BIOINFORMATICS ORIGINAL PAPER

Vol. 30 no. 24 2014, pages 3590–3597
doi:10.1093/bioinformatics/btu516

Gene expression

Advance Access publication August 30, 2014

Amplicon identification using SparsE representation of multiplex PYROsequencing signal (AdvISER-M-PYRO): application to bacterial resistance genotyping

Jérôme Ambroise[1,*], Yann Deccache[2], Leonid Irenge[1,2], Encho Savov[3], Annie Robert[4] and Jean-Luc Gala[1,2]

[1]Center for Applied Molecular Technologies (CTMA), Institut de Recherche Expérimentale et Clinique (IREC), Université catholique de Louvain, Clos Chapelle-aux-Champs 30, 1200 Bruxelles, [2]Defence Laboratories Department, Belgian Armed Forces, 1800 Vilvoorde, Belgium, [3]Department of Epidemiology and Hygiene, Military Medical Academy of Sofia, 1000 Sofia, Bulgaria and [4]Epidemiology and Biostatistics Department (EPID), Institut de Recherche Expérimentale et Clinique (IREC), Université catholique de Louvain, Clos Chapelle-aux-Champs 30, 1200 Bruxelles, Belgium

Associate Editor: John Hancock

ABSTRACT

Motivation: Pyrosequencing is a cost-effective DNA sequencing technology that has many applications, including rapid genotyping of a broad spectrum of bacteria. When molecular typing requires to genotype multiple DNA stretches, several pyrosequencing primers could be used simultaneously but this would create overlapping primer-specific signals, which are visually uninterpretable. Accordingly, the objective was to develop a new method for signal processing (AdvISER-M-PYRO) to automatically analyze and interpret multiplex pyrosequencing signals. In parallel, the nucleotide dispensation order was improved by developing the SENATOR ('SElecting the Nucleotide dispensATion Order') algorithm.

Results: In this proof-of-concept study, quintuplex pyrosequencing was applied on eight bacterial DNA and targeted genetic alterations underlying resistance to β-lactam antibiotics. Using SENATOR-driven dispensation order, all genetic variants (31 of 31; 100%) were correctly identified with AdvISER-M-PYRO. Among nine expected negative results, there was only one false positive that was tagged with an 'unsafe' label.

Availability and implementation: SENATOR and AdvISER-M-PYRO are implemented in the AdvISER-M-PYRO R package (http://sites.uclouvain.be/md-ctma/index.php/softwares) and can be used to improve the dispensation order and to analyze multiplex pyrosequencing signals generated in a broad range of typing applications.

Contact: jerome.ambroise@uclouvain.be

Received on January 30, 2014; revised on July 23, 2014; accepted on July 24, 2014

1 INTRODUCTION

Pyrosequencing is a DNA sequencing technology based on pyrophosphate release during nucleotide incorporation. The four possible nucleotides are sequentially dispensed in a pre-determined order. The first chemiluminescent signal produced during nucleotide incorporation is detected in the pyrosequencer and displayed as a pyrosequencing signal (also known as *pyrogram*^{TM}). The number of incorporated nucleotides at each position is computed from the corresponding peak height in the pyrosequencing signal. It has to be emphasized that the current pyrosequencing method should not be confused with high-throughput sequencing platforms, which also use pyrosequencing reactions. This article is clearly restricted to the analysis of pyrosequencing signals obtained with the pyrosequencing PyroMark technology commercialized by QIAGEN (Hilden, Germany).

Pyrosequencing has many applications, including rapid genotyping of a broad spectrum of bacteria (Ronaghi, 2001; Ronaghi and Elahi, 2002) and human single nucleotide polymorphism (SNP) genotyping (Aquilante *et al*., 2006; Butoescu *et al*., 2014; Koontz *et al*., 2009; Langaee and Ronaghi, 2005). Despite the increased use of next-generation sequencing for studying genomic diversity, pyrosequencing as defined here remains a cost-effective solution for genotyping short DNA stretches within bacterial genomes. This has previously been demonstrated by detecting SNPs associated with quinolone resistance and clustered in short DNA sequences known as quinolone resistance-determining region (Deccache *et al*., 2011). Considering that antibiotic resistance mechanisms are often linked to several hotspots located a small distance apart in the nucleotide sequence of a single gene as well as in several distinct genes (Hawkey, 1998), extended DNA-based identification of antibiotic resistance requires therefore analyzing successively, or at best, in parallel, resistance determinants spread at different locations.

Anyway, pyrosequencing reactions have to be conducted one by one, a procedure which substantially increases analytical expenses and technician workload. The alternative is to use several pyrosequencing primers simultaneously (Aquilante *et al*., 2006; Pourmand *et al*., 2002). However, the unavoidable consequence of this solution is that overlapping primer-specific signals are created, which are not visually interpretable. Another complementary issue is the need to select the nucleotide dispensation order for taking maximum advantage of all nucleotide sequence differences between the respective genetic targets.

Regarding the first issue related to multiplex signal interpretation, it is worth stating that complementary mPSQed and the

[*]To whom correspondence should be addressed.

