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

(19) World Intellectual Property Organization
International Bureau

(43) International Publication Date
23 March 2006 (23.03.2006)

PCT

(10) International Publication Number
WO 2006/031204 A1

(51) International Patent Classification[7]: C12Q 1/68, C12N 15/00, 15/66

(21) International Application Number: PCT/SG2005/000281

(22) International Filing Date: 17 August 2005 (17.08.2005)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:
10/939,592 13 September 2004 (13.09.2004) US

(71) Applicant *(for all designated States except US)*: **AGENCY FOR SCIENCE, TECHNOLOGY AND RESEARCH** [SG/SG]; 20 Biopolis Way, #07-01 Centros, Singapore 138668 (SG).

(72) Inventors; and

(75) Inventors/Applicants *(for US only)*: **SUNG, Wing Kin Ken** [—/SG]; Kent Vale, Block B, Clementi Road 105, #12-03, Singapore 129789 (SG). **RUAN, Yijun** [US/SG]; 123 Verde Crescent, Singapore 688461 (SG).

(74) Agent: **AXIS INTELLECTUAL CAPITAL PTE LTD**; 21A Duxton Road, Singapore 089487 (SG).

(81) Designated States *(unless otherwise indicated, for every kind of national protection available)*: AE, AG, AL, AM, AT, AU, AZ, BA, BB, BG, BR, BW, BY, BZ, CA, CH, CN, CO, CR, CU, CZ, DE, DK, DM, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, HR, HU, ID, IL, IN, IS, JP, KE, KG, KM, KP, KR, KZ, LC, LK, LR, LS, LT, LU, LV, MA, MD, MG, MK, MN, MW, MX, MZ, NA, NG, NI, NO, NZ, OM, PG, PH, PL, PT, RO, RU, SC, SD, SE, SG, SK, SL, SM, SY, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, YU, ZA, ZM, ZW.

(84) Designated States *(unless otherwise indicated, for every kind of regional protection available)*: ARIPO (BW, GH, GM, KE, LS, MW, MZ, NA, SD, SL, SZ, TZ, UG, ZM, ZW), Eurasian (AM, AZ, BY, KG, KZ, MD, RU, TJ, TM), European (AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HU, IE, IS, IT, LT, LU, LV, MC, NL, PL, PT, RO, SE, SI, SK, TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, ML, MR, NE, SN, TD, TG).

Declaration under Rule 4.17:

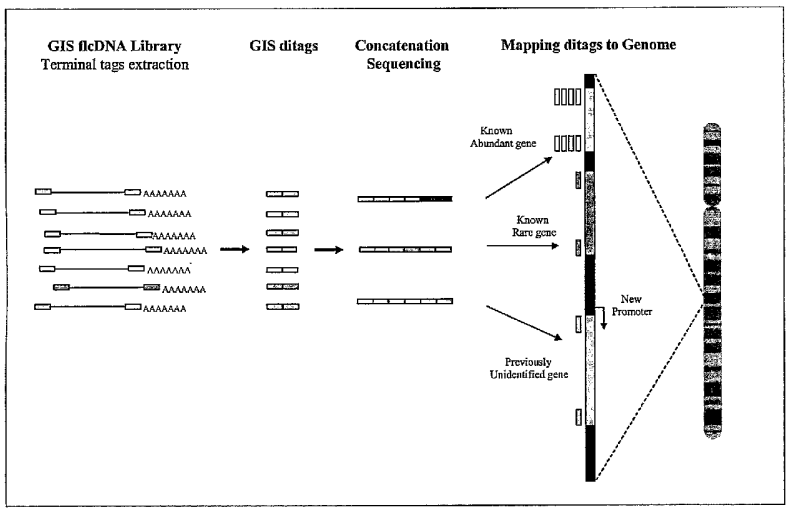
— *as to the applicant's entitlement to claim the priority of the earlier application (Rule 4.17(iii)) for all designations*

Published:

— *with international search report*

For two-letter codes and other abbreviations, refer to the "Guidance Notes on Codes and Abbreviations" appearing at the beginning of each regular issue of the PCT Gazette.

(54) **Title:** GENE IDENTIFICATION SIGNATURE (GIS) ANALYSIS FOR TRANSCRIPT MAPPING



(57) **Abstract:** A transcript mapping method according to an embodiment of the invention is described hereinafter and combines short tag based (SAGE and MPSS) efficiency with the accuracy of full-length cDNA (flcDNA) for comprehensive characterization of transcriptomes. This method is also referred to as Gene Identification Signature (GIS) analysis. In this method, the 5' and 3' ends of full-length cDNA clones are initially extracted into a ditag structure, with the ditag concatemers of the ditag being subsequently sequenced in an efficient manner, and finally mapped to the genome for defining the gene structure. As a GIS ditag represents the 5' and 3' ends of a transcript, it is more informative than SAGE and MPSS tags. Segment lengths between 5' and 3' tag pairs are obtainable including orientation, ordering and chromosome family for efficient transcript mapping and gene location identification. Furthermore, a compressed suffix array (CSA) is used for indexing the genome sequence for improve mapping speed and to reduce computational memory requirements.

WO 2006/031204 A1

GENE IDENTIFICATION SIGNATURE (GIS) ANALYSIS FOR TRANSCRIPT MAPPING

Field Of Invention

The present invention relates generally to a transcript mapping method. In particular, the invention relates to a transcript mapping method for mapping from a transcript to a compressed suffix array indexed genome sequence.

Background

Since the completion of the genome sequences for human and several other organisms, attention has been drawn towards annotation of genomes for functional elements including gene coding transcript units and regulatory cis-acting elements that modulate gene expression levels.

Currently there are three main approaches for genome annotation. The first approach uses existing transcript data to identify gene-coding regions in the genomes, the second approach uses computational algorithms to statistically predict genes and regulatory elements and the third approach compares genomic sequences from other vertebrates for conserved regions based on the view that functional elements in genomes are conserved during evolution.

Despite considerable success, these approaches are unsatisfactory for determining the complete and precise content of all functional elements in the human genome. As a result, a complete list of genes in the human genome is still unavailable. In particular, all the low abundant and cell specific genes have not been identified. Many gene models suggest that the current genome annotation is incorrect, particularly regarding where the transcription starts and ends.

All the gene predictions have to be validated by experimental means and the prospective genes are required to be cloned in full-length for further functional studies. It is therefore clear that many challenges surround the field of human genome annotation.

One of the challenges is the identification of all genes and all transcripts expressed from the genes in human and model organisms. In the annotation of genes, full-length cDNA cloning and sequencing is the most conclusive and is viewed as the gold standard for the analysis of transcripts. However, this approach is expensive and slow when applied to a large number of transcripts across a large number of species and biological conditions. There are short tag based approaches such as SAGE (serial analysis of gene expression) and MPSS (massively parallel signature sequence). These short tag based methods extract a 14-20bp signature for representing each transcript. Though this approach is efficient in tagging and counting transcripts in a given transcriptome, the specificity of the tags is often poor and the information yielded regarding transcript structures are frequently incomplete and ambiguous.

