METHOD AND SYSTEM FOR UNBIASED GENOME AMPLIFICATION

Field of the Invention

[0001] The present invention generally relates to the technology of DNA amplification, and more particularly to methods and systems for unbiased amplification of genomic DNA by optimal primers.

Background of the Invention

[0002] Genomic DNA with adequate quantity and quality is a prerequisite for the application of many genetic and genomic techniques. Currently, the applications are diverse such as preimplantation and prenatal diagnosis, “immortalization” of clinical samples, genotyping, molecular tumor pathology, microbial diversity, forensic genetics, bioterrorism genome detection, and molecular anthropology. The common genetic and genomic techniques requiring genomic DNA include comparative genomic hybridization (CGH), fluorescence in situ hybridization (FISH), polymerase chain reaction (PCR), single strand conformation polymorphism (SSCP), DNA sequence, "loss of heterozygosity (LOH), fingerprint, and restriction fragment length polymorphism (RFLP).

[0003] In practice, genomic DNA has to be obtained from biological samples e.g. fresh frozen or fixed tissues, forensic specimens, cells, blood and body fluids. Unfortunately, biological samples collected in many circumstances make the amount of available genomