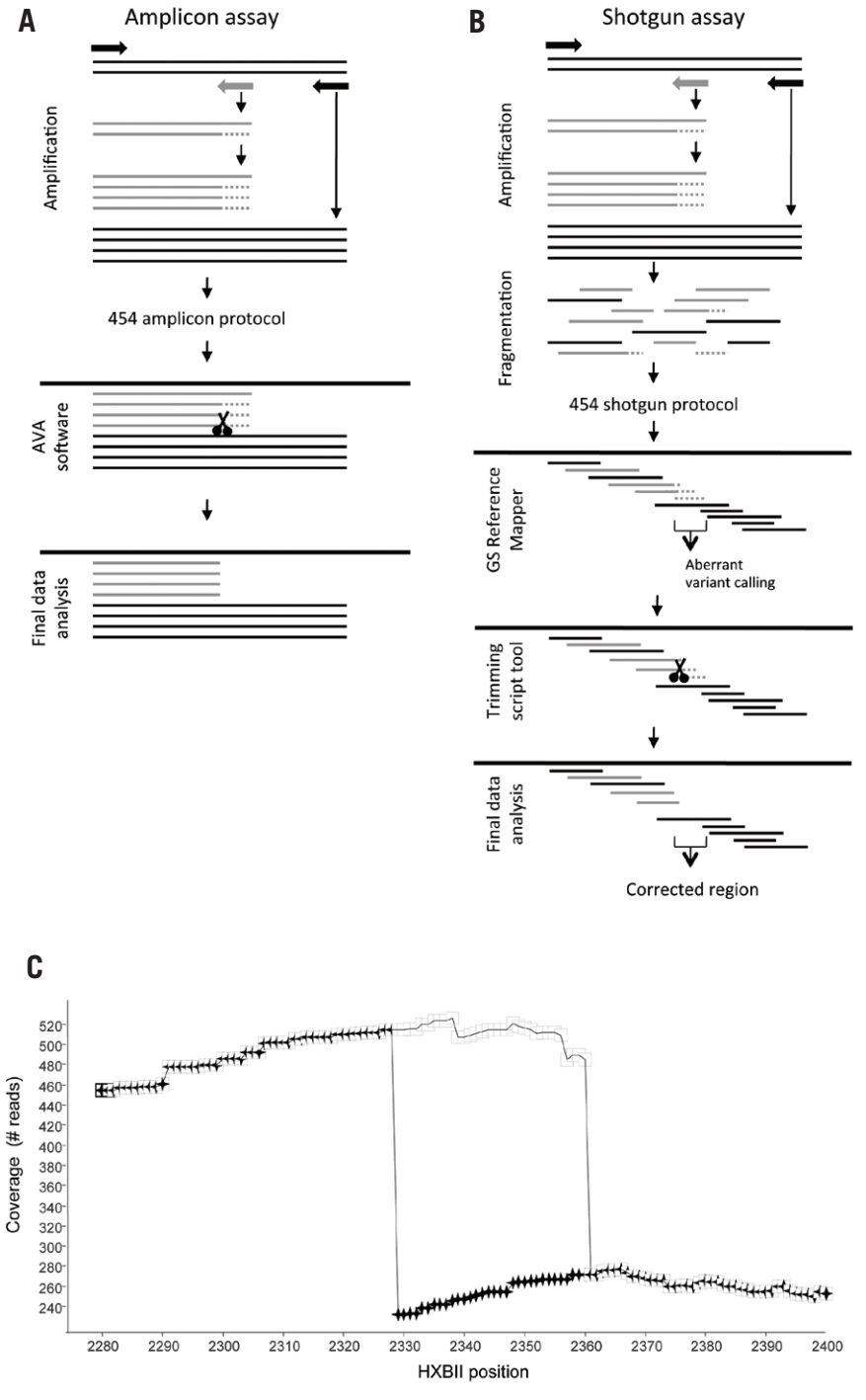

Figure 1. Nonspecific primer binding during PCR amplification may lead to errors in sequence data analysis. Representation of the consequences for sequence data analysis of nonspecific primer binding during PCR amplification for (A) amplicon or (B) shotgun 454 assays. Specific primer binding is shown as a black arrow; gray arrows represent nonspecific primer binding. AVA software (used for amplicon 454 assays) is able to correct for nonspecific primer binding, whereas nonspecific primer binding is not controlled for in shotgun assays and might cause aberrant sequence variant calling. As a solution to this problem, an algorithm has been developed that trims the primer sequences from the reads that are situated in a region of nonspecific primer binding. (C) Detail of a coverage plot (read counts/position) of a shotgun experiment of the HIV-1 PR-RT region, aligned to the HXBII reference viral strain. Using the algorithm outlined in panel B, the primer sequences were removed from the reads ending or starting within the nonspecific priming region. White squares, read count before cleaning; black crosses, read count after cleaning for nonspecific priming.

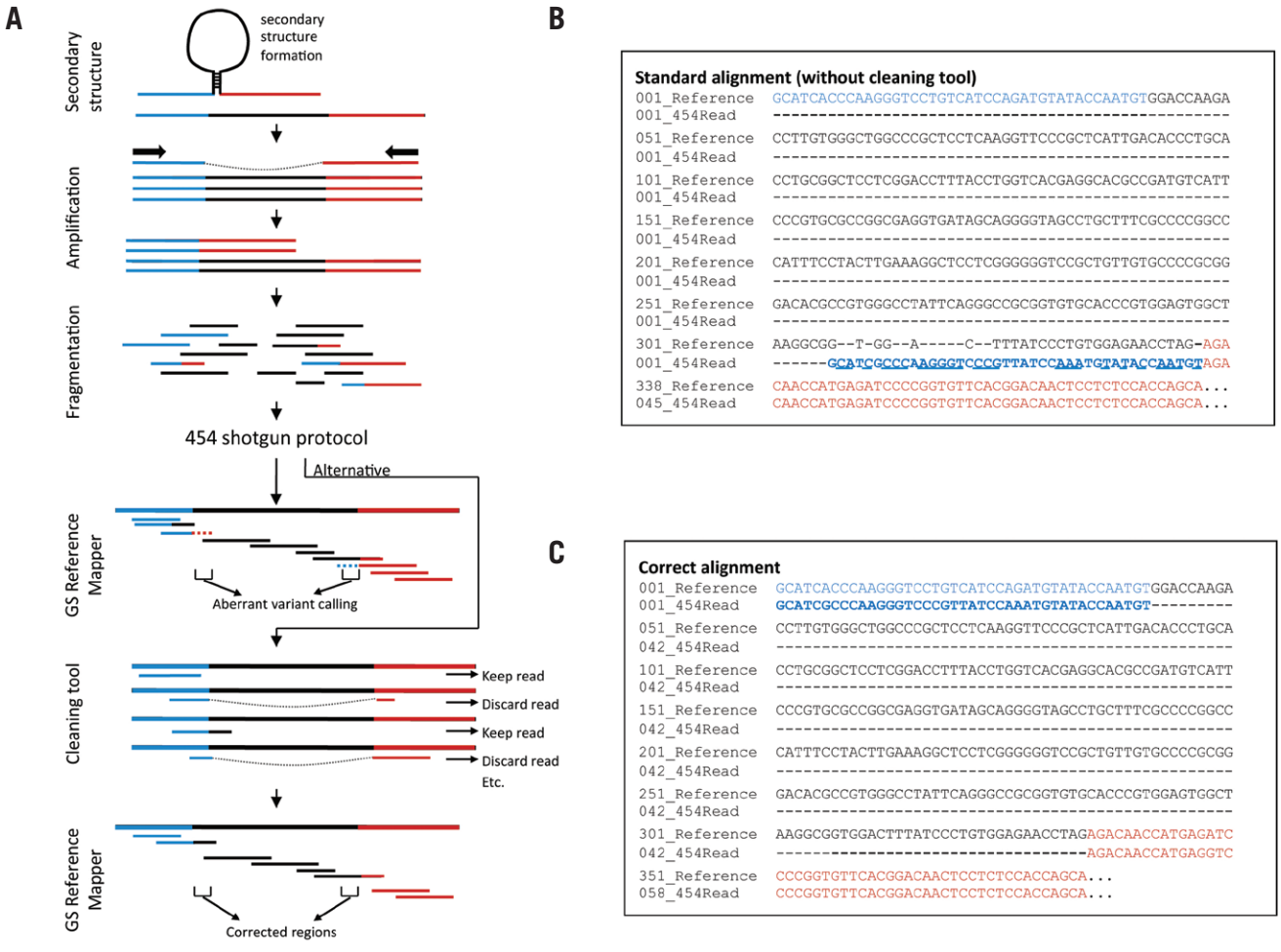

Figure 2. Formation of secondary structures during PCR amplification and its effects on data analysis following UDS using the 454 shotgun protocol. The representation (A) also proposes a solution to the problem through the introduction of a cleaning algorithm before final data analysis in GS Reference Mapper. (B and C) A real-life example of a 454 read comprising a large deletion that was erroneously introduced during amplification, aligned versus a reference sequence. Using the standard Roche analysis software, proper alignment of the sequence read is impaired resulting in aberrant variant calling (panel B, blue underlined sequence). Panel C shows the correct alignment of this sequence read. The cleaning tool algorithm identifies and discards the sequence reads with large gaps (at least 10 nucleotides) to prevent aberrant variant calling.

thereby avoiding the formation of primer dimers and nonspecific primer binding, is a prerequisite toward high-quality amplicon-based deep sequencing.

Primer selection

The source material for PCR amplification is often a heterogeneous or mixed population of DNA molecules (e.g., reflecting viral strain variability in a clinical sample). In order to capture all the variability and obtain a good representation of the diversity in the initial sample, PCR primers must be designed toward conserved regions. This avoids the selection of a specific subpopulation by preferential annealing of the primers to only one or part of the templates present in a sample. Consequently, optimal primer design requires precise and reliable a priori knowledge of the variable regions of the target DNA to be amplified. Large sequence databases are of great help for adequate primer design, but can be biased toward specific geographical regions, patients, specimen subtypes, or genomic regions. Furthermore, no or very few sequences might be available in these databases for some organisms or targeted DNA regions. As an example, we developed an assay for the hypervariable V3 region in the envelope gene of HIV-1, which is the major determinant for coreceptor usage (20,21). Accurate prediction of coreceptor usage is essential to establish patient eligibility to treatment with coreceptor antagonists (22,23). The variability in and around this V3 region necessitated us to design a reverse primer located 330 bp downstream of the target region, which was the closest conserved region. In case such solution is not possible because of a variable target region without proximity of a conserved nucleotide stretch, the mixing of different primers or the use of degenerate nucleotides could be an alternative. For example, for the development of a sensitive HIV-1 resistance test for the region of the reverse transcriptase gene that covers amino acids 59–190, an in-house database of aligned Sanger sequences from 266,781 samples was interrogated to achieve a good primer design. This in silico analysis revealed numerous variable positions in the region that was targeted for primer selection. The best performing primer pair, out of several that were tested in the laboratory, was successful in only 70% of a set of 57 samples. Complementing this primer pair with two partially overlapping primers (one forward and one reverse) in the PCR mixture increased the amplification success of this assay to 83%.

