Published in final edited form as:
Electrophoresis. 2006 August ; 27(15): 3042–3047. doi:10.1002/elps.200500834.

Large-scale Pyrosequencing of synthetic DNA: A comparison with results from Sanger dideoxy sequencing

Baback Gharizadeh[1], **Zelek S. Herman**[1,*], **Robert G. Eason**[1,**], **Olufisayo Jejelowo**[2], and **Nader Pourmand**[1]

[1]Stanford Genome Technology Center, Stanford University, Palo Alto, CA, USA

[2]Texas Southern University, Houston, TX, USA

Abstract

Pyrosequencing is a relatively recent method for sequencing short stretches of DNA. Because both Pyrosequencing and Sanger dideoxy sequencing were recently used to characterize and validate DNA molecular barcodes in a large yeast gene-deletion project, a meta-analysis of those data allow an excellent and timely opportunity for evaluating Pyrosequencing against the current gold standard, Sanger dideoxy sequencing. Starting with yeast genomic DNA, parallel PCR amplification methods were used to prepared 4747 short barcode-containing constructs from 6000 *Saccharomyces cerevisiae* gene-deletion strains. Pyrosequencing was optimized for average read lengths of 25–30 bases, which included in each case a 20-mer barcode sequence. Results were compared with sequence data obtained by the standard Sanger dideoxy chain termination method. In most cases, sequences obtained by Pyrosequencing and Sanger dideoxy sequencing were of comparable accuracy, and the overall rate of failure was similar. The DNA in the barcodes is derived from synthetic oligonucleotide sequences that were inserted into yeast-deletion-strain genomic DNA by homologous recombination and represents the most significant amount of DNA from a synthetic source that has been sequenced to date. Although more automation and quality control measures are needed, Pyrosequencing was shown to be a fast and convenient method for determining short stretches of DNA sequence.

Keywords

DNA sequencing; Pyrosequencing technology; Sanger dideoxy sequencing; Synthetic oligonucleotides; Yeast deletion strains

1 Introduction

The ability to sequence DNA has had crucial impacts on biological and biomedical sciences and, since its advent in 1977 [1], it has assumed an important role in many scientific fields. DNA sequencing is used in such diverse areas as molecular biology, genetics, biotechnology, pharmacology, forensics, archeology, and anthropology. Given the new discoveries it continues to generate, DNA sequencing will continue to revolutionize the conceptual foundations of many sciences [2].

Correspondence: Dr. Nader Pourmand, Stanford Genome Technology Center, Stanford University, 855 California Ave, Palo Alto, CA 94304 USA, pourmand@stanford.edu, **Fax:** +1-650-812-1975.

*Current address: Herman Scientific Consulting, 521 Del Medio Avenue, no. 107, Mountain View, CA 94040, USA

**Current address: Applied Biosystems, 850 Lincoln Center Drive, Foster City, CA 94404, USA

The leading method for DNA sequencing is the elegant Sanger dideoxy chain termination reaction [1]. While this approach has been constantly improved and refined during the nearly three decades since its introduction, it is still limited by the cost of sequencing equipment and time to prepare samples. Pyrosequencing technology [3] is emerging as an alternate method for sequencing short segments of DNA. Although the Pyrosequencing technique has had read-length limitations [4], it is a fast method with real-time readout that is highly suitable for sequencing short stretches of DNA. Unlike Sanger sequencing, which incurs a reading gap of roughly 20 bases from the primer, Pyrosequencing has the significant advantage that it can be used to read DNA immediately adjacent to the primer. Sample preparation prior to DNA Pyrosequencing is also relatively rapid. Finally, the initial capital costs are lower and reagents are considerably less expensive for Pyrosequencing than for Sanger dideoxy sequencing for short stretches of DNA.

Despite the apparent advantages of Pyrosequencing for various applications, the two methods have not been rigorously compared. Thus in the present work, we compare Pyrosequencing technology and the Sanger dideoxy method for base-calling accuracy, sequence resolution, and convenience using 4747 synthetic DNA fragments [5]. We demonstrate here that Pyrosequencing is optimal for short stretches of DNA.

2 Materials and methods

2.1 DNA samples

Synthetic oligonucleotides inserted in the yeast gene-deletion strains were those originally acquired from Research Genetics Corp. (Huntsville, AL, USA) for our earlier study [5]. Culture conditions and DNA extraction were performed as described in that report.

2.2 Oligonucleotides and PCR

The synthetic DNA primer oligonucleotides were obtained from Qiagen (Alameda, CA, USA) and MWG Biotech (High Point, NC, USA). A complete set of 4747 yeast ORF-specific PCR primers was obtained from Illumina (San Diego, CA, USA) or from the oligonucleotide synthesis facility at the Stanford Genome Technology Center. Yeast deletion primer sequences for the complete set of strains are listed at http://www-sequence.stanford.edu/group/yeast_deletion_project/Deletion_primers_PCR_sizes.txt. Outer PCR (common) primers are: UPTAG, TAG1_FPCR2 (5’-TCATGCCCCTGAGCTGCGCACGT-3’); DNTAG, TAG2_FPCR2 (5’-TCGCCTCGACATCATCTGCCCAGA-3’). Inner PCR primers: UPTAG, TAG1_FPCR3 (5’-biotin-GAGCTGCGCACGTCAAGACTGTC-3’) and TAG1_RPCR1 (5’-GATGTCCACGAGGTCTCT-3’); DNTAG, TAG2_FPCR3 (5’-biotin-GACATCATCTGCCCAGATGCGAAG-3’) and TAG2_RPCR1 (5’-ACGGTGTCGGTCTCGTAG-3’). Sequencing primers: UPTAG, TAG1_RSEQ1 (5’-GATGTCCACGAGGTCT-3’); DNTAG, TAG2_RSEQ1 (5’-ACGGTGTCGGTCTCGT-3’). Amplification of the molecular barcodes in the yeast deletion strains is illustrated schematically in Fig. 1. Amplification of DNA for barcode Pyrosequencing was a two-stage process. Initially, each of the two barcode-containing segments (UPTAG and DNTAG) was amplified separately, using one common primer from the gene-deletion cassette approximately 150 nucleotides downstream of the tag (UPTAG TAG1_FPCR2, or DNTAG TAG2_FPCR2, 20 pmol per 15-μL reaction), and one strain-specific primer from a region approximately 300 nucleotides upstream of the tag (yeast ORF-specific UPTAG primer A or DNTAG primer D, 20 pmol). Each 15-μL PCR sample contained 10 mM Tris-HCl pH 8.0, 50 mM KCl, 3.3 mM $MgCl_2$, 0.2 mM dNTPs and 0.5 U *Taq* polymerase. PCR was run for 35 cycles with extension at 55°C. The actual sizes of the amplicons varied between 400 and 500 nucleotides, depending on the strain. These products (5 μL of a 1000-fold dilution, 2–5 ng DNA) were used as templates for nested PCR along with two common primers, one of 40–45 nucleotides downstream from the tag (UPTAG TAG1_FPCR1, or DNTAG TAG2_FPCR1, 5 pmol), and the other located in an 18-nucleotide common region immediately upstream of the tag (UPTAG TAG1_RPCR1, or DNTAG TAG2_RPCR1, 5 pmol). One inner PCR primer (FPCR1) was 5'-biotinylated to enable subsequent strand separation. Each 50-μL PCR sample contained 10 mM Tris-HCl pH 8.0, 50 mM KCl, 2.5 mM $MgCl_2$, 0.2 mM dNTPs, 1.5 U *Taq* polymerase, and 0.25 μL tetramethylenesulfone. The inner PCR was run with a touch-down program, 64-57°C over 14 cycles, then 26 cycles at 57°C. The resulting products were either 78 or 82 nucleotides in length. Approximately 1 pmol DNA was obtained per reaction. After removal of 10 μL for gel analysis, the remaining 40-μL product volume was used for Pyrosequencing.

2.3 Pyrosequencing

Biotinylated PCR product (40 μL) was immobilized onto streptavidin-coated super-paramagnetic beads (Dyna-beads M-280-streptavidin; Dynal AS, Oslo, Norway) by incubation at 43°C for 30 min. Single-stranded DNA was obtained by incubating the immobilized PCR product in 50 μL 0.1 M NaOH for 5 min. After a washing step, the immobilized strand was resuspended in 35 μL H_2O plus 4 μL annealing buffer (1 M Tris acetate pH 7.75, 200 mM magnesium acetate). Single-stranded DNA corresponding to 40 μL PCR product was hybridized to 10 pmol sequencing primer at 70°C for 3 min followed by incubation at room temperature for 5 min.

The Pyrosequencing reaction was performed at 28°C in a final volume of 50 μL on an automated PSQ 96 System according to the manufacturer's instructions (Biotage AB, Uppsala, Sweden). Briefly, single-stranded DNA (prepared as described above) was used for each Pyrosequencing reaction. The sequencing procedure was carried out by stepwise elongation of the primer strand upon sequential addition of the different deoxynucleoside triphosphates and simultaneous degradation of nucleotides by apyrase. The identity and quantity of nucleotide extension were determined by automated measurement of the amount of light generated after addition of a dNTP. Raw data were interpreted manually, and in a relatively small number of cases, questionable calls were verified by a second person.

2.4 Sanger dideoxy sequencing

The universal primer approach, FPCR3_TAG1 (UPTAG) and FPCR3_TAG2 (DNTAG) general sequencing primers (Fig. 1), for tag amplicons was used in Sanger DNA sequencing on an ABI 3700DNA Analyzer (Applied Biosystems) using BigDye terminator chemistry (Ver. 3.0) according to manufacturer's instructions. The PCR products were run in a cycle sequencing reaction with thermocycling conditions as follows: 25 cycles of 96°C for 10 s, 50°C for 5 s and 60°C for 4 min. The Sanger dideoxy sequencing base calls were done using phred [6]. Associated with each base call is a quality value (q) given by the log-transformed probability p of the base call being incorrect according to the equation [7]

$$q=-10\log_{10}p \tag{1}$$

Thus, a base call having a probability of 10^{-6} of being incorrect is assigned a q of 60, that of 10^{-3} has a q of 30, and a random call (1 in 4 probability of being correct) has a q of 1.25.

Point mutations having a q of <20 are usually regarded as unreliable, because this corresponds to an error rate of 1%. However, in the Sanger sequencing of the barcodes reported in [5], a large number (~500) of sequences were redone independently two, or sometimes three times. In about 20% of these cases, results were obtained with a q of <20 for a given point mutation, whereas in the independent sequencing of the same barcode, the same mutation was found to have *q* of >40 in subsequent runs. Moreover, Pyrosequencing confirmed these point mutations. In this report, therefore a more liberal approach has been adopted regarding Sanger point mutations; namely, point mutations having *q* of >15, corresponding to a base-call error probability of 3.16×10^{-3} (97% confidence) are considered acceptable for sequencing applications in order not to miss any possibly significant point mutations found by the Sanger technique. Unfortunately, there is no quality assessment for the base calls in Pyrosequencing.

3 Results

As shown in Table 1, the sequencing results fall into seven general categories. The first two categories indicate that dideoxy sequencing and Pyrosequencing are in complete agreement with each other in 3837 (80.8%) of the samples, of which 3044 (64.1%) are in accordance with the putative or expected sequence, and 793 (16.7%) differ from that sequence, but agree with the Sanger dideoxy sequencing results. In a slightly less straightforward category, 198 sequences (4.2%) gave a failing quality measure ($q \leq 15$) under Sanger sequencing, while Pyrosequencing results matched the putative sequence.

