



- (51) **International Patent Classification:**
G06F 19/00 (2011.01)
- (21) **International Application Number:**
PCT/EP2014/058793
- (22) **International Filing Date:**
30 April 2014 (30.04.2014)
- (25) **Filing Language:** English
- (26) **Publication Language:** English
- (30) **Priority Data:**
1307913.2 2 May 2013 (02.05.2013) GB
- (71) **Applicant:** UNIVERSITE CATHOLIQUE DE LOUVAIN [BE/BE]; Place de l'Université 1, B-1348 Louvain-la-neuve (BE).
- (72) **Inventors:** AMBROISE, Jerome; rue des Carillonneurs 44, B-1348 Louvain-la-Neuve (BE). GALA, Jean-Luc; Countrylaan 24, B-1932 Zaventem (BE).
- (74) **Agents:** PECHER, Nicolas et al.; Centre Monnet, avenue Jean Monnet 1, B-1348 Louvain-la-Neuve (BE).
- (81) **Designated States** (unless otherwise indicated, for every kind of national protection available): AE, AG, AL, AM,

AO, AT, AU, AZ, BA, BB, BG, BH, BN, BR, BW, BY, BZ, CA, CH, CL, CN, CO, CR, CU, CZ, DE, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IR, IS, JP, KE, KG, KN, KP, KR, KZ, LA, LC, LK, LR, LS, LT, LU, LY, MA, MD, ME, MG, MK, MN, MW, MX, MY, MZ, NA, NG, NI, NO, NZ, OM, PA, PE, PG, PH, PL, PT, QA, RO, RS, RU, RW, SA, SC, SD, SE, SG, SK, SL, SM, ST, SV, SY, TH, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, ZA, ZM, ZW.

- (84) **Designated States** (unless otherwise indicated, for every kind of regional protection available): ARIPO (BW, GH, GM, KE, LR, LS, MW, MZ, NA, RW, SD, SL, SZ, TZ, UG, ZM, ZW), Eurasian (AM, AZ, BY, KG, KZ, RU, TJ, TM), European (AL, AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HR, HU, IE, IS, IT, LT, LU, LV, MC, MK, MT, NL, NO, PL, PT, RO, RS, SE, SI, SK, SM, TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, KM, ML, MR, NE, SN, TD, TG).

Published:

— without international search report and to be republished upon receipt of that report (Rule 48.2(g))

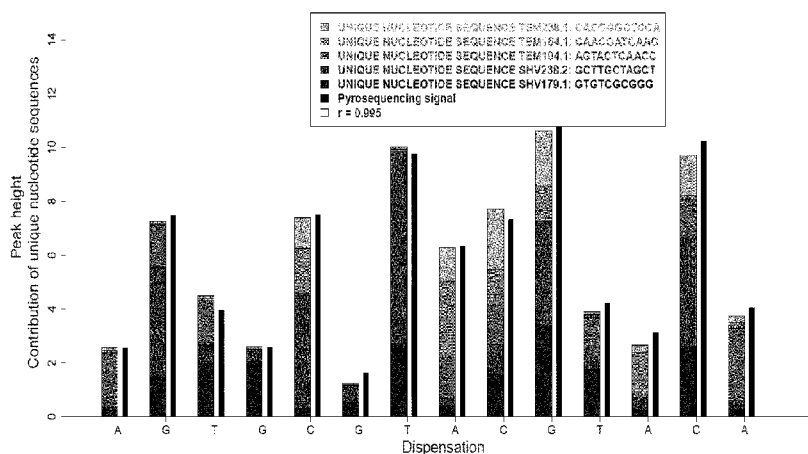
(54) **Title:** METHOD FOR ANALYSING A PYRO-SEQUENCING SIGNAL

Fig. 8

(57) **Abstract:** A method for analyzing a pyrosequencing signal that enables an automated, fast and reliable analyzes of combined pyrosequencing signals from homogeneous or heterogeneous samples based on an optimized order of dispensation. The method allows analyzing a pyrosequencing signal even when signal intensities are low or when there is more than one signal contribution. The method comprises a step of determining the respective contribution of standardized pyrosequencing signals in the analyzed pyrosequencing signal, the standardized pyrosequencing signals being comprised within a dictionary.

Method for analysing a pyro-sequencing signal

Field of the invention

5 **[0001]** The invention relates to a new approach for the analysis of pyro-sequencing signals. It comprises a method for analyzing pyrosequencing signals. The invention permits to identify and/or quantify each unique nucleotide sequence contribution in a global pyro-sequencing signal. The invention also relates to the creation of a dictionary of standardized pyro-sequencing signals
10 for its use within the method. The invention also relates to the determination of a nucleotide dispensation order that enhances the analysis of a pyro-sequencing signal by the method.

Description of prior art

15 **[0002]** Pyro-sequencing is a sequencing-by-synthesis method allowing the identification of single nucleotide polymorphisms (SNP) and other sequence variations between different cells, individuals and species. Pyro-sequencing is based on pyrophosphate release during nucleotide incorporation. The four
20 possible nucleotides are sequentially dispensed in a pre-determined order. The first chemi-luminescent signal produced during nucleotides incorporation is detected in the pyro-sequencer and displayed in pyro-sequencing signal (also known as *pyrogram*TM). The number of incorporated nucleotides at each position is computed from the corresponding peak height in the pyro-sequencing signal.
25

[0003] Pyro-sequencing is a method for genotyping a portion of a micro-organism genome that allows its rapid identification as well as an assessment of its antibiotic resistance. Identifying the specific composition of microorganisms inside a microbial population, i.e., type of strain and their respective proportion,
30 may help to choose the most specific treatment in case of illness due to micro-organism infection.

[0004] Pyro-sequencing is also a useful method when assessing the presence of oncogenic mutation(s) in a heterogeneous sample containing

tumour cells of interest mixed with background cells. The identification and/or the quantification of somatic mutations in specific genomic region of interest is of primary significance when making a diagnosis, determining a prognosis and/or choosing therapeutic strategies. Pyro-sequencing is also a useful method for genotyping specific genetic polymorphism(s) (*e.g.* Single Nucleotide Polymorphism(s) or SNP(s)) with pharmaco-genetic implications.

[0005] Pyrosequencing signals are generated from multiple amplicons produced from samples collected for numerous applications. A first one is dedicated to multiplex pyrosequencing. In this case, several pyrosequencing primers are used simultaneously, which leads to overlapping of primer-specific pyrosequencing signals. Hence, pyro-sequencing signals must be analyzed using data analysis methods. For example, a specific method was recently developed to analyze a multiplex pyrosequencing signal in Dabrowski *et al.* (PLoS ONE 8(3):e60055). This method was successfully applied to analyse duplex pyrosequencing signals generated from viral (orthopox) samples with two primers. Virus type differentiation was based on variations or SNPs within two genomic regions of interest. However, this method is based on the estimation of multiple models that integrate theoretical pyrosequencing signals and the method consequently does not permit an accurate and reliable analysis when signals are generated with a high number of pyrosequencing primers because the contribution of each primer-specific pyrosequencing signal becomes then very low. The method described here is dedicated to the analysis of a pyrosequencing signal generated from microorganisms with a single copy of each allele. This method is therefore also unable to analyse multiplex pyrosequencing signals generated from human, animals and vegetal diploid or polyploid cells.

[0006] A second application generating pyrosequencing signals from multiple amplicons is found in clinical molecular diagnostic laboratories testing somatic mutations in KRAS, NRAS, BRAF, PIK3CA, and EGFR genes. In such application, the pyrosequencing dispensation order is usually selected so that the mutated cells produce a peak at a position where the non-mutated cells produce no peak. The percentage of mutated cells is therefore estimated from a single peak height. In some situations, the signal from mutated cells and non-

mutated cells differs at more than at one single peak position. In such case, the method must be used to estimate the percentage of mutated cells. The interpretation of such pyrosequencing signals was recently addressed by Shen *et. al.* (*Diagnostic Pathology* 2012, 7:56) who developed a pyrosequencing data analysis method to detect EGFR, KRAS, and BRAF mutation analysis by taking into account multiple peak heights. However, this method requires a built-in formula defined specifically for each mutation and is not based on objective parameter computation exploiting a statistical method. Furthermore, this method estimates the percentage of mutated cells from a reduced subset of the signal by ignoring the information contained in the remaining unexploited signals and produces therefore suboptimal precision.

[0007] Current methods do not always allow an automated and reliable interpretation of pyro-sequencing signals, especially in the following specific situations:

- When a sample contains a very low DNA target concentration, which induces a signal with peak heights close to the noise level, rendering the current analyzes methods unable to identify correctly the DNA target (i.e., the unique nucleotide sequence) present in the sample. In such situations, the only option left is a cumbersome, time-consuming and usually very inefficient visual interpretation of the pyro-sequence signal.
- When a pyro-sequencing signal is generated from multiple amplicons. In that case, the global pyro-sequencing signal reflects the integrated contribution of each specific unique nucleotide sequence. Current analysis method does not permit to accurately identify and/or estimate the specific contribution of each specific and unique nucleotide sequence in a single pyro-sequencing signal, irrespective of the origin of said amplicons.

Summary of the invention

[0008] To solve at least partially the problems of the state of the art, it is an object of the invention to provide a method that permits the identification and/or the quantification of each unique nucleotide sequence contribution in a

pyro-sequencing signal irrespective of the presence of one or more than one unique nucleotide sequence within the sample. It is an object of the invention to provide a more accurate method to analyze pyro-sequencing signals with low signal-to-noise ratio. The inventors have developed a method which enables
5 also to analyze a pyro-sequencing signal when a single specific genomic region of interest with mutated and wild-type genetic variants is amplified after DNA extraction from a human tumor sample containing a variable proportion of mutated cells within a background of wild-type cells.

[0009] The method also enables to analyze a pyro-sequencing signal of a
10 single and unique consensus sequence presenting microorganism-specific DNA variations when this signal has been generated from a heterogeneous sample that contains various microorganisms. The method also enables the interpretation of a multiplex pyro-sequencing signal generated when several pyro-sequencing primers are used simultaneously in order to genotype at the
15 same time in a single reaction multiple genomic regions of interest. Hence, within only one pyro-sequencing reaction, several genomic regions of interest may be analyzed.

[0010] The invention is defined by the independent claims. The dependent claims define advantageous embodiments.

[0011] According to the invention there is provided a method for
20 analyzing a pyro-sequencing signal comprising the steps of:

a) Providing a simplex or multiplex genetic amplification product comprising at least one genomic region of interest generated from a DNA or RNA sample,

b) Acquiring one tested pyro-sequencing signal (Y) by the pyro-sequencing
25 reaction of said product,

c) Providing a dictionary of standardized pyro-sequencing signals (X_j), comprising at least one standardized pyro-sequencing signal (X_j) for each unique nucleotide sequence that can potentially be found within each genomic region of interest and wherein the nucleotide dispensation order used for
30 generating the standardized pyro-sequencing signals (X_j) contained within the dictionary and for generating the tested pyro-sequencing signal (Y) is the same,

d) Building a first penalized multiple regression model having an intercept set to zero, wherein the tested pyro-sequencing signal (Y) is a response variable and

all standardized pyro-sequencing signals (X_j) contained within the dictionary are predictor variables, wherein a non-negativity constraint is applied on the regression coefficients (β_j), and wherein a L_1 -norm and/or a L_2 -norm penalty is applied for estimating the regression coefficients (β_j) of the model,

5 e) Using the estimated regression coefficients ($\hat{\beta}_j$) of the penalized multiple regression model in order to estimate the respective contribution in the tested pyro-sequencing signal (Y) of each pyro-sequencing signal (X_j) contained in the standardized dictionary,

10 f) if a unique nucleotide sequence comprises more than one standardized pyro-signal (X_j) within the dictionary, then summing the estimated regression coefficients ($\hat{\beta}_j$) of all the standardized pyro-sequencing signals (X_j) corresponding to this unique nucleotide sequence for estimating the contribution of this unique nucleotide sequence.

[0012] Indeed, with the present method, it is possible to analyze a pyro-sequencing signal with a high accuracy because a dictionary of standardized pyro-sequencing signals (X_j) is provided, which may comprise more than one standardized signal for each unique nucleotide sequence, increasing the reliability of the method. Moreover, because a single penalized multiple regression model is used in relation with the dictionary of standardized pyro-sequencing signals (X_j), there is no need to create a multiplicity of models in order to select the best one, which can lead to the use of an erroneous model lacking of sparsity.

[0013] It is also an object of the invention to enhance the accuracy of the method by providing a specific dictionary of standardized pyro-sequencing signals (X_j) that may be adapted to the specific nature of the simplex or multiplex genetic amplification product that has to be analyzed with the method of the invention.

[0014] Accordingly, a method for constructing a dictionary of standardized pyro-sequencing signals for its use in the analysis method comprises the steps of:

30 A) Obtaining a list of all unique nucleotide sequences expected to be found in an at least one genomic region of interest,

B) Generating at least one pyro-sequencing signal for each unique nucleotide sequence of each genomic region of interest, all pyro-sequencing signals being generated with an identical nucleotide dispensation order,

C) Standardizing each pyro-sequencing signal relatively to an height of at least one peak present in the pyro-sequencing signal, the at least one peak height being representative of the incorporation of a given number of nucleotide, this given number being identical for every pyro-sequencing signal of the dictionary,

D) Generating a standardized pyro-sequencing signal dictionary by associating to each unique nucleotide sequence expected to be found in an at least one genomic region of interest at least one corresponding standardized pyro-sequencing signal.

[0015] Indeed, with this dictionary of standardized pyro-sequencing signals, standardization is individually performed for each unique nucleotide sequence expected to be found in an at least one genomic region of interest.

The dictionary may comprise more than one standardized signal for a unique nucleotide sequence, increasing the reliability of such dictionary. Moreover, specific dictionaries may be constructed for specific genetic polymorphisms, somatic mutations, or microorganisms consensus sequences analysis within a single pyro-sequencing reaction.

[0016] It is also an object of the invention to provide a nucleotide dispensation order to be used for constructing a dictionary of standardized pyro-sequencing signals and during a pyro-sequencing reaction of a sample, irrespective of the presence of one or more than one unique nucleotide sequence within the sample, said dispensation order permits to further enhance the accuracy and the reliability of the method.

[0017] Accordingly, a nucleotide dispensation order to be used during the construction of a dictionary of standardized pyro-sequencing signals and/or during a method to pyro-sequence at least one genomic region of interest is determined by the following steps:

- Obtaining a list of all unique nucleotide sequences expected to be found in an at least one genomic region of interest,
- Generating a plurality of nucleotide dispensation orders susceptible to be used during a pyro-sequencing of the at least one genomic region of interest,

- Generating a theoretical pyro-sequencing signal for each unique nucleotide sequence of the list, said pyro-sequencing signals being generated with each nucleotide dispensation order contained in the plurality of nucleotides dispensation orders,

- 5 - Determining within the plurality of nucleotide dispensation orders, the one that produces the less correlated pyro-sequencing signals for the nucleotide sequences expected to be found in the at least one genomic region of interest.

[0018] Such dispensation order may allow a more accurate estimation of the contribution to a global pyro-sequencing signal of each unique nucleotide
10 sequence presents in the dictionary of standardized pyro-sequencing signals.

Short description of the drawings

[0019] These and further aspects of the invention will be explained in
15 greater detail by way of example and with reference to the accompanying drawings in which:

Fig.1 shows the representation of theoretical pyrosequencing signals generated for the exon 20 of the EGFR oncogenic target with (a) a
20 prior art nucleotide dispensation order; (b) an optimized dispensation order according to one aspect of the invention. The horizontal axis represents the nucleotide dispensation order. The vertical axis represents the peak height.

