(12) INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(19) World Intellectual Property Organization
International Bureau

(43) International Publication Date
09 November 2017 (09.11.2017)

WIPO | PCT

(10) International Publication Number
WO 2017/191068 A1

(51) International Patent Classification:
G06F 19/18 (2011.01) *G06F 19/24* (2011.01)

(21) International Application Number:
PCT/EP2017/060279

(22) International Filing Date:
28 April 2017 (28.04.2017)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:
1607628.3 01 May 2016 (01.05.2016) GB

(71) Applicant: **GENOME RESEARCH LIMITED** [GB/GB]; Wellcome Trust Genome Campus, Hinxton, Cambridge Cambridgeshire CB10 1SA (GB).

(72) Inventors: **NIK-ZAINAL, Serena**; c/o Genome Research Limited, Wellcome Trust Genome Campus, Hinxton, Cambridge Cambridgeshire CB10 1SA (GB). **STRATTON, Mike**; c/o Genome Research Limited, Wellcome Trust Genome Campus, Hinxton, Cambridge Cambridgeshire CB10 1SA (GB). **GLODZIK, Dominik**; c/o Genome Research Limited, Wellcome Trust Genome Campus, Hinxton, Cambridge Cambridgeshire CB10 1SA (GB).

(74) Agent: **HODSDON, Stephen** et al.; Mewburn Ellis LLP, City Tower, 40 Basinghall Street, London Greater London EC2V 5DE (GB).

(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, DJ, 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, KH, KN, KP, KR, KW, KZ, LA, LC, LK, LR, LS, 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, ST, 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).

(54) Title: METHOD OF DETECTING A MUTATIONAL SIGNATURE IN A SAMPLE

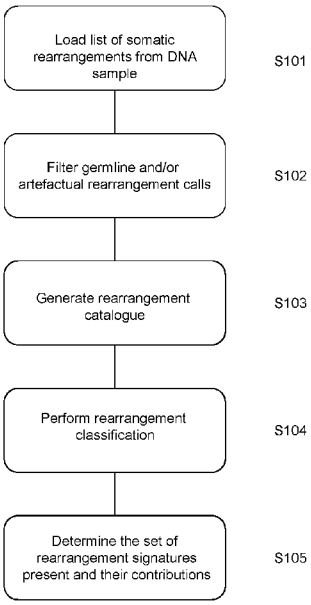


Figure 1

(57) Abstract: The present invention provides a method of detecting mutational signatures in a DNA sample. The invention relates to method of detecting signatures arising from rearrangements in the DNA in the sample and determining the contributions of known rearrangement signatures to said rearrangements. In particular embodiments, the contributions are determined by computing the cosine similarity between the rearrangement mutations in said catalogue and the rearrangement mutational signatures. The rearrangement signatures are classified based on whether they are clustered or not, whether they are tandem duplications, deletions, inversions or translocations and on the basis of their size.

WO 2017/191068 A1

[Continued on next page]

Declarations under Rule 4.17:

— *of inventorship (Rule 4.17(iv))*

Published:

— *with international search report (Art. 21(3))*

— *before the expiration of the time limit for amending the claims and to be republished in the event of receipt of amendments (Rule 48.2(h))*

METHOD OF DETECTING A MUTATIONAL SIGNATURE IN A SAMPLE

FIELD OF INVENTION

The present invention relates to a method for detecting mutational signatures in a DNA sample. It is particularly concerned with a method for detecting rearrangement signatures in a DNA sample.

BACKGROUND TO THE INVENTION

Somatic mutations are present in all cells of the human body and occur throughout life. They are the consequence of multiple mutational processes, including the intrinsic slight infidelity of the DNA replication machinery, exogenous or endogenous mutagen exposures, enzymatic modification of DNA and defective DNA repair. Different mutational processes generate unique combinations of mutation types, termed “Mutational Signatures”.

In the past few years, large-scale analyses have revealed many mutational signatures across the spectrum of human cancer types.

The mutational theory of cancer proposes that changes in DNA sequence, termed “driver” mutations, confer proliferative advantage upon a cell, leading to outgrowth of a neoplastic clone [1]. Some driver mutations are inherited in the germline, but most arise in somatic cells during the lifetime of the cancer patient, together with many “passenger” mutations not implicated in cancer development [1]. Multiple mutational processes, including endogenous and exogenous mutagen exposures, aberrant DNA editing, replication errors and defective DNA maintenance, are responsible for generating these mutations [1-3].

Over the past five decades, several waves of technology have advanced the characterisation of mutations in cancer genomes. Karyotype analysis revealed rearranged chromosomes and copy number alterations. Subsequently, loss of heterozygosity analysis, hybridisation of cancer-derived DNA to microarrays and other approaches provided higher resolution insights into copy number changes [4-8]. Recently, DNA sequencing has enabled systematic characterisation of the full repertoire of mutation types including base substitutions, small insertions/deletions, rearrangements and copy number changes [9-13], yielding substantial insights into the mutated cancer genes and mutational processes operative in human cancer.

Mutational processes generating somatic mutations imprint particular patterns of mutations on cancer genomes, termed signatures [2, 15, 16]. Applying a mathematical approach [15] to extract mutational signatures previously revealed five base substitution signatures in breast cancer; signatures 1, 2, 3, 8 and 13 [2,14].

Whilst base substitution signatures have been investigated and methods for their detection proposed, signatures of rearrangement mutational processes have not previously been formally investigated and in particular no methods proposed for the characterisation of rearrangement mutational signatures and identification of the presence of one or more rearrangement signatures in a DNA sample taken from a single patient.

A method of identifying the presence of rearrangement signatures in a DNA sample taken from a single patient would provide for considerable benefit as it may provide a potential route for diagnosis of possible cancer types in that patient or may provide identification of an underlying defect and therefore allow selection of patients for particular types of therapy.

STATEMENTS OF INVENTION

An exemplary embodiment of the present invention provides a method of detecting rearrangement signatures in a previously obtained DNA sample, the method including the steps of: cataloguing the somatic mutations in said sample to produce a rearrangement catalogue for that sample which classifies identified rearrangement mutations in the sample into a plurality of categories; determining the contributions of known rearrangement signatures to said rearrangement catalogue by computing the cosine similarity between the rearrangement mutations in said catalogue and the rearrangement mutational signatures.

A further exemplary embodiment of the present invention provides a computer program product containing non-transitory memory storing a computer program which, when run on a computer, performs the steps of: cataloguing the somatic mutations in said sample to produce a rearrangement catalogue for that sample which classifies identified rearrangement mutations in the sample into a plurality of categories; determining the contributions of known rearrangement signatures to said rearrangement catalogue by computing the cosine similarity between the rearrangement mutations in said catalogue and the rearrangement mutational signatures.