Primer dimers

Primer molecules with (partially) complementary sequences can hybridize and form

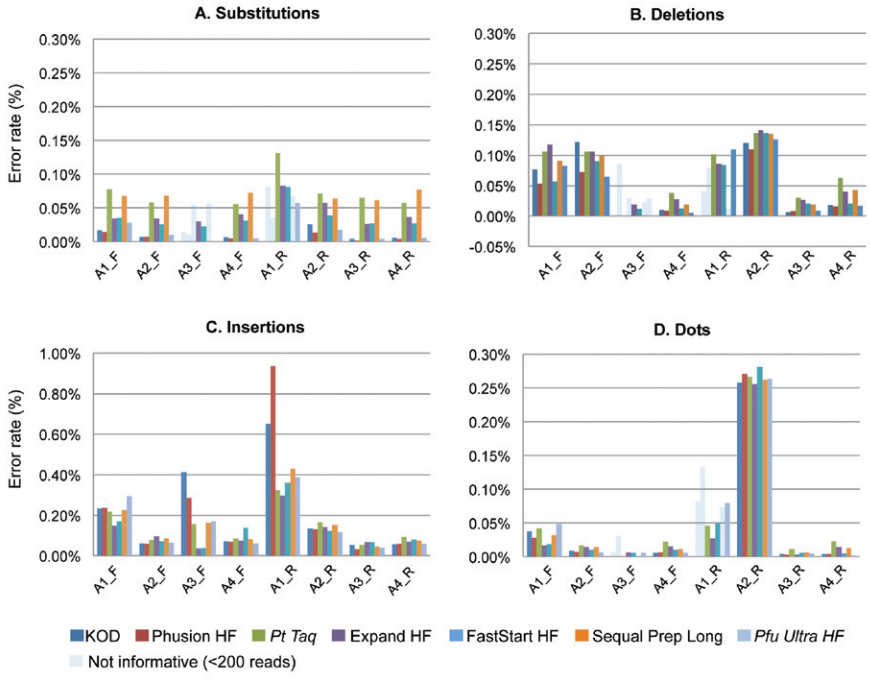

Figure 3. Assessment of the error profiles of seven different DNA polymerases as determined by 454 sequencing. Each of the seven enzymes was used to amplify four different exons from the human *TP53* oncogene, starting from plasmid clones comprising a cloned *TP53* exon. (A) Error rate for nucleotide substitutions: number of substitutions divided by the number of bases. (B) Error rate for deletions: number of deleted bases divided by the number of bases. Deletions of more than one base at a certain position were counted as one deletion. The faded out bars represent samples of amplicons with a sequence coverage much lower (<200 reads) than from the run design (average coverage of 4111 reads) and should not be considered representative. (C) Error rate for insertions: number of inserted bases divided by the number of bases. Insertions of more than one base at a certain position were counted as one insertion. (D) Error rate for dots: number of dots (N) divided by the number of bases (dot = three successive negative flows during 454 sequencing). A1_F, amplicon 1 forward read; A2_F, amplicon 2 forward read; A3_F, amplicon 3 forward read; A4_F, amplicon 4 forward read; A1_R, amplicon 1 reverse read; A2_R, amplicon 2 reverse read; A3_R, amplicon 3 reverse read; A4_R, amplicon 4 reverse read.

dimers that can be further amplified during PCR and result in very short amplicons. Fusion of the 454-specific adaptors (called A and B) and barcodes (multiplex identifiers or MIDs) to the gene-specific primers increases the length of the primer with an extra 29 bp, resulting in primers of 50 bp or longer. These lengthy primers make primer dimer formation more likely to occur. Normally, these short byproducts are removed during the small fragment removal step (Agencourt AMPure XP bead purification; Beckman-Coulter, Brea, CA, USA) of the 454 protocol. This step, however, can be insufficient when large amounts of dimers are present. Moreover, these short products are preferentially amplified during emulsion PCR in the 454 process, leading to an overrepresentation of short sequence reads (containing mainly primer sequences) and reducing the number of useful sequence reads obtained. As an example, we observed a nonneglectable peak of small fragments upon DNA LabChip (Agilent Technologies, Santa Clara, CA, USA) analysis of a sample that had undergone PCR amplification of the HIV-1 V3 region. The sample was subsequently purified using AMPure beads, which, as expected, significantly reduced the primer dimer peak, although a very minor peak remained visible. Further processing of this sample in the 454 workflow and subsequent data analysis ($n = 10{,}687$ reads) revealed that 37% ($n = 3951$) of the reads were derived from primer dimers. Next, this amplicon generation protocol was optimized by introducing a first-round PCR, using naked gene-specific primers (no A/B and MID adaptors), followed by a second-round PCR using the same gene-specific primers fused to the A/B and MID adaptors. As a result, no primer dimers were detected anymore upon gel electrophoresis analysis, and virtually all 454 sequencing reads were over 200 bp in length and mapped uniquely to the target region (Supplementary Figure S1). Over the years, we experienced effective primer dimer reduction for several other assays using this two-round PCR approach (Vandenbroucke and Verhasselt, unpublished data). In addition, in all cases where primer dimers remained visible upon gel electrophoresis quality check after PCR, we applied gel extraction for small fragment removal.

Nonspecific binding

The specificity of DNA amplification during PCR is determined by the ratio of the primers' capability and affinity to recognize and bind the intended target DNA sequence as compared with nontarget DNA sequences. The formation of nonspecific PCR products that are completely off-target usually poses relatively minor problems during data analysis, because their sequences do not align with the target sequence and can easily be filtered. Their formation, however, will result in loss of sequencing capacity from a 454 run, and hence lower the coverage of the desired amplicon.

Nonspecific binding of one of the primers either within, up- or downstream of the region of interest, will lead to shorter or larger amplicons as compared with the intended region. If present at low quantity, these side-products might remain undetectable during gel-based fragment size control. In UDS, however, this can lead to the erroneous calling of single-nucleotide polymorphism (SNPs) or mutations in the region where the primer was bound nonspecifically (Figure 1). Such nonspecific binding can be easily detected in amplicon-based 454 assays as it becomes apparent after alignment of the reads to the reference. The Amplicon Variant Analyzer software, used for the alignment of reads and the variant calling, automatically corrects for this type of errors by trimming the reads (Figure 1A). However this is not the case for larger amplicons that are processed using the 454 shotgun procedure (as was used in the HIV-1 PR-RT assay). This protocol includes a random fragmentation step to generate a library of appropriate length for 454 sequencing (Figure 1B), resulting in a more challenging data analysis. After alignment of the reads to the sequence of the HIV-1 reference strain HXBII, the Reference Mapper software (Roche Applied Science) cannot distinguish between nonspecific primer binding and genuine variability. Consequently, the primer-template mismatches are falsely reported as variations (Figure 1B). If reads are ending in a nonspecifically bounded primer and contain only a few nucleotides of the primer, it is even more challenging to distinguish between genuine sequence variations and nonspecific priming. In order to clean the data from variations introduced by nonspecific primer annealing, we developed a filtering algorithm that searches for reads that start or end with a primer sequence and are located within the amplicon of interest. Next, the primer sequences are trimmed from these reads in order to avoid the erroneous calling of variations within this sequence region, resulting in a more correct alignment (Figure

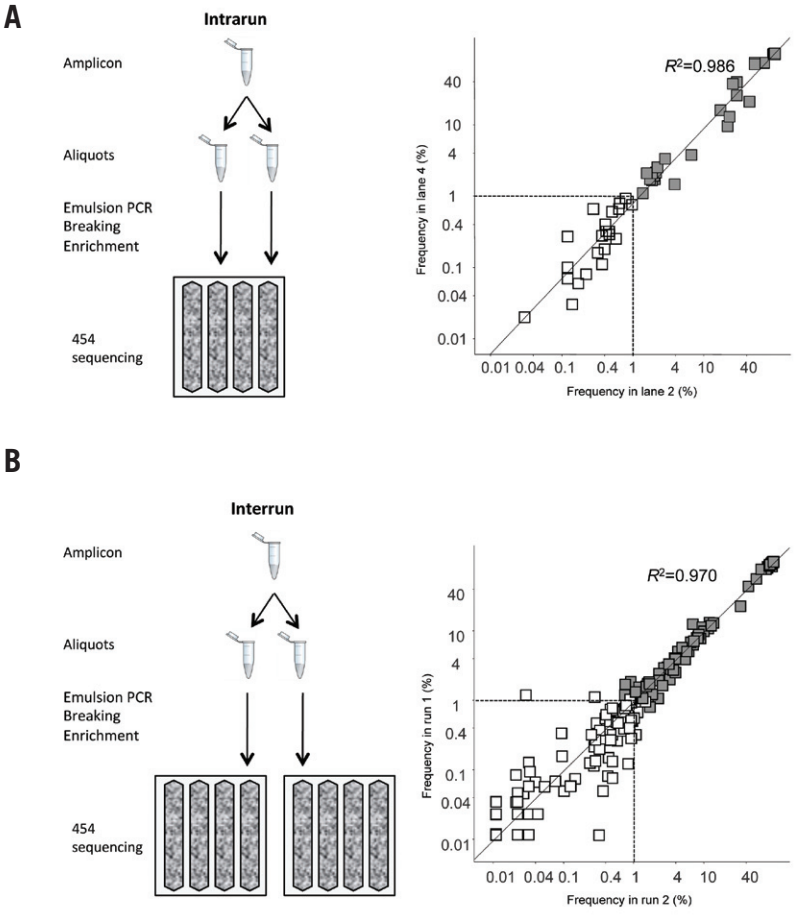

Figure 4. Reproducibility of minor variant detection in UDS amplicon 454 sequencing. Representation of the experimental setup and scatter plots derived from representative experiments assessing (A) intrarun (interlane) variability and (B) interrun variability. An in-house designed assay for the detection of resistance mutations in HIV-1 reverse transcriptase gene was used for this analysis. The 454 run was designed to allow identification of minor variants at 1% frequency (i.e., 5000 reads/sample). The markers on the scatter plot indicate variations from the reference HIV-1 reverse transcriptase gene sequence. White squares, variants frequencies <1%; black squares, variant frequencies >1%.

1B). An illustrative example is given in Figure 1C that shows part of the coverage plot of a shotgun sequencing experiment covering the HIV-1 protease and reverse transcriptase genes. Nonspecific binding of the reverse primer within the region of interest (position 2329 to 2356) generated an additional smaller product, albeit at a low amount. This amplicon remained undetectable after size confirmation by agarose electrophoresis, but resulted in erroneous variant calling in the region of interest. Reads ending or starting within the region of nonspecific binding were subsequently trimmed using our algorithm and further used for the final data analysis (Figure 1C).