Fig. 2 shows the representation of pyrosequencing signals of NRAS
25 oncogenic target, disclosing all the unique nucleotide sequences that can potentially be found within the target, all the pyrosequencing signals being generated with a single dispensation order according to one aspect of the invention. The horizontal axis represents the nucleotide dispensation order. The vertical axis represents the peak
30 height.

Fig. 3 shows the representation of theoretical pyrosequencing signals generated respectively with the Monkeypox virus and the Camelpox
CMS virus targets with (a) a prior art nucleotide dispensation order; (b)

an optimized dispensation order according to one aspect of the invention. The horizontal axis represents the nucleotide dispensation order. The vertical axis represents the peak height.

Fig. 4 shows a representative illustration of a dictionary of standardized pyrosequencing signals. Fig. 4(a) is a single pyrosequencing signal extracted from a dictionary with the nucleotide dispensation order on the horizontal axis. Fig. 4(b) is a sample from a dictionary. For clarity, only 9 pyrosequencing signals are represented, but ninety nine standardized pyrosequencing signals corresponding to the 33 distinct unique nucleotide sequences that can potentially be found in the genomic region of interest (i.e., the upstream-p34region of *mycobacterium*) were generated. The horizontal axis represents the nucleotide dispensation order. The vertical axis represents the peak height.

Fig. 5 shows a pyro-sequencing signal generated from a single genomic region of interest generated from homogeneous sample containing a pure population of the same microorganisms, and analyzed with a method according to the invention. The tested pyrosequencing signal is modeled as a linear combination of a single pyrosequencing signal found in the dictionary corresponding to the unique nucleotide sequence n°32. The horizontal axis represents the nucleotide dispensation order. The vertical axis represents the peak height.

Fig. 6 shows a pyrosequencing signal generated from a single genomic region of interest generated from a heterogeneous sample containing a mixed population of distinct microorganisms, and analyzed with a method according to the invention. The tested pyrosequencing signal is modeled as a linear combination of four pyrosequencing signals found in the dictionary, two corresponding to the unique nucleotide sequence n°19, and the other two corresponding to the unique nucleotide sequence n°18. The horizontal axis represents the nucleotide dispensation order. The vertical axis represents the peak height.

- Fig. 7 shows a pyrosequencing signal generated from a single genomic region of interest generated from a heterogeneous sample containing human tumour cells (i.e., the target cells) and background (i.e., irrelevant) cells, and analyzed with a method according to the invention. The tested pyrosequencing signal is modeled as a linear combination of three pyrosequencing signals found in the dictionary, two corresponding to the unique nucleotide sequence n°1, and the remaining signal corresponding to the unique nucleotide sequence n°5. The horizontal axis represents the nucleotide dispensation order. The vertical axis represents the peak height.
- Fig. 8 shows a pyrosequencing signal generated from multiple genomic regions of interest generated from a homogenous sample containing a pure population of the same micro-organism, and analyzed with a method according to the invention. The tested pyrosequencing signal is modeled as a linear combination of seven pyrosequencing signals of the dictionary, one corresponding to the unique nucleotide Sequence TEM238.1, two corresponding to the unique nucleotide sequence TEM164.1, one corresponding to the unique nucleotide sequence TEM104.1, two corresponding to the unique nucleotide sequence SHV238.2, and one corresponding to the unique nucleotide sequence SHV179.1. The horizontal axis represents the nucleotide dispensation order. The vertical axis represents the peak height.
- Fig. 9 a pyrosequencing signal generated from multiple genomic regions of interest generated from homogenous sample containing a quasi-pure population of human cells of interest, and analyzed with a method according to the invention. The tested pyrosequencing signal is modeled as a linear combination of six pyrosequencing signals found in the dictionary, one corresponding to the unique nucleotide sequence rs1016343 – C/T, two corresponding to the unique nucleotide sequence rs10896449 – A/A, two corresponding to the unique nucleotide sequence rs5945572 – A, and one corresponding to the unique nucleotide sequence rs5945619 – C. The horizontal axis

represents the nucleotide dispensation order. The vertical axis represents the peak height.

[0020] The drawings of the figures are neither drawn to scale nor proportioned. Generally, identical components are denoted by the same reference numerals in the figures.

Definitions

[0021] As used herein, the term “pyro-sequencing” refers to a method of nucleic acid sequencing based on a sequencing by synthesis principle, where a single strand of DNA to be sequenced is taken and its complementary strand is synthesized enzymatically. The pyro-sequencing relies on the detection of a pyrophosphate release upon nucleotide incorporation. Since the detection is performed sequentially following a dispensation order, analysis by the present method of the pyro-sequencing signal allows determination of all the nucleotide sequences of a genomic region of interest present in the DNA or RNA sample.

[0022] A genomic region of interest is defined as the genomic region extending beyond 3' end of the pyro-sequencing primer and that can represent either multiple distinct genetic variants including homozygous or heterozygous alleles (e.g., single nucleotide polymorphisms, deletion, insertion, duplication...) or multiple distinct nucleotide sequences (e.g. micro-organism-specific variations in a consensus sequence). Irrespective of the application (i.e., either identifying distinct genetic variants or distinct nucleotide sequences), both types of pyro-sequences will be hereafter named as “unique nucleotide sequence”.

[0023] The term pyro-sequencing primer refers to the single primer needed for initiating a pyro-sequencing reaction.

[0024] A simplex amplification product refers to genetic amplification of one single genomic region of interest.

[0025] A multiplex amplification product refers to the genetic amplification of more than one single region of interest. A multiplex pyro-sequencing signal is produced when multiple distinct genomic targets have simultaneously been amplified, hence are present in the amplification product after the processing of a homogeneous sample. For example, a multiplex pyro-sequencing signal is

generated when a multiplex analysis of constitutional Single Nucleotide Polymorphisms (SNPs) carried by different genes is performed using human DNA (e.g., constitutional pharmacogenetic testing of multiple SNPs in the same patient DNA), or when a multiplex analysis is carried out on DNA extracted from a pure culture of microorganisms (e.g. distinct genetic determinants of bacterial resistance in a single bacterial species...).

[0026] A L_1 -norm of a vector is defined as the sum of the absolute values of the elements of this vector.

[0027] A L_2 -norm of a vector is defined as the square root of the sum of the squared values of the elements of this vector.

[0028] An intercept of a linear regression model is defined as the expected mean value of the response variable when all predictor variable are equal to 0.

[0029] A response variable (Y) in a multiple linear regression model is defined as the variable which is modelled as a linear combination of the predictor variables.

[0030] A predictor variable (X_j) in a multiple linear regression model is defined as one of the variable which is used in a linear combination to model the response variable.

[0031] A regression coefficient (β_j) in a multiple linear model is defined as the factor which multiplies a predictor variable and that can be estimated from the data (the tested pyrosequencing signal (Y) and the standardized pyrosequencing signals (X_j) from the dictionary). An estimated regression coefficient ($\hat{\beta}_j$) is defined as the value of the factor after its estimation.

[0032] Building a penalized multiple regression model is defined as estimating the regression coefficients (β_j) of the model by imposing penalties on the L_p -norm (p being a positive integer) of these regression coefficients (β_j).

[0033] A Pearson product-moment correlation coefficient is defined as a coefficient which measures the linear dependence between two variables.

[0034] A nucleotide dispensation order refers to the sequential manner in which nucleotides are added during a pyro-sequencing reaction in order to determine the nucleotide sequences.

[0035] A pyro-sequencing signal is defined as a succession of positive numbers each of them representing a light intensity corresponding to the number of incorporated nucleotides at each dispensation step of a pyro-sequencing reaction. The pyro-sequencing signal obtained at the end of the pyrosequencing reaction is the global pattern resulting from the overall nucleotide incorporation according to all unique nucleotide sequences that are present in the simplex or multiplex amplification products and have matched dispensed nucleotides at each dispensation step.

10 Detailed description of embodiments of the invention

[0036] According to a first aspect of the invention, a method for analyzing a pyro-sequencing signal is provided, said method comprising the steps of:

- a) Providing a simplex or multiplex genetic amplification product comprising at least one genomic region of interest generated from a DNA or RNA sample,
- b) Acquiring one tested pyro-sequencing signal (Y) by the pyro-sequencing reaction of said product,
- c) Providing a dictionary of standardized pyro-sequencing signals (X_j), comprising at least one standardized pyro-sequencing signal (X_j) for each unique nucleotide sequence that can potentially be found within each genomic region of interest and wherein the nucleotide dispensation order used for generating the standardized pyro-sequencing signals (X_j) contained within the dictionary and for generating the tested pyro-sequencing signal (Y) is the same,
- d) Building a first penalized multiple regression model having an intercept set to zero, wherein the tested pyro-sequencing signal (Y) is a response variable and all standardized pyro-sequencing signals (X_j) contained within the dictionary are predictor variables, wherein a non-negativity constraint is applied on the regression coefficients (β_j), and wherein a L_1 -norm and/or a L_2 -norm penalty is applied for estimating the regression coefficients (β_j) of the model,
- e) Using the estimated regression coefficients ($\hat{\beta}_j$) of the penalized multiple regression model in order to estimate the respective contribution in the tested pyro-sequencing signal (Y) of each pyro-sequencing signal (X_j) contained in the standardized dictionary,

f) if a unique nucleotide sequence comprises more than one standardized pyro-signal (X_j) within the dictionary, then summing the estimated regression coefficients ($\hat{\beta}_j$) of all the standardized pyro-sequencing signals (X_j) corresponding to this unique nucleotide sequence for estimating the contribution of this unique nucleotide sequence.

[0037] The amplification product may be generated from any known method to amplify RNA and/or DNA sequences, like any kind of Polymerase Chain Reactions, allele-specific-PCR, assembly-PCR, asymmetric-PCR, hot-start-PCR, inverse-PCR, ligation-mediated-PCR, methylation-specific-PCR, multiplex-ligation-PCR, nested-PCR, quantitative-PCR, reverse-transcription-PCR, TAIL-PCR, all well known in the art.

[0038] DNA and/or RNA sample may be any kind of sample that allows a further amplification of at least a part of the genetic material, like a DNA sample extracted from a heterogeneous sample that contains the mutated cells or the micro-organism of interest within a background of wild-type cells or of other microorganisms. It can also be a DNA sample extracted from a homogeneous sample which includes a single and unique type of cell or a pure culture of micro-organism.

[0039] The tested pyro-sequencing signal (Y) is the overall signal generated after a pyro-sequencing reaction of the at least one genomic region of interest representing signal intensity corresponding to the number of incorporated nucleotides at each dispensation step. For example, the tested pyro-sequencing signal (Y) may be generated with a pyro-sequencer *PSQTM 96 M A* (Biotage, AB, Sweden) following successive dispensation of the nucleotides.

[0040] The dictionary of standardized pyro-sequencing signals (X_j) comprises at least one standardized pyro-sequencing signal (X_j) for each unique nucleotide sequence that can potentially be found within each genomic region of interest. A genomic region of interest may comprise a plurality of sequence variants, because of genetic polymorphism and/or mutation for example. The plurality of unique nucleotide sequences may be determined from genomic databases known from the man skilled in the art like NCBI for

example, or from experimental sequencing of each genomic sequence of interest.

[0041] The nucleotide dispensation order used for generating the standardized pyro-sequencing signals (X_j) of the dictionary has to be the same as the one used to generate the tested pyro-sequencing signal (Y) from the sample of interest. The dictionary also has to comprise at least one pyro-sequencing signals (X_j) for each unique nucleotide sequence expected to be found in the sample of interest comprising the at least one genomic region of interest.

[0042] A first penalized multiple regression model comprises the pyro-sequence signal (Y) as response variable, all signals from a dictionary of standardized pyro-sequencing signals (X_j) as predictors variable. In this model, the sum of all estimated regression coefficients ($\hat{\beta}_j$) corresponding to each unique nucleotide sequence is computed and recorded as the specific contribution to the pyro-sequencing signal (Y) of a unique nucleotide sequence. When the length of the pyro-sequencing signal (Y) is smaller than the number of pyro-sequencing signals present in the standardized dictionary, a L_1 and/or a L_2 norm penalty is applied for estimating the regression coefficients (β_j). When the length of the pyro-sequencing signal (Y) is not smaller than the number of pyro-sequencing signals present in the standardized dictionary, L_1 and/or a L_2 norm penalty may be set to zero. As the signal contribution from each standardized pyrosequencing signal (X_j) presents in the dictionary should not be negative, an additional constraint imposing a non-negativity on the regression coefficients (β_j) is implemented. The intercept of the model is also set to 0.

[0043] Any kind of penalized multiple regression models may be used. In a preferred embodiment, the penalized multiple regression model of the corresponding R package (Goeman, 2008) may be used.

[0044] The L_1 and the L_2 norm are hyper parameters of the model. The determination of their respective value may be performed using a validation dataset. Different values for the L_1 and for the L_2 norms are generated, and for each combination of said values, all pyro-sequencing signals of the validation dataset with known contributions are analysed with the method. The combination of values for the L_1 and the L_2 norms that permits to obtain the

highest rate of correct identification is selected. In certain contexts, the unique nucleotide sequences that can potentially be found in the genomic region of interest have different epidemiologic prevalences, some of them being very rare while some others are very frequent. In these cases, it is possible to apply
 5 distinct penalties on L1- and/or L2 – norm of the regression coefficients β_j . Higher penalties on L1 and/or L2 – norm may be applied for the rare mutations to increase the specificity, while lower penalties on L1 and/or L2 – norm may be applied for the common mutations to increase the sensitivity.

[0045] In a preferred embodiment, the model is defined by:

$$Y = \sum_{j=1}^p X_j \beta_j + \varepsilon$$

And the estimation of the regression coefficients in the penalized multiple regression model is defined by:

$$\hat{\beta} = \underset{\beta}{\operatorname{argmin}} \left\{ \sum_{i=1}^N \left(y_i - \sum_{j=1}^p x_{ij} \beta_j \right)^2 + \lambda_1 \|\beta_j\|_1 + \lambda_2 \|\beta_j\|_2 \right\}$$

or

$$\hat{\beta} = \underset{\beta}{\operatorname{argmin}} \left\{ \sum_{i=1}^n (y_i - \sum_{j=1}^p \beta_j x_{ij})^2 + \lambda_{1j} \|\beta_j\|_1 + \lambda_{2j} \|\beta_j\|_2 \right\}$$

when different penalties are used on the regression coefficients,

where Y is the pyrosequencing signal, y_i is the i^{th} element of the Y vector, X_j is the j^{th} pyrosequencing signal of the dictionary of standardized pyro-sequencing signal that corresponds to a unique nucleotide sequence, x_{ij} is the i^{th} element of the j^{th} pyrosequencing signal, each β_j is a regression coefficient associated to the X_j pyro-sequencing signal that corresponds to a unique nucleotide sequence, $\|\beta_j\|_1$ is the L_1 – norm of the vector of the regression coefficients β_j , $\|\beta_j\|_2$ is the L_2 – norm of the vector of the regression coefficients β_j , λ_{1j} are penalty factors on the L_1 -norm of the regression coefficients (β_j) which are determined with a validation dataset as explained before, λ_{2j} are penalty factor on the L_2 -norm of the regression coefficients (β_j) which are determined with a validation dataset as explained before, ε is a vector of random error with a normal distribution.

[0046] The pyro-sequencing signal (Y) is therefore modeled as a linear combination of the pyro-sequencing signals present in the standardized dictionary (X_j): the estimated regression coefficients ($\hat{\beta}_j$) of the penalized multiple regression model are used to determine the respective contribution in the tested pyro-sequencing signal (Y) of each pyro-sequencing signal (X_j) contained in the standardized dictionary.