Gene Identification Signature (GIS) ditag sequences, obtained by extracting interlinked 5' and 3' ends of full-length cDNA clones into a ditag structure, provide substantial tag specificity. However, there are no existing computer algorithms that are readily applicable for mapping the GISditag sequences to genome. In the past, SAGE and MPSS tags were analyzed using a two-step approach. The tags were first matched to cDNA sequences and then to the genome. In this approach, novel transcripts that did not exist in cDNA databases would not be mapped. The two most often used sequence alignment tools, BLAST (basic local alignment search tool) and BLAT (BLAST-like alignment tool), are not designed for short tag sequences and often leads to poor or incorrect results.

Hence, this clearly affirms a need for an improved transcript mapping method.

Summary

A transcript mapping method according to an embodiment of the invention is described hereinafter and combines short tag based (SAGE and MPSS) efficiency with the accuracy of full-length cDNA (flcDNA) for comprehensive characterization of transcriptomes. This method is also referred to as Gene Identification Signature (GIS) analysis. In this method, the 5' and 3' ends of full-length cDNA clones are initially extracted into a ditag structure, with the ditag concatemers of the ditag being subsequently sequenced in an efficient manner, and finally mapped to the genome for

defining the gene structure. In the GIS analysis, each sequence read reveals approximately 15 ditags representing 15 transcripts. This approach increases efficiency by at least 30-folds for identifying and quantifying full-length transcripts compared to the current flcDNA cloning and sequencing approaches. Because each GISditag sequence contains 36 base pairs (bp) to represent the beginning and the end of a transcript, the specificity of mapping tag-to-genome is greatly increased compared to the 14-21bp SAGE and MPSS tags. In addition, as a GISditag represents the 5' and 3' ends of a transcript, it is more informative than SAGE and MPSS tags.

To accommodate the GISditag data, a Suffix Array based Tag to Genome (SAT2G) algorithm is used for mapping the GISditag sequences to a genome sequence built and indexed on an advanced data structure Compressed Suffix Array (CSA).

Therefore, in accordance with a first aspect of the invention, there is disclosed a transcript mapping method comprising the steps of:

obtaining a 5' terminal tag and a 3' terminal tag from a transcript of a gene;

matching the 5' terminal tag to at least a portion of a genome sequence to thereby identify at least one 5' site therefrom, each of the at least one 5' site having a sequence matching the 5' terminal tag;

matching the 3' terminal tag to at least a portion of the genome sequence to thereby identify at least one 3' site therefrom, each of the at least one 5' site having a sequence matching the 5' terminal tag;

identifying at least one occurring segment, each of the at least one occurring segment being a sequence segment along the genome sequence extending from one of the at least one 5' site to one of the at least one 3' site, each of the at least one occurring segment having a sequence length; and

identifying at least one feasible gene location, each of the feasible gene location being one of the at least one occurring segment having a sequence length not exceeding that of a predefined gene length.

In accordance with a second aspect of the invention, there is disclosed a transcript mapping system comprising:

means for obtaining a 5' terminal tag and a 3' terminal tag from a transcript of a gene;

means for matching the 5' terminal tag to at least a portion of a genome sequence to thereby identify at least one 5' site therefrom, each of the at least one 5' site having a sequence matching the 5' terminal tag;

means for matching the 3' terminal tag to at least a portion of the genome sequence to thereby identify at least one 3' site therefrom, each of the at least one 3' site having a sequence matching the 3' terminal tag;

means for identifying at least one occurring segment, each of the at least one occurring segment being a sequence segment along the genome sequence extending from one of the at least one 5' site to one of the at least one 3' site, each of the at least one occurring segment having a sequence length; and

means for identifying at least one feasible gene location, each of the feasible gene location being one of the at least one occurring segment having a sequence length not exceeding that of a predefined gene length.

In accordance with a third aspect of the invention, there is disclosed a transcript mapping method comprising the steps of:

obtaining a 5' terminal tag and a 3' terminal tag from a transcript of a gene;

matching the 5' terminal tag to at least a portion of a genome sequence to thereby identify at least one 5' site therefrom, each of the at least one 5' site having a sequence matching the 5' terminal tag;

matching the 3' terminal tag to at least a portion of the genome sequence to thereby identify at least one 3' site therefrom, each of the at least one 5' site having a sequence matching the 5' terminal tag;

identifying at least one occurring segment, each of the at least one occurring segment being a sequence segment along the genome sequence extending from one of the at least one 5' site to one of the at least one 3' site, each of the at least one occurring segment having a sequence length; and

identifying at least one feasible gene location from the at least one occurring segment, each of the at least one feasible gene location being one of the at least one occurring segment with at least one of the sequence length thereof not exceeding that of the predefined gene length, the sequence order thereof and of the at least one 5' site

and one of the at least one 3' site corresponding thereto in accordance with a 5'-occurring segment-3' structure matching the sequence order of the corresponding portion of the genome sequence, the 5' site and one of the at least one 5' site and one of the at least one 3' site corresponding thereto having a 5'-3' orientation, and one of the at least one 5' site and one of the at least one 3' site corresponding to each of the occurring segment being located within the same chromosome.

Brief Description Of The Drawings

Embodiments of the invention are described hereinafter with reference to the following drawings, in which:

FIG. 1 shows a schematic diagram of a 5' and 3' terminal tags SAGE technique for use in genome annotation;

FIG. 2 shows a process flow chart of a transcript mapping method according to an embodiment of the invention;

FIG. 3 shows a schematic diagram of a GIS ditag for application of the transcript mapping technique of FIG. 2 thereto;

FIG. 4 shows a pseudo code "Find_Sites" of the transcript mapping method of FIG. 2 for forward and reverse searching of 5' sites and 3' sites from a genome sequence; and

FIG. 5 shows a pseudo code "Match_sites_1" of the transcript mapping method of FIG. 2 for identifying the sequence length of an occurring segment, the sequence length being subsequently compared with a predefined length for identifying of a feasible gene location therefrom; and

FIG. 6 shows a pseudo code "Match_sites_2" of the transcript mapping method of FIG. 2 for identifying an occurring segment when a disparity condition is met wherefrom a feasible gene location is subsequently obtained.

Detailed Description

A transcript mapping method is described hereinafter for addressing the foregoing problems.

Complete genome annotation relies on precise identification of transcription units bounded by a transcription initiation site (TIS) and a polyadenylation site (PAS). To facilitate this, a pair of complementary methods, namely 5'LongSAGE (long serial analysis of gene expression) and 3'LongSAGE, was developed. These methods are based on the original SAGE (serial analysis of gene expression) and LongSAGE methods that utilize typical full-length cDNA cloning technologies to enable high-throughput extraction of the first and the last 20 base pairs (bp) of each transcript. Mapping of 5' and 3' LongSAGE tags to the genome allows the localization of the TIS and the PAS.