Intrinsic PCR characteristics

PCR bias

Quantification of variation in PCR-amplified samples can differ significantly from the genuine composition of a sample. This bias in template-to-product ratio is known as PCR drift and is caused by stochastic variation in the early cycles of the reaction, leading to a deviating value of the original template frequency. This can result in different outcomes in replicate reactions, especially for low frequency variants (24). We assessed this effect in a study of four HIV-infected plasma samples from unrelated clinical cases (25). Here, seven viral RNA aliquots (per patient sample) were each individually reverse-transcribed, amplified, and sequenced, without amplicon pooling. Analysis of the quasispecies variability indicated that the range of variants per sample varied considerably depending on the RT-PCR experiment. To minimize this effect, we implemented in all amplicon-based assays a strategy of pooling seven RT-PCRs performed in parallel and demonstrated that this strategy reduced intra-assay variability (25).

Secondary structures

Watson-Crick base pairing of nucleotides within the same strand can lead to the formation of secondary structures of the template, especially in GC-rich regions. During sequence extension of the DNA strand by the polymerase, a region with a strong secondary structure might be skipped due to the looping-out of this region, which will then be excluded from further amplification in the next PCR cycles. If this occurs at a low frequency, it remains unnoticed after quality check of the amplicon, and sequencing of this amplicon will lead to the erroneously reporting of a deletion in those reads passing the breakpoint (Figure 2). If a large gap was created, proper alignment of the read will be impaired, and incorrect fusion of different regions of the genome will result in the aberrant calling of variants (Figure 2). This can result in erroneous calling of variants (up to 10%; data not shown). Unfortunately, correction for these errors is not possible by means of the current Roche Applied Science analysis software. Trimming of the reads at the breakpoint or splitting the read in two separate reads (before and after the breakpoint) might be envisaged during analysis, however this strategy is impaired as the exact position of the breakpoint is uncertain. As a solution to this problem, a software scripting tool was developed that specifically recognizes the reads containing secondary structure artifacts and removes these from the analysis. This preprocessing tool identifies large gaps by pairwise alignment of each read to the reference, using a small gap extension penalty. Reads containing a sufficiently large gap (at least 10 nucleotides) are subsequently removed from the data set, and the cleaned data can be further processed for full alignment and variant calling (see Supplementary materials for more details). It should be noted, however, that this tool cannot distinguish between genuine deletions and deletions caused by secondary structures.

Influence of DNA polymerases

It is well known that DNA polymerases introduce random errors at low frequency in nucleotide sequences due to base misincorporation during amplification. Such incorrect bases are present in individual DNA strands as a small minority in the PCR product. Although such errors are of little concern in conventional Sanger sequencing analysis, they will be observed in UDS applications using technologies that involve a PCR amplification step (such as 454 sequencing). More specifically, such errors are further amplified on a bead during the 454 sequencing preparation because of the clonal nature of the emulsion

PCR, and these individual molecules will be sequenced. If UDS of a heterogeneous population is envisaged, these errors will be reported as variations, and unfortunately, it is impossible to distinguish between DNA polymerase misincorporations and genuine mutations present in subpopulations of DNA molecules in the sample. The use of a high-fidelity enzyme with proofreading activity can reduce the error rate considerably. The PCR error rate can also vary according to the nucleotide sequence context of the target DNA fragment that is amplified. Today, little is known about the differences between polymerases from various sources and the impact of such errors introduced during PCR on 454 sequencing outcome. We assessed and compared the error frequency and the type of errors introduced by seven different polymerases which were tested in parallel: KOD, Phusion HF, Pt *Taq*, Expand HF, FastStart HF, SequalPrep Long, and *PfuUltra* HF DNA polymerase (see Table 1 for an overview of enzyme characteristics). Thereto, four plasmid clones, each containing one exon of the human P53 oncogen (*TP53*) were used as starting template for PCR amplification of each of the four exons (see Supplementary Tables S1 and S3 for experimental details). Hence, all variants detected after sequencing analysis can be considered errors introduced during amplicon synthesis and/or GS FLX sequencing. The inclusion of different exons in our experimental setup allowed us to assess the influence of the sequence context on the error profile. Overall, the error rate varied between 0.11% and 0.34%, consisting of insertions (0.07%–0.14%), deletions (0.02%–0.08%), substitutions (0.01%–0.07%), and dots (0.01%–0.05%; dot: three successive negative flows during 454 sequencing) (Table 2). Our results demonstrate that the DNA polymerase, as well as the sequence context, influence the type of errors observed (Figure 3). More specifically, the substitution error rate was primarily associated with the polymerase rather than the sequencing context (Figure 3A). All polymerase mixtures containing *Taq* DNA polymerase (Pt *Taq*, Expand HF, and FastStart HF) showed a higher substitution rate than the other polymerases tested. Moreover, these enzymes generated far more A/T to G/C transitions—and to a lesser extent the reverse transitions G/C to A/T—than other substitutions. On the other hand, insertions and deletions were primarily associated with the sequence context (interamplicon variability), rather than with the polymerase (interpolymerase variability within the same amplicon) (Figure 3, B and C). The rate and position of insertions and deletions were very similar for the different polymerases and correlated with the presence of homopolymers (data not shown). It is well known that homopolymeric DNA stretches are the major source of erroneous base calls when using the 454 technology due to resolution difficulties and the appearance of carry forward incomplete extension (CAFIE) errors. Hence, insertions and deletions presumably are more linked to the 454 sequencing technology than to the amplicon synthesis. Together, our results demonstrate that careful selection of an appropriate polymerase with low error rate is highly recommended for UDS applications to detect nucleotide substitutions at frequencies below 1%. Of note, UDS of viral samples often also implies an additional initial step of reverse transcription of viral RNA into cDNA (e.g., plasma HIV). Although high-fidelity reverse transcriptase enzymes are used in this step, it cannot be excluded that errors can be introduced, albeit that it is difficult to account for. This calls for a somewhat higher cut-off value than the error threshold calculated based on plasmid sequences for distinguishing authentic variants from errors in viral RNA plasma samples.