A further exemplary embodiment of the present invention provides a computer having a processor, wherein the processor is configured to: catalogue the somatic mutations in said

sample to produce a rearrangement catalogue for that sample which classifies identified rearrangement mutations in the sample into a plurality of categories; determine the contributions of known rearrangement signatures to said rearrangement catalogue by computing the cosine similarity between the rearrangement mutations in said catalogue and the rearrangement mutational signatures.

BRIEF DESCRIPTION OF THE FIGURES & TABLE

Figure 1 is a flow diagram showing, in schematic form, a method of detecting a rearrangement signature in the DNA of a single patient according to an embodiment of the present invention; and

Figure 2 is a diagram showing seven major subgroups exhibiting distinct associations with other genomic, histological or gene expression features, along with the six rearrangement signatures extracted from the data.

Table 1 shows a quantitative definition of a number of rearrangement signatures.

DETAILED DESCRIPTION

A first aspect of the present invention provides a method of detecting rearrangement signatures in a previously obtained DNA sample, the method including the steps of: cataloguing the somatic mutations in said sample to produce a rearrangement catalogue for that sample which classifies identified rearrangement mutations in the sample into a plurality of categories; and determining the contributions of known rearrangement signatures to said rearrangement catalogue by computing the cosine similarity between the rearrangement mutations in said catalogue and the rearrangement mutational signatures.

Preferably the method includes the further step of, prior to said step of determining, filtering the mutations in said catalogue to remove either residual germline structural variations or known sequencing artefacts or both. Such filtering can be highly advantageous to remove rearrangements from the catalogue which are known to arise from mechanisms other than somatic mutation, and may therefore cloud or obscure the contributions of the rearrangement signatures, or lead to false positive results.

For example, the filtering may use a list of known germline rearrangement or copy number polymorphisms and remove somatic mutations resulting from those polymorphisms from the catalogue prior to determining the contributions of the rearrangement signatures.

As a further example, the filtering may use BAM files of unmatched normal human tissue sequenced by the same process as the DNA sample and discards any somatic mutation which is present in at least two well-mapping reads in at least two of said BAM files. This approach can remove artefacts resulting from the sequencing technology used to obtain the sample.

The classification of the rearrangement mutations may include identifying mutations as being clustered or non-clustered. This may be determined by a piecewise-constant fitting ("PCF") algorithm which is a method of segmentation of sequential data. In particular embodiments, rearrangements may be identified as being clustered if the average density of rearrangement breakpoints within a segment is a certain factor greater than the whole genome average density of rearrangements for an individual patient's sample. For example the factor may be at least 8 times, preferably at least 9 times and in particular embodiments is 10 times. The inter-rearrangement distance is the distance from a rearrangement breakpoint to the one immediately preceding it in the reference genome. For any given breakpoint, this measurement is already known.

The classification of the rearrangement mutations may include identifying rearrangements as one of: tandem duplications, deletions, inversions or translocations. Such classifications of rearrangement mutations are already known.

The classification of the rearrangement mutations may further include grouping mutations identified as tandem duplications, deletions or inversions by size. For example, the mutations may be grouped into a plurality of size groups by the number of bases in the rearrangement. Preferably the size groups are logarithmically based, for example 1-10kb, 10-100kb, 100kb-1Mb, 1Mb-10Mb and greater than 10Mb. Translocations cannot be classified by size.

In particular embodiments, in each DNA sample the number of rearrangements E_i associated with the *i*th mutational signature $\vec{S}_i$ is determined as proportional to the cosine similarity ($\vec{C}_i$) between the catalogue of this sample $\vec{M}$ and $\vec{S}_i$:

$$\vec{C}_i = \frac{\vec{S}_i \cdot \vec{M}}{\| \vec{S}_i \| \| \vec{M} \|}$$

wherein:

$$E_i = \frac{\vec{C}_i}{\sum_{i=1}^{q} \vec{C}_i} \sum_{j=1}^{36} \vec{M}^j$$

wherein $\overrightarrow{S_i}$ and $\overrightarrow{M}$ are equally-sized vectors with nonnegative components being, respectively, a known rearrangement signature and the mutational catalogue and q is the number of signatures in said plurality of known rearrangement signatures.

The method may further include the step of filtering the number of rearrangements determined to be assigned to each signature by reassigning one or more rearrangements from signatures that are less correlated with the catalogue to signatures that are more correlated with the catalogue. Such filtering can serve to reassign rearrangements from a signature which has only a few rearrangements associated with it (and so is probably not present) to a signature which has a greater number of rearrangement associated with it. This can have the effect of reducing "noise" in the assignment process.

In one embodiment, the step of filtering uses a greedy algorithm to iteratively find an alternative assignment of rearrangements to signatures that improves or does not change the cosine similarity between the catalogue $\vec{M}$ and the reconstructed catalogue $\vec{M}' = S \times \vec{E}'_{ij}$, wherein $\vec{E}'_{ij}$ is the version of the vector $\vec{E}$ obtained by moving the mutations from the signature i to signature j, wherein, in each iteration, the effects of all possible movements between signatures are estimated, and the filtering step terminates when all of these possible reassignments have a negative impact on the cosine similarity.

The subject may be a cancer patient or a suspected cancer patient. For example, the method may be used in the determination or identification of a rearrangement sequence to predict whether the subject has cancer or not or what type of cancer a patient has, or to select the subject for a particular form of treatment.

The method may further include the step of determining if the number or proportion of rearrangements in the rearrangement catalogue which are determined to be associated with one or more of said rearrangement signatures each or in combination exceeds a predetermined threshold and, if so, determining that said rearrangement signature is present in the sample.

The present inventors have determined that, by classifying rearrangement mutations by clustered/non-clustered, type and size (where appropriate), clear rearrangement signatures can be identified in a number of tumours. Accordingly, these classifications, in conjunction

with the method of the present embodiment can provide an ability to identify the presence of particular rearrangement signatures and therefore determine a likelihood that a sample from a patient is indicative of the presence of a tumour and/or the form of cancer causing the tumour. As different forms of cancer are known to react different to particular treatments, the identification of the likely form of cancer present in a sample can guide the selection of the treatment for the subject.

The present inventors have also identified clear links between the rearrangement signatures and the underlying mechanisms contributing to a cancer. Accordingly, the presence (or absence) of a particular rearrangement signature (or collection of rearrangement signatures) can alternatively or additionally be used to determine the underlying mechanisms that are contributing to the tumour from which the sample is taken.

The method of the present aspect may include any combination of some, all or none of the above described preferred and optional features.

Further aspects of the present invention include computer programs for running on computer systems which carry out the method of the above aspect, including some, all or none of the preferred and optional features of that aspect.

A further aspect of the present invention provides a computer program product containing non-transitory memory storing a computer program which, when run on a computer, performs the steps of: cataloguing the somatic mutations in said sample to produce a rearrangement catalogue for that sample which classifies identified rearrangement mutations in the sample into a plurality of categories; determining the contributions of known rearrangement signatures to said rearrangement catalogue by computing the cosine similarity between the rearrangement mutations in said catalogue and the rearrangement mutational signatures.