[0047] Because the dictionary used in the method may comprise more than one standardized pyro-sequencing signal for a unique nucleotide sequence, if a unique nucleotide sequence comprises more than one standardized pyro-sequencing signal (X_j) within the dictionary, the contribution of this unique nucleotide sequence is finally estimated by summing the

estimated regression coefficients ($\hat{\beta}_j$) of all the standardized pyro-sequencing signals within the dictionary corresponding to this unique nucleotide sequence. The sum of the estimated regression coefficients ($\hat{\beta}_j$) corresponding to each unique nucleotide sequence are then computed and recorded as the unique nucleotide sequence-specific contribution.

[0048] With the method, it may be possible to detect the contribution of a unique nucleotide sequence within a plurality of unique nucleotide sequences when the unique nucleotide sequence contributes only to 10%, more preferably only to 5%, still more preferably only to 1% of the tested pyro-sequencing signals (Y).

[0049] In some experiments, the method may be improved by determining a single unique nucleotide sequence contribution for each genomic region of interest and performing a subsequent analysis. For example, when a sample includes a single and unique type of cell or microorganism, these further steps allow a more accurate analysis of a tested pyrosequencing signal.

[0050] Accordingly, in a preferred embodiment, the method of the invention also comprises the steps of:

g) Determining for each genomic region of interest, the unique nucleotide sequence with the highest contribution in the tested pyro-sequencing signal (Y),

h) Creating a reduced dictionary by selecting for each genomic region of interest the standardized pyro-sequencing signals (X_j) corresponding to the determined unique nucleotide sequence,

i) Building a second penalized multiple regression model having an intercept set to zero, wherein the tested pyro-sequencing signal (Y) is a response variable and all standardized pyro-sequencing signals (X_j') contained within the reduced dictionary are predictor variables, wherein a non-negativity constraint is applied on the regression coefficients (β_j'), and wherein a L_1 -norm and/or a L_2 -norm penalty is applied for estimating the regression coefficients (β_j') of the second model,

j) Using the estimated regression coefficients ($\hat{\beta}_j'$) of the penalized multiple regression model in order to estimate the respective contribution in the resulting tested pyro-sequencing signal (Y) of each pyro-sequencing signal (X_j') present in the reduced dictionary,

k) if a unique nucleotide sequence comprises more than one standardized pyro-signal (X_j') within the dictionary, then summing the estimated regression coefficients ($\hat{\beta}_j'$) of all the standardized pyro-sequencing signals (X_j') corresponding to this unique nucleotide sequence for estimating the contribution of this unique nucleotide sequence,

these steps being performed after the step of estimation of contribution of each unique nucleotide sequence of each genomic region of interest.

[0051] In a preferred embodiment, these steps are performed when the product is generated from a homogeneous sample including a single type of cell or a single type of microorganism (i.e. a pure strain). A homogeneous sample is defined as a sample that comprises at least 95%, preferably at least 99%, ideally 100%, of a single cell type or microorganism. The determination of sample homogeneity may be performed by cell counting or genetic screening for example.

[0052] The second penalized multiple regression model used may be the same as the one used before. L_1 and L_2 norms may be the same as the one used in the previous step. L_1 and L_2 norms may also be different. In that case, a determination of their values may be defined with a validation dataset as explained before.

[0053] When a genomic region of interest comprises two known homozygous variants and one known heterozygous variant, it is possible to further enhance the analysis method by determining the unique nucleotide sequence from the genomic region of interest with the highest contribution to the tested pyro-sequencing signal.

[0054] Accordingly, in a preferred embodiment, a step of selecting for each genomic region of interest the unique nucleotide sequence with the highest contribution in the tested pyro-sequencing signal (Y) comprises the further steps of:

l) Identifying the minimum contribution of both homozygous variants of the genomic region of interest,

m) Adding twice the estimated minimum value to the contribution of the heterozygous variant of the genomic region of interest,

n) Subtracting the estimated minimum value from the contribution of both homozygous variant of the genomic region of interest,

o) Determining the unique nucleotide sequence in the genomic region of interest with the estimated highest contribution in the tested pyro-sequencing signal (Y),

these steps being performed when the simplex or multiplex genetic amplification product is generated from DNA or RNA extracted from a diploid cell and the genomic region of interest presents two known homozygous variants and one known heterozygous variant.

In some experiments, determining a single unique nucleotide sequence contribution for each genomic region of interest may be improved by iteratively eliminating at least one pyro-sequencing signal from the dictionary, and iteratively build reduced pyro-sequencing signals dictionaries to be used for building subsequent penalized multiple regression models, until a single pyro-sequencing signal for each genomic region of interest is present in a last dictionary. Accordingly, the step of selecting for each genomic region of interest the unique nucleotide sequence with the highest contribution in the tested pyro-sequencing signal (Y) may comprise the further steps of:

ei) determining for at least one genomic region of interest the standardized pyro-sequencing signal (X_i) corresponding to the less contributive signal in the tested pyro-sequencing signal (Y),

eii) removing the corresponding pyro-sequencing signal (X_i) from the dictionary, thereby creating a reduced dictionary,

eiii) building a new additional penalized multiple regression model with the reduced dictionary created at the previous step,

are performed iteratively after each step of estimation of the respective contribution in the tested pyro-sequencing signal (Y) until a unique nucleotide sequence is obtained for each genomic region of interest may be performed. These steps are performed after each step of estimation of the respective contribution of each pyro-sequencing signal comprised in the dictionary. This iterative process may be performed for each genomic region, by starting with the one which is associated with the highest contribution, and finishing with the one with the lowest contribution.

Accordingly, the step ei), ei)i and eiii) are applied on each genomic region of interest, by starting with the one which is associated with the highest contribution and finishing with the one which is associated with the lowest contribution.

- 5 **[0055]** When the genomic region of interest can present a large number of different unique nucleotide sequences, it may be possible to select significant contributors in order to have a clearer and workable analysis.

[0056] Accordingly, in a preferred embodiment, the method comprises the further steps of:

- 10 p) Determining the unique nucleotide sequences whose contributions are above a determined Significant Contribution Threshold, and when no contribution is above the Significant Contribution Threshold, determining the most contributing unique nucleotide sequence,
- q) Selecting the standardized pyro-sequencing signals (X_j'') corresponding to
15 the determined unique nucleotide sequences from the standardized dictionary, in order to create a reduced dictionary,
- r) Using an additional penalized multiple regression model having an intercept set to zero, wherein the tested pyro-sequencing signal (Y) is a response variable and all standardized pyro-sequencing signals (X_j'') contained within the
20 reduced dictionary are predictor variables, wherein a non-negativity constraint is applied on the regression coefficients (β_j'') and wherein a L_1 -norm and/or a L_2 -norm penalty is applied for estimating the regression coefficients (β_j'') of the additional model,
- s) Using the estimated regression coefficients ($\hat{\beta}_j''$) of the penalized multiple
25 regression model in order to estimate the respective contribution in the resulting tested pyro-sequencing signal (Y) of each pyro-sequencing signal (X_j'') present in the reduced dictionary,
- t) if a unique nucleotide sequence comprises more than one standardized pyro-signal (X_j'') within the dictionary, then summing the estimated regression
30 coefficients ($\hat{\beta}_j''$) of all the standardized pyro-sequencing signals (X_j'') corresponding to this unique nucleotide sequence for estimating the contribution of this unique nucleotide sequence.

Determining the unique nucleotide sequences whose contributions are above a determined Significant Contribution Threshold may be improved by iteratively eliminating at least one pyro-sequencing signal from the dictionary, and iteratively build reduced pyro-sequencing signals dictionaries to be used for building subsequent penalized multiple regression models, until each unique nucleotide sequences in the reduced dictionary has a contribution above the Significant Contribution Threshold. Accordingly, the step of determining the unique nucleotide sequences whose contributions are above a determined Significant Contribution Threshold comprises the steps ei), eii) and eiii) until each unique nucleotide sequence of the reduced dictionary has a contributions above the determined Significant Contribution Threshold.

[0057] A Significant Contribution Threshold is a defined threshold depending on the specificity of the product. In a preferred embodiment, the Significant Contribution is defined as a hyper parameter of the model. Like the determination of the L_1 and L_2 norms, it may be determined with a validation dataset. Different values for the Significant Contribution Threshold are generated, and for each one, all pyrosequencing signals from the validation dataset with known contributions are analysed with the method. The value that permits to obtain the highest rate of correct identification is selected as the Significant Contribution Threshold of the model.

[0058] The additional penalized multiple regression model used may be the same as the model used before in the method. L_1 norm and L_2 norm may be the same as the norms used in the previous step. L_1 and L_2 norms may also be different. In that case, a determination of their values may be defined with a validation dataset, as explained before.

[0059] In a preferred embodiment, it may be possible to generate a confidence factor of the analysis. Accordingly, a step of computing a correlation coefficient between the tested pyrosequencing signal (Y) and the fitted values ($\hat{Y}$) of the regression model may be performed after any step of estimation of the contribution in the tested pyro-sequencing signal (Y) of each unique nucleotide sequence of each genomic region of interest.

[0060] The fitted values of the regression model ($\hat{Y}$) are defined as the linear combination of the standardized pyrosequencing signals (X_i) present in a

dictionary in which each signal (X_j) is multiplied by the estimated regression coefficient ($\hat{\beta}_j$).

[0061] In a complementary embodiment, steps for constructing a dictionary of standardized pyro-sequencing signals for its use in any one of the previous methods may be present.

[0062] Accordingly, A method for constructing a dictionary of standardized pyro-sequencing signals for its use in any method according to the invention for analyzing a pyrosequencing signal (Y) comprising the steps of:

A) Obtaining a list of all unique nucleotide sequences expected to be found in an at least one genomic region of interest,

B) Generating at least one pyro-sequencing signal for each unique nucleotide sequence of each genomic region of interest, all pyro-sequencing signals being generated with an identical nucleotide dispensation order,

C) Standardizing each pyro-sequencing signal relatively to an height of at least one peak present in the pyro-sequencing signal, the at least one peak height being representative of the incorporation of a given number of nucleotide, this given number being identical for every pyro-sequencing signal of the dictionary,

D) Generating a standardized pyro-sequencing signal dictionary by associating to each unique nucleotide sequence expected to be found in an at least one genomic region of interest at least one corresponding standardized pyro-sequencing signal.

[0063] A list of all unique nucleotide sequences expected to be found in an at least one genomic region of interest may be obtained from genomic databases known from the man skilled in the art like NCBI for example.

[0064] The step of generating at least one pyro-sequencing signal for each unique nucleotide sequence of each genomic region of interest may be performed by generating a theoretical pyro-sequencing signal, i.e. estimating for a given nucleotide dispensation order each peak height in the pyro-sequencing signal.

[0065] In a preferred embodiment, the step for generating at least one pyro-sequencing signal is performed by a simplex pyro-sequencing of a genetic amplification product comprising a genomic region of interest that contains a single unique nucleotide sequence. In this preferred embodiment, the pyro-

sequencing signal is generated from a pure biological sample. Hence, the pyro-sequencing signal is an experimental signal instead of a theoretical signal, i.e. closer to the reality because theoretical signal does not always fit to the experimental one.

- 5 **[0066]** A standardizing step is present in order to normalize each signal relatively to an height of at least one peak corresponding to a given number of nucleotide incorporation (i.e. 1 incorporation, 2, 3, 4 or 5 incorporations) in any number of peak representing a nucleotide incorporation during a dispensation step (one peak may represent the incorporation of a single nucleotide, the
10 incorporation of two, three, four or five identical nucleotides).

[0067] A step of associating each pyro-sequencing signal to each specific sequence is finally performed.

[0068] In an alternative embodiment, a dictionary of standardized pyro-sequencing signals is provided wherein the dictionary is obtained by:

- 15 A) Obtaining a list of all unique nucleotide sequences expected to be found in an at least one genomic region of interest,
B) Generating at least one pyro-sequencing signal for each unique nucleotide sequence, all pyro-sequencing signals being generated with an identical nucleotide dispensation order,
20 C) Standardizing each pyro-sequencing signal relatively to an height of at least one peak present in the pyro-sequencing signal, the at least one peak height being representative of the incorporation of a given number of nucleotide, this given number being identical for every pyro-sequencing signal of the dictionary,
D) Generating a standardized pyro-sequencing signal dictionary by associating
25 to each unique nucleotide sequence expected to be found in an at least one genomic region of interest at least one corresponding standardized pyro-sequencing signal.

- [0069]** In a more preferred embodiment, the provided dictionary is obtained by performing the subsequent step of generating at least one pyro-sequencing signal for each unique nucleotide sequence is performed by a
30 simplex pyro-sequencing of a genetic amplification product comprising a genomic region of interest that contains a single unique nucleotide sequence.

[0070] It is also an object of the invention to provide a nucleotide dispensation order which enables a more accurate method for analysing a pyro-sequencing signal (Y), and for constructing a dictionary of standardized pyro-sequencing signals (X_j) to be used with the method. The dispensation order is chosen to minimize the correlations between pyrosequencing signals generated for each distinct unique nucleotide sequence that can be potentially be found in each genomic region of interest.

[0071] Accordingly, the dictionary and/or the pyro-sequencing reaction are generated with a pre-determined nucleotide dispensation order determined by:

- Obtaining a list of all unique nucleotide sequences expected to be found in an at least one genomic region of interest,
- Generating a plurality of nucleotide dispensation orders susceptible to be used during a pyro-sequencing of the at least one genomic region of interest
- Generating a theoretical pyro-sequencing signal for each unique nucleotide sequence of the list, said pyro-sequencing signals being generated with each nucleotide dispensation order contained in the plurality of nucleotides dispensation orders,
- Determining within the plurality of nucleotide dispensation orders, the one that produces the less correlated pyro-sequencing signals for the nucleotide sequences expected to be found in the at least one genomic region of interest.

[0072] A list of all unique nucleotide sequences expected to be found in an at least one genomic region of interest may be obtained from genomic databases known from the man skilled in the art like NCBI for example.

[0073] The generation of a plurality of nucleotide dispensation order may be performed by a random process.

[0074] The determination of the nucleotide dispensation order which produces the less correlated pyro-sequencing signals may be performed by calculating a correlation coefficient between each pyrosequencing signal pair generated. The purpose of the nucleotide dispensation order is to generate highly different pyrosequencing signal for each unique nucleotide sequence that can be found in an at least one genomic region of interest.

[0075] In a more preferred embodiment, the step of generating the standardized pyrosequencing signals (X_j) of the dictionary or a step of generating a tested pyrosequencing signal (Y) may comprise an additional step wherein determining, within the plurality of nucleotide dispensation orders, the order that produces the less correlated pyro-sequencing signals for the nucleotide sequences expected to be found in the at least one genomic region of interest, comprises the steps of:

- Calculating for each nucleotide dispensation order, a vector that includes the correlation coefficients between each possible pair of pyro-sequencing signals generated with this nucleotide dispensation order,
- Computing for each nucleotide dispensation order the maximum value of the vector that includes the correlation coefficients,
- Selecting the nucleotide dispensation order with the lowest maximum value.

[0076] In a more preferred embodiment, the step of computing for each nucleotide dispensation order the maximum value of the vector also comprises a step of computing the average value of the vector that includes the correlation coefficients. Hence, when two dispensation orders have the same maximum value of the vector, then the one with the lowest average value is selected.

[0077] In a more preferred embodiment, the correlation coefficients between each possible pair of pyro-sequencing signals are calculated by a Pearson product-moment correlation.

Examples

[0078] Example 1a: Determination of a dispensation order to identify the presence of mutated alleles in the exon 20 of the EGFR oncogenic target within human tumour cells from a sample containing a mixture of tumour and background cells

[0079] Two nucleotide sequences can potentially be found for the genomic region of interest which is the exon 20 of the EGFR oncogenic target.

The wild-type nucleotide sequence corresponds to the nucleotide sequence 'CGCAGC' while the mutant nucleotide sequence corresponds to the nucleotide sequence 'TGCAGC'.