However, matching of 5' and 3' tags derived from same transcripts in genome sequences are not always straightforward and can sometimes be very ambiguous. A solution is to clone the 5' and 3' tags of the same transcript by inter-linking the 5' and 3' tags. To achieve this, a specially designed device comprising cloning adapters and a vector links the 5' tag and the 3' tag derived from the same transcript into a ditag.

A plurality of ditags can be concatenated for cloning and sequencing, with each ditag representing an individual transcript. Unlike single tag sequences, the paired ditag sequences can be specifically multiplied with a frame of transcripts being precisely definable when being mapped to the genome sequences. This approach, named Gene Identification Signature (GIS) analysis, can accurately map the 5' and 3' ends of transcription units encoded by genes as shown in FIG. 1.

In the GIS analysis, the conventional cap-trapper method is applied to enrich a full-length cDNA and incorporated adapter sequences that bear an MmeI restriction site at each end of the cDNA fragments. The cDNA fragments are then cloned in a cloning vector to construct a GIS flcDNA (full-length cDNA) library. The plasmid prepared from the library is digested by MmeI (a type II restriction enzyme) and cleaved 20bp

downstream of its binding site. After digestion, the flcDNA inserts of the library were dropped from the plasmid to leave 18bp signatures of 5' and 3' ends with the learned cloning vector. Re-circling the vector would create a GIS single ditag library. The ditags of the library were then sliced out and purified for concatenating and cloning to generate the final GIS ditag library for sequence analysis. Typically, each sequence read of the GIS ditag clones reveals 15 ditags. Each unit of the ditag sequence contains 18bp of 5' and 18bp of 3' signatures with a 12bp spacer to separate one ditag sequence from another.

Mapping ditags to the genome is akin to searching occurrences of a pattern in the genome sequence. Approaches for pattern searching include the conventional BLAST (basic local alignment search tool) and BLAT (BLAST-like alignment tool) method. Both the BLAST and BLAT methods are very slow because each thereof requires a pattern to be searched by scanning through the whole genome. Moreover, conventional full-text indexing is usually employed if exact occurrences of a pattern with a small mismatch margin are required. Efficient full-text indexing data-structures include a suffix tree and a suffix array.

A suffix tree is a tree-like data-structure having branches stemming from a root with each branch terminating at a leaf that encodes a suffix of the genome sequence. The suffix array is a sorted sequence of all suffices of the genome according to lexicographic order. The suffix array is represented as an array SA[i] where i=1...n and that SA[i] = j means that the j-suffix (suffix starting from character j) is the i-th smallest suffix in the lexicographic order.

Both the suffix tree and the suffix array allow for fast pattern searching. Given a pattern of length x, its occurrences in the genome G[1...n] can be reported in O(x) time and O(x log n) time for the suffix tree and the suffix array respectively. Although the query time is fast, it is not always feasible to build the suffix tree or the suffix array due to large space requirements thereof. For example, for a mouse genome, the suffix tree and the suffix array require 40 Gigabytes (GB) and 13 GB respectively. Such memory requirement far exceeds the memory space capacity of ordinary computers. To solve the memory space problem, we apply the space-

efficient compressed suffix array (CSA) indexing data structure. CSA is a compressed form of the suffix array. It can be built efficiently without need for enormous memory requirements using known algorithms. Also, the built CSA is very small. For example, a CSA for the mouse genome (mm3) occupies approximately 1.3 GB. Additionally, CSA is also able to support searching efficiently. Searching a pattern of length x requires only O(x log n) time.

A first embodiment of the invention, a transcript mapping method 20 is described with reference to FIG. 2, which shows a process flow chart of the transcript mapping method 100. The transcript mapping method 100 is for application to a transcript obtained from a gene. The transcript mapping method 100 is preferably implemented using a computer-based system. In a step 110 of the transcript mapping method 100, a 5' terminal tag 24 and a 3' terminal tag 26 are obtained from the transcript.

In combination, the 5' terminal tag 24 and the 3' terminal tag 26 forms a GIS ditag 30 as described above and as shown in FIG. 3. The GIS ditag 30 has a ditag length 32 of 36bp with 18bp nucleotide sequence being derived from the 5' terminal tag 24 and another 18bp of nucleotide sequence being derived from the 3' terminal tag 26. Due to some enzymatic variations during molecular cloning, the ditag length 32 of the GIS ditag 30 may vary from 34bp to 38bp.

This variation often occurs proximate to the extremities of the 5' terminal tag 24 and the 3' terminal tag 26 with the internal nucleotides remaining structurally conserved. In the 3' terminal tag 26, two residual nucleotides 34 (AA) are retained during poly-A tail removal therefrom. The AA residual nucleotides 34 are subsequently for use as an orientation indicator. Therefore, only 16 bp of the 3' terminal tag 26 in the GIS ditag 30 is usuable for mapping to a genome sequence 36.

Following the step 110, each of the 5' terminal tag 24 and the 3' terminal tag 26 is matched to the genome sequence 36 in a step 112. In the step 112, 5' sites 38 and 3' sites 40 are identified when the 5' terminal tag 24 and the 3' terminal tag 26 are respectively matched to the genome sequence 36. Each of the 5' sites 38 and each of

the 3' sites 40 is a portion of the genome sequence 36 that has a sequence that substantially matches the 5'terminal tag 24 and the 3' terminal tag 26 respectively.

In a step 114, at least one occurring segment 42 is identified from the genome sequence 36. Each of the at least one occurring segment 42 is a sequence segment along the genome sequence 36 situated between one 5' site 38 and one 3' site 40. Each of the at least one occurring segment 42 has a sequence length 44.

Given the GIS ditag 30 (P) for the transcript (R), the computational problem of locating R in the genome sequence 36 (G) is referred to as a transcript location identification problem. Therefore, given G[1...n] and P[1...m], the occurring segment 42 is identified as being a feasible gene location of P when: the sequence length 44 (j-i) is smaller than the predefined gene length (*maxlength*), which is typically less than 1 million base pairs in length for known genes; the 5' terminal tag 24 and the 3' terminal tag 26 are longer than predefined $minlength_5$ and $minlength_3$ respectively (where $minlength_5$ = 16 bp and $minlength_3$ = 14 bp); and the 5' terminal tag 24 and the 3' terminal tag 26 of R are the substrings of P[1...$boundary_5$] and P [$boundary_3$...m] respectively (where $boundary_5$ = 19 and $boundary_3$ = 18).

The genome sequence 36 is preferably indexed using a compressed suffix array (CSA). The 5' terminal tag 24 and the 3' terminal tag are matched to the genome sequence 36 preferably by applying binary search to the compressed suffix array. The binary search for matching the 5' terminal tag 24 and the 3' terminal tag 26 are dependent on two lemmas, namely, lemma 1 for performing a forward search on the compressed suffix array and lemma 2 for performing a reverse search on the compressed suffix array.