MultiPSQ softwares were recently developed to help researchers designing and analyzing multiplex pyrosequencing assays (Dabrowski and Nitsche, 2012; Dabrowski *et al.*, 2013). The mPSQed software is used to define pyrosequencing primers, which generate unique uniplex pyrosequencing fingerprints and prevent the occurrence of competing signals from SNPs at different locations. The MultiPSQ software enables the analysis of multiplex *pyrogram*^{TM} originating from various pyrosequencing primers. MultiPSQ computes the similarity between multiplex pyrosequencing raw data and fingerprints resulting from any combination of theoretical uniplex pyrosequencing signals generated from a known sequence. The goal is to select the combination leading to the highest similarity. While MultiPSQ was claimed to suit the analysis of multiplex pyrosequencing signals generated by an unrestricted number of sequencing primers for identifying an unlimited number of polymorphisms, the application was practically validated on duplex signals (i.e. pyrosequencing reactions performed with two pyrosequencing primers).

In this study, the new SENATOR method was developed to improve the nucleotide dispensation order to be used all along the multiplex pyrosequencing experiment. Irrespective of the location of the pyrosequencing primer, the dispensation order is a crucial feature for avoiding similarities between uniplex pyrosequencing signals. The SENATOR function considers all unique nucleotide sequences (UNS) expected to be found within each genomic region of interest, hence improving the selection of a dispensation order that produces uncorrelated uniplex pyrosequencing signals. The global multiplex pyrosequencing signal is then interpreted using a new signal processing method based on a sparse representation of the pyrosequencing signal. The problem solved by sparse representation consists in creating a compact signal representation in terms of a linear combination of signals in an over-complete dictionary [i.e. a dictionary including a number of signals (p) that exceeds the dimension of the signal space (n)].

Sparse representation was recently used to develop AdvISER-PYRO, a new method dedicated to the analysis of low pyrosequencing signal intensities and complex signals from several target amplicons (Ambroise *et al.*, 2013). While used for rapidly identifying mycobacterial species-specific signals after pyrosequencing, AdvISER-PYRO showed high prediction performances. In this article, AdvISER-M-PYRO is used as a modified and updated version of original AdvISER-PYRO. AdvISER-M-PYRO is designed to enable the analysis and interpretation of multiplex pyrosequencing signals. Considering the steady increase of extended spectrum beta-lactamase producing bacteria (ESBL) worldwide and the occurrence of large ESBL-related outbreaks, a specific, high-throughput and multiplex DNA-based identification methods would have outstanding clinical relevance. This new analytical application was therefore selected as a proof-of-concept for SENATOR and AdvISER-M-PYRO.

In this study, AdvISER-M-PYRO was used to interpret pyrosequencing signals generated during quintuplex pyrosequencing experiments (i.e. pyrosequencing reactions performed with five specific pyrosequencing primers) to genotype SNPs within blaTEM and blaSHV beta-lactamase resistance genes [9]. To the best of our knowledge, it is the first time that quintuplex pyrosequencing signals are produced and reliably translated in sequences corresponding to each of their five respective targets.

2 METHODS

As a proof-of-concept, the new procedure for improving the selection of a nucleotide dispensation order and the new signal processing method were both applied to DNA extracted from four confirmed ESBL clinical bacteria of three different species (i.e. *Escherichia coli*, *Klebsiella pneumoniae*, *Enterobacter cloacae*), and four reference strains (ATCC700603, ATCC35218, DSM22313 and DSM22314), including three antibiotic-resistant and one antibiotic-susceptible strains. The method was used to generate and analyze quintuplex pyrosequencing signals resulting from simultaneous pyrosequencing of 11 unique SNPs distributed through three and two genomic regions of blaTEM and blaSHV beta-lactamase resistance genes (Cohen Stuart *et al.*, 2010), respectively. A multiplex polymerase chain reaction (PCR) was carried out in 50 μl of a reaction mixture containing the extracted DNA as template, primers and Power SYBR ® Green reagents (Applied Biosystems, Nieuwerkerk, The Netherlands). Amplification was performed on a 7900HT Fast Real-Time PCR System (Applied Biosystems, Nieuwerkerk, The Netherlands). Pyrosequencing was then carried out with a pyrosequencer PyroMark Q96 ID Sequencer from Qiagen (Hilden, Germany) on PCR products by using a mixture of the five pyrosequencing primers. Dispensed nucleotides produced distinct *pyrogram*^{TM} peaks, each peak height being proportional to the number of identical nucleotides consecutively incorporated.

2.1 SENATOR: SElecting nucleotide dispensATion ORder

Dispensation of selected nucleotides is carried out all along a pyrosequencing experiment and is absolutely mandatory to produce a pyrosequencing signal. For all uniplex pyrosequencing applications, this order of dispensation is determined using commercial software. As said above, this is, however, not applicable when designing multiplex applications. Consequently, the new SENATOR method was developed for selecting a suitable nucleotide dispensation order.