Reproducibility of the 454 amplicon sequencing process

Although it is generally assumed that the major source of variability resides in the amplicon preparation steps as described above, erroneous variations in the final DNA sequence results might also originate from the 454 sequencing process itself, which de facto is a fixed protocol. We evaluated the variability of the 454 UDS amplicon protocol, thereby excluding errors introduced during the upfront PCR, by means of an experiment designed to detect resistance mutations in the HIV-1 reverse transcriptase gene. All factors contributing to the fidelity, quality, and outcome of amplicon-based UDS, as described in this article, were carefully implemented during amplicon generation. Amplicons were split in two aliquots and processed in parallel either on different lanes on the same run (intrarun) or on different runs (interrun) (Figure 4). Intra- and interrun reproducibility was assessed for three and two independent samples, respectively. The scatter plots in Figure 4 show the data for one sample, which is representative for all examined samples. Comparison of the observed sequence mutation frequencies of the parallel experiments demonstrated highly reproducible results, with both excellent intrarun variability (Pearson correlation >0.98; concordance >0.97 for each of the three samples tested) as well as interrun variability (Pearson correlation >0.98; concordance >0.98 for each of the two

Table 1. Characteristics of DNA polymerases used for PCR amplification in ultra-deep 454 sequencing applications

DNA polymerase	Full name	Supplier	Composition[a]	Fidelity[a] (1/error rate) according to the manufacturer
KOD	KOD hot start DNA polymerase	Novagen	- KOD (*Thermococcus kodakaraensis*) DNA polymerase - Two monoclonal antibodies	50 times higher than *Taq* DNA polymerase (using an *rpsL*+ assay)
Phusion HF	Phusion High-Fidelity DNA Polymerase	Finnzymes	- *Pyrococcus*-like polymerase - a dsDNA binding domain	50 times higher than *Taq* DNA polymerase (using an *LacI*-based assay)
Pt *Taq*	Platinum *Taq* DNA Polymerase High Fidelity	Invitrogen	- *Taq* (*Thermus aquaticus*) DNA polymerase (recombinant) - *Pyrococcus* species GB-D polymerase - Platinum *Taq* Antibody	3–4 times higher than *Taq* DNA polymerase (using an *rpsL*+ assay)
Expand HF	Expand High Fidelity PCR System	Roche Applied Science	- *Taq* (*Thermus aquaticus*) DNA polymerase - *Tgo* (*Thermococcus gorgonarius*) DNA polymerase	3 times higher than *Taq* DNA polymerase (using an *LacI*-based assay)
FastStart HF	FastStart High Fidelity PCR System	Roche Applied Science	- FastStart *Taq* (chemically modified *Taq*) DNA Polymerase - a proofreading protein	4 times higher than *Taq* DNA polymerase (using an *LacI*-based assay)
Sequal Prep Long	SequalPrep Long PCR Kit	Invitrogen	- No information in the public domain	3–4 times higher than *Taq* DNA polymerase (using an *rpsL*+ assay)
PfuUltra HF	*PfuUltra* High-Fidelity DNA polymerase	Stratagene	- Genetically engineered mutant of *Pfu* (*Pyrococcus furiosus*) DNA polymerase - ArchaeMaxx polymerase-enhancing factor	19 times higher than *Taq* DNA polymerase (using an *LacI*-based assay)

[a]Source: Technical bulletins from the commercial providers. The error rate of *Taq* DNA polymerase is 8.0 (± 3.9) × 10^{-6} (Reference 31).

Table 2. Error rates determined by 454 sequencing for seven different DNA polymerases after PCR amplification of four different exons from the human *TP53* oncogene, using a clonal *TP53* plasmid as starting template

	KOD (%)	Phusion HF (%)	Pt *Taq* (%)	Expand HF (%)	FastStart HF (%)	Sequal Prep Long (%)	*Pfu Ultra* HF (%)
Overall error rate[a]	0.21	0.11	0.34	0.25	0.23	0.29	0.23
Insertions	0.10	0.07	0.14	0.11	0.11	0.11	0.12
Deletions	0.06	0.02	0.08	0.07	0.05	0.06	0.05
Substitutions	0.01	0.01	0.07	0.04	0.03	0.07	0.01
Dots[b]	0.04	0.01	0.05	0.04	0.04	0.05	0.05

[a]Error rate, number of errors (miscalled bases, inserted or deleted bases) divided by total number of bases.
[b]Dot, three successive negative flows during 454 sequencing.

samples tested in this assay) (Figure 4). For mutations identified at frequencies below 1%, the intra- and interrun reproducibility was somewhat lower (intrarun: correlation >0.77 and concordance >0.66; interrun: correlation >0.83 and concordance >0.74). As the sequencing depth of the samples in these runs was designed to reach 1% (i.e., 5000 reads/sample), it was expected that reproducibility would decrease at mutation frequencies below this threshold level (1% = 50 reads, according to the Roche specifications). Besides, a higher variability at lower frequencies is also well described in other genomic technologies, such as microarray analysis (26).

Conclusions

UDS of amplicons using the GS FLX system enables the identification of rare variants by sequencing massive amounts of clonally amplified single molecules. Although the 454 sequencing process itself proves to be highly reproducible, the generation of accurate and reliable UDS results depends on many variables in the different steps of sample and data analysis processing. Awareness of the sources of potential sequencing errors helps to identify and distinguish erroneous from genuine sequence variations. The described case studies demonstrate that many of the sequencing errors sources can be avoided, solved, or controlled, but continuous attention in all UDS experiments is recommended, especially when newly developed amplicon assays are used.