Lemma 1 (forward search): given the CSA for the genome G[1..n] and a set of occurrences of a pattern Q in G, for any base c∈ {adebine (A), cytosine (C), guanine (G), thymine (T)}, a set of occurrences of the pattern Qc is obtainable in O(log n) time. A forward binary search is achieved by modifying a conventional binary search algorithm to use values in the compressed suffix array and suffix array instead of

explicit text for the suffixes in the genome sequence 36 when comparing with pattern Q in the binary search.

Lemma 2 (reverse search): given the CSA for the genome G[1...n] and a set of occurrences of a pattern Q in G, for any base $c \in \{A, C, G, T\}$, we can find the set of occurrences of the pattern cQ using O(log n) time.

The pseudo code "Find_Sites" for both the forward search and the reverse search is shown in Fig. 4. Instead of applying both the forward search and the reverse search in tandem in the step 114, an alternative approach is to apply either only the forward search using lemma 1 or only the reverse search using lemma 2 to the genome sequence 36 for identifying the at least one occurring segment 42.

The GIS ditag 30 may appear in the genome sequence 36 in sense or anti-sense. To address this issue, an index is created for each of the sense genome sequence and the anti-sense genome sequence. Instead of creating two separated indexing arrays, an anti-sense GIS ditag is created. The suffix array is searched twice in the step 110 for each of the 5' terminal tag 24 and the 3' terminal tag 26, once using the sense GIS ditag 30 and a second time using the anti-sense GIS ditag (not shown).

Additionally, the genome sequence 36 is naturally partitioned into chromosomes. This enables a compressed suffix array to be created for the sequence segment of each chromosome. By doing so, 5' sites 38 and 3' sites 40 are obtainable for specific chromosomes instead of the entire genome sequence 36.

Besides the compressed suffix array, a suffix array, a suffix tree, a binary or the like indexing data structure is usable for indexing the genome sequence 36 as abovementioned.

Following the step 114, the 5' sites 38 and the 3' sites 40 undergo a series of checks to identify a feasible gene location. The checks comprises of length, locality, orientation and ordering checks.

In a step 116, the length check is performed by comparing the sequence length 44 of each of the at least one occurring segment 42 with a predefined gene length in a step 116. Initially, the 5' sites 38 and 3' sites 40 are sorted preferably in an ascending order. Next, each of the at least one occurring segment 42 has a sequence length 44 that does not exceed that of the predefined gene length (*maxlength*) is identified as a potential feasible gene location. The pseudo code "Match_sites_1" for step 116 is shown in Fig. 5.

In a step 118, the locality check is performed whereby the 5' site 38 and the 3' site 40 corresponding to each of the at least one occurring segment 42 are analysed to identify which chromosome each of them are located within. The occurring segment 42 identifies a potential feasible gene location only when the 5' site 38 and the 3' site 40 thereof belongs to the same chromosome.

In a step 120, the orientation check is performed by identifying the orientation of the 5' site 38 and the 3' site 40 that corresponds to each occurring segment 42. The orientation of the 5' site 38 and the 3' site is identifiable by locating the position of the residual nucleotide 34. Preferably, the 5' site 38 and the 3' site 40 should have a 5'-3' orientation for the occurring segment 42 thereof to identify a potential feasible gene location.

In a step 122, the ordering check is performed by comparing each of the occurring segments 42 and the corresponding 5' site 38 and 3' site 40 to the genome sequence 36. Preferably, the ordering of each of the occurring segments 42 and its corresponding 5' site 38 and 3' site 40 should follow a 5'-occurring segment-3' structure for it to be a potential feasible site.

Steps 116-122 of the transcript mapping method can occur in any sequence in combination or independently.

In a situation where the feasible gene location is not found from the GIS ditag 30, the constraints are relaxed to allow at least one mismatch when matching the 3' terminal tag 26 to the genome sequence 36 in the step 112.

Alternatively, the quantity of the 5' sites 38 and the quantity of the 3' sites 40 are initially obtained before the 5' sites 38 and the 3' sites 40 are matched to the genome sequence 36 in the step 112. This enables identification of quantity disparity between the 5' sites 38 and the 3' sites 40, for example, when there only exist less than ten of the 5' sites 38 and more than tens of thousand of the 3' sites 40, or vice versa.

When large quantity disparity between the 5' sites 38 and the 3' sites 40 exists, the transcript mapping method 20 undergoes multiple iterations of redundant mapping to the genome sequence 36. Therefore, a modified approach is required for the transcript mapping method 100 when a large quantity disparity arises. To identify the quantity disparity, a disparity condition is established as:

$$\frac{1}{threshold_{5,3}} \geq \frac{count_5}{count_3} \geq threshold_{5,3}$$

where $count_5$ is the quantity of 5' sites 38, $count_3$ is the quantity of 3' sites 40, and $threshold_{5,3}$ is a pre-defined threshold, for example $threshold_{5,3} = 10{,}000$, for limiting the quantitative disparity between $count_5$ and $count_3$. The CSA enables both $count_5$ and $count_3$ to be obtained without enumerating either any of the 5' sites 38 or any of the 3' sites.

The method described in the pseudo code "Match_sites_2" of FIG. 6 is applied when the above disparity condition is met. In the pseudo code "Match_sites_2", the number of iterations required for mapping to the genome sequence 36 is determined by the smaller one of $count_5$ and $count_3$. For example, should there be only two 5' sites 38, the mapping to or traversal along the genome sequence 36 for obtaining the corresponding one of the 3' sites 40 is only iterated twice, once for each of the two 5' sites 38, for obtaining the occurring segments 42 therefrom.

However, should the above disparity condition be unmet, the quantity disparity between $count_5$ and $count_3$ is not large and therefore the transcript mapping method

100 reverts to the method described in "Match_sites_1" for obtaining the occurring segments 42.

In the foregoing manner, a transcript mapping method is described according to one embodiment of the invention for addressing the foregoing disadvantages of conventional mapping methods. Although only one embodiment of the invention are disclosed, it will be apparent to one skilled in the art in view of this disclosure that numerous changes and/or modification can be made without departing from the scope and spirit of the invention.

Claims

1. A transcript mapping method comprising the steps of:

obtaining a 5' terminal tag and a 3' terminal tag from a transcript of a gene;

matching the 5' terminal tag to at least a portion of a genome sequence to thereby identify at least one 5' site therefrom, each of the at least one 5' site having a sequence matching the 5' terminal tag;

matching the 3' terminal tag to at least a portion of the genome sequence to thereby identify at least one 3' site therefrom, each of the at least one 5' site having a sequence matching the 5' terminal tag;

identifying at least one occurring segment, each of the at least one occurring segment being a sequence segment along the genome sequence extending from one of the at least one 5' site to one of the at least one 3' site, each of the at least one occurring segment having a sequence length; and

identifying at least one feasible gene location, each of the feasible gene location being one of the at least one occurring segment having a sequence length not exceeding that of a predefined gene length.

2. The transcript mapping method as in claim 1, the step of identifying a 5' terminal tag and a 3' terminal tag comprising the step of:

providing a nucleotide sequence with at least 16 base pairs for forming the 5' terminal tag; and

providing a nucleotide sequence with at least 16 base pairs for forming the 3' terminal tag.