A further aspect of the present invention provides a computer having a processor, wherein the processor is configured to: catalogue the somatic mutations in said sample to produce a rearrangement catalogue for that sample which classifies identified rearrangement mutations in the sample into a plurality of categories; determine the contributions of known rearrangement signatures to said rearrangement catalogue by computing the cosine similarity between the rearrangement mutations in said catalogue and the rearrangement mutational signatures.

The computer program and the processor of the above two aspects may also carry out some or all of the optional or preferred steps described above in relation to the first aspect.

These and other aspects of the invention are described in further detail below.

IDENTIFICATION OF REARRANGEMENT SIGNATURES LINKED TO CANCER

The complete genomes of 560 breast cancers and non-neoplastic tissue from each individual (556 female and four male) were sequenced. 3,479,652 somatic base substitutions, 371,993 small indels and 77,695 rearrangements were detected, with substantial variation in the number of each between individual samples.

To enable investigation of signatures of rearrangement mutational processes, a rearrangement classification was adopted incorporating 32 subclasses.

In many cancer genomes, large numbers of rearrangements are regionally clustered, for example in zones of gene amplification. Therefore, the rearrangements were first classified into those that occurred as clusters or were dispersed, further sub-classified into deletions, inversions and tandem duplications, and then according to the size of the rearranged segment. The final category in both groups was inter-chromosomal translocations.

Application of the mathematical framework used for base substitution signatures [2, 14, 15] extracted six rearrangement signatures. Unsupervised hierarchical clustering on the basis of the proportion of rearrangements attributed to each signature in each breast cancer yielded seven major subgroups exhibiting distinct associations with other genomic, histological or gene expression features as shown in Figure 2.

Rearrangement Signature 1 (9% of all rearrangements) and Rearrangement Signature 3 (18% rearrangements) were characterised predominantly by tandem duplications. Tandem duplications associated with Rearrangement Signature 1 were mostly >100kb, and those with Rearrangement Signature 3 <10kb. More than 95% of Rearrangement Signature 3 tandem duplications were concentrated in 15% of cancers, many with several hundred rearrangements of this type. Almost all cancers (91%) with BRCA1 mutations or promoter hypermethylation were in this group, which was enriched for basal-like, triple negative cancers and copy number classification of a high Homologous Recombination Deficiency (HRD) index [17-19]. Thus, inactivation of BRCA1, but not BRCA2, may be responsible for the Rearrangement Signature 3 small tandem duplication mutator phenotype.

More than 35% of Rearrangement Signature 1 tandem duplications were found in just 8.5% of the breast cancers and some cases had hundreds of these. The cause of this large tandem duplication mutator phenotype is unknown. Cancers exhibiting it are frequently TP53-mutated, relatively late diagnosis, triple-negative breast cancers, showing enrichment for base substitution signature 3 and a high Homologous Recombination Deficiency (HRD) index but do not have BRCA1/2 mutations or BRCA1 promoter hypermethylation.

Rearrangement Signature 5 (accounting for 14% rearrangements) was characterised by deletions <100kb. It was strongly associated with the presence of BRCA1 mutations or promoter hypermethylation (Figure 2, Cluster D), BRCA2 mutations (Figure 2, Cluster G) and with Rearrangement Signature 1 large tandem duplications (Figure 2, Cluster F).

Rearrangement Signature 2 (accounting for 22% rearrangements) was characterised by non-clustered deletions (>100kb), inversions and interchromosomal translocations, was present in most cancers but was particularly enriched in ER positive cancers with quiet copy number profiles (Figure 2, Cluster E, GISTIC Cluster 3). Rearrangement Signature 4 (accounting for 18% of rearrangements) was characterised by clustered interchromosomal translocations while Rearrangement Signature 6 (19% of rearrangements) by clustered inversions and deletions (Figure 2, Clusters A, B, C).

The methods according to embodiments of the invention set out below determine the presence or absence of a rearrangement signature in DNA samples obtained from a single patient. Preferably, these are whole genome samples and the presence or absence of mutational signatures may be determined by whole genome sequencing.

The DNA samples are preferably obtained from both tumour and normal tissues obtained from the patient, e.g. blood sample from the patient and breast tumour tissue obtained by a biopsy. Somatic mutations in the tumour sample are detected, standardly, by comparing its genomic sequences with the one of the normal tissue.

METHOD OF DETECTION OF REARRANGEMENT SIGNATURES IN A SINGLE PATIENT

In embodiments of the present invention, detection of a rearrangement signature in the DNA obtained from a single patient is performed. In these embodiments, this detection is performed by a computer-implemented method or tool that examines a list of somatic mutations

generated through high-coverage or low-pass sequencing of nucleic acid material obtained from fresh-frozen derived DNA, circulating tumour DNA of formalin-fixed paraffin-embedded (FFPE) DNA representative of a suspected or known tumour from a patient. The steps of this method are illustrated schematically in Figure 1.

The list of somatic mutations for these embodiments can be provided in variety of different formats (including, VCF, BEDPE, text etc.) but at the very minimum needs to contain the following information: genome assembly version, lower breakpoint chromosome, lower breakpoint coordinate, higher breakpoint chromosome, higher breakpoint coordinate and either rearrangement class (inversion, tandem duplication deletion, translocation) or strand information of lower and higher breakpoints to enable orientation of rearrangement breakpoints in order to correctly classify them.

In broad terms, after loading the list of somatic mutations from the DNA sample (S101) the tool firstly filters out any known germline and/or artifactual somatic mutations (S102), then generates the rearrangement catalogue of the sample, then classifies the rearrangements based on the classification described below (S103), then evaluates the contributions of known consensus rearrangement mutational signatures to this sample (S104) and finally determines the set of signatures of rearrangement processes, and their respective contributions, that are operative in the sample (S105).

By default, the patterns of the consensus rearrangement signatures are those shown in Table 1, but these patterns of mutational signatures could be also user provided and the method is not limited to known signatures and can be readily applied to new or modified signatures which are discovered in the future.

Filtering initial data

Prior to analysing the data, the input list of somatic rearrangements is extensively filtered to remove any residual germline mutations as well as technology specific sequencing artefacts.

Germline rearrangements or copy number polymorphisms are filtered out from the lists of reported somatic mutations using the complete list of germline mutations from dbSNP [21], 1000 genomes project [22], NHLBI GO Exome Sequencing Project [23] and 69 Complete Genomics panel (http://www.completegenomics.com/public-data/69-Genomes/).

Technology specific sequencing artefacts (related to library-making or sequencing chemistry) and mapping-related artefacts caused by errors or biases in the reference genome, are filtered out by using panels of BAM files of unmatched normal human tissues containing at least 100

normal whole-genomes. The remaining somatic mutations are used to construct the mutational catalogue of the examined sample.

Generating the mutational catalogue for a sample

The list of remaining (*i.e.,* post-filtered) somatic rearrangements is used to generate the rearrangement mutational catalogue of a sample.