In 226 sequences (4.8%), the Sanger and putative sequences agree with each other but not with the Pyrosequences. In 57 sequences (1.2%), the Pyrosequencing and the putative sequences are in agreement with each other but not with the Sanger dideoxy sequences having point mutations with q >15. In 176 sequences (3.7%), the Sanger and the Pyrosequence do not agree with each other, nor does either agree with the putative sequence. Finally, there are 253 sequences (5.3%) for which Pyrosequencing yielded ambiguous results in the middle or at the end of the sequence.

Figure 2 provides an example of a definitive Pyrosequence that agrees with the putative sequence but not with the Sanger dideoxy sequence. In Fig. 2b, the Pyrogram sequence for the tag is shown in the shaded rectangle with one additional trailing base. The sequence derived from the pyrogram is in complete agreement with the putative sequence (Fig. 2a). In Fig. 2c, the Sanger dideoxy result is shown for the same tag. The Sanger result, which is in the opposite read direction from the Pyrosequence, indicates that there is a substitution of C to G, with a *q* of 39, at position 3 of the tag.

4 Discussion

In the present study, we analyzed in detail the sequence data of a large number of amplicons (barcode targets) from our earlier study [5] using two techniques (Pyrosequencing and Sanger dideoxy sequencing). These data were originally generated to validate barcodes in a yeast deletion project. Hence, the putative sequences of the barcodes were known, but the actual sequence was being tested in the earlier study. Although we have no reason to believe that either method is more accurate, to understand the current data, we made the assumptions that (i) the putative sequence may not be the actual sequence, and that (ii) Sanger sequencing with a good quality measure ($q > 15$) is likely to be correct.

The most straightforward case occurred when the sequences obtained by Pyrosequencing and by Sanger dideoxy sequencing were in complete agreement (categories 1 and 2). This occurred in nearly 81% of the amplicons. As expected, these sequences did not match the putative sequence in some cases due to misproduction of the barcode (category 2, ~17%). In a less straightforward case (category 3), the Sanger sequencing failed (q <15), but Pyrosequencing agreed with the putative sequence. This occurred in ~4% of cases. Assuming that the rate of misproduction was roughly equal throughout, then it is possible that up to 3.5% of these amplicons were correctly sequenced by Pyrosequencing.

Outright failure of Pyrosequencing is assumed to be likely in category 4, where the results of Sanger sequencing agree with the putative sequence but Pyrosequencing results differ. This occurred in 4.8% of cases. Similarly, in cases where Pyrosequencing differed from Sanger sequencing, but agreed with the putative sequence, we assume that the Sanger method was correct when attended by good quality control (q >15, category 5). This occurred in 1.2% of all cases, bringing the probable failure rate of Pyrosequencing close to 6%. Category 6 represents the situation where Pyrosequencing results were ambiguous. This occurred in 5.3% of all reads, a rate that is comparable with the dideoxy failure rate (q ≤15) of 4.2% cited above. These ambiguous cases represent areas of most probable improvement for the method (see below).

The least interpretable case occurred when Pyrosequencing, Sanger sequencing, and the putative sequence all mutually disagreed with each other (category 7, 3.7% of all cases). It would be tempting to attribute some of these discrepancies to variations in incorporation efficiency due to template base composition of a sequencing reaction or errors made by polymerase enzyme used in PCR. For example, in Sanger dideoxy sequencing, GC content is known to be problematic due to its propensity for forming secondary structures, while in Pyrosequencing, homopolymer string (mainly, homopolymeric T) regions can decrease uniform sequence peak heights and read-length. However, the barcodes were specifically designed to have a low GC content and to never have stretches of more than three homo-nucleotides [8].