3. The transcript mapping method as in claim 1, the step of matching the 5' terminal tag to at least a portion of a genome sequence comprising the step of:

matching the 5' terminal tag to a chrosmosome sequence.

4. The transcript mapping method as in claim 3, the step of matching the 3' terminal tag to at least a portion of the genome sequence comprising the step of:

matching the 3' terminal tag to a chrosmosome sequence.

5. The transcript mapping method as in claim 1, further comprising the step of generating a data structure for indexing the genome sequence.

6. The transcript mapping method as in claims 1, further comprising the step of generating at least one of a tree structure and an ordered array for indexing the genome sequence.

7. The transcript mapping method as in claims 1, further comprising the step of generating at least one of a suffix array, a suffix tree, a binary tree and a compressed suffix array for indexing the genome sequence.

8. The transcript mapping method as in claim 7, the step of the matching the 5' terminal tag to at least a portion of a genome sequence comprising the step of:

at least one of forward traversing and reverse traversing the genome sequence for comparing the 5' terminal tag to at least a portion of the genome sequence to obtain the at least one 5' site.

9. The transcript mapping method as in claim 8, the step of the matching the 3' terminal tag to at least a portion of a genome sequence comprising the step of:

at least one of forward traversing and reverse traversing the genome sequence for comparing the 3' terminal tag to at least a portion of the genome sequence to obtain the at least one 3' site.

10. The transcript mapping method as in claim 1, the step of identifying at least one feasible gene location comprising the step of comparing sequence order of each of the at least one occurring segment and one of the at least one 5' site and one of the at least one 3' site corresponding thereto to at least a portion of the genome sequence for obtaining the at least one feasible gene location therefrom.

11. The transcript mapping method as in claim 10, the step of comparing sequence order of each of the at least one occurring segment and one of the at least one 5' site and one of the at least one 3' site corresponding thereto comprising the step of comparing the sequence order of each of the at least one occurring segment and one of the at least one 5' site and one of the at least one 3' site corresponding thereto being in accordance with a 5'-occurring segment-3' structure.

12. The transcript mapping method as in claim 1, the step of identifying at least one feasible gene location comprising the step of identifying the 5'-3' orientation of each of the at least one occurring segment for obtaining the at least one feasible gene location therefrom.

13. The transcript mapping method as in claim 12, the step of identifying the 5'-3' orientation comprising the step of identifying a residual AA neucleotide, the residual AA neucleotide constituting a portion of the 3' terminal tag.

14. The transcript mapping method as in claim 1, the step of identifying at least one feasible gene location comprising the step of:

identifying the chromosome wherein each of one of the at least one 5' site and one of the at least one 3' site corresponding to each of the occurring segment is located for identifying the at least one feasible gene location therefrom.

15. The transcript mapping method as in claim 1, the step of matching the 5' terminal tag to at least a portion of a genome sequence comprising the step of:

identifying quantity of the at least one 5' site, and

the step of matching the 3' terminal tag to at least a portion of a genome sequence comprising the step of:

identifying quantity of the at least one 3' site.

16. The transcript mapping method as in claim 15, the step of identifying at least one occurring segment comprising the step of:

traversing along the genome sequence towards one of the extremities thereof from each of the at least one 5' site for identifying at least one of the at least one 3' site.

17. The transcript mapping method as in claim 16, the step of identifying the at least one feasible gene location comprising the step of:

terminating traversal along the genome sequence in response to one of the at least one feasible gene location being identified for each of the at least one 5' site.

18. The transcript mapping method as in claim 15, the step of identifying at least one occurring segment comprising the step of:

traversing along the genome sequence towards one of the extremities thereof from each of the at least one 3' site for identifying at least one of the at least one 5' site.

19. The transcript mapping method as in claim 18, the step of identifying the at least one feasible gene location comprising the step of:

terminating traversal along the genome sequence in response to one of the at least one feasible gene location being identified for each of the at least one 3' site.

20. A transcript mapping system comprising:

means for obtaining a 5' terminal tag and a 3' terminal tag from a transcript of a gene;

means for matching the 5' terminal tag to at least a portion of a genome sequence to thereby identify at least one 5' site therefrom, each of the at least one 5' site having a sequence matching the 5' terminal tag;

means for matching the 3' terminal tag to at least a portion of the genome sequence to thereby identify at least one 3' site therefrom, each of the at least one 3' site having a sequence matching the 3' terminal tag;

means for identifying at least one occurring segment, each of the at least one occurring segment being a sequence segment along the genome

sequence extending from one of the at least one 5' site to one of the at least one 3' site, each of the at least one occurring segment having a sequence length; and

means for identifying at least one feasible gene location, each of the feasible gene location being one of the at least one occurring segment having a sequence length not exceeding that of a predefined gene length.

21. The transcript mapping system as in claim 20, the means for identifying a 5' terminal tag and a 3' terminal tag comprising:

means for providing a nucleotide sequence with at least 16 base pairs for forming the 5' terminal tag; and

means for providing a nucleotide sequence with at least 16 base pairs for forming the 3' terminal tag.

22. The transcript mapping system as in claim 20, the means for matching the 5' terminal tag to at least a portion of a genome sequence comprising:

means for matching the 5' terminal tag to a chrosmosome sequence.

23. The transcript mapping system as in claim 22, the means for matching the 3' terminal tag to at least a portion of the genome sequence comprising:

means for matching the 3' terminal tag to a chrosmosome sequence.

24. The transcript mapping system as in claim 20, further comprising:

means for generating a data structure for indexing the genome sequence.

25. The transcript mapping system as in claims 20, further comprising:

means for generating at least one of a tree structure and an ordered array for indexing the genome sequence.

26. The transcript mapping system as in claims 20, further comprising:

means for generating at least one of a suffix array, a suffix tree, a binary tree and a compressed suffix array for indexing the genome sequence.

27. The transcript mapping system as in claim 26, the means for matching the 5' terminal tag to at least a portion of a genome sequence comprising:

means for at least one of forward traversing and reverse traversing the genome sequence for comparing the 5' terminal tag to at least a portion of the genome sequence to obtain the at least one 5' site.

28. The transcript mapping system as in claim 27, the means for matching the 3' terminal tag to at least a portion of a genome sequence comprising:

means for at least one of forward traversing and reverse traversing the genome sequence for comparing the 3' terminal tag to at least a portion of the genome sequence to obtain the at least one 3' site.

29. The transcript mapping system as in claim 20, the means for identifying at least one feasible gene location comprising:

means for comparing sequence order of each of the at least one occurring segment and one of the at least one 5' site and one of the at least one 3' site corresponding thereto to at least a portion of the genome sequence for obtaining the at least one feasible gene location therefrom.

30. The transcript mapping system as in claim 29, the means for comparing sequence order of each of the at least one occurring segment and one of the at least one 5' site and one of the at least one 3' site corresponding thereto comprising the means for comparing the sequence order of each of the at least one occurring segment and one of the at least one 5' site and one of the at least one 3' site corresponding thereto being in accordance with a 5'-occurring segment-3' structure.