(1) Clustered vs non-clustered

The first classification applied to the mutations is whether they are clustered (closely-grouped) or not.

To distinguish collections of rearrangements that are clustered or close together in a patient's cancer genome from other rearrangements that are distributed or dispersed throughout the genome, the data is parsed through a PCF-based algorithm. The PCF (Piecewise-Constant-Fitting) algorithm is a method of segmentation of sequential data.

Before applying PCF, a number of steps are performed on the rearrangement data.

Unlike substitutions or indels that have a single genomic coordinate to signify their position, rearrangements have two coordinates or "breakpoints" that identify two distant genomic loci that have been brought together by a large structural mutation event.

First, both breakpoints of each rearrangement are treated independently. The breakpoints are then sorted according to reference genomic coordinate in each sample. The intermutation distance (IMD), defined as the number of base pairs from one rearrangement breakpoint to the one immediately preceding it in the reference genome, is calculated for each breakpoint. The calculated IMD is then fed to the PCF algorithm.

To identify regions of "clustered" rearrangements from "non-clustered" rearrangements, a set of rearrangements was required to have an average density of rearrangement breakpoints that was at least 10 times greater than the whole genome average density of rearrangements for an individual patient's sample. Additionally, a gamma parameter (a measure of smoothness of segmentation) was stipulated, $\gamma = 25$, and required that a minimum of 10 breakpoints were present in each region, before it could be classified as a cluster of rearrangements. Biologically, the respective partner breakpoint of any rearrangement involved in a clustered region is likely to have arisen at the same mechanistic instant and so can be considered as being involved in the cluster even if located at a distant genomic site according to the reference genome.

Thus rearrangements are first classified as “clustered” or “non-clustered.

(2) Type and Size

In both clustered and non-clustered categories, rearrangements are then classified based on the information provided into the main classes of rearrangements:

- tandem duplications
- deletions
- inversions
- translocations

Tandem duplications, deletions and inversions can then be categorised into the following 5 size groups where the size of a rearrangement is obtained through subtracting the lower breakpoint coordinate from the higher one.

- 1-10kb
- 10-100kb
- 100kb-1Mb
- 1Mb-10Mb
- >10Mb

Translocations are the exception and cannot be classified by size.

In all, there will be 16 subgroups of clustered and 16 subgroups of non-clustered rearrangements and thus 32 categories altogether. These are listed in Table 1.

The outcome of this classification can then be fed into a latent variable analysis such as NNMF, to obtain a non-negative vector of 32 elements describing each rearrangement signature.

Evaluating the numbers of somatic mutations attributed to re-arrangement signatures in the mutational catalogue of the examined sample

Calculating the contributions of all mutational signatures is performed by estimating the number of mutations associated to the consensus patterns of the signatures of all operative mutational processes in the sample. Below a method of estimating this using non-negative

matrix factorisation (NNMF) is set out, although alternative methods such as EMU or a hierarchical Dirichlet process (HDP) may equally be used.

More specifically, all consensus rearrangement signatures are examined as a set P containing s vectors $P=\left\{\begin{bmatrix} p_1^1 \\ \vdots \\ p_1^{32} \end{bmatrix},\begin{bmatrix} p_2^1 \\ \vdots \\ p_2^{32} \end{bmatrix}\ldots\begin{bmatrix} p_{s-1}^1 \\ \vdots \\ p_{s-1}^{32} \end{bmatrix},\begin{bmatrix} p_s^1 \\ \vdots \\ p_s^{32} \end{bmatrix}\right\}$, where each of the vectors is a discrete probability density function reflecting a consensus rearrangement signature. For the currently known rearrangement signatures, these vectors are set out in the respective columns of Table 1. Here, s refers to the number of known consensus rearrangement signatures (currently 6) and the 32 nonnegative components of each vector correspond to the different categories of rearrangements (*i.e.*, clustered/non-clustered, type & size) of these consensus rearrangement signatures.

The contributions of all consensus rearrangement signatures are estimated independently for the mutational catalogue of the examined sample. The estimation algorithm consists of computing the cosine similarity between each signature and examined sample. For a set of vectors $S_{1..q}, q \leq s,$ the cosine similarity $\vec{C}_i$ is given by:

$$\vec{C}_i = \frac{\vec{S}_i \cdot \vec{M}}{\| \vec{S}_i \| \| \vec{M} \|}$$

The number of rearrangements E_i associated with the *i*th mutational signature $\vec{S}_i$ is proportional to the cosine similarity ($\vec{C}_i$):

$$E_i = \frac{\vec{C}_i}{\sum_{i=1}^{q} \vec{C}_i} \sum_{j=1}^{36} \vec{M}^j$$

wherein $\overrightarrow{S_i}$ and $\overrightarrow{M}$ are equally-sized vectors with nonnegative components being, respectively, a known rearrangement signature and the mutational catalogue and *q* is the number of signatures in said plurality of known rearrangement signatures.

In the above equation, $\overrightarrow{S_i}$ and $\overrightarrow{M}$ represent vectors with 32 nonnegative components (corresponding to the clustered/non-clustered characteristic and the type and size of the rearrangements) reflecting, respectively, a consensus mutational signature and the mutational catalogue of the examined sample. Hence, $\overrightarrow{S_i} \in \Re_+^{32}$ while $\overrightarrow{M} \in \mathrm{N}_0^{32}$. Further, both vectors have known numerical values either from the consensus mutational signatures (*i.e.*, $\overrightarrow{S_i}$) or

from generating the original mutational catalogue of the sample (*i.e.,* $\overrightarrow{M}$). In contrast, E_i corresponds to an unknown scalar reflecting the number of rearrangements contributed by signature $\overrightarrow{S_i}$ in the mutational catalogue $\overrightarrow{M}$.

The above equation is universally constrained in regards to the parameter E_i . More specifically, the number of somatic rearrangements contributed by a rearrangement signature in a sample must be nonnegative and it must not exceed the total number of somatic mutations in that sample. Furthermore, the mutations contributed by all signatures in a sample must equal the total number of somatic mutations of that sample. These constraints can be mathematically expressed as $0 \leq E_i \leq \left\|\overrightarrow{S_i}\right\|_1, i = 1..q,$ and $\sum_{i=1}^{q} E_i = \left\|\overrightarrow{S_i}\right\|_1$.

When no prior biological knowledge is available the whole set Q of signatures is used in the determination of E_i ., and a filter step is used to move the mutations from the least correlated signatures the ones that best explain the considered sample (signature highly correlated). Given the catalogue $\vec{M}$ and given all $\| Q^Q \|$ possible movements between two signatures i and j $(i \neq j \text{ and } i, j = 1, \ldots, Q)$, the filtering step uses a greedy algorithm to iteratively choose the movement that improves or does not change the cosine similarity between the catalogue $\vec{M}$ and the reconstructed catalogue $\vec{M}' = S \times \vec{E}'_{ij}$. ($\vec{E}'_{ij}$ is the version of the vector $\vec{E}$ obtained by moving the mutations from the signature i to signature j). The filtering step terminates when all the movement between signatures have a negative impact on the cosine similarity.