In a first step, a list was compiled with all UNS expected to be found within each genomic region of interest (i.e. three well-defined regions in blaTEM and two in blaSHV genes). Then, a series of dispensation order candidates was randomly generated and each dispensation order was evaluated. Accordingly, all theoretical UNS-based pyrosequencing signals were generated using an internal R function, and pairwise correlation coefficients between theoretical pyrosequencing signals were computed to identify the maximum correlation value. At the end of this analysis, the dispensation order producing the smallest maximum correlation was selected. Parameters of the SENATOR function include the length of the dispensation order as well as the number of dispensation order candidates to be evaluated. The length of the dispensation order can be chosen by trial and error but depends on the number of UNSs. When this number increases, a longer dispensation order is usually required to avoid the occurrence of perfectly correlated signals between distinct UNSs. If adequately selected, a longer dispensation order should indeed produce larger differences between highly similar UNS signals. On the other hand, a longer dispensation order increases the turnaround time and cost of the pyrosequencing reaction. Increasing the number of dispensation order candidates increases the probability of identifying an optimal dispensation order but at the expense of longer computational time. On an Intel(R) Core(TM) i7-2640 M CPU @ 2.80 GHz computer, the computation times for a list of 27 UNS and a dispensation order including 15 nt were 23 and 46 s for 1000 and 2000 candidates, respectively. Considering that SENATOR computation time is low compared with PCR and pyrosequencing time, it is therefore recommended to test a sufficient number of candidates.

2.2 AdvISER-M-PYRO

In the first step, of the AdvISER-M-PYRO algorithm development, a standardized learning dictionary was created that included theoretical pyrosequencing signals corresponding to all UNSs expected to be found within each genomic region of interest. For some UNSs, available experimental pyrosequencing signals (i.e. a signal obtained after carrying out uniplex pyrosequencing analysis of a gene of interest) were used to enrich the dictionary. Each experimental pyrosequencing signal was standardized by dividing the whole signal (i.e. the global pattern integrating all successive peak heights) by the first peak height that was representative of the incorporation of a single nucleotide.

In a second step, the multiplex pyrosequencing signal was modeled as a sparse representation of the signals from the dictionary. For the y testing signal of length n, the issue for sparse representation was to find a vector β $(\beta_j, j = 1, \ldots, p)$ minimizing the following function:

$$\sum_{i=1}^{n}(y_i - \sum_{j=1}^{p}\beta_j x_{ij})^2 + \lambda\|\beta_j\|_1, \tag{1}$$

where x_{ij} is *i-th* element of the *j-th* signal, $\|\beta_j\|_1$ is the L_1-norm of vector β_j and λ is a shrinkage parameter. In this study, the L_1- penalty was set at 0.05 because this value delivered high performance in a previous application of AdvISER-PYRO (Ambroise *et al.*, 2013). In this study, all penalized regression models were built using the penalized function of the corresponding R package (Goeman, 2008). Considering that the signal contribution of each UNS should have a positive value, an additional constraint imposing this prerequisite was implemented through the 'positive' parameter of the penalized function. The sum of regression coefficients corresponding to each UNS was computed and recorded as the UNS contribution to the signal. The contributions of each genomic region to the signal were also computed by summing the corresponding UNS contributions.

While not implemented with the previous AdvISER-PYRO version, the third step was a key feature of current AdvISER-M-PYRO. The main reason is related to the hierarchical structure of the UNS list in case of multiplex application. Considering that all UNSs belong to a restricted number of genomic regions, it is therefore required to select a single UNS for each genomic region. This selection was carried out by iteratively removing pyrosequencing signals from non-contributive UNSs from the standardized dictionary. This iterative process was sequentially applied on each genomic target, starting first with those with the highest contributions. At each iteration, a new penalized regression model was built and the UNS with the lowest contribution was identified and removed to obtain a reduced dictionary. This iterative process was stopped when a single UNS was obtained for each genomic target.

Finally, UNSs with a contribution lower than the 'Significant Contribution Threshold' (SCT) parameter were also removed from the reduced dictionary. The SCT parameter was fixed at 1 to avoid false-negative results. However, this low SCT value is likely to result in false-positive value and UNS contribution <2 were therefore tagged with an 'unsafe' label.

The ultimate step was to build a last penalized regression model with the final reduced dictionary. A correlation coefficient (r) was computed between predicted values of the penalized regression model and peak heights of the observed multiplex pyrosequencing signal. Considering that a low correlation coefficient is indicative of a significant difference between the observed multiplex pyrosequencing signal and the selected combination of uniplex pyrosequencing signals from the dictionary, this coefficient was used to assess the global confidence of the predicted UNS combination.

3 RESULTS

3.1 Nucleotide dispensation order selection

A list of all UNS expected to be found for each genomic region of interest was first compiled (Table 1). Because each genomic region can either be present or absent and because of the number of possible variants in each genomic region, the number of distinct genetic profiles in each bacteria can be as high as 5250 in total (i.e. 5*14*3*5*5).