[0080] Various dispensation orders are first generated. A Pearson correlation coefficient between each pyrosequencing signal pair is computed for each dispensation order. Because only two unique nucleotide sequences can be found in the genomic region of interest, a single and unique correlation coefficient is computed. Then, the nucleotides dispensation orders are sorted: the nucleotide dispensation order with the lowest Pearson correlation coefficient is selected and used during the pyrosequencing reaction of the exon 20 of the EGFR oncogenic target. As we can see in **fig.1**, the selected nucleotide dispensation order allows generating two highly uncorrelated ($R=-0.79$) pyrosequencing signals, with 8 different peaks in 9 nucleotides dispensation steps while the nucleotide dispensation order of the prior art generates correlated ($R=0.5$) pyrosequencing signals with only 2 different peaks.

[0081] Although the illustrated case has only two different nucleotide sequences in the genomic region of interest, the method can be applied with samples that can potentially include a high number of different nucleotide sequences (with no theoretical limits).

[0082] Example 1b: Determination of a dispensation order to identify the presence of a mutation in the NRAS oncogenic target within human tumour cells in a sample containing mixed tumour (i.e., the target) and background (i.e., irrelevant) cells

[0083] Ten distinct nucleotide sequences can potentially be found for the genomic region of interest which is the NRAS oncogenic target.

[0084] Various dispensation orders are first generated. A Pearson correlation coefficient between each pyrosequencing signal pair is computed for each dispensation order. According to these ten distinct unique nucleotide sequences of the NRAS genomic region, a vector that includes the 45 correlation coefficients between each possible pair of pyrosequencing signals is computed for each dispensation order. The maximum value of this vector is then computed for each dispensation order. The dispensation order with the lowest maximum value is selected. As several nucleotide dispensation orders

produce the same lowest maximum value of the vector, the nucleotide dispensation order with the lowest average value of the vector is finally selected. As we can see in **fig.2**, the selected nucleotide dispensation order allows generating ten highly uncorrelated pyrosequencing signals.

- 5 The vector of the correlation coefficients generated with this selected dispensation order is defined by:

(-0.34 ; -0.50 ; -0.14 ; 0.31 ; -0.04 ; 0.44 ; -0.40 ; -0.53 ; 0.54 ; 0.11 ; -0.63 ; 0.61 ; 0.48 ; 0.20 ; -0.04 ; -0.05 ; -0.53 ; -0.03 ; -0.19 ; 0.55 ; -0.29 ; 0.58 ; -0.34 ; 0.19 ; 0.66 ; -0.05 ; -0.34 ; -0.40 ; -0.23 ; -0.53 ; 0.35 ; 0.11 ; 0.85 ; -0.04 ; 0.85 ; -0.23 ; 0.79 ; -0.34 ; -0.27 ; 0.31 ; -0.23 ; -0.63 ; -0.40 ; 0.79 ; -0.40)

The maximum value of this vector is equal to 0.85 while the average is equal to 0.012.

[0085] Example 1c: Determination of a dispensation order to distinguish Monkeypox from Camelpox viruses

- 15 **[0086]** The selected genomic region of interest enables to distinguish the Monkeypox and Camelpox CMS viruses. In this genomic region, the Monkeypox virus is indeed characterized by the unique nucleotide sequence 'TATTAGTAA' while the Camelpox CMS virus is characterized by the unique nucleotide sequence 'CATTAGTAA'.

- 20 **[0087]** Various dispensation orders are first generated. A Pearson correlation coefficient between each pyrosequencing signal pair is computed for each dispensation order. Because only two unique nucleotide sequences can be found in the genomic region of interest, a single and unique correlation coefficient is computed. Nucleotide dispensation orders are then sorted: the nucleotide dispensation order with the lowest Pearson correlation coefficient is selected and used during the pyrosequencing reaction. As we can see in **fig.3** the selected nucleotide dispensation order allows generating two highly uncorrelated ($R=-0.71$) pyrosequencing signals, with 8 different peaks in 10 nucleotides dispensation steps while the nucleotide dispensation order of the prior art generates correlated ($R=0.76$) pyrosequencing signals with only 2 different peaks.

[0088] Example 2: Construction of a dictionary of standardized pyrosequencing signals

[0089] First, a nucleotide dispensation order is determined according to the method described in the previous example. Then, a theoretical pyrosequencing signal is generated with the nucleotide dispensation order previously selected. In this example, the genomic region of interest is the upstream-p34region of *mycobacterium*. 33 different unique nucleotide sequences are known and susceptible to be found in the genomic region of interest. Accordingly, for each unique nucleotide sequence, at least one experimental pyrosequencing signal is generated. An example of a single experimental pyrosequencing signal is illustrated on **fig. 4(a)**. 99 experimental pyrosequencing signals are generated. In **fig. 4(b)**, only 9 pyrosequencing signals corresponding to 9 distinct nucleotide sequences are illustrated for clarity. Each pyrosequencing signal is standardized: each peak of the signal is divided by a single peak height that corresponds to a single nucleotide incorporation. After standardization, all pyrosequencing signals are associated to their unique nucleotide sequence inside a dictionary, illustrated on **fig. 4(b)**.

[0090] Although the illustrated case is performed by generating an experimental pyrosequencing signal, it may be possible to construct a dictionary by generating a theoretical pyrosequencing signal of at least one unique nucleotide sequence.

[0091] Example 3a: Analysis of a low pyrosequencing signal generated from a single genomic region of interest generated from a homogeneous sample containing a pure population of the same microorganisms:

[0092] The genomic region of interest is the upstream-p34 region of the *mycobacterium* genome. After selecting a dispensation order and creating a dictionary of standardized pyrosequencing signals (which is the same as the dictionary created in the second example), a pyrosequencing reaction is performed on a genetic amplification product and a tested pyrosequencing signal is generated. The pyrosequencing signal Y, of length N=26, is analysed by the method.

[0093] In a first step, the method enables to compute the contribution of each unique nucleotide sequence by summing the estimated regression coefficients ($\hat{\beta}_j$) of all the standardized pyro-sequencing signals (X_j)

corresponding to this unique nucleotide sequence. This first step produces the following contributions:

unique nucleotide sequence	contribution	unique nucleotide sequence	contribution	unique nucleotide sequence	contribution
unique nucleotide sequence 1	0.011	unique nucleotide sequence 12	0.000	unique nucleotide sequence 23	0.000
unique nucleotide sequence 2	0.000	unique nucleotide sequence 13	0.000	unique nucleotide sequence 24	0.000
unique nucleotide sequence 3	0.000	unique nucleotide sequence 14	0.000	unique nucleotide sequence 25	0.000
unique nucleotide sequence 4	0.000	unique nucleotide sequence 15	0.000	unique nucleotide sequence 26	0.000
unique nucleotide sequence 5	0.000	unique nucleotide sequence 16	0.000	unique nucleotide sequence 27	0.000
unique nucleotide sequence 6	0.000	unique nucleotide sequence 17	0.000	unique nucleotide sequence 28	0.000
unique nucleotide sequence 7	0.000	unique nucleotide sequence 18	0.000	unique nucleotide sequence 29	0.000
unique nucleotide sequence 8	0.001	unique nucleotide sequence 19	0.000	unique nucleotide sequence 30	0.000
unique nucleotide sequence 9	0.000	unique nucleotide sequence 20	0.000	unique nucleotide sequence 31	0.000
unique nucleotide sequence 10	0.000	unique nucleotide sequence 21	0.001	unique nucleotide sequence 32	1.620
unique nucleotide sequence 11	0.000	unique nucleotide sequence 22	0.000	unique nucleotide sequence 33	0.000

5 The Significant Contribution Threshold has been selected from a validation dataset and is equal to 2. As the contributions of all unique nucleotide sequences are below the Significant Contribution Threshold, a reduced dictionary is constructed by selecting only the most contributing Unique Nucleotide Sequence (in this case, the unique nucleotide sequence 32).

10 In a subsequent step, the reduced dictionary is used and, consecutively, a single unique nucleotide sequence contribution is correctly identified from the pyrosequencing signal despite its very low signal-to-noise ratio (see **fig. 5**).

[0094] A very high correlation coefficient ($r=1$) is computed between the fitted values of the model ($\hat{Y}$) and the pyrosequencing signal Y , evidencing a
15 highly reliable result (see **fig. 5**).

[0095] Example 3b: Analysis of a pyrosequencing signal generated from a single genomic region of interest generated from an heterogeneous sample containing a mixed population of distinct microorganisms:

[0096] The genomic region of interest is the upstream-p34 region of the
20 *mycobacterium* genome. After selecting a dispensation order and creating a dictionary of standardized pyrosequencing signals (which is the same as the dictionary created in the second example), a pyrosequencing reaction is performed on a genetic amplification product and a tested pyrosequencing

signal is generated. The pyrosequencing signal Y , of length $N=26$, is analysed by the method.

[0097] In a first step, the method enables to compute the contribution of each unique nucleotide sequence by summing the estimated regression coefficients ($\hat{\beta}_j$) of all the standardized pyro-sequencing signals (X_j) corresponding to this unique nucleotide sequence. This first step produces the following contributions:

Unique Nucleotide Sequence	contribution	amplicons	contribution	amplicons	contribution
Unique Nucleotide Sequence 1	0.099	Unique Nucleotide Sequence 12	0.000	Unique Nucleotide Sequence 23	0.000
Unique Nucleotide Sequence 2	0.000	Unique Nucleotide Sequence 13	0.000	Unique Nucleotide Sequence 24	0.000
Unique Nucleotide Sequence 3	0.000	Unique Nucleotide Sequence 14	0.000	Unique Nucleotide Sequence 25	0.000
Unique Nucleotide Sequence 4	0.000	Unique Nucleotide Sequence 15	0.000	Unique Nucleotide Sequence 26	0.000
Unique Nucleotide Sequence 5	0.000	Unique Nucleotide Sequence 16	0.000	Unique Nucleotide Sequence 27	0.000
Unique Nucleotide Sequence 6	0.000	Unique Nucleotide Sequence 17	0.000	Unique Nucleotide Sequence 28	0.000
Unique Nucleotide Sequence 7	0.010	Unique Nucleotide Sequence 18	15.884	Unique Nucleotide Sequence 29	0.000
Unique Nucleotide Sequence 8	0.108	Unique Nucleotide Sequence 19	22.435	Unique Nucleotide Sequence 30	0.000
Unique Nucleotide Sequence 9	0.000	Unique Nucleotide Sequence 20	0.000	Unique Nucleotide Sequence 31	0.000
Unique Nucleotide Sequence 10	0.000	Unique Nucleotide Sequence 21	0.000	Unique Nucleotide Sequence 32	0.000
Unique Nucleotide Sequence 11	0.000	Unique Nucleotide Sequence 22	0.257	Unique Nucleotide Sequence 33	0.000

The Significant Contribution Threshold has been selected from a validation dataset and is equal to 2. The contributions of 2 unique nucleotide sequences (18 and 19) are above the Significant Contribution Threshold. A reduced dictionary is constructed by selecting all pyrosequencing signals corresponding to both unique nucleotide sequences.

In a subsequent step, the reduced dictionary is used and, consecutively, two unique nucleotide sequence contributions are correctly identified from the pyrosequencing signal (see **fig. 6**).

[0098] A very high correlation coefficient ($r = 0.9988$) is computed between the fitted values of the model ($\hat{Y}$) and the pyrosequencing signal Y , traducing a highly reliable result (see **fig. 6**).

[0099] Example 3c: Analysis of a pyrosequencing signal generated from a single genomic region of interest generated from a heterogeneous human sample containing human tumour cells (i.e., the target cells) and background (i.e., irrelevant) cells

[0100] The genomic region of interest is the NRAS oncogenic target. After selecting a dispensation order (which is the same as the order selected in example 1b) and creating a dictionary of standardized pyrosequencing signals, a pyrosequencing reaction is performed on the amplification product and a tested pyrosequencing signal is generated. The pyrosequencing signal Y, of length N=11, is analysed by the method.

[0101] In a first step, the method enables to compute the contribution of each unique nucleotide sequence by summing the estimated regression coefficients ($\hat{\beta}_j$) of all the standardized pyro-sequencing signals (X_j) corresponding to this unique nucleotide sequence. This first step produces the following contributions:

Unique Nucleotide Sequence	contribution	Unique Nucleotide Sequence	contribution	Unique Nucleotide Sequence	contribution
Unique Nucleotide Sequence 1	2.483	Unique Nucleotide Sequence 5	0.393	Unique Nucleotide Sequence 8	0.000
Unique Nucleotide Sequence 2	0.000	Unique Nucleotide Sequence 6	0.026	Unique Nucleotide Sequence 9	0.166
Unique Nucleotide Sequence 3	0.000	Unique Nucleotide Sequence 7	0.000	Unique Nucleotide Sequence 10	0.035
Unique Nucleotide Sequence 4	0.000				

The Significant Contribution Threshold has been selected from a validation dataset and is equal to 0.2. The contributions of 2 unique nucleotide sequences (1 and 5) are above the Significant Contribution Threshold. A reduced dictionary is constructed by selecting all pyrosequencing signals corresponding to both unique nucleotide sequences.

In a second step, the reduced dictionary is used, and subsequently two unique nucleotide sequence contributions (2.69 for unique nucleotide sequence 1 and 0.34 for unique nucleotide sequence 5) are finally correctly identified from the pyrosequencing signal (see **fig. 7**). As the unique nucleotide sequence 1 corresponds to the Wild-type allele while unique nucleotide sequence 5 corresponds to a mutated allele, the percentage of mutated allele is estimated at 11.2 %.

[0102] A very high correlation coefficient ($r = 0.999$) is computed between the fitted values of the model ($\hat{Y}$) and the pyrosequencing signal Y, traducing a highly reliable result (see **fig. 7**).

[0103] Example 3d: Analysis of a pyrosequencing signal generated from multiple genomic regions of interest generated from homogenous sample containing a pure microbial population

[0104] The 5 genomic regions of interest belong to the SHV and TEM gene from bacteria genomes (SHV.179, SHV.238, TEM.104, TEM164, TEM238). After selecting a dispensation order and creating a dictionary of standardized pyrosequencing signals, a pyrosequencing reaction is performed on the amplification product and a tested pyrosequencing signal is generated. The pyrosequencing signal Y, of length N=14, is analysed by the method.

[0105] In a first step, the method enables to compute the contribution of each unique nucleotide sequence by summing the estimated regression coefficients ($\hat{\beta}_j$) of all the standardized pyro-sequencing signals (X_j) corresponding to this unique nucleotide sequence. This first step produces the following contributions:

Genomic Region	contribution	Genomic Region	contribution	Genomic Region	contribution	Genomic Region	contribution	Genomic Region	contribution
SHV179		SHV238		TEM104		TEM164		TEM238	
SHV179.1	1.545	SHV238.1	0.944	TEM104.1	0.836	TEM164.1	1.951	TEM238.1	0.614
		SHV238.2	3.279	TEM104.2	0.664	TEM164.2	0.000	TEM238.2	0.000
		SHV238.3	0.000					TEM238.3	0.000
		SHV238.4	0.000						

As the amplification product was produced from a pure microbial population consisting of a single type of bacteria, a single unique nucleotide sequence is expected to be found in each genomic region of interest.

A reduced dictionary is therefore built by selecting, for each genomic region of interest, the most contributing unique nucleotide sequence.

In a subsequent step, the reduced dictionary is used and, consecutively, five unique nucleotide sequence contributions are correctly identified from the pyrosequencing signal (see **fig. 8**).

[0106] A very high correlation coefficient ($r = 0.995$) is computed between the fitted values of the model ($\hat{Y}$) and the pyrosequencing signal Y, traducing a highly reliable result (see **fig. 8**).