The filtering step can thus reduce the “noise” in the DNA sample which may initially result in the attribution of a small number of rearrangements to a signature which is not in fact present. The filtering allows such rearrangement to be reassigned to a signature which is more prevalent.

It is then possible to determine whether the sample exhibits one or more of the rearrangement signatures from the known rearrangement signatures from the number of rearrangements which are present in the sample and which are associated with a particular signature. Different thresholds for this determination may be set depending on the context and the desired certainty of the outcome. Generally the threshold will combine the total number of rearrangements detected in the sample (to ensure that the analysis is representative) along with a proportion of the rearrangements which are associated with a particular signature as determined by the above method.

For example, for data obtained from genomes sequenced to 30-40 fold depth, the requirements for detection may be that there are at least 20, preferably at least 40, more preferably at least 50 rearrangements and a signature is deemed to be present if a proportion of at least 10%, preferably at least 20%, more preferably at least 30% of the rearrangements are associated with it. As indicated below, the proportional thresholds may be adjusted depending on the number of other signatures which make up a significant portion of the rearrangements found in the sample (e.g., if 4 signatures are present each with 25% of the rearrangements, then it may be determined that all 4 are present, rather than no signatures at all are present, even if the general requirement for detection is set higher than 25%).

The rearrangement signatures are generally "additive" with respect to each other (i.e. a tumour may be affected by the underlying mutational processes associated with more than one signature and, if this is the case, a sample from that tumour will generally display a higher overall number of rearrangements (being the sum of the separate rearrangements associated with each of the underlying processes), but with the proportion of rearrangements spread over the signatures which are present). As a result, in determining the presence or absence of a particular signature, attention may be paid to the absolute number of rearrangements associated with a particular signature in the sample (as calculated by the method above). Such alternative requirements for detection can better account for the situation where multiple signatures are present. Under this approach, a signature may be determined to be present if at least 10 and preferably at least 20 rearrangements are associated with it.

The systems and methods of the above embodiments may be implemented in a computer system (in particular in computer hardware or in computer software) in addition to the structural components and user interactions described.

The term "computer system" includes the hardware, software and data storage devices for embodying a system or carrying out a method according to the above described embodiments. For example, a computer system may comprise a central processing unit (CPU), input means, output means and data storage. Preferably the computer system has a monitor to provide a visual output display (for example in the design of the business process). The data storage may comprise RAM, disk drives or other computer readable media. The computer system may include a plurality of computing devices connected by a network and able to communicate with each other over that network.

The methods of the above embodiments may be provided as computer programs or as computer program products or computer readable media carrying a computer program which is arranged, when run on a computer, to perform the method(s) described above.

The term "computer readable media" includes, without limitation, any non-transitory medium or media which can be read and accessed directly by a computer or computer system. The media can include, but are not limited to, magnetic storage media such as floppy discs, hard disc storage media and magnetic tape; optical storage media such as optical discs or CD-ROMs; electrical storage media such as memory, including RAM, ROM and flash memory; and hybrids and combinations of the above such as magnetic/optical storage media.

REFERENCES

1 Stratton, M. R., Campbell, P. J. & Futreal, P. A. The cancer genome. *Nature* **458**, 719-724, doi:10.1038/nature07943 (2009).

2 Nik-Zainal, S. *et al.* Mutational processes molding the genomes of 21 breast cancers. *Cell* **149**, 979-993, doi:10.1016/j.cell.2012.04.024 (2012).

3 Nik-Zainal, S. *et al.* The life history of 21 breast cancers. *Cell* **149**, 994-1007, doi:10.1016/j.cell.2012.04.023 (2012).

4 Hicks, J. *et al.* Novel patterns of genome rearrangement and their association with survival in breast cancer. *Genome research* **16**, 1465-1479, doi:10.1101/gr.5460106 (2006).

5 Bergamaschi, A. *et al.* Extracellular matrix signature identifies breast cancer subgroups with different clinical outcome. *The Journal of pathology* **214**, 357-367, doi:10.1002/path.2278 (2008).

6 Ching, H. C., Naidu, R., Seong, M. K., Har, Y. C. & Taib, N. A. Integrated analysis of copy number and loss of heterozygosity in primary breast carcinomas using high-density SNP array. *International journal of oncology* **39**, 621-633, doi:10.3892/ijo.2011.1081 (2011).

7 Fang, M. *et al.* Genomic differences between estrogen receptor (ER)-positive and ER-negative human breast carcinoma identified by single nucleotide polymorphism array comparative genome hybridization analysis. *Cancer* **117**, 2024-2034, doi:10.1002/cncr.25770 (2011).

8 Curtis, C. *et al.* The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. *Nature* **486**, 346-352, doi:10.1038/nature10983 (2012).

9 Pleasance, E. D. *et al.* A comprehensive catalogue of somatic mutations from a human cancer genome. *Nature* **463**, 191-196, doi:10.1038/nature08658 (2010).

10 Pleasance, E. D. *et al.* A small-cell lung cancer genome with complex signatures of tobacco exposure. *Nature* **463**, 184-190, doi:10.1038/nature08629 (2010).

11 Banerji, S. *et al.* Sequence analysis of mutations and translocations across breast cancer subtypes. *Nature* **486**, 405-409, doi:10.1038/nature11154 (2012).

12 Ellis, M. J. *et al.* Whole-genome analysis informs breast cancer response to aromatase inhibition. *Nature* **486**, 353-360, doi:10.1038/nature11143 (2012).

13 Shah, S. P. *et al.* The clonal and mutational evolution spectrum of primary triple-negative breast cancers. *Nature* **486**, 395-399, doi:10.1038/nature10933 (2012).

14 Alexandrov, L. B. *et al.* Signatures of mutational processes in human cancer. *Nature* **500**, 415-421, doi:10.1038/nature12477 (2013).

15 Alexandrov, L. B., Nik-Zainal, S., Wedge, D. C., Campbell, P. J. & Stratton, M. R. Deciphering signatures of mutational processes operative in human cancer. *Cell reports* **3**, 246-259, doi:10.1016/j.celrep.2012.12.008 (2013).

16 Helleday, T., Eshtad, S. & Nik-Zainal, S. Mechanisms underlying mutational signatures in human cancers. *Nature reviews. Genetics* **15**, 585-598, doi:10.1038/nrg3729 (2014).

17 Birkbak, N. J. *et al.* Telomeric allelic imbalance indicates defective DNA repair and sensitivity to DNA-damaging agents. *Cancer discovery* **2**, 366-375, doi:10.1158/2159-8290.CD-11-0206 (2012).

18 Abkevich, V. *et al.* Patterns of genomic loss of heterozygosity predict homologous recombination repair defects in epithelial ovarian cancer. *British journal of cancer* **107**, 1776-1782, doi:10.1038/bjc.2012.451 (2012).