31. The transcript mapping system as in claim 20, the means for identifying at least one feasible gene location comprising:

means for identifying the 5'-3' orientation of each of the at least one occurring segment for obtaining the at least one feasible gene location therefrom.

32. The transcript mapping system as in claim 31, the means for identifying the 5'-3' orientation comprising:

means for identifying a residual AA neucleotide, the residual AA neucleotide constituting a portion of the 3' terminal tag.

33. The transcript mapping system as in claim 20, the means for identifying at least one feasible gene location comprising:

means for identifying the chromosome wherein each of one of the at least one 5' site and one of the at least one 3' site corresponding to each of the occurring segment is located for identifying the at least one feasible gene location therefrom.

34. The transcript mapping system as in claim 20, the means for matching the 5' terminal tag to at least a portion of a genome sequence comprising:

means for identifying quantity of the at least one 5' site, and

the means for matching the 3' terminal tag to at least a portion of a genome sequence comprising:

means for identifying quantity of the at least one 3' site.

35. The transcript mapping system as in claim 34, the means for identifying at least one occurring segment comprising:

means for traversing along the genome sequence towards one of the extremities thereof from each of the at least one 5' site for identifying at least one of the at least one 3' site.

36. The transcript mapping system as in claim 35, the means for identifying the at least one feasible gene location comprising:

means for terminating traversal along the genome sequence in response to one of the at least one feasible gene location being identified for each of the at least one 5' site.

37. The transcript mapping system as in claim 34, the means for identifying at least one occurring segment comprising:

means for traversing along the genome sequence towards one of the extremities thereof from each of the at least one 3' site for identifying at least one of the at least one 5' site.

38. The transcript mapping system as in claim 37, the means for identifying the at least one feasible gene location comprising:

means for terminating traversal along the genome sequence in response to one of the at least one feasible gene location being identified for each of the at least one 3' site.

39. A transcript mapping method comprising the steps of:

obtaining a 5' terminal tag and a 3' terminal tag from a transcript of a gene;

matching the 5' terminal tag to at least a portion of a genome sequence to thereby identify at least one 5' site therefrom, each of the at least one 5' site having a sequence matching the 5' terminal tag;

matching the 3' terminal tag to at least a portion of the genome sequence to thereby identify at least one 3' site therefrom, each of the at least one 5' site having a sequence matching the 5' terminal tag;

identifying at least one occurring segment, each of the at least one occurring segment being a sequence segment along the genome sequence extending from one of the at least one 5' site to one of the at least one 3' site, each of the at least one occurring segment having a sequence length; and

identifying at least one feasible gene location from the at least one occurring segment, each of the at least one feasible gene location being one of the at least one occurring segment with at least one of the sequence length thereof not exceeding that of the predefined gene length, the sequence order thereof and of the at least one 5' site and one of the at least one 3' site corresponding thereto in accordance with a 5'-occurring segment-3' structure matching the sequence order of the corresponding portion of the genome sequence, the 5' site and one of the at least one 5' site and one of the at least

one 3' site corresponding thereto having a 5'-3' orientation, and one of the at least one 5' site and one of the at least one 3' site corresponding to each of the occurring segment being located within the same chromosome.

5

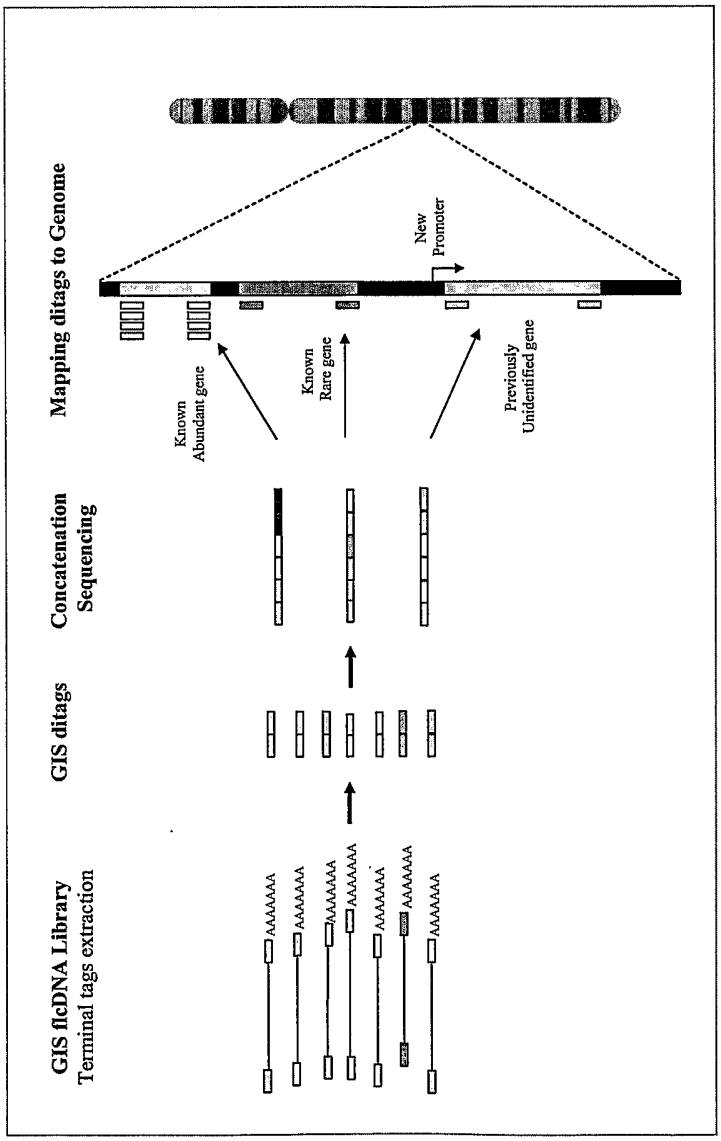

GIS flcDNA Library
Terminal tags extraction
GIS ditags
Concatenation
Sequencing
Mapping ditags to Genome
Known
Abundant gene
Known
Rare gene
New
Promoter
Previously
Unidentified gene