[0107] Example 3e: Analysis of a pyrosequencing signal generated from multiple genomic regions of interest generated from homogenous sample containing a quasi-pure population of human cells of interest

[0108] Four genomic regions of interest, representing four different SNPs that are associated with prostate cancer, are selected: rs1016343 (Chromosome 8), rs10896449 (Chromosome 11), rs5945572 (Chromosome X), rs5945619 (Chromosome X). After selecting a dispensation order and creating
 5 a dictionary of standardized pyrosequencing signals, a pyrosequencing reaction is performed on the amplification product and a tested pyrosequencing signal is generated. The pyrosequencing signal Y, of length N=14, is analysed by the method.

[0109] In a first step, the method enables to compute the contribution of
 10 each unique nucleotide sequence (i.e. each genotype of each SNPs) by summing the estimated regression coefficients ($\hat{\beta}_j$) of all the standardized pyro-sequencing signals (X_j) corresponding to this unique nucleotide sequence. This first step produces the following contributions:

Genomic region SNP rs1016343	contribution	Genomic region SNP rs10896449	Genomic region SNP rs5945572	Genomic region SNP rs5945619
rs1016343 - C/C	3.326	rs10896449 - A/A	rs5945572 - A	rs5945619 - C
rs1016343 - T/T	2.879	rs10896449 - G/G	rs5945572 - G	rs5945619 - G
rs1016343 - C/T	2.796	rs10896449 - A/G		

15 As the amplification product was produced from a blood sample from a single patient, a single unique nucleotide sequence is expected to be found in each genomic region of interest. SNPs rs1016343 and rs10896449 present two homozygous variants (C/C and T/T for rs1016343; A/A and G/G for
 20 rs10896449) and one heterozygous variant (C/T for rs1016343; A/G for rs10896449). SNPs on the Chromosome X present only two genotype variants because the patient is a man.

For SNPs rs1016343 and rs10896449, the minimum contribution of homozygous variants is identified for each SNP. For each SNP, the minimum
 25 value is then added twice to the contribution of the heterozygous variant of the SNP and subtracted from the contribution of homozygous variants of the SNP. This step produces the following contributions:

Genomic region SNP rs1016343		Genomic region SNP rs10896449		Genomic region SNP rs5945572		Genomic region SNP rs5945619	
rs1016343 - C/C	0.448	rs10896449 - A/A	7.516	rs5945572 - A	10.076	rs5945619 - C	11.803
rs1016343 - T/T	0.000	rs10896449 - G/G	0.000	rs5945572 - G	0.000	rs5945619 - G	0.000
rs1016343 - C/T	8.554	rs10896449 - A/G	0.000				

A reduced dictionary is built by selecting the most contributing unique nucleotide sequence for each genomic region of interest.

In a subsequent step, the reduced dictionary is used and, consequently, four
 5 unique nucleotide sequence contributions are finally correctly identified from the pyrosequencing signal (see **fig. 9**).

[0110] A very high correlation coefficient ($r = 0.997$) is computed between the fitted values of the model ($\hat{Y}$) and the pyrosequencing signal Y , traducing a highly reliable result (see **fig. 9**).

10 **[0111]** The present invention has been described in terms of specific embodiments, which are illustrative of the invention and not to be construed as limiting. More generally, it will be appreciated by persons skilled in the art that the present invention is not limited by what has been particularly shown and/or described hereinabove.

15 Reference numerals in the claims do not limit their protective scope.

Use of the verbs "to comprise", "to include", "to be composed of", or any other variant, as well as their respective conjugations, does not exclude the presence of elements other than those stated.

20 Use of the article "a", "an" or "the" preceding an element does not exclude the presence of a plurality of such elements.

[0112] The invention may also be described as follows:

A method for analyzing a pyrosequencing signal that enables an automated, fast and reliable analyzes of combined pyrosequencing signals from homogeneous or heterogeneous samples based on an optimized order of
 25 dispensation. The method allows analyzing a pyrosequencing signal even when signal intensities are low or when there is more than one signal contribution. The method comprises a step of determining the respective contribution of standardized pyrosequencing signals in the analyzed pyrosequencing signal, the standardized pyrosequencing signals being comprised within a dictionary.

Claims

1. A method for analyzing a pyro-sequencing signal (Y) comprising the steps of:

- 5 a) Providing a simplex or multiplex genetic amplification product comprising at least one genomic region of interest generated from a DNA or RNA sample,
- b) Acquiring one tested pyro-sequencing signal (Y) by the pyro-sequencing reaction of said product,
- c) Providing a dictionary of standardized pyro-sequencing signals (X_j),
10 comprising at least one standardized pyro-sequencing signal (X_j) for each unique nucleotide sequence that can potentially be found within each genomic region of interest and wherein the nucleotide dispensation order used for generating the standardized pyro-sequencing signals (X_j) contained within the dictionary and for generating the tested pyro-sequencing signal (Y) is the same,
- 15 d) Building a first penalized multiple regression model having an intercept set to zero, wherein the tested pyro-sequencing signal (Y) is a response variable and all standardized pyro-sequencing signals (X_j) contained within the dictionary are predictor variables, wherein a non-negativity constraint is applied on the regression coefficients (β_j), and wherein a L_1 -norm and/or a L_2 -norm penalty is
20 applied for estimating the regression coefficients (β_j) of the model,
- e) Using the estimated regression coefficients ($\hat{\beta}_j$) of the penalized multiple regression model in order to estimate the respective contribution in the tested pyro-sequencing signal (Y) of each pyro-sequencing signal (X_j) contained in the standardized dictionary,
- 25 f) if a unique nucleotide sequence comprises more than one standardized pyro-signal (X_j) within the dictionary, then summing the estimated regression coefficients ($\hat{\beta}_j$) of all the standardized pyro-sequencing signals (X_j) corresponding to this unique nucleotide sequence for estimating the contribution of this unique nucleotide sequence.

30

2. The method of claim 1, wherein the further steps of:

- g) Determining for each genomic region of interest, the unique nucleotide sequence with the highest contribution in the tested pyro-sequencing signal (Y),
- h) Creating a reduced dictionary by selecting for each genomic region of interest the standardized pyro-sequencing signals (X_j) corresponding to the
- 5 determined unique nucleotide sequence,
- i) Building a second penalized multiple regression model having an intercept set to zero, wherein the tested pyro-sequencing signal (Y) is a response variable and all standardized pyro-sequencing signals (X_j) contained within the reduced dictionary are predictor variables, wherein a non-negativity constraint is applied
- 10 on the regression coefficients (β_j), and wherein a L_1 -norm and/or a L_2 -norm penalty is applied for estimating the regression coefficients (β_j) of the second model,
- j) Using the estimated regression coefficients ($\hat{\beta}_j$) of the penalized multiple regression model in order to estimate the respective contribution in the resulting
- 15 tested pyro-sequencing signal (Y) of each pyro-sequencing signal (X_j) present in the reduced dictionary,
- k) if a unique nucleotide sequence comprises more than one standardized pyro-signal (X_j) within the dictionary, then summing the estimated regression coefficients ($\hat{\beta}_j$) of all the standardized pyro-sequencing signals (X_j)
- 20 corresponding to this unique nucleotide sequence for estimating the contribution of this unique nucleotide sequence,
- are performed after the step of estimation of contribution of each unique nucleotide sequence of each genomic region of interest.
- 25 3. The method according to claim 2, wherein the step of selecting for each genomic region of interest the unique nucleotide sequence with the highest contribution in the tested pyro-sequencing signal (Y) comprises the further steps of:
- l) Identifying the minimum contribution of both homozygous variants of the
- 30 genomic region of interest,
- m) Adding twice the estimated minimum value to the contribution of the heterozygous variant of the genomic region of interest,

n) Subtracting the estimated minimum value from the contribution of both homozygous variants of the genomic region of interest,

o) Determining the unique nucleotide sequence in the genomic region of interest with the estimated highest contribution in the tested pyro-sequencing signal (Y),

are performed when the simplex or multiplex genetic amplification product is generated from DNA or RNA extracted from a diploid cell and the genomic region of interest presents two known homozygous variants and one known heterozygous variant.

4. The Method according to any one of claims 1 to 3, wherein the further steps of:

p) Determining the unique nucleotide sequences whose contributions are above a determined Significant Contribution Threshold, and when no contribution is above the Significant Contribution Threshold, determining the most contributing unique nucleotide sequence,

q) Selecting the standardized pyro-sequencing signals (X') corresponding to the determined unique nucleotide sequences from the standardized dictionary, in order to create a reduced dictionary,

r) Using an additional penalized multiple regression model having an intercept set to zero, wherein the tested pyro-sequencing signal (Y) is a response variable and all standardized pyro-sequencing signals (X_j'') contained within the reduced dictionary are predictor variables, wherein a non-negativity constraint is applied on the regression coefficients (β_j'') and wherein a L_1 -norm and/or a L_2 -norm penalty is applied for estimating the regression coefficients (β_j'') of the additional model,

s) Using the estimated regression coefficients ($\hat{\beta}_j''$) of the penalized multiple regression model in order to estimate the respective contribution in the resulting tested pyro-sequencing signal (Y) of each pyro-sequencing signal (X_j'') present in the reduced dictionary,

t) if a unique nucleotide sequence comprises more than one standardized pyro-signal (X_j'') within the dictionary, then summing the estimated regression coefficients ($\hat{\beta}_j''$) of all the standardized pyro-sequencing signals (X_j'')

corresponding to this unique nucleotide sequence for estimating the contribution of this unique nucleotide sequence, are performed.

5 5. The method of claim 2 or 4, wherein the step of selecting for each genomic region of interest the unique nucleotide sequence with the highest contribution in the tested pyro-sequencing signal (Y) comprises the further steps of:

10 ei) determining for at least one genomic region of interest the standardized pyro-sequencing signal (Xi) corresponding to the less contributive signal in the tested pyro-sequencing signal (Y),

eii) removing the corresponding pyro-sequencing signal (Xi) from the dictionary, thereby creating a reduced dictionary,

15 eiii) building a new additional penalized multiple regression model with the reduced dictionary created at the previous step,

are performed iteratively after each step of estimation of the respective contribution in the tested pyro-sequencing signal (Y) until a unique nucleotide sequence is obtained for each genomic region of interest.

20 6. The method of claim 5, wherein the steps ei), eii) and eiii) are applied on each genomic region of interest, by starting with the one which is associated with the highest contribution and iteratively in descending order of contribution.

25 7. The method of claim 4, wherein the step of determining the unique nucleotide sequences whose contributions are above the determined Significant Contribution Threshold comprises,

ei) determining for at least one genomic region of interest the standardized pyro-sequencing signal (Xi) corresponding to the less contributive signal in the tested pyro-sequencing signal (Y),

30 eii) removing the corresponding pyro-sequencing signal (Xi) from the dictionary, thereby creating a reduced dictionary,

eiii) building a new additional penalized multiple regression model with the reduced dictionary created at the previous step,

until each unique nucleotide sequence of the reduced dictionary has a contributions above the determined Significant Contribution Threshold.

8. The method of any one of claims 1 to 7, wherein a step of computing a correlation coefficient between the tested pyro-sequencing signal (Y) and a fitted values ($\hat{Y}$) of the regression model is performed after any step of estimation of the contribution in the tested pyro-sequencing signal (Y) of each unique nucleotide sequence of each genomic region of interest.

9. The method according to any one of claim 1 to 8, wherein the built penalized multiple regression model is defined by:

$$Y = \sum_{j=1}^p X_j \beta_j + \varepsilon$$

and the estimation of the regression coefficients in the penalized multiple regression model is defined by:

$$\hat{\beta} = \underset{\beta}{\operatorname{argmin}} \left\{ \sum_{i=1}^N \left(y_i - \sum_{j=1}^p x_{ij} \beta_j \right)^2 + \lambda_1 \|\beta_j\|_1 + \lambda_2 \|\beta_j\|_2 \right\}$$

or

$$\hat{\beta} = \underset{\beta}{\operatorname{argmin}} \left\{ \sum_{i=1}^n (y_i - \sum_{j=1}^p \beta_j x_{ij})^2 + \lambda_{1j} \|\beta_j\|_1 + \lambda_{2j} \|\beta_j\|_2 \right\}$$

when different penalties are used on the regression coefficients,

where Y is the pyrosequencing signal, y_i is the i^{th} element of the Y vector, X_j is the j^{th} pyrosequencing signal of the dictionary of standardized pyro-sequencing signal that corresponds to a unique nucleotide sequence, x_{ij} is the i^{th} element of the j^{th} pyrosequencing signal, each β_j is a regression coefficient associated to the X_j pyro-sequencing signal that corresponds to a unique nucleotide sequence, $\|\beta_j\|_1$ is the L_1 – norm of the vector of the regression coefficients β_j , $\|\beta_j\|_2$ is the L_2 – norm of the vector of the regression coefficients β_j , λ_1 is penalty factor on the L_1 -norm of the regression coefficients (β_j), λ_2 is penalty factor on the L_2 -norm of the regression coefficients (β_j), ε is a vector of random error with a normal distribution.

10. A method for constructing a dictionary of standardized pyro-sequencing signals for its use in the method of claims 1 to 9, comprising the steps of:
 - A) Obtaining a list of all unique nucleotide sequences expected to be found in an at least one genomic region of interest,
 - B) Generating at least one pyro-sequencing signal for each unique nucleotide sequence of each genomic region of interest, all pyro-sequencing signals being generated with an identical nucleotide dispensation order,
 - C) Standardizing each pyro-sequencing signal relatively to an height of at least one peak present in the pyro-sequencing signal, the at least one peak height being representative of the incorporation of a given number of nucleotide, this given number being identical for every pyro-sequencing signal of the dictionary,

D) Generating a standardized pyro-sequencing signal dictionary by associating to each unique nucleotide sequence expected to be found in an at least one genomic region of interest at least one corresponding standardized pyro-sequencing signal.

5

11. The method according to claim 10 wherein the step of generating at least one pyro-sequencing signal for each unique nucleotide sequence is performed by a simplex pyro-sequencing of a genetic amplification product comprising a genomic region of interest that contains a single unique nucleotide sequence.

10

12. A dictionary of standardized pyro-sequencing signals (X_j) wherein the dictionary is obtained by:

A) Obtaining a list of all unique nucleotide sequences expected to be found in an at least one genomic region of interest,

15

B) Generating at least one pyro-sequencing signal for each unique nucleotide sequence, all pyro-sequencing signals being generated with an identical nucleotide dispensation order,

C) Standardizing each pyro-sequencing signal relatively to an height of at least one peak present in the pyro-sequencing signal, the at least one peak height being representative of the incorporation of a given number of nucleotide, this given number being identical for every pyro-sequencing signal of the dictionary,

20

D) Generating a standardized pyro-sequencing signal dictionary by associating to each unique nucleotide sequence expected to be found in an at least one genomic region of interest at least one corresponding standardized pyro-sequencing signal.

25

13. The dictionary according to claim 12 wherein the step of generating at least one pyro-sequencing signal for each unique nucleotide sequence is performed by a simplex pyro-sequencing of a genetic amplification product comprising a genomic region of interest that contains a single unique nucleotide sequence.

30

14. The dictionary according to claim 12 or 13, wherein the standardized pyro-sequencing signals are generated with a nucleotide dispensation order determined by:

- Obtaining a list of all unique nucleotide sequences expected to be found in an at least one genomic region of interest,
- Generating a plurality of nucleotide dispensation orders susceptible to be used during a pyro-sequencing of the at least one genomic region of interest,
- Generating a theoretical pyro-sequencing signal for each unique nucleotide sequence of the list, said pyro-sequencing signals being generated with each nucleotide dispensation order contained in the plurality of nucleotides dispensation orders,
- Determining within the plurality of nucleotide dispensation orders, the one that produces the less correlated pyro-sequencing signals for the nucleotide sequences expected to be found in the at least one genomic region of interest.

15. The dictionary according to claim 14, wherein the step of determining within the plurality of nucleotide dispensation order, the one that produces the less correlated pyro-sequencing signals, comprises the steps of:

- Calculating for each nucleotide dispensation order, a vector that includes the correlation coefficients between each possible pair of pyro-sequencing signals generated with this nucleotide dispensation order,
- Computing for each nucleotide dispensation order the maximum value of the vector that includes the correlation coefficients,
- Selecting the nucleotide dispensation order with the lowest maximum value.

16. The method according to claim 1-9, wherein the nucleotides are dispensed with a specific order determined by the following steps:

- Obtaining a list of all unique nucleotide sequence expected to be found in an at least one genomic region of interest,
- Generating a plurality of nucleotide dispensation orders susceptible to be used during a pyro-sequencing of the at least one genomic region of interest,

- Generating a theoretical pyro-sequencing signal for each unique nucleotide sequence of the list, said pyro-sequencing signal being generated with each nucleotide dispensation order contained in the plurality of nucleotides dispensation orders,

- 5 - Determining within the plurality of nucleotide dispensation order, the one that produces the less correlated pyro-sequencing signals for the nucleotide sequences expected to be found in the at least genomic region of interest.

10 17. The method according to claim 16 wherein the step of determining within the plurality of nucleotide dispensation order, the one that produces the less correlated pyro-sequencing signals, comprises the steps of:

- Calculating for each nucleotide dispensation order, a vector that includes the correlation coefficients between each possible pair of pyro-sequencing signals generated with this nucleotide dispensation order,
- 15 - Computing for each nucleotide dispensation order the maximum value of the vector that includes the correlation coefficients,
- Selecting the nucleotide dispensation order with the lowest maximum value.

20 18. The method according to claim 17, wherein the correlation coefficients between each possible pair of pyro-sequencing signals are calculated by a Pearson product-moment correlation.

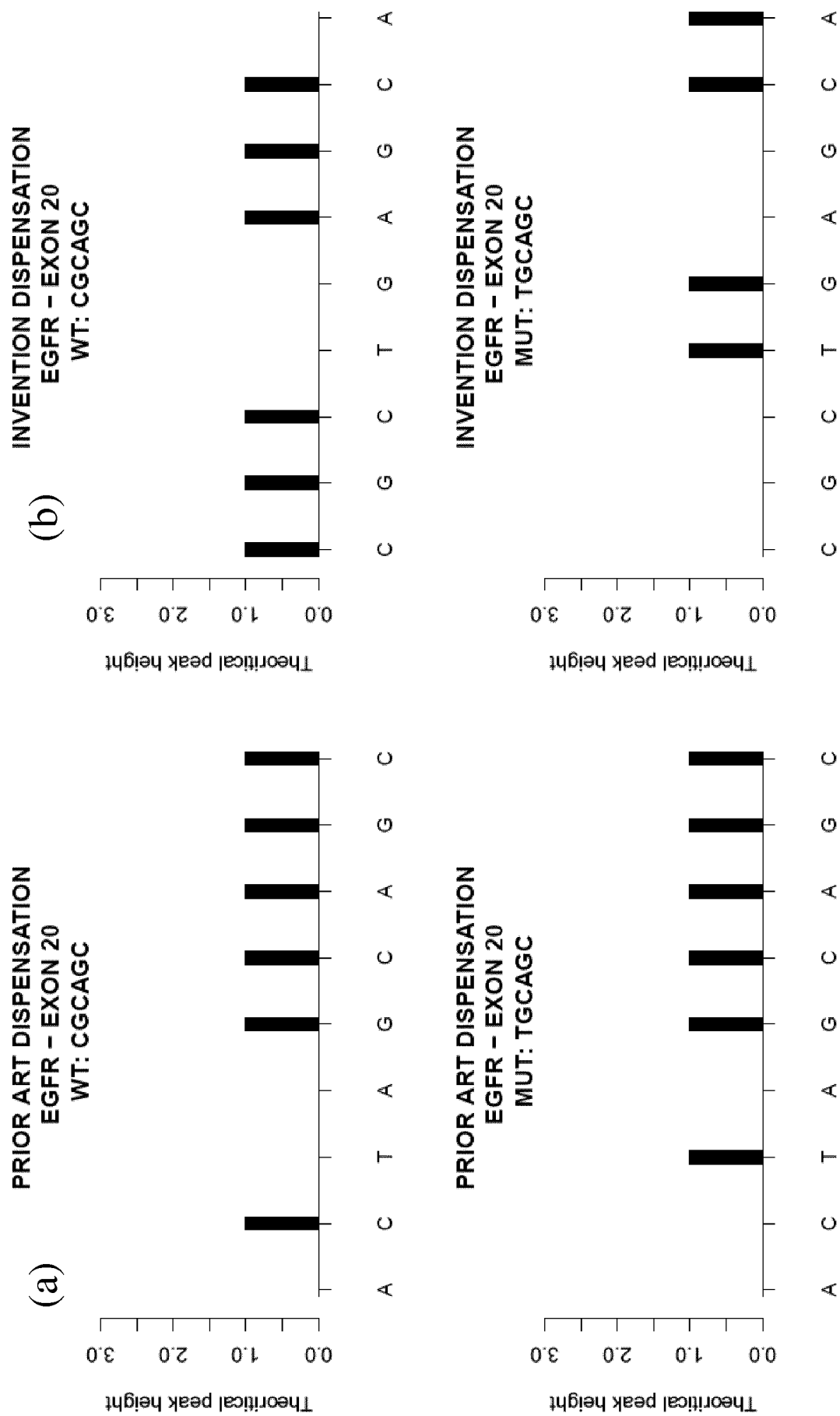


Fig. 1

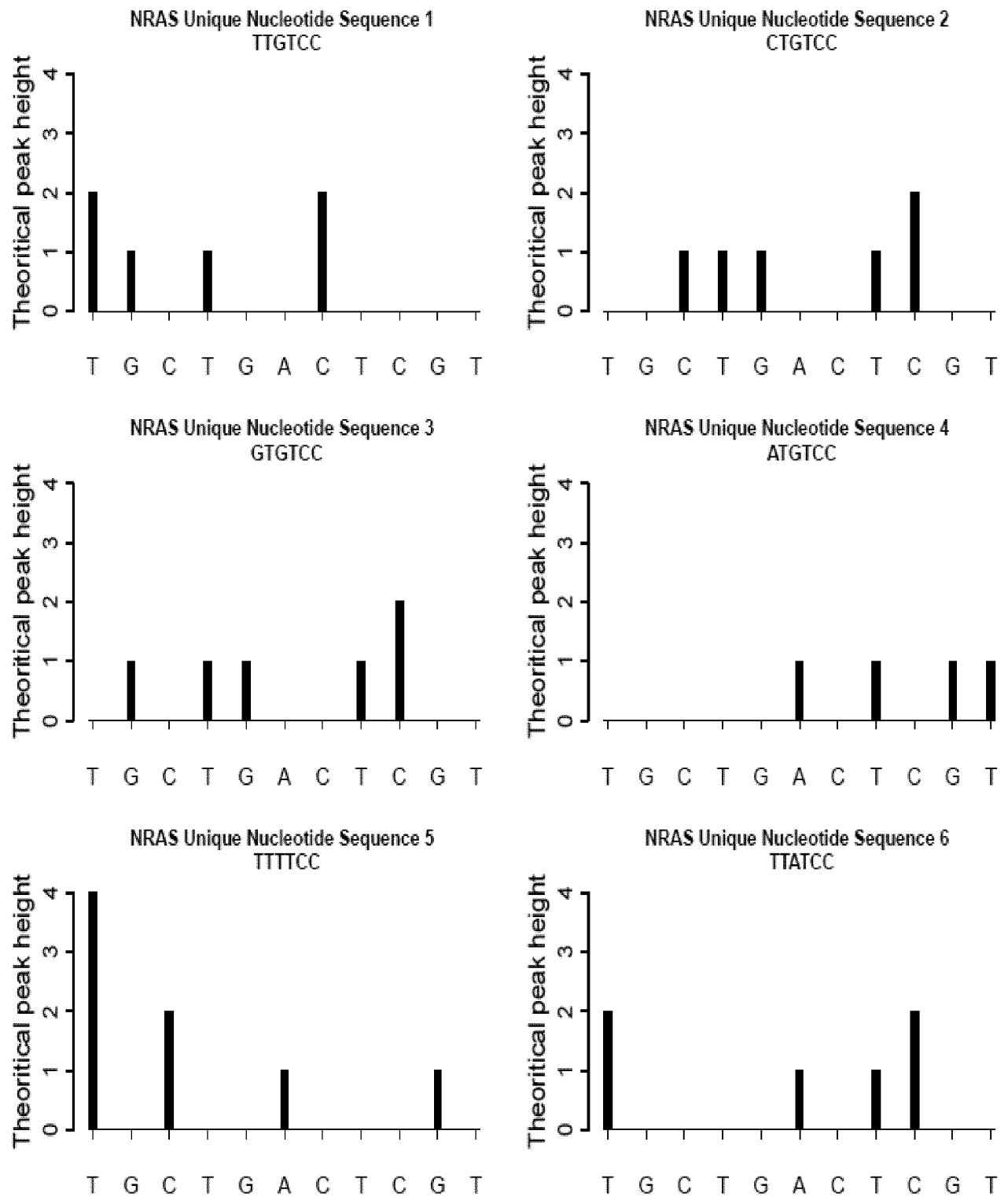


Fig. 2 (1/2)

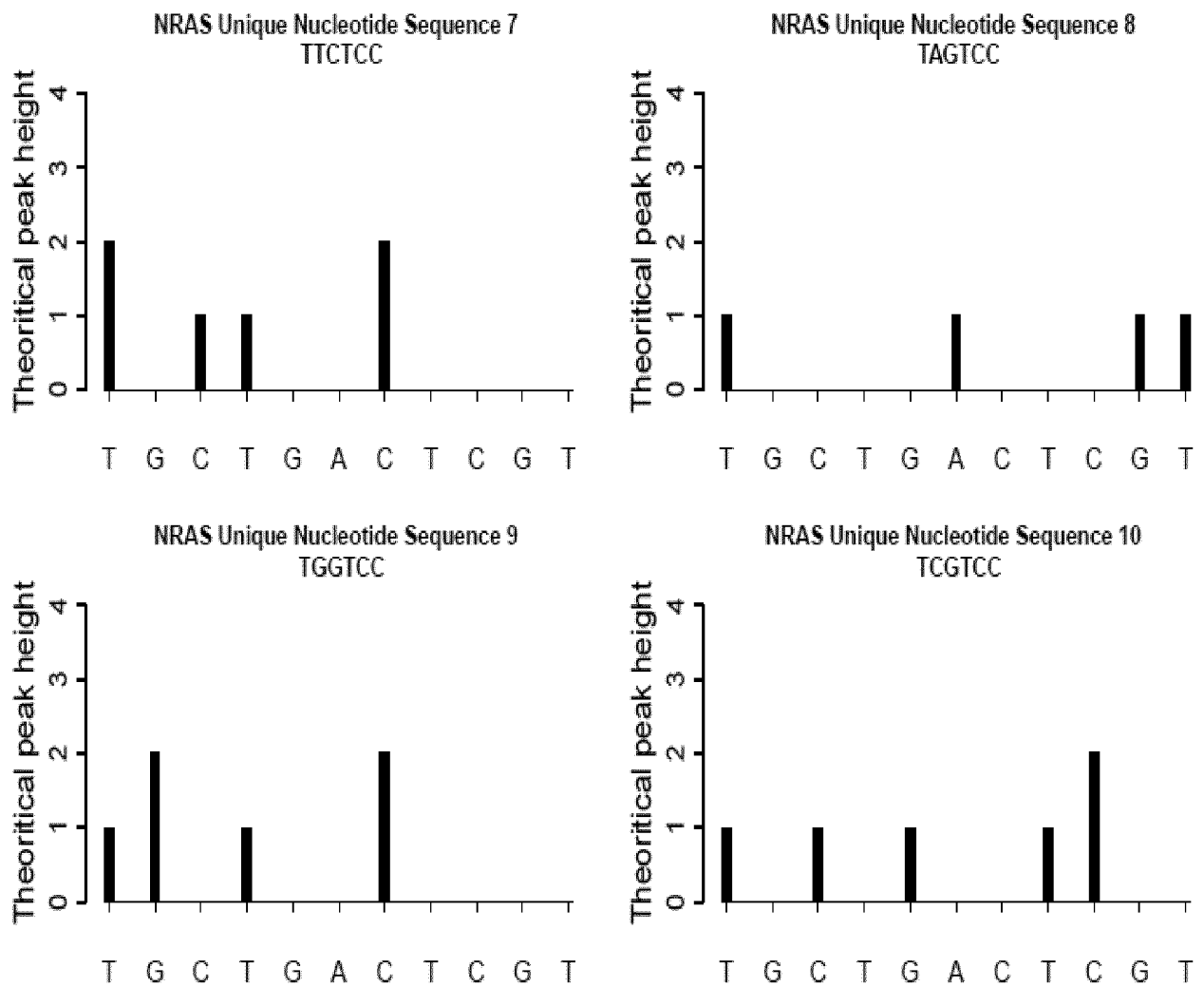


Fig. 2 (2/2)

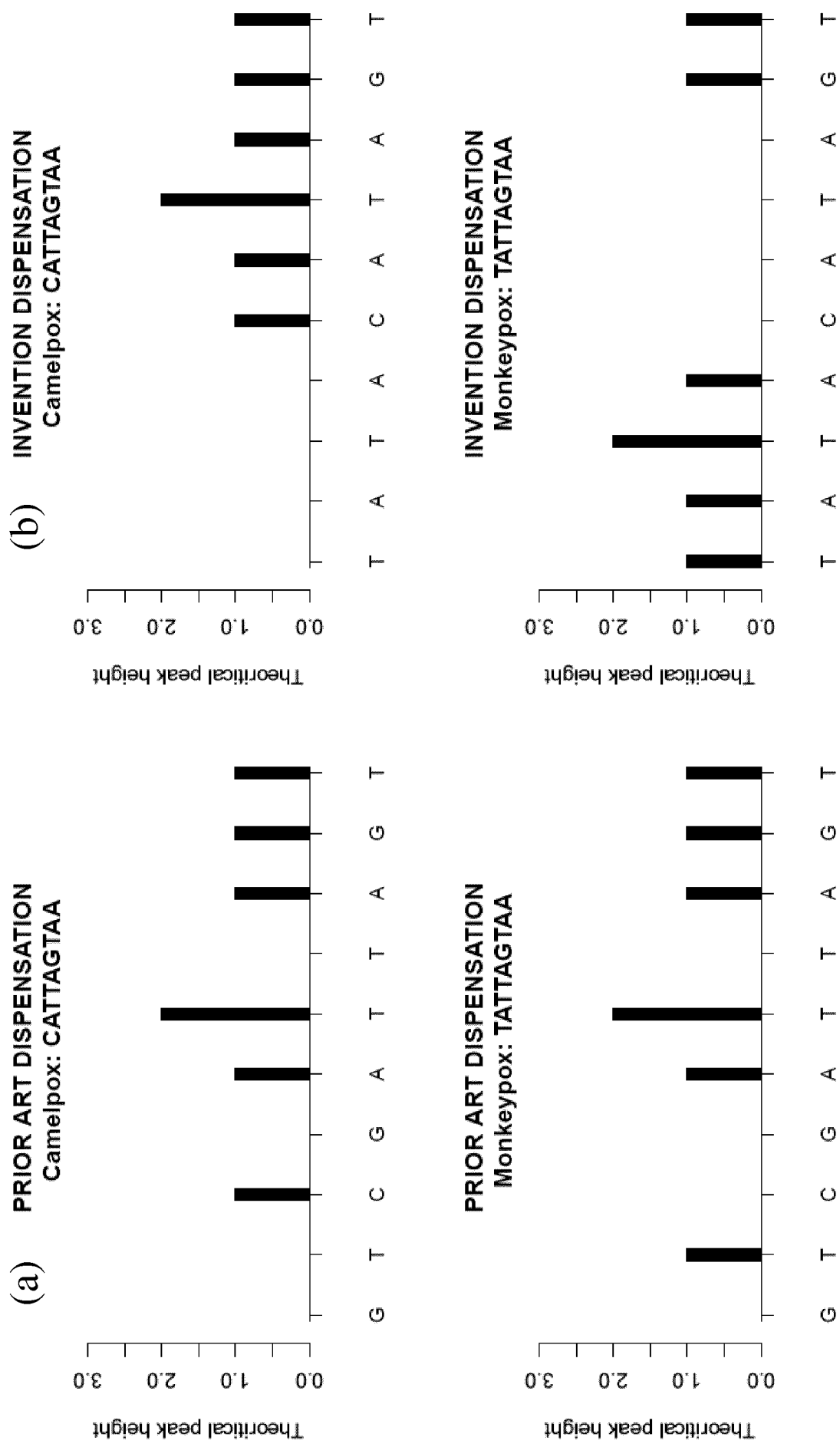


Fig. 3

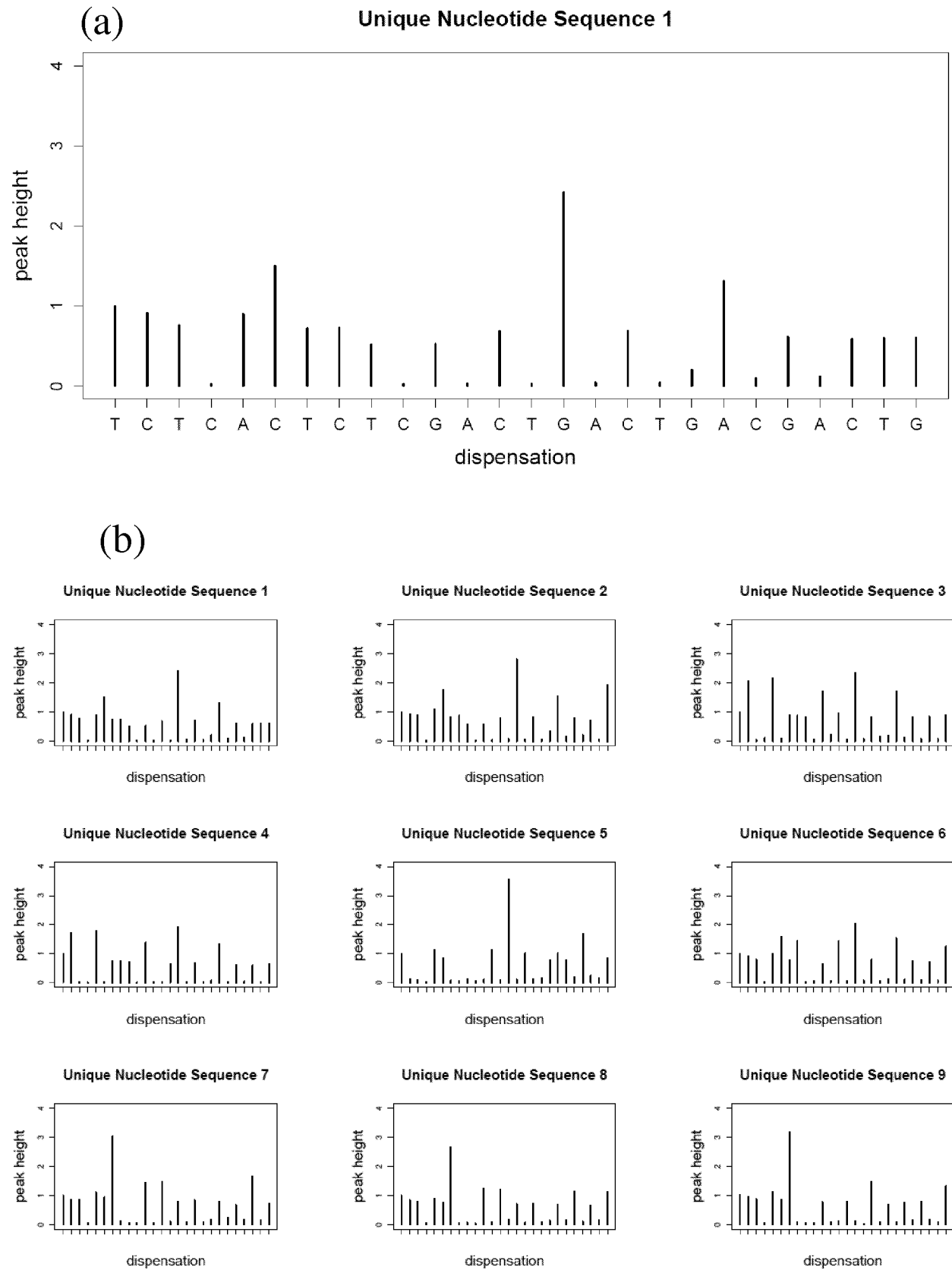


Fig. 4

(a)

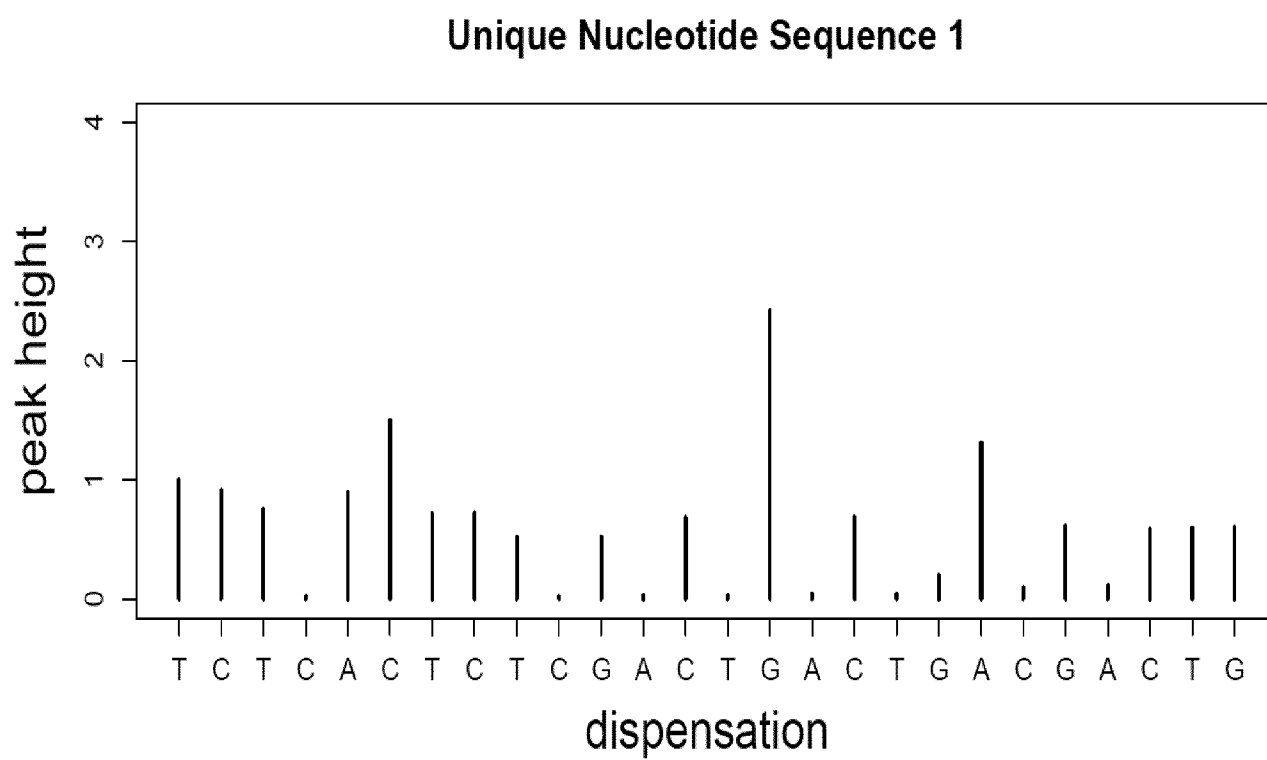
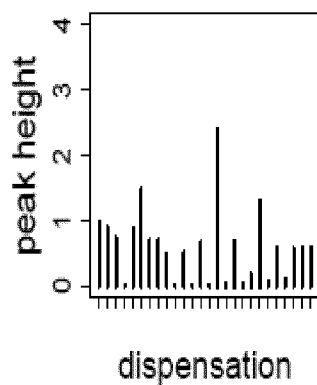


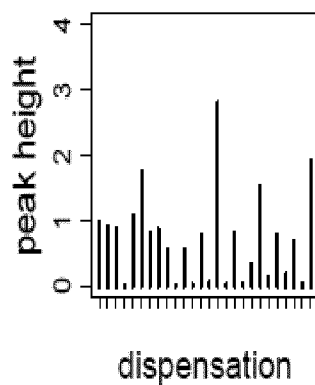
Fig. 4 (1/2)

(b)

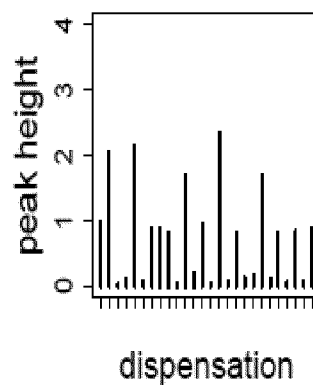
Unique Nucleotide Sequence 1



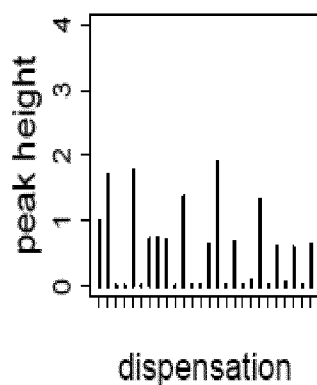
Unique Nucleotide Sequence 2



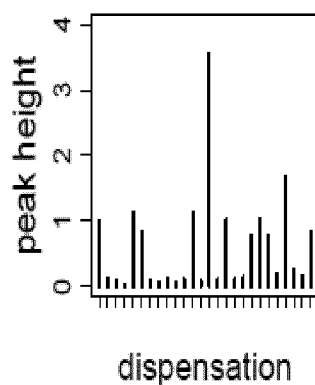
Unique Nucleotide Sequence 3



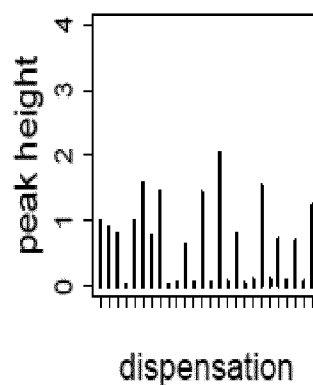
Unique Nucleotide Sequence 4



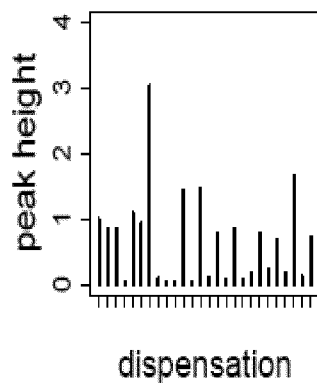
Unique Nucleotide Sequence 5



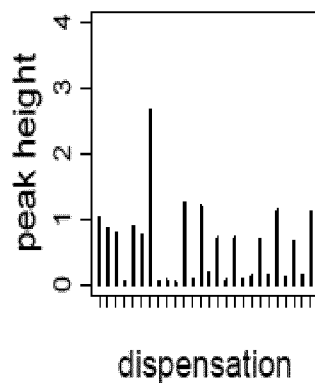
Unique Nucleotide Sequence 6



Unique Nucleotide Sequence 7



Unique Nucleotide Sequence 8



Unique Nucleotide Sequence 9

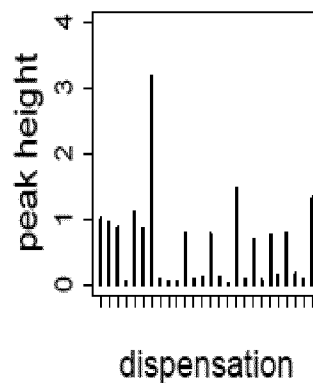


Fig. 4 (2/2)