Even triplet homopolymeric T, however, could account for some Pyrosequencing error; the incorporation of dATP-α-S nucleotide (2’-deoxyadenosine-5’-*O*’-1-thiotriphosphate) in T-homopolymeric regions results in uneven sequence signals and reduced data quality immediately downstream of such homopolyers [4]. During the development of the Pyrosequencing method, the natural dATP was replaced by dATP-α-S to increase the signal-to-noise ratio [9]. The natural dATP is a substrate for luciferase, giving rise to false-positive signals. Inefficient dATP-α-S incorporation by the exonuclease-deficient Klenow DNA polymerase causes the sequence traces to go out of phase, and result in asynchronous and ambiguous sequence data that can be manifested as uniformly reduced sequence signal peak-height. Therefore, efforts were made to compensate for this in the reading of those sections. This is an area where objective reading software with careful mathematical corrections could be of use in improving the method.

Since the time of this study, determining the sequence of these poly-T tracts has been improved significantly by use of sequenase [10], an exonuclease deficient T7 DNA polymerase, as it generates more high-resolution and uniform peak heights downstream of homopolymer stretches. Sequenase is not yet commercially available in Pyrosequencing kits, however. Most of the 5.3% sequence ambiguities observed and asynchronous extensions could be addressed using a more efficient DNA polymerase

Notably, discrepancies between the Sanger sequence and the putative sequence tended to indicate misincorporation (insertions) in either the readout or the actual sequence. On the other hand, there were no apparent insertions in Pyrosequencing sequence results.

In terms of convenience, Pyrosequencing showed itself to be highly adaptable to rapid parallelization as 96-samples could be sequenced in 40–50 min (sequencing 25 bases in average) with generated sequence data in real time. The current sample preparation method (single-strand separation and sequencing primer annealing) is relatively rapid (about 15 min for a 96-well microplate). In general, Sanger sequencing produced longer reads. Although dideoxy sequences having point mutations with low quality scores (q <20) often are found to have high quality (q >50) when the sequence is repeated, sample preparation takes approximately 4 h for Sanger sequencing (60 min for PCR clean-up, 3–4 h for cyclic amplification and 15 min for dye clean-up). On the other hand, sequence analysis of the Pyrosequencing results was performed manually for maximum accuracy, making data analysis slow. The Biotage AB is improving a new version of the present software, SQA 1.2, to make it suitable for automated analysis of longer sequences.

One of the main reasons that longer sequence reads (over 25 bases) were not generated using Pyrosequencing was because researchers at the time used low-cost SNP kits specifically designed for SNP analysis. Another limiting factor was the high cost of single-strand binding protein (SSB) [11], which precluded its use in these studies. However, SSB does improve sequence reads in general, and has recently been included in commercially available Pyrosequencing kits (Pyro Gold kit from Biotage). At present, the Pyrosequencing technology is limited to reads of up to 100 nucleotides [4]; however, efforts are under way to increase the read length. In contrast, the Sanger dideoxy sequencing method has average read lengths of 400–600 bases [12], and there are reports of Sanger sequencing reads of over 1000 bases [2].

While the completion of the first human genome draft has been solely achieved through the use of Sanger DNA sequencing, Pyrosequencing is clearly emerging as a viable alternative for whole-genome sequencing. Indeed, an adaptation of Pyrosequencing, 454 Life Science (www.454.com), has been used toward genomic level sequencing, such that 200 000 DNA fragments with an average read length of 105 bases are sequenced in a picoliter-format plate, and an average of 20 million bases are generated per 4-h run [13]. Despite its obvious advantages in reading length, the Sanger technique has not been successfully adapted to a highly multiplexed platform that would allow it to compete in rapidly sequencing large quantities of DNA for whole genome sequencing applications.