As all pyrosequencing experiments were designed with reverse primers only, the reverse complement sequence was computed for each UNS. SENATOR was used to select a suitable dispensation order by integrating the list of all possible UNSs. Various lengths (between 10 and 20 with incremental steps of 2) of the dispensation order were tested and candidates (n = 1000) were generated and compared for each length. A nucleotide dispensation order (AGTGCGTACGTACA) of 14 nt was finally selected as it produced non-correlated uniplex pyrosequencing signals,

Table 1. List of all UNS expected to be found for each genomic region of interest

SHV179	SHV238	TEM104	TEM164	TEM238
ACGCCCGC**GA**C	AGCT**GG**C**G**A**G**C	GGTT**G**AGTACT	CTTGAT**CG**TTG	TGGAGCC**G**GT**G**
ACGCCCGC**GC**C	AGCT**AG**C**G**A**G**C	GGTT**A**AGTACT	CTTGAT**AG**TTG	TGGAGCC**A**GT**G**
ACGCCCGC**AA**C	AGCT**GC**C**G**A**G**C		CTTGAT**TG**TTG	TGGAGCC**G**GT**A**
ACGCCCGC**GG**C	AGCT**GG**C**G**A**A**C		CTTGAT**CA**TTG	TGGAGCC**A**GT**A**
	AGCT**AG**C**G**A**A**C			
	AGCT**CC**C**G**A**A**C			
	AGCT**GG**C**A**A**A**C			
	AGCT**AG**C**A**A**A**C			
	AGCT**CC**C**A**A**A**C			
	AGCT**GG**C**A**A**G**C			
	AGCT**AG**C**A**A**G**C			
	AGCT**CC**C**A**A**G**C			
	AGCT**GC**C**A**A**A**C			

Note. In each UNS, SNPs are shown with bold black letters.

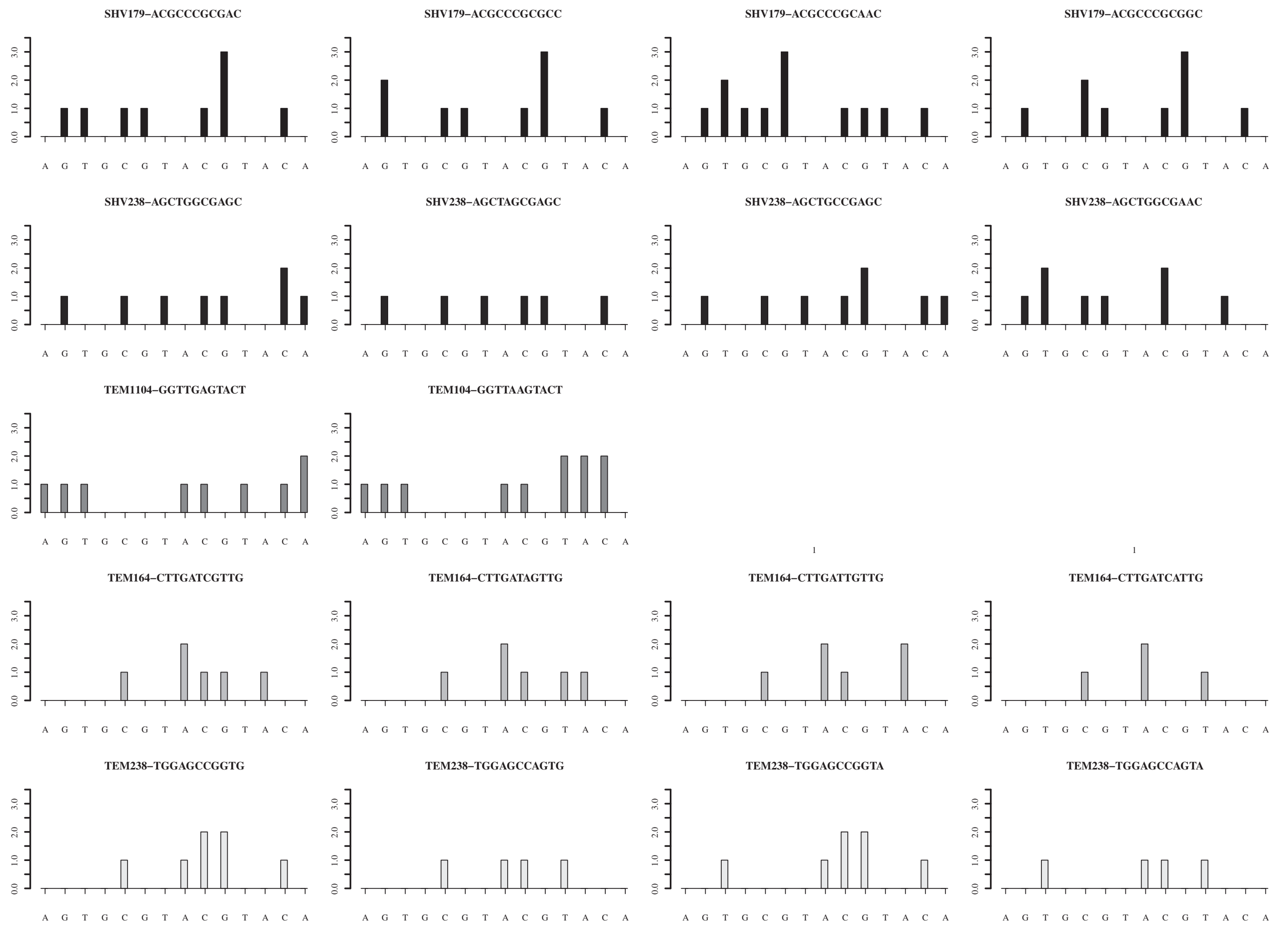

Fig. 1. Theoretical pyrosequencing signals obtained with UNS expected to be found in each genomic region of interest. For SHV238, only 4 of the 13 theoretical pyrosequencing signals are shown to preserve readability

while also limiting the length of the dispensation order. Theoretical pyrosequencing signals generated with the selected dispensation order showed an average pairwise correlation coefficient of 0.12 (range: –0.54–0.93), avoiding collinearity between signals contained in the dictionary and used as predictors in the regression models. Based on the selected dispensation order, all pyrosequencing signals are therefore defined as the global pattern integrating 14 successive peak heights generated by the pyrosequencer (Fig. 1).

3.2 AdvISER-M-PYRO

Five genomic regions of interest within blaTEM and blaSHV were characterized for eight selected bacteria (Table 2). Uniplex pyrosequencing signals were generated and were used to enrich the dictionary, as explained in the Methods section. Sanger sequencing analysis was used as a gold standard to ensure correct UNS-signal matching in the reference dictionary. In this proof-of-concept study, the dictionary included 65 pyrosequencing signals (i.e. 27 theoretical and 38 experimental signals). The dictionary consisted therefore of a matrix with 65 columns (i.e. one column for each signal) and 14 rows (i.e. one row for each dispensed nucleotide). All signals from the dictionary were processed the same way through the subsequent steps of AdvISER-M-PYRO, irrespective of their theoretical or experimental nature.

Multiplex pyrosequencing signals were then analyzed and automatically converted into their respective contributing genetic target, using AdvISER-M-PYRO. As shown in Table 3, a high global confidence of the predicted UNS combination was obtained for each multiplex pyrosequencing signal ($R > 0.99$), irrespective of corresponding signal-to-noise ratio. All genetic variants (31 of 31; 100%) were identified with AdvISER-M-PYRO among which 29 of 31 (93.5%) were tagged with a 'safe' label (i.e. with a contribution >2) and 2 of 31 (6.5%) with an 'unsafe' label (i.e. with a contribution <2). All these identifications were perfectly concordant (31 of 31; 100%) with Sanger sequencing and uniplex pyrosequencing results.

Among the nine expected negative results, there was only one false-positive UNS, which corresponded to strain ATCC700603 and was tagged as 'unsafe' label. The contribution of this false-positive result to the signal was small (1.14). It is of note that increasing the SCT parameter to a value >1.14 (but <1.34) would lead to a perfect concordance between Sanger Sequencing and

Table 2. UNS within the five genomic regions of the eight bacteria, as determined using Sanger Sequencing and uniplex pyrosequencing

Strain	SHV179	SHV238	TEM104	TEM164	TEM238
ATCC700603	ACGCCCGCGAC	AGCTGCCAAAC	–	–	–
ATCC35218	–	–	GGTTGAGTACT	CTTGATCGTTG	TGGAGCCGGTG
DSM22313	–	–	GGTTAAGTACT	CTTGATCGTTG	TGGAGCCAGTG
DSM22314	–	–	GGTTAAGTACT	CTTGATAGTTG	TGGAGCCGGTA
R021	ACGCCCGCGAC	AGCTAGCAAGC	GGTTGAGTACT	CTTGATCGTTG	TGGAGCCGGTG
MMA55	ACGCCCGCGAC	AGCTAGCAAGC	GGTTGAGTACT	CTTGATCGTTG	TGGAGCCGGTG
BS031	ACGCCCGCGAC	AGCTGGCGAAC	GGTTGAGTACT	CTTGATCGTTG	TGGAGCCGGTG
BS035	ACGCCCGCGAC	AGCTGGCGAGC	GGTTGAGTACT	CTTGATCGTTG	TGGAGCCGGTG

multiplex pyrosequencing but such tuning is likely to result in overoptimistic results and was therefore not used in the present proof-of-concept study. While reducing the number of pyrosequencing from 40 (uniplex) pyrosequencing reactions down to 11 reactions (i.e. eight quintuplex and three uniplex for confirming ‘unsafe’ tagged results), AdvISER-M-PYRO produced high-quality results.

3.3 Illustration of AdvISER-M-PYRO

Figure 2 illustrates the results obtained with AdvISER-M-PYRO when applied on two distinct pyrosequencing signals generated with resistance genetic determinants from strains DSM22314 and R021, respectively. When the over-complete dictionary was used, the pyrosequencing signal generated with former and latter strains was correctly converted into three (DSM22314, upper left panel) and five (R021, upper right panel) corresponding UNSs. In both cases, the correlation coefficient between predicted values of the penalized regression model and the 14 values of the pyrosequencing signal was 0.999.