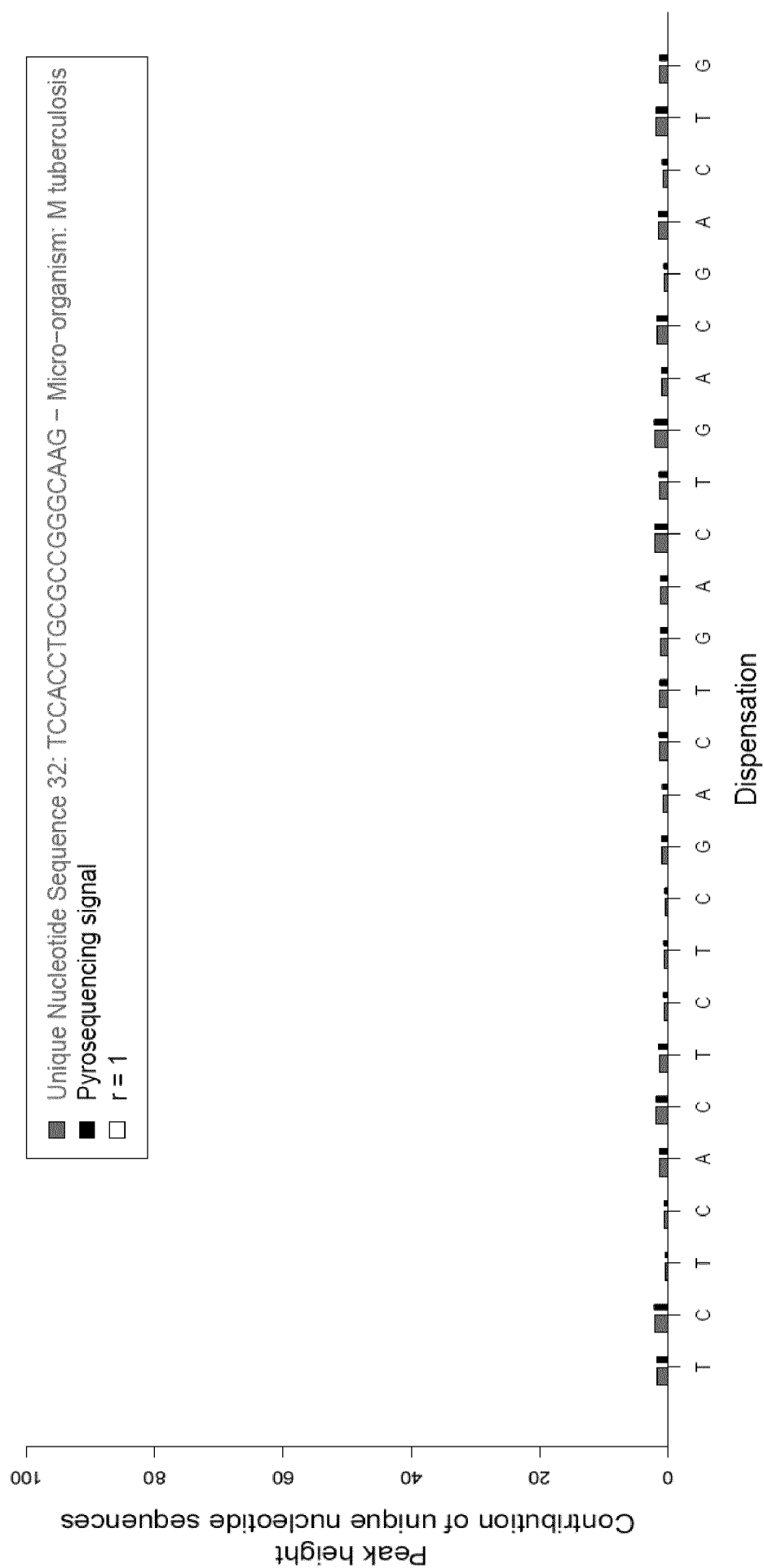


Fig. 5

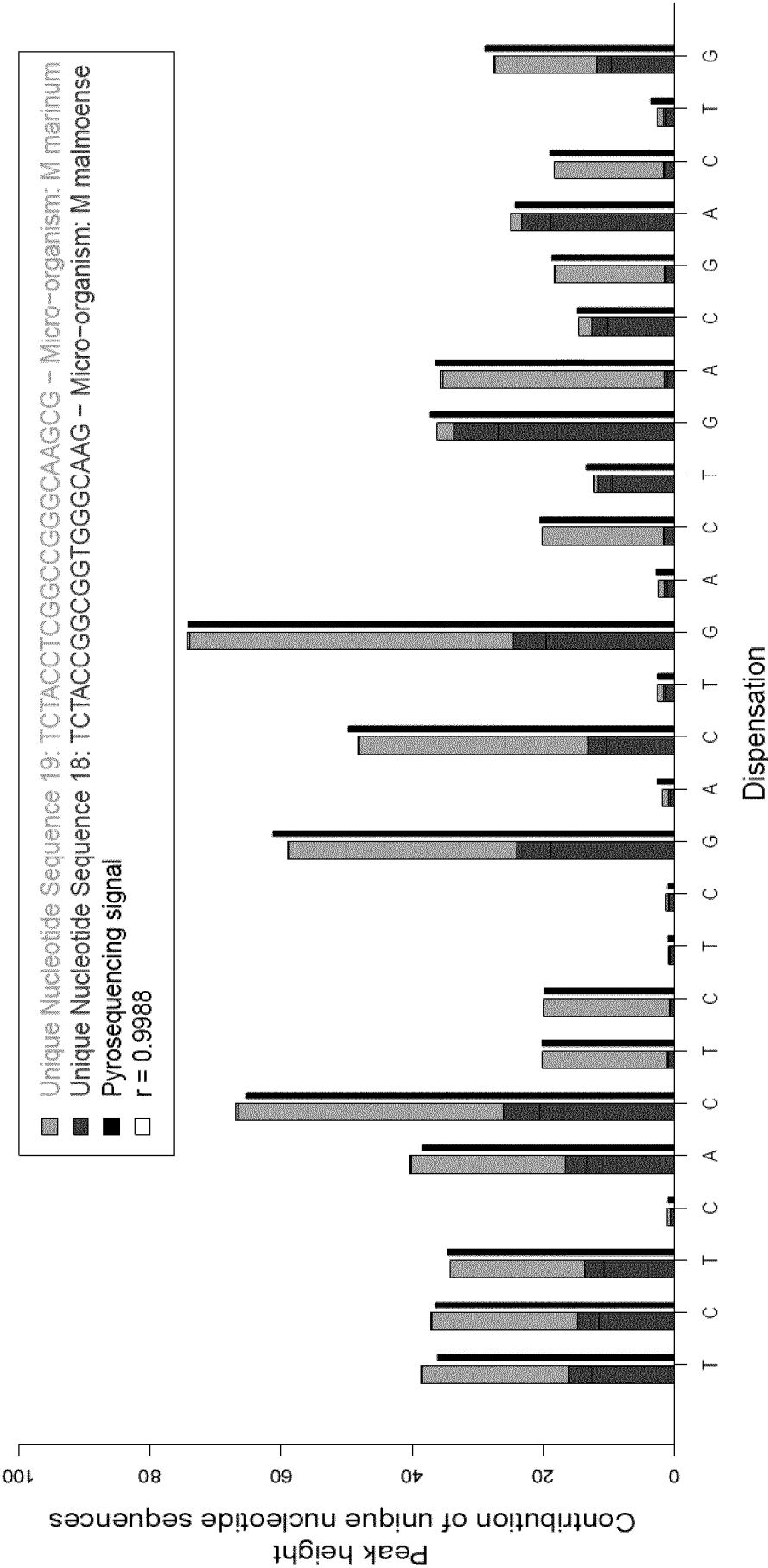


Fig. 6

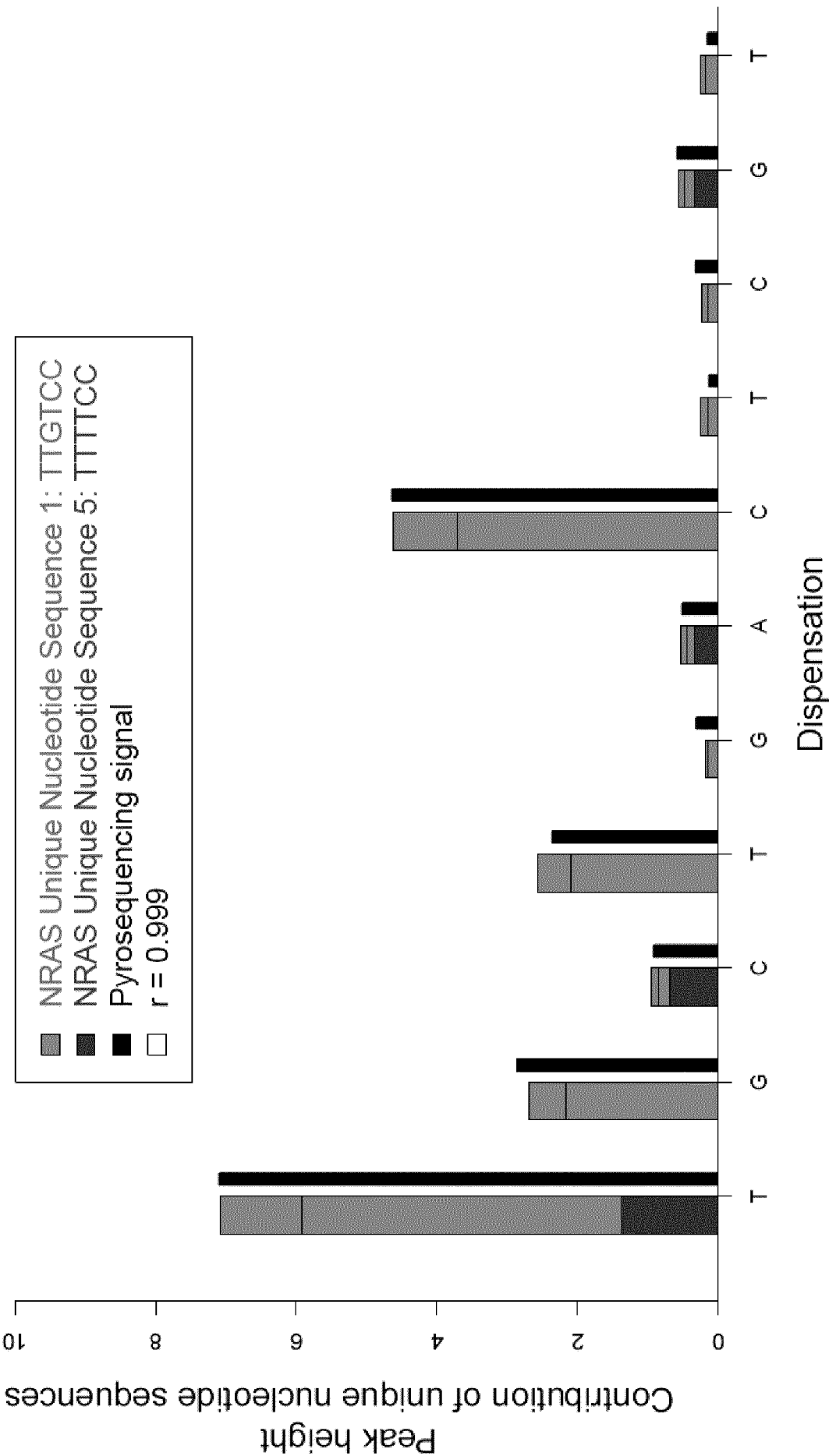


Fig. 7

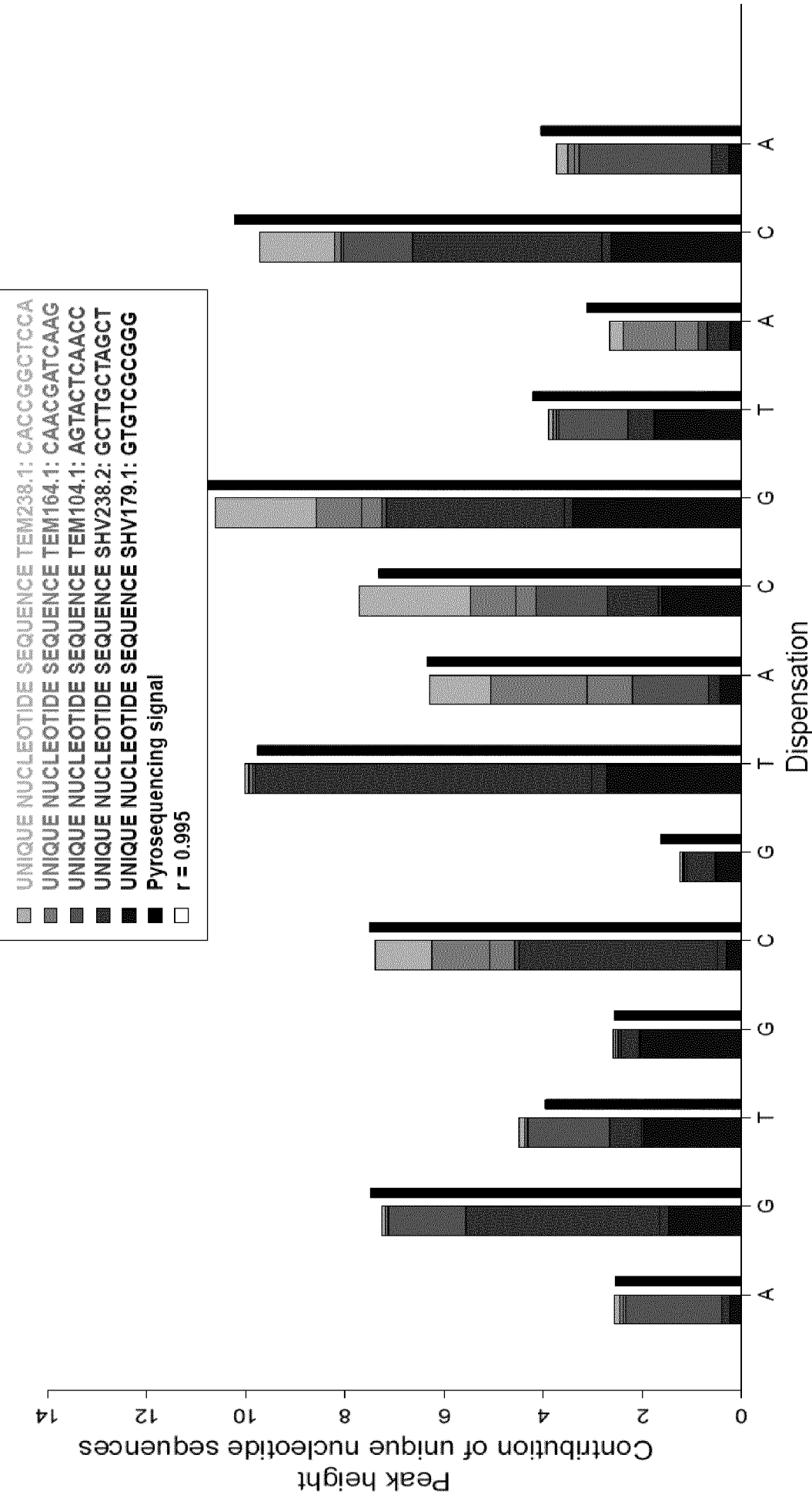


Fig. 8

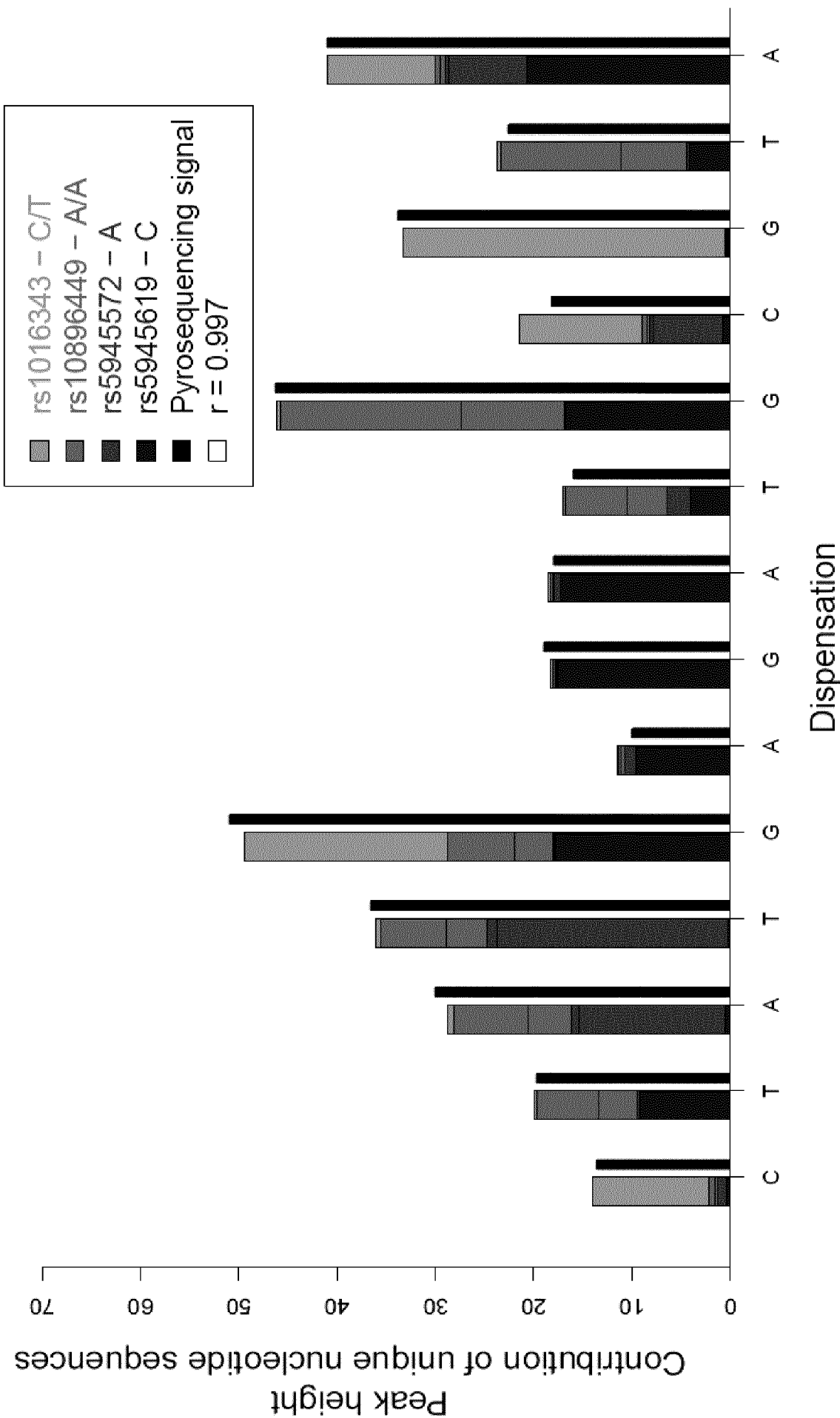


Fig. 9