19 Popova, T. *et al.* Ploidy and large-scale genomic instability consistently identify basal-like breast carcinomas with BRCA1/2 inactivation. *Cancer research* **72**, 5454-5462, doi:10.1158/0008-5472.CAN-12-1470 (2012).

20 Fischer A, Illingworth CJ, Campbell PJ, Mustonen V.; EMu: probabilistic inference of mutational processes and their localization in the cancer genome *Genome Biol.* 2013 Apr 29;14(4):R39. doi: 10.1186/gb-2013-14-4-r39.

21 Sherry, S. T. et al. dbSNP: the NCBI database of genetic variation. Nucleic acids research 29, 308-311 (2001).

22 Abecasis, G. R. et al. An integrated map of genetic variation from 1,092 human genomes. Nature 491, 56-65, doi:10.1038/nature11632 (2012).

23 Fu, W. et al. Analysis of 6,515 exomes reveals the recent origin of most human protein-coding variants. Nature 493, 216-220, doi:10.1038/nature11690 (2013).

All of the above references are hereby incorporated by reference.

TABLE 1

			Probability					
Type	**Class**	**Size**	**Signature 1**	**Signature 2**	**Signature 3**	**Signature 4**	**Signature 5**	**Signature 6**
clustered	deletion	1-10kb	0%	0%	0%	1%	0%	1%
clustered	deletion	10-100kb	0%	0%	0%	1%	0%	1%
clustered	deletion	100kb-1Mb	0%	0%	0%	2%	0%	3%
clustered	deletion	1Mb-10Mb	0%	0%	0%	3%	0%	7%
clustered	deletion	>10Mb	0%	0%	0%	1%	0%	7%
clustered	tandem duplication	1-10kb	0%	0%	0%	0%	0%	0%
clustered	tandem duplication	10-100kb	0%	0%	0%	1%	0%	1%
clustered	tandem duplication	100kb-1Mb	1%	0%	0%	1%	0%	3%
clustered	tandem duplication	1Mb-10Mb	0%	0%	0%	3%	0%	7%
clustered	tandem duplication	>10Mb	0%	0%	0%	1%	0%	7%
clustered	inversion	1-10kb	0%	0%	0%	3%	0%	2%
clustered	inversion	10-100kb	0%	0%	0%	2%	0%	2%
clustered	inversion	100kb-1Mb	0%	0%	0%	3%	0%	5%
clustered	inversion	1Mb-10Mb	0%	0%	0%	6%	0%	15%
clustered	inversion	>10Mb	0%	0%	0%	2%	0%	14%
clustered	translocation		0%	0%	0%	56%	0%	0%
non-clustered	deletion	1-10kb	0%	2%	2%	0%	32%	3%
non-clustered	deletion	10-100kb	1%	1%	0%	0%	22%	2%
non-clustered	deletion	100kb-1Mb	4%	5%	0%	0%	5%	2%

			Probability					
Type	**Class**	**Size**	**Signature 1**	**Signature 2**	**Signature 3**	**Signature 4**	**Signature 5**	**Signature 6**
non-clustered	deletion	1Mb-10Mb	1%	6%	0%	1%	1%	2%
non-clustered	deletion	>10Mb	0%	6%	1%	0%	1%	2%
non-clustered	tandem duplication	1-10kb	0%	0%	53%	0%	1%	0%
non-clustered	tandem duplication	10-100kb	16%	0%	22%	0%	12%	0%
non-clustered	tandem duplication	100kb-1Mb	54%	0%	1%	0%	1%	0%
non-clustered	tandem duplication	1Mb-10Mb	17%	2%	0%	1%	0%	1%
non-clustered	tandem duplication	>10Mb	0%	5%	1%	0%	1%	1%
non-clustered	inversion	1-10kb	1%	5%	1%	1%	5%	1%
non-clustered	inversion	10-100kb	2%	2%	0%	0%	3%	1%
non-clustered	inversion	100kb-1Mb	2%	4%	0%	0%	0%	1%
non-clustered	inversion	1Mb-10Mb	0%	10%	0%	1%	0%	4%
non-clustered	inversion	>10Mb	1%	12%	1%	0%	2%	3%
non-clustered	translocation		1%	39%	16%	7%	13%	1%

CLAIMS

1. A method of detecting rearrangement signatures in a previously obtained DNA sample, the method including the steps of:

cataloguing the somatic mutations in said sample to produce a rearrangement catalogue for that sample which classifies identified rearrangement mutations in the sample into a plurality of categories; and

determining the contributions of known rearrangement signatures to said rearrangement catalogue by computing the cosine similarity between the rearrangement mutations in said catalogue and the rearrangement mutational signatures.

2. The method according to claim 1 wherein the method includes the further step of, prior to said step of determining, filtering the mutations in said catalogue to remove one or more of: residual germline mutations; copy number polymorphisms; and known sequencing artefacts.

3. The method according to claim 2 wherein the filtering uses a list of known germline polymorphisms.

4. The method according to claim 2 wherein the filtering uses BAM files of unmatched normal human tissue sequenced by the same process as the DNA sample and discards any somatic mutation which is present in at least two well-mapping reads in at least two of said BAM files.

5. The method according any one of the preceding claims wherein the classification of the rearrangement mutations includes identifying mutations as being clustered or non-clustered.

6. The method according to claim 5 wherein mutations are identified as being clustered if they have an average density of rearrangement breakpoints that is at least 10 times greater the whole genome average density of rearrangements for an individual patient's sample.

7. The method according to any one of the preceding claims wherein the classification of the rearrangement mutations includes identifying mutations as one of: tandem duplications, deletions, inversions or translocations.

8. The method according to claim 7 wherein the classification of the rearrangement mutations includes grouping mutations identified as tandem duplications, deletions or inversions by size.

9. The method according to any one of the preceding claims further including the step of determining the number of rearrangements E_i in the rearrangement catalogue associated with the *i*th known mutational signature $\vec{S}_i$, which is proportional to the cosine similarity ($\vec{C}_i$) between the catalogue of this sample $\vec{M}$ and $\vec{S}_i$:

$$\vec{C}_i = \frac{\vec{S}_i \cdot \vec{M}}{\| \vec{S}_i \| \| \vec{M} \|}$$

wherein:

$$E_i = \frac{\vec{C}_i}{\sum_{i=1}^{q} \vec{C}_i} \sum_{j=1}^{36} \vec{M}^j$$

wherein $\vec{S}_i$ and $\vec{M}$ are equally-sized vectors with nonnegative components being, respectively, the known rearrangement signature and the rearrangement catalogue and *q* is the number of signatures in said plurality of known rearrangement signatures, and wherein E_i are further constrained by the requirements that

$$0 \leq E_i \leq \left\| \vec{S}_i \right\|_1, i = 1..q, \text{ and } \sum_{i=1}^{q} E_i = \left\| \vec{S}_i \right\|_1 .$$

10. The method according to claim 9 wherein the step of determining the number of rearrangements further includes the step of filtering the number of rearrangements determined to be assigned to each signature by reassigning one or more rearrangements from signatures that are less correlated with the catalogue to signatures that are more correlated with the catalogue.

11. The method according to claim 10 wherein the step of filtering uses a greedy algorithm to iteratively find an alternative assignment of rearrangements to signatures that improves or does not change the cosine similarity between the catalogue $\vec{M}$ and the reconstructed catalogue $\vec{M}' = S \times \vec{E}'_{ij}$, wherein $\vec{E}'_{ij}$ is the version of the vector $\vec{E}$

obtained by moving the mutations from the signature i to signature j, wherein, in each iteration, the effects of all possible movements between signatures are estimated, and the filtering step terminates when all of these possible reassignments have a negative impact on the cosine similarity.

12. The method according to any one of the preceding claims further including the step of determining if the number or proportion of rearrangements in the rearrangement catalogue which are determined to be associated with one of said rearrangement signatures exceeds a predetermined threshold and, if so, determining that said rearrangement signature is present in the sample.

13. A computer program product containing non-transitory memory storing a computer program which, when run on a computer, performs the steps of:

cataloguing the somatic mutations in said sample to produce a rearrangement catalogue for that sample which classifies identified rearrangement mutations in the sample into a plurality of categories;

determining the contributions of known rearrangement signatures to said rearrangement catalogue by computing the cosine similarity between the rearrangement mutations in said catalogue and the rearrangement mutational signatures.

14. A computer having a processor, wherein the processor is configured to:

catalogue the somatic mutations in said sample to produce a rearrangement catalogue for that sample which classifies identified rearrangement mutations in the sample into a plurality of categories;

determine the contributions of known rearrangement signatures to said rearrangement catalogue by computing the cosine similarity between the rearrangement mutations in said catalogue and the rearrangement mutational signatures.

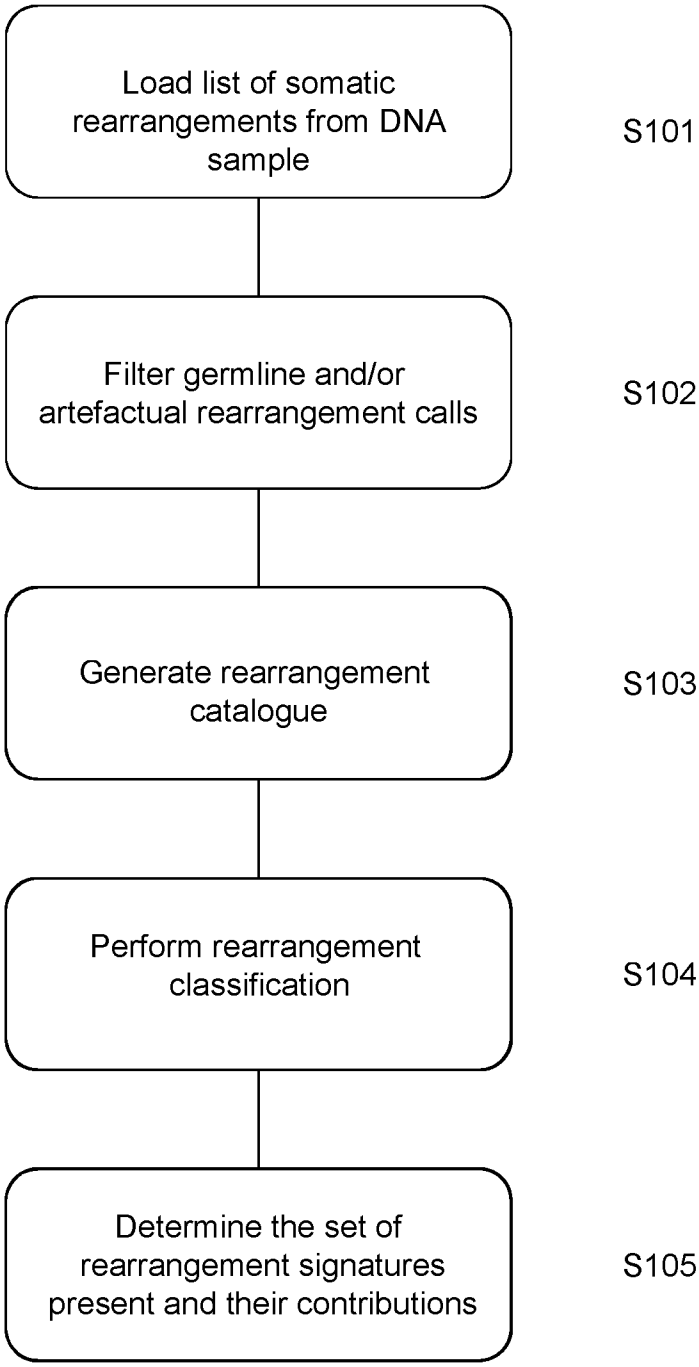
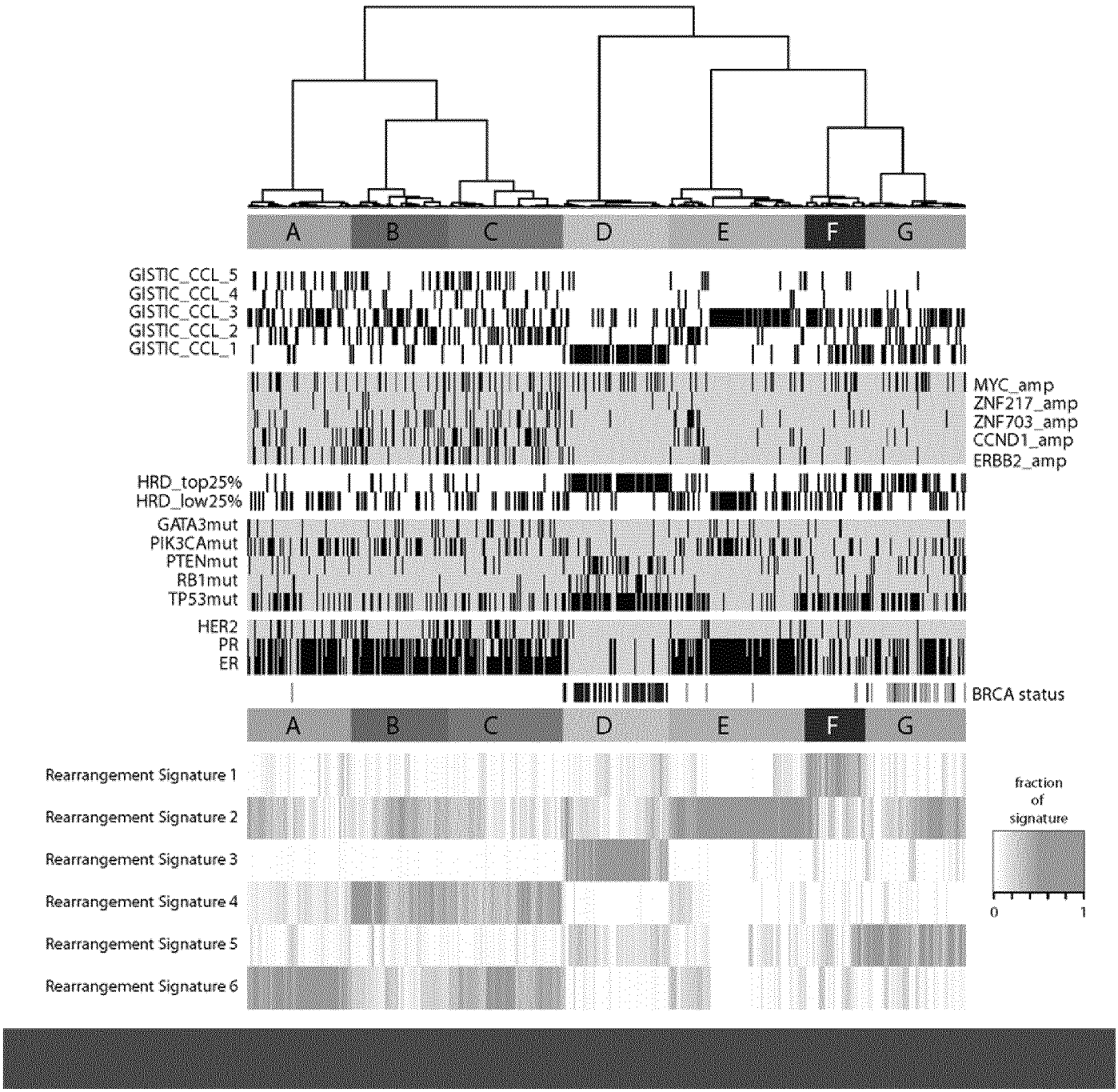