The presence of a new mutation was simulated by removing all signals corresponding to UNSs TEM238-TGGAGCCGGTG and TEM238-TGGAGCCGGTA from the dictionary. In both cases, AdvISER-M-PYRO was prompted to create, from the library, an irrelevant combination of UNS signals to match the multiplex pyrosequencing signals (bottom panels of Fig. 2). Considering the high number of UNSs in each genomic region characterizing this proof-of-concept study, the irrelevant combination matched closely the pyrosequencing signal. Consequently, the correlation coefficient between predicted values of the penalized regression model and the 14 values of the pyrosequencing signal remained high (0.994 for DSM22314 and 0.997 for R021). As these confidence values were higher than those found with some correctly interpreted signals (i.e. 0.991 and 0.996 obtained with correct identification of ATCC700603 and MMA55, respectively), the correlation coefficient appeared to be totally inappropriate to detect such misinterpretation in the present application.

3.4 Impact of the AdvISER-M-PYRO enriched dictionary

To assess the impact of having an enriched dictionary, all multiplex signals were analyzed with a dictionary excluding all experimental pyrosequencing signals generated by genuine sample analyses. This reduced dictionary restricted, therefore, the analysis to a comparison of sample signals with 27 theoretical pyrosequencing signals. With this limited library, the analysis produced a high number of uncorrect (15 of 31; 48.4%) and false-positive identifications (6 of 9; 66.6%). In addition, all global confidence indexes were significantly reduced (0.848 <R <0.987).

4 DISCUSSION

Pyrosequencing is a cost-effective DNA sequencing technology that can be used for genotyping short DNA stretches within bacterial genomes. When the genotyping application requires analyzing multiple DNA stretches for diagnostic relevance, each pyrosequencing reaction must be carried out successively or in parallel, which increases reagent costs and technician work load. An alternative would consist in performing multiplex PCR followed by a multiplex use of pyrosequencing primers but visual interpretation of the resulting multiplex signals is tedious, time-consuming and mostly unreliable.

In this study, AdvISER-M-PYRO was, therefore, developed and used to test the feasibility of analyzing and automatically interpreting signals resulting from quintuplex pyrosequencing of SNPs located within blaTEM and blaSHV beta-lactamase resistance genes, all being first amplified by multiplex PCR. The huge (n = 5250) number of possible genetic profiles characterizing each bacterial strain made this case study highly challenging. Nonetheless, it enabled us to demonstrate the feasibility and strengths of the new analytical approach while also clarifying its potential limitations.

Dealing with a huge number of potential genetic profiles creates the possibility that any new mutation, being by definition not yet included in the original dictionary, produces a multiplex pyrosequencing signal, which may coincidentally match an irrelevant combination of UNS signals already present in the library (as illustrated in bottom panels of Fig. 2). This would unavoidably lead to misleading interpretation of the *pyrogram*^{TM}, resulting in genetic sequences assigned to wrong resistance determinants. Moreover, the high correlation coefficient between the recorded signal and model predictions would prevent the operator from detecting such misinterpretation, but would also lead to overlooking this yet unidentified mutation. Occurrence of

Table 3. Results obtained with AdvISER-M-PYRO on multiplex pyrosequencing signals from eight bacterial DNA

Strain	SHV179 UNS	SHV179 Contribution	SHV238 UNS	SHV238 Contribution	TEM104 UNS	TEM104 Contribution	TEM164 UNS	TEM164 Contribution	TEM238 UNS	TEM238 Contribution	R
ATCC700603	ACGCCCGCGAC	3.81	AGCTGCCAAAC	11.71	GGTTAAGTACT	1.14	–	0.00	–	0.00	0.991
ATCC35218	–	0.00	–	0.00	GGTTGAGTACT	19.21	CTTGATCGTTG	6.99	TGGAGCCGGTG	5.58	0.997
DSM22313	–	0.00	–	0.00	GGTTAAGTACT	18.14	CTTGATCGTTG	7.33	TGGAGCCAGTG	7.27	0.999
DSM22314	–	0.00	–	0.00	GGTTAAGTACT	17.69	CTTGATAGTTG	6.97	TGGAGCCGGTA	3.66	0.999
R021	ACGCCCGCGAC	3.00	AGCTAGCAAGC	5.81	GGTTGAGTACT	12.76	CTTGATCGTTG	5.04	TGGAGCCGGTG	2.61	0.999
MMA55	ACGCCCGCGAC	2.91	AGCTAGCAAGC	8.28	GGTTGAGTACT	11.52	CTTGATCGTTG	2.87	TGGAGCCGGTG	1.38	0.996
BS031	ACGCCCGCGAC	2.93	AGCTGGCGAAC	4.74	GGTTGAGTACT	14.04	CTTGATCGTTG	5.29	TGGAGCCGGTG	1.88	0.999
BS035	ACGCCCGCGAC	3.45	AGCTGGCGAGC	4.87	GGTTGAGTACT	14.72	CTTGATCGTTG	5.87	TGGAGCCGGTG	2.47	0.999

Note. For each genomic region, the first column corresponds to the identified UNS, while the second column corresponds to the contribution of this UNS to the global multiplex pyrosequencing signal. UNS contributions lower than two were tagged as 'unsafe'.

those issues is, however, less probable when the multiplex pyrosequencing application targets a limited number of well-defined, highly stable and reproducible genetic alterations, such as these found in human multiple SNP genotyping.