FIG. 1

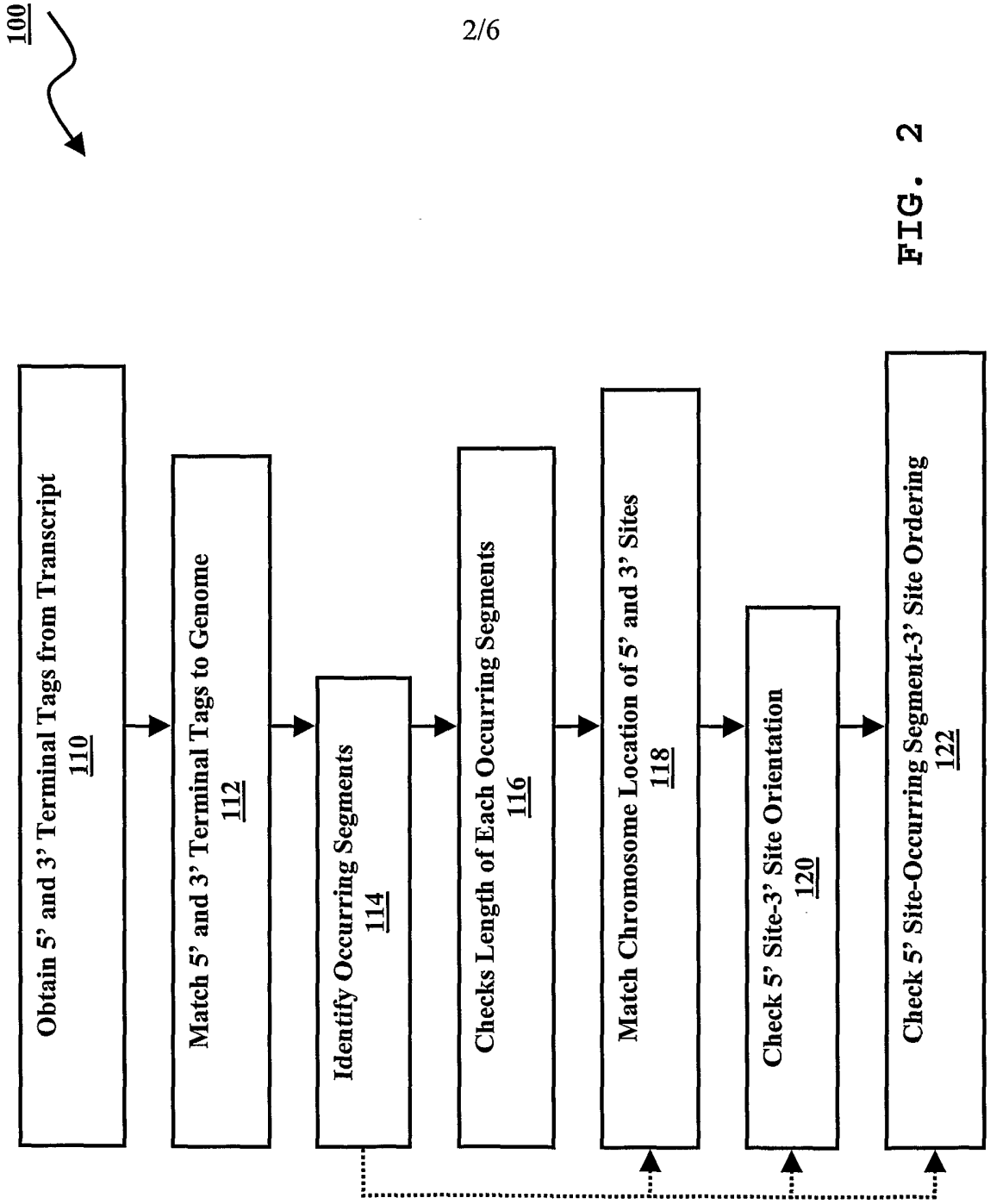

100
Obtain 5' and 3' Terminal Tags from Transcript
110
Match 5' and 3' Terminal Tags to Genome
112
Identify Occurring Segments
114
Checks Length of Each Occurring Segments
116
Match Chromosome Location of 5' and 3' Sites
118
Check 5' Site-3' Site Orientation
120
Check 5' Site-Occurring Segment-3' Site Ordering
122

FIG. 2

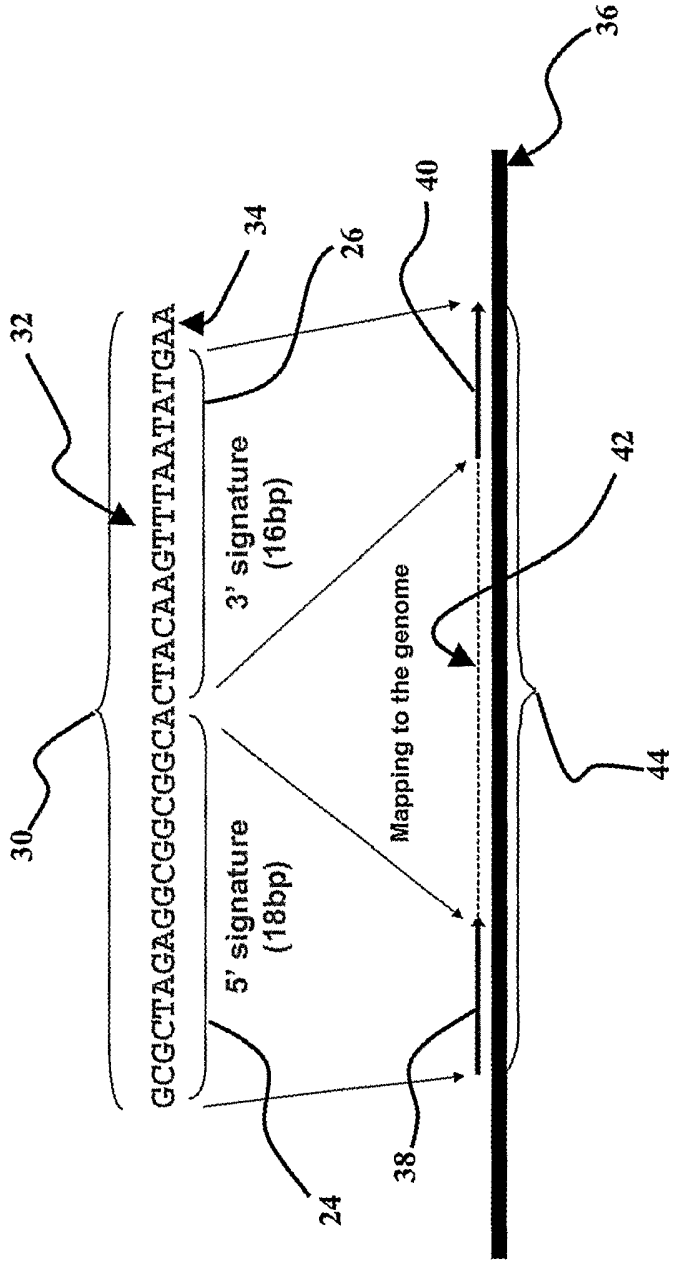

GCGCTAGAGGCGGCGGCACTACAAGTTTAATATATGAA
3' signature (16bp)
5' signature (18bp)
Mapping to the genome
30
32
34
26
40
36
42
44
24
38

FIG. 3

Algorithm Find_Sites (5' sites)
Input: A DNA segment P[1..m], a genome G[1..n], and a threshold $minlength_5$ of minimum contiguous exact match.
Output: A set of 5' sites.

for i = $minlength_5$ downto 1 **do**
Given the occurrence range of P[i+1..$minlength_5$], we find the occurrence range of P[i..$minlength_5$] by Backward search (Lemma 2).
Extend and report all the occurrences of P[1..$minlength_5$]
for i = 2 to $boundary_5$ - $minlength_5$ + 1 **do**
for j= $minlength_5$+1 to i+$minlength_5$-1 **do**
Given the occurrence range of P[i..j-1], we find the occurrence range of P[i..j] by Forward search (Lemma 1).
Extend and report all occurrences of P[i..i+$minlength_5$-1]