Load list of somatic rearrangements from DNA sample
S101
Filter germline and/or artefactual rearrangement calls
S102
Generate rearrangement catalogue
S103
Perform rearrangement classification
S104
Determine the set of rearrangement signatures present and their contributions
S105

Figure 1

Figure 2


A
B
C
D
E
F
G
GISTIC_CCL_5
GISTIC_CCL_4
GISTIC_CCL_3
GISTIC_CCL_2
GISTIC_CCL_1
MYC_amp
ZNF217_amp
ZNF703_amp
CCND1_amp
ERBB2_amp
HRD_top25%
HRD_low25%
GATA3mut
PIK3CAmut
PTENmut
RB1mut
TP53mut
HER2
PR
ER
BRCA status
A
B
C
D
E
F
G
Rearrangement Signature 1
Rearrangement Signature 2
Rearrangement Signature 3
Rearrangement Signature 4
Rearrangement Signature 5
Rearrangement Signature 6
fraction of signature
0
1

INTERNATIONAL SEARCH REPORT

International application No
PCT/EP2017/060279

A. CLASSIFICATION OF SUBJECT MATTER
INV. G06F19/18 G06F19/24
ADD.

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)
G06F C12Q

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)
EPO-Internal

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	Ludmil b. Alexandrov ET AL: "Deciphering Signatures of Mutational Processes Operative in Human Cancer", Cell Reports, vol.3, no.1, 1 January 2013 (2013-01-01), pages 246-259, XP55355545, DOI: 10.1016/j.celrep.2012.12.008 Retrieved from the Internet: URL:http://api.elsevier.com/content/article/PII:S2211124712004330?httpAccept=text/plain [retrieved on 2017-03-16]	1,4,9, 10,12-14
Y	whole document, in particular p. 247, 248,	2,3,5-8
A	249, 251, 254, 255, 258, fig. 1 and 2.	11

-/--

[X] Further documents are listed in the continuation of Box C.

[X] See patent family annex.

* Special categories of cited documents :

"A" document defining the general state of the art which is not considered to be of particular relevance

"E" earlier application or patent but published on or after the international filing date

"L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)

"O" document referring to an oral disclosure, use, exhibition or other means

"P" document published prior to the international filing date but later than the priority date claimed

"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention

"X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone

"Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art

"&" document member of the same patent family

Date of the actual completion of the international search	Date of mailing of the international search report
21 July 2017	31/08/2017

Name and mailing address of the ISA/	Authorized officer
European Patent Office, P.B. 5818 Patentlaan 2 NL - 2280 HV Rijswijk Tel. (+31-70) 340-2040, Fax: (+31-70) 340-3016	Lüdemann, Susanna

2

Form PCT/ISA/210 (second sheet) (April 2005)

C(Continuation). DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
Y	ERIN D. PLEASANCE ET AL: "A small-cell lung cancer genome with complex signatures of tobacco exposure", NATURE, vol. 463, no. 7278, 16 December 2009 (2009-12-16), pages 184-190, XP055392740, ISSN: 0028-0836, DOI: 10.1038/nature08629 whole doc, in particular p. 185, left and right col., fig. 1a and supplementary fig. 1	2,3
A	WO 2013/190441 A2 (UNIV HONG KONG CHINESE [CN]; CHIU WAI KWUN ROSSA [CN]) 27 December 2013 (2013-12-27) whole doc, in particular [0083]. [0111]	2,3
Y	A. MALHOTRA ET AL: "Breakpoint profiling of 64 cancer genomes reveals numerous complex rearrangements spawned by homology-independent mechanisms", GENOME RESEARCH, vol. 23, no. 5, 14 February 2013 (2013-02-14), pages 762-776, XP055128483, ISSN: 1088-9051, DOI: 10.1101/gr.143677.112 whole doc, in particular abstract and fig. 3	5-8
A	KIN CHAN ET AL: "Clusters of Multiple Mutations: Incidence and Molecular Mechanisms", ANNUAL REVIEW OF GENETICS., vol. 49, no. 1, 23 November 2015 (2015-11-23), pages 243-267, XP055392498, US ISSN: 0066-4197, DOI: 10.1146/annurev-genet-112414-054714 the whole document	1-14

2

INTERNATIONAL SEARCH REPORT

Information on patent family members

International application No
PCT/EP2017/060279

Patent document cited in search report	Publication date	Patent family member(s)	Publication date
WO 2013190441 A2	27-12-2013	AU 2013278994 A1	29-01-2015
		CA 2876327 A1	27-12-2013
		CN 104662168 A	27-05-2015
		EA 201500027 A1	29-05-2015
		EP 2864501 A2	29-04-2015
		HK 1204013 A1	06-11-2015
		HK 1205533 A1	18-12-2015
		JP 2015527057 A	17-09-2015
		KR 20150032708 A	27-03-2015
		SG 11201408113Q A	29-01-2015
		TW 201403066 A	16-01-2014
		WO 2013190441 A2	27-12-2013