In other applications, a regular update of the dictionary is compulsory for integrating newly discovered UNSs. In the current application, the dictionary was therefore updated to reflect exactly the state of the art, as for today. Defining the upper threshold for the number of various genetic profiles minimizing the risk of misinterpretation would be highly valuable and extremely useful when designing new multiplex applications. This is unfortunately hardly achievable, except on a case-by-case basis, as too many factors influence directly the quality of genotyping results, among which the quality of the *pyrogram*^{TM} (i.e. the signal-to-noise ratio), the number of genomic regions targeted in the application, the length of the signals (i.e. the length of the target sequence that determines the number of dispensed nucleotides), and the pairwise correlation between uniplex pyrosequencing signals referenced in the dictionary. The same factors also influence the optimal number of genomic regions to be considered in the updated dictionary for ensuring a reliable multiplex identification of SNPs. Including more than five genomic regions in a multiplex application is perfectly feasible but with an increased risk of 'unsafe' results. The respective contribution of each genomic region in sparse representation of the multiplex result decreases when additional genomic regions are included. Likewise, the latter would increase the total number of UNSs, hence the probability of finding UNSs that are so similar that they cannot reliably be discriminated from one another, irrespective of the selected dispensation order.

In addition to its regular update on the basis of new data from the literature, the dictionary needs also to integrate experimental uniplex pyrosequencing signals generated by the analysis of genuine target samples. In the present study, using a dictionary only based on theoretical signals, decreased significantly the performances of AdvISER-M-PYRO. To be able to include experimental pyrosequencing signals appears therefore to be a clear and significant benefit of AdvISER-M-PYRO, compared with the MultiPSQ software. Moreover, the SCT parameter of AdvISER-M-PYRO enables to discard less contributing UNS, hence avoiding false-positive results, which can not be prevented by MultiPSQ.

In summary, an essential prerequisite of AdvISER-M-PYRO is to build an over-complete dictionary based on a list of all UNSs expected to characterize the genotyping application, and to update it regularly if necessary. When a compiled over-complete dictionary is available, multiplex pyrosequencing produces highly reliable results, as shown in the current proof-of-concept study. Most UNSs were correctly identified and tagged with a 'safe label', enabling to reduce substantially the total number of pyrosequencing reactions from 40 to 11. This novel multiplex pyrosequencing approach, which integrates the selection of the nucleotide dispensation order with SENATOR and the signals interpretation with AdvISER-M-PYRO reading software, enables therefore to lower the global turnaround time of genotyping and to decrease substantially analytical reagent costs while providing reliable target-specific results.