Our results demonstrate that for short DNA sequences, Pyrosequencing excels in reducing sample preparation time, providing ease-of-use and cost and labor savings. Error rates for Sanger sequencing and Pyrosequencing were comparable. Advantages of dideoxy sequencing are greater suitability for reading longer stretches of DNA, and availability of advanced software for analysis. Drawbacks of Pyrosequencing are rapidly being addressed with the use of SSB, sequenase, and automated software. Base calling software for Pyrosequencing now exists, but could be improved by optimizing base-calling software that eliminates out-of-phase reading following homopolymeric regions. Based on the analysis presented here, we see no reason that the accuracy of Pyrosequencing cannot be brought to >90%. Combined with its convenience, this level of accuracy would make Pyrosequencing an attractive technique for use in many fields.

Acknowledgments

This study was supported by grants from the National Institute of Health (1R21 A1059499-01 and PO1-HG000205). The other funding sources were from NASA Cooperative Agreements NCC-9-165, NCC-5-647 and NCC1-02038 (BG, OJ).

References

1. Sanger F, Nicklen S, Coulson AR. Proc. Natl. Acad. Sci. USA 1977;74:5463–5467. [PubMed: 271968]
2. Franca LT, Carrilho E, Kist TB. Q. Rev. Biophys 2002;35:169–200. [PubMed: 12197303]
3. Ronaghi M, Uhlen M, Nyren P. Science 1998;281:363–365. [PubMed: 9705713]
4. Gharizadeh B, Nordstrom T, Ahmadian A, Ronaghi M, Nyren P. Anal. Biochem 2002;301:82–90. [PubMed: 11811970]
5. Eason RG, Pourmand N, Tongprasit W, Herman ZS, et al. Proc. Natl. Acad. Sci. USA 2004;101:11046–11051. [PubMed: 15258289]
6. Ewing B, Hillier L, Wendl MC, Green P. Genome Res 1998;8:175–185. [PubMed: 9521921]

7. Ewing B, Green P. Genome Res 1998;8:186–194. [PubMed: 9521922]

8. Shoemaker DD, Lashkari DA, Morris D, Mittmann M, Davis RW. Nat. Genet 1996;14:450–456. [PubMed: 8944025]

9. Ronaghi M, Karamohamed S, Pettersson B, Uhlen M, Nyren P. Anal. Biochem 1996;242:84–89. [PubMed: 8923969]

10. Gharizadeh B, Eriksson J, Nourizad N, Nordstrom T, Nyren P. Anal. Biochem 2004;330:272–280. [PubMed: 15203333]

11. Ronaghi M. Anal. Biochem 2000;286:282–288. [PubMed: 11067751]

12. Marziali A, Akeson M. Annu. Rev. Biomed. Eng 2001;3:195–223. [PubMed: 11447062]

13. Margulies M, Egholm M, Altman WE, Attiya S, et al. Nature 2005;437:376–380. [PubMed: 16056220]

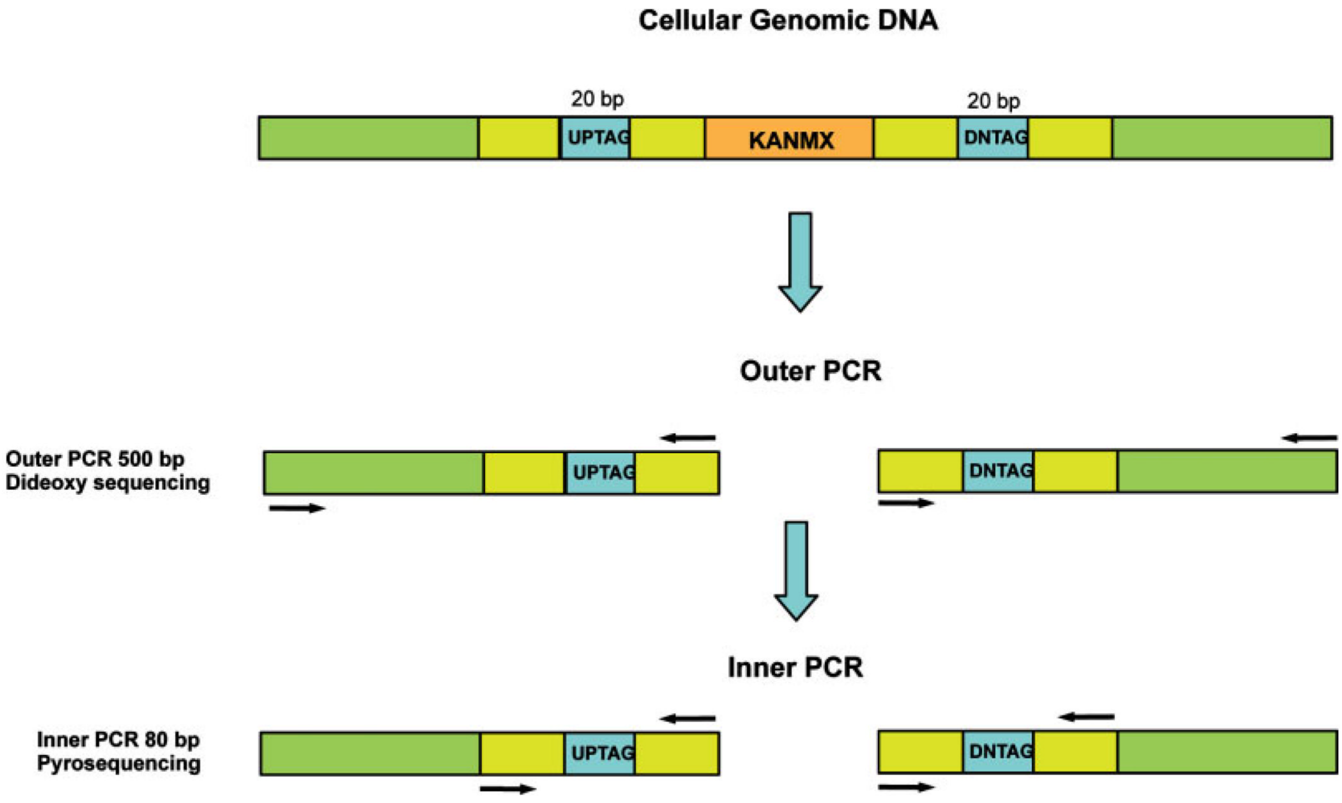

Figure 1.

Schematic illustration of the yeast gene-deletion cassette sequencing strategy. Outer and inner amplification of molecular DNA tags from cellular genomic DNA were performed for dideoxy sequencing and Pyrosequencing, respectively. The first round of PCR generated DNA fragments of 400–500 bases (for dideoxy sequencing), and the second PCR yielded 78–82-bp barcode-containing fragments for Pyrosequencing.

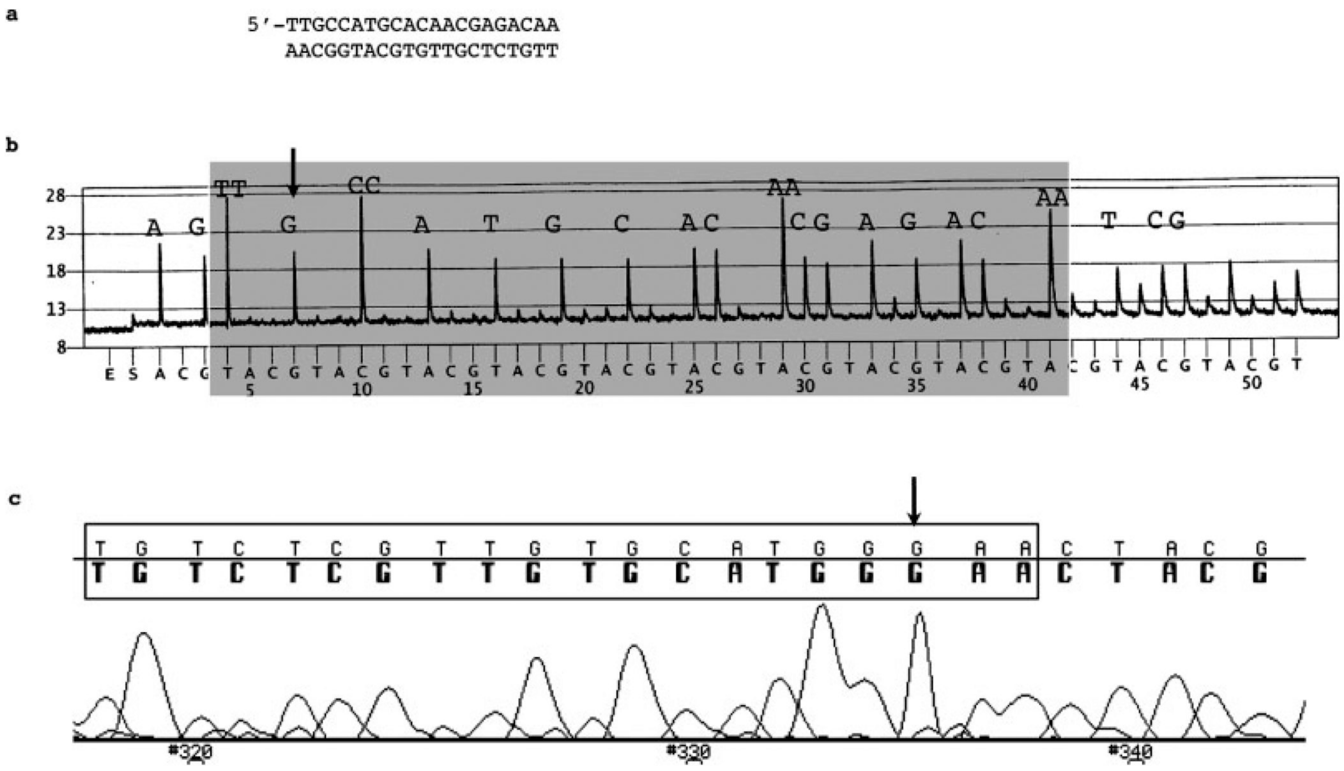

Figure 2.

Example of (ORF YOR054C DNTAG) sequence where Pyrosequencing confirms the putative sequence, which is not confirmed by Sanger dideoxy sequencing. (a) The putative sequence, (b) pyrogram from Pyrosequencing. The shadowed rectangle marks the region with the correct tag sequence (20 bases). The Pyrogram sequence agrees with putative sequence, (c) electropherogram from Sanger dideoxy sequencing. Arrow indicates where pyrogram agrees with the putative sequence, but not with the Sanger-derived sequence. The Sanger dideoxy sequencing result has a substitution of C to G (q = 39) at the position of the tag.

Table 1

Concordance of Pyrosequencing and Sanger sequencing: results of Pyrosequencing and Sanger dideoxy sequencing on the set of 4747 yeast deletion tags, as well as the expected putative sequences

Category	Result	No. of sequences	%	Cumulative totals
1	Pyro = Sanger = putative	3044	64.12	85.00%
2	(Pyro = Sanger) ≠ putative	793	16.71	
3	Pyro = putative and Sanger ≠ putative (*q* ≤15)	198	4.17	
4	Pyro ≠ (Sanger = Putative)	226	4.76	5.96%[a]
5	Pyro = putative and Sanger ≠ putative (*q* >15)	57	1.20	
6	Pyro ambiguous and Sanger definitive	253	5.33	5.33%[a]
7	Pyro ≠ Sanger ≠ putative	176	3.71	3.71%[a][b]
	Total	4747	100	100%

[a] Pyrosequencing improvable with SSB, sequenase, and better base calling software.

[b] Not easily interpretable with respect to accuracy.