It is clear from the presented cases that a thorough data analysis is a prerequisite for high-quality results. Rather than developing our own algorithms and software to enable data analysis tailored to specific applications of deep sequencing, as has been proposed and described by several investigators as an attractive solution (27–30), our data analyses have been largely based on the alignments derived from the standard data analysis software (Amplicon Variant Analysis; AVA) that is developed and supported by the technology provider (454 Life Sciences, Roche Applied Science). However, as we identified some specific types of 454 sequencing errors in our data, we developed complementary software tools that can resolve these issues (available for download in the Supplementary materials). Although our experiences and results show that our data analysis approach works for deep sequencing applications in general, there might be specific applications of 454-based deep sequencing that could benefit from the implementation of more tailored/developed algorithms and software

(27–30), but these were not considered in the scope of this manuscript.

Our analyses suggest that prudence is called for in UDS experiments targeting the identification of sequence variations at frequencies far below 0.5%, as the overall error rate inherent to PCR amplification was found at 0.11%–0.34%, depending on the enzyme used. A similar analysis of error sources in amplicon-based UDS applications by other novel generation sequencing technologies would be of interest and would allow for comparisons.

Competing interests

The authors declare no competing interests.

References

1. **Sanger, F., S. Nicklen, and A.R. Coulson.** 1977. DNA sequencing with chain-terminating inhibitors. Proc. Natl. Acad. Sci. USA *74*:5463-5467.
2. **Schuurman, R., L. Demeter, P. Reichelderfer, J. Tijnagel, T. de Groot, and C. Boucher.** 1999. Worldwide evaluation of DNA sequencing approaches for identification of drug resistance mutations in the human immunodeficiency virus type 1 reverse transcriptase. J. Clin. Microbiol. *37*:2291-2296.
3. **Van Laethem, K., K. Van Vaerenbergh, J.C. Schmit, S. Sprecher, P. Hermans, V. De Vroey, R. Schuurman, T. Harrer, et al.** 1999. Phenotypic assays and sequencing are less sensitive than point mutation assays for detection of resistance in mixed HIV-1 genotypic populations. J. Acquir. Immune Defic. Syndr. *22*:107-118.
4. **Milbury, C.A., J. Li, and G.M. Makrigiorgos.** 2009. PCR-based methods for the enrichment of minority alleles and mutations. Clin. Chem. *55*:632-640.
5. **Margulies, M., M. Egholm, W.E. Altman, S. Attiya, J.S. Bader, L.A. Bemben, J. Berka, M.S. Braverman, et al.** 2005. Genome sequencing in microfabricated high-density picolitre reactors. Nature *437*:376-380.
6. **Hirsch, M.S., H.F. Gunthard, J.M. Schapiro, F. Brun-Vezinet, B. Clotet, S.M. Hammer, V.A. Johnson, D.R. Kuritzkes, et al.** 2008. Antiretroviral drug resistance testing in adult HIV-1 infection: 2008 recommendations of an International AIDS Society—USA panel. Clin. Infect. Dis. *47*:266-285.
7. **Vandamme, A.M., A. Sonnerborg, M. Ait-Khaled, J. Albert, B. Asjo, L. Bacheler, D. Banhegyi, C. Boucher, et al.** 2004. Updated European recommendations for the clinical use of HIV drug resistance testing. Antivir. Ther. *9*:829-848.
8. **Balduin, M., M. Oette, M.P. Daumer, D. Hoffmann, H.J. Pfister, and R. Kaiser.** 2009. Prevalence of minor variants of HIV strains at reverse transcriptase position 103 in therapy-naive patients and their impact on the virological failure. J. Clin. Virol. *45*:34-38.
9. **Hoffmann, C., N. Minkah, J. Leipzig, G. Wang, M.Q. Arens, P. Tebas, and F.D. Bushman.** 2007. DNA bar coding and pyrosequencing to identify rare HIV drug resistance mutations. Nucleic Acids Res. *35*:e91.
10. **Le, T., J. Chiarella, B.B. Simen, B. Hanczaruk, M. Egholm, M.L. Landry, K. Dieckhaus, M.I. Rosen, and M.J. Kozal.** 2009. Low-abundance HIV drug-resistant viral variants in treatment-experienced persons correlate with historical antiretroviral use. PLoS One *4*:e6079.
11. **Metzner, K.J., S.G. Giulieri, S.A. Knoepfel, P. Rauch, P. Burgisser, S. Yerly, H.F. Gunthard, and M. Cavassini.** 2009. Minority quasispecies of drug-resistant HIV-1 that lead to early therapy failure in treatment-naive and -adherent patients. Clin. Infect. Dis. *48*:239-247.
12. **Mitsuya, Y., V. Varghese, C. Wang, T.F. Liu, S.P. Holmes, P. Jayakumar, B. Gharizadeh, M. Ronaghi, et al.** 2008. Minority human immunodeficiency virus type 1 variants in antiretroviral-naive persons with reverse transcriptase codon 215 revertant mutations. J. Virol. *82*:10747-10755.