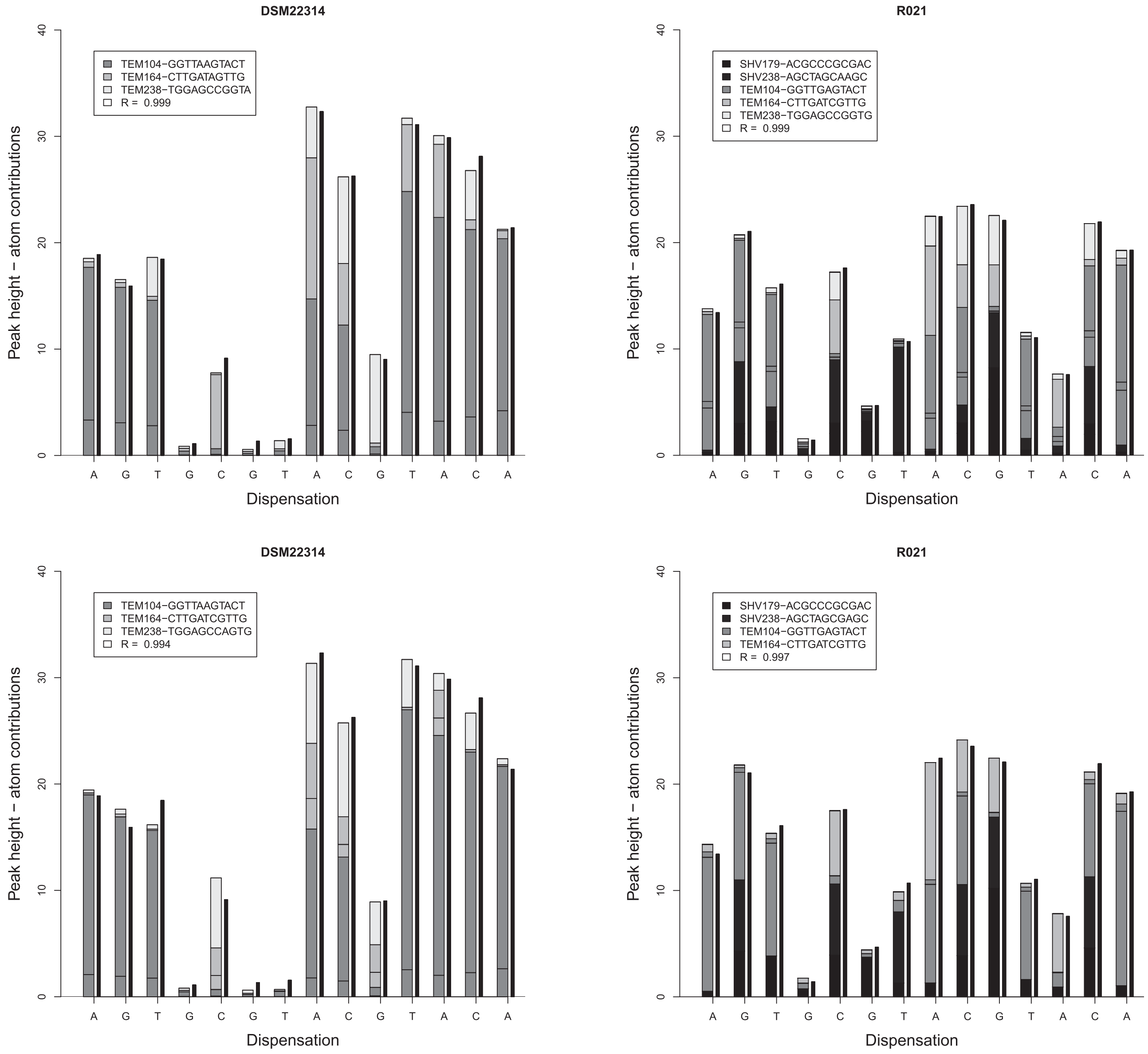

Fig. 2. Four examples of multiplex pyrosequencing signal identification with AdvISER-M-PYRO. The pyrosequencing signal is represented by vertical black lines. The contribution of each atom is represented with gray boxes stacked on top of the other. In both upper panels, an over-complete dictionary was used and signals were correctly converted into corresponding UNS. In both bottom panels, some signals were intentionally removed from the dictionary to demonstrate that an incomplete dictionary can lead to signals misinterpretation, hence resulting in misleading translation into non-corresponding UNS

ACKNOWLEDGEMENTS

The authors gratefully acknowledge Dan Marinescu from Bucharest Clinical Emergency Hospital, Romania for providing some of the strains used in this study.

Funding: This project was funded by the Department Management of Scientific and Technological Research of Defence (IRSD-RSTD; Royal High Institute for Defence) supporting research and development (grants MED-20). J.A. is funded by a grant from Institut de Recherche Expérimentale et Clinique (IREC).

Conflict of interest: none declared.

REFERENCES

Ambroise,J. *et al.* (2013) Adviser-pyro: amplicon identification using sparse representation of pyrosequencing signal. *Bioinformatics*, **29**, 1963–1969.

Aquilante,C.L. *et al.* (2006) Multiplex pcr-pyrosequencing assay for genotyping CYP3A5 polymorphisms. *Clinica Chimica Acta*, **372**, 195–198.

Butoescu,V. *et al.* (2014) Does genotyping of risk-associated single nucleotide polymorphisms improve patient selection for prostate biopsy when combined with a prostate cancer risk calculator? *Prostate*, **74**, 365–371.

Cohen Stuart,J. *et al.* (2010) Rapid detection of tem, shv and ctx-m extended-spectrum beta-lactamases in enterobacteriaceae using ligation-mediated amplification with microarray analysis. *J. Antimicrob. Chemother.*, **65**, 1377–1381.

Dabrowski,P. and Nitsche,A. (2012) mpsqed: a software for the design of multiplex pyrosequencing assays. *PloS One*, **7**, e38140.

Dabrowski,P.W. *et al.* (2013) Multipsq: a software solution for the analysis of diagnostic n-plexed pyrosequencing reactions. *PloS One*, **8**, e60055.

Deccache,Y. *et al.* (2011) Development of a pyrosequencing assay for rapid assessment of quinolone resistance in acinetobacter baumannii isolates. *J. Microbiol. Methods*, **86**, 115–118.

Goeman,J. (2008) Penalized: l1 (lasso) and l2 (ridge) penalized estimation in glms and in the cox model. R package version 09-21 2008. (Available from http://cran.rproject.org/web/packages/penalized/index.html.).

Hawkey,P.M. (1998) Action against antibiotic resistance: no time to lose. *Lancet*, **351**, 1298–1299.

Koontz,D.A. *et al.* (2009) Rapid detection of the cyp2a6* 12 hybrid allele by pyrosequencing® technology. *BMC Med. Genet.*, **10**, 80.

Langaee,T. and Ronaghi,M. (2005) Genetic variation analyses by pyrosequencing. *Mutat. Res.*, **573**, 96–102.

Pourmand,N. *et al.* (2002) Multiplex pyrosequencing. *Nucleic Acids Res.*, **30**, e31–e31.

Ronaghi,M. (2001) Pyrosequencing sheds light on dna sequencing. *Genome Res.*, **11**, 3–11.

Ronaghi,M. and Elahi,E. (2002) Pyrosequencing for microbial typing. *J. Chromatogr. B*, **782**, 67–72.