End Find_Sites (5' sites)

FIG. 4

```
Algorithm Match_Sites_1
Input: A set of 5' sites and 3' sites.
Output: A set of occurring segments (feasible gene locations).

Sort the 5' sites in ascending order and let A[1], A[2], ..., A[p] be the list of 5' sites.
Sort the 3' sites in ascending order and let B[1], B[2], ..., B[q] be the list of 3' sites.
Let i=1 and j=1.
while (i ≤ p and j ≤ q) do
if (A[i]>B[j]) then
j = j + 1;
elseif (B[j]-A[i] ≤ maxlength)
report (A[i], B[j]) as an occurring segment and j=j+1;
else
i = i + 1;

End Match_Sites_1
```

FIG. 5

Algorithm Match_Sites_2
Input: A set of 5' sites and 3' sites.
Output: A set of occurring segments.

Let i=1 and j=1.

if ($count_5$ << $count_3$) **then**
while (i ≤ p) **do**
find 3' site located within *maxlength* downstream from A[i]
report found 3' site and A[i] as an occurring segment
i = i + 1;;
elseif ($count_5$ >> $count_3$)
while (j ≤ q) **do**
find 5' site located within *maxlength* upstream from B[j]
report found 5' site and B[j] as an occurring segment
j = j + 1;;
else
try Match_Sites_1

End Match_Sites_2

FIG. 6

INTERNATIONAL SEARCH REPORT

International application No.
PCT/SG2005/000281

A. CLASSIFICATION OF SUBJECT MATTER

Int. Cl. [7]: C12Q 1/68, C12N 15/00, C12N 15/66

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)

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)
WPIDS, MEDLINE, CAPLUS, BIOTECHABS (5', 3', ditag, di tag, map?, annotat?, chromosom?, transcriptom?)

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
P, X	WO 2005026391 A1 (AGENCY FOR SCIENCE, TECHNOLOGY AND RESEARCH) 24 March 2005. See whole document.	1 – 4, 10 – 19, 39
P, X	Ng, P. *et al.* 2005. Gene identification signature (GIS) analysis for transcriptome characterization and genome annotation. Nature Methods. 2(2): 105 – 111. See whole document.	1 – 4, 10 – 19, 39
X	Wei, C-L. *et al.* 2004. 5' long serial analysis of gene expression (LongSAGE) and 3' LongSAGE for transcriptome characterization and genome annotation. PNAS 101(32): 11701 – 11706. See pg's 11701 – 11702, Fig. 1.	1 – 4, 10 – 19, 39

[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
7 October 2005	14 OCT 2005
Name and mailing address of the ISA/AU AUSTRALIAN PATENT OFFICE PO BOX 200, WODEN ACT 2606, AUSTRALIA E-mail address: pct@ipaustralia.gov.au Facsimile No. (02) 6285 3929	Authorized officer **CHRIS LUTON** Telephone No : (02) 6283 2256

INTERNATIONAL SEARCH REPORT

International application No.
PCT/SG2005/000281

C (Continuation). DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
A	WO 2004099445 A1 (IWATE PREFECTURAL GOVERNMENT) 18 November 2004. See whole document.	1 – 19, 39
A	WO 2004050918 A1 (AGENCY FOR SCIENCE, TECHNOLOGY AND RESEARCH) 17 June 2004. See whole document.	1 – 19, 39
A	US 20040096892 A1 (WANG ET AL) 20 May 2004. See whole document.	1 – 19, 39
A	Gowda, M. *et al.* 2004. Robust-LongSAGE (RL-SAGE): A Substantially Improved LongSAGE Method for Gene Discovery and Transcriptome Analysis. Plant Physiology. 134(3): 890 – 897. See whole document.	1 – 19, 39
A	Ruan, Y. *et al.* 2004. Interrogating the transcriptome. TRENDS in Biotechnology. 22(1): 23 – 30. See whole document.	1 – 19, 39
A	WO 2002010438 A2 (THE JOHN HOPKINS UNIVERSITY) 7 February 2002. See whole document.	1 – 19, 39

INTERNATIONAL SEARCH REPORT

International application No.
PCT/SG2005/000281

Box No. II Observations where certain claims were found unsearchable (Continuation of item 2 of first sheet)

This international search report has not been established in respect of certain claims under Article 17(2)(a) for the following reasons:

1. ☐ Claims Nos.:
because they relate to subject matter not required to be searched by this Authority, namely:

2. ☒ Claims Nos.: **20 – 38**
because they relate to parts of the international application that do not comply with the prescribed requirements to such an extent that no meaningful international search can be carried out, specifically:

See Supplemental Box.

3. ☐ Claims Nos.:
because they are dependent claims and are not drafted in accordance with the second and third sentences of Rule 6.4(a)

Box No. III Observations where unity of invention is lacking (Continuation of item 3 of first sheet)

This International Searching Authority found multiple inventions in this international application, as follows:

1. ☐ As all required additional search fees were timely paid by the applicant, this international search report covers all searchable claims.
2. ☐ As all searchable claims could be searched without effort justifying an additional fee, this Authority did not invite payment of any additional fee.
3. ☐ As only some of the required additional search fees were timely paid by the applicant, this international search report covers only those claims for which fees were paid, specifically claims Nos.:

4. ☐ No required additional search fees were timely paid by the applicant. Consequently, this international search report is restricted to the invention first mentioned in the claims; it is covered by claims Nos.:

Remark on Protest

☐ The additional search fees were accompanied by the applicant's protest.

☐ No protest accompanied the payment of additional search fees.

Supplemental Box
(To be used when the space in any of Boxes I to VIII is not sufficient)

Continuation of Box No: II

Claims 20 – 38 do not comply with PCT Rule 6.3(a). Said claims are not defined in terms of the technical features of the invention. The claims are directed towards a transcript mapping system comprising *means for* performing the method steps of claims 1 – 19. Said claims appear to define a collocation of various known integers which are not limited to performing the methods of the invention. As such, said claims are indeterminate in scope and no meaningful search could be conducted.

INTERNATIONAL SEARCH REPORT

Information on patent family members

International application No.
PCT/SG2005/000281

This Annex lists the known "A" publication level patent family members relating to the patent documents cited in the above-mentioned international search report. The Australian Patent Office is in no way liable for these particulars which are merely given for the purpose of information.

Patent Document Cited in Search Report		Patent Family Member					
WO	2005026391	EP	1533386	JP	2005118036	US	2005059022
WO	2004099445						
WO	2004050918	AU	2003291610	EP	1567666	GB	2409457
US	2004096892	AU	2003290933	WO	2004046370		
WO	2002010438	AU	80875/01	US	6498013	US	2003008290

Due to data integration issues this family listing may not include 10 digit Australian applications filed since May 2001.

END OF ANNEX