13. **Simen, B.B., J.F. Simons, K.H. Hullsiek, R.M. Novak, R.D. Macarthur, J.D. Baxter, C. Huang, C. Lubeski, et al.** 2009. Low-abundance drug-resistant viral variants in chronically HIV-infected, antiretroviral treatment-naive patients significantly impact treatment outcomes. J. Infect. Dis. *199*:693-701.
14. **Varghese, V., R. Shahriar, S.Y. Rhee, T. Liu, B.B. Simen, M. Egholm, B. Hanczaruk, L.A. Blake, et al.** 2009. Minority variants associated with transmitted and acquired HIV-1 nonnucleoside reverse transcriptase inhibitor resistance: implications for the use of second-generation nonnucleoside reverse transcriptase inhibitors. J. Acquir. Immune Defic. Syndr. *52*:309-315.
15. **Wang, C., Y. Mitsuya, B. Gharizadeh, M. Ronaghi, and R.W. Shafer.** 2007. Characterization of mutation spectra with ultra-deep pyrosequencing: application to HIV-1 drug resistance. Genome Res. *17*:1195-1201.
16. **Thomas, R.K., E. Nickerson, J.F. Simons, P.A. Janne, T. Tengs, Y. Yuza, L.A. Garraway, T. LaFramboise, et al.** 2006. Sensitive mutation detection in heterogeneous cancer specimens by massively parallel picoliter reactor sequencing. Nat. Med. *12*:852-855.
17. **Ding, L., G. Getz, D.A. Wheeler, E.R. Mardis, M.D. McLellan, K. Cibulskis, C. Sougnez, H. Greulich, et al.** 2008. Somatic mutations affect key pathways in lung adenocarcinoma. Nature *455*:1069-1075.
18. **Qin, W., J.A. Chan, H.V. Vinters, G.W. Mathern, D.N. Franz, B.E. Taillon, P. Bouffard, and D.J. Kwiatkowski.** 2010. Analysis of TSC cortical tubers by deep sequencing of TSC1, TSC2 and KRAS demonstrates that small second-hit mutations in these genes are rare events. Brain Pathol. *20*:1096-1105.
19. **Campbell, P.J., E.D. Pleasance, P.J. Stephens, E. Dicks, R. Rance, I. Goodhead, G.A. Follows, A.R. Green, et al.** 2008. Subclonal phylogenetic structures in cancer revealed by ultra-deep sequencing. Proc. Natl. Acad. Sci. USA *105*:13081-13086.
20. **Cormier, E.G. and T. Dragic.** 2002. The crown and stem of the V3 loop play distinct roles in human immunodeficiency virus type 1 envelope glycoprotein interactions with the CCR5 coreceptor. J. Virol. *76*:8953-8957.
21. **Sander, O., T. Sing, I. Sommer, A.J. Low, P.K. Cheung, P.R. Harrigan, T. Lengauer, and F.S. Domingues.** 2007. Structural descriptors of gp120 V3 loop for the prediction of HIV-1 coreceptor usage. PLOS Comput. Biol. *3*:e58.
22. **Hughes, A. and M. Nelson.** 2009. HIV entry: new insights and implications for patient management. Curr. Opin. Infect. Dis. *22*:35-42.
23. **Wilkin, T.J., Z. Su, D.R. Kuritzkes, M. Hughes, C. Flexner, R. Gross, E. Coakley, W. Greaves, et al.** 2007. HIV type 1 chemokine coreceptor use among antiretroviral-experienced patients screened for a clinical trial of a CCR5 inhibitor: AIDS Clinical Trial Group A5211. Clin. Infect. Dis. *44*:591-595.
24. **Becker-Pergola, G., J.L. Mellquist, J.R. Eshleman, J. Brooks Jackson, and S.H. Eshleman.** 1999. Improved detection of human immunodeficiency virus type 1 variants by analysis of replicate amplification reactions: relevance to studies of human immunodeficiency virus type 1 vertical transmission. Mol. Diagn. *4*:261-268.
25. **Vandenbroucke, I., H. Van Marck, W. Mostmans, V. Van Eygen, E. Rondelez, K. Thys, K. Van Baelen, K. Fransen, et al.** 2010. HIV-1 V3 envelope deep sequencing for clinical plasma specimens failing in phenotypic tropism assays. AIDS Res. Ther. *7*:4.
26. **Huber, W., A. von Heydebreck, H. Sultmann, A. Poustka, and M. Vingron.** 2002. Variance stabilization applied to microarray data calibration and to the quantification of differential expression. Bioinformatics *18(Suppl 1)*:S96-S104.
27. **Wang, C., Y. Mitsuya, B. Gharizadeh, M. Ronaghi, and R.W. Shafer.** 2007. Characterization of mutation spectra with ultra-deep pyrosequencing: application to HIV-1 drug resistance. Genome Res. *17*:1195-1201.
28. **Eriksson, N., L. Pachter, Y. Mitsuya, S.Y. Rhee, C. Wang, B. Gharizadeh, M. Ronaghi, R.W. Shafer, and N. Beerenwinkel.** 2008. Viral population estimation using pyrosequencing. PLOS Comput. Biol. *4*:e1000074.
29. **Prosperi, M.C., L. Prosperi, A. Bruselles, I. Abbate, G. Rozera, D. Vincenti, M.C. Solmone, M.R. Capobianchi, and G. Ulivi.** 2011. Combinatorial analysis and algorithms for quasispecies reconstruction using next-generation sequencing. BMC Bioinformatics *12*:5.
30. **Zagordi, O., R. Klein, M. Däumer, and N. Beerenwinkel.** 2010. Error correction of next-generation sequencing data and reliable estimation of HIV quasispecies. Nucleic Acids Res. *38*:7400-7409.
31. **Cline, J., J.C. Braman, and H.H. Hogrefe.** 1996. PCR fidelity of pfu DNA polymerase and other thermostable DNA polymerases. Nucleic Acids Res. *24*:3546-3551.

Received 6 March 2011; accepted 9 August 2011.

Address correspondence to Jeroen Aerssens, Department of Translational Genomics & Genetics, Janssen Pharmaceutical Companies of Johnson & Johnson, Turnhoutseweg 30, 2340 Beerse, Belgium. e-mail: jaerssen@its.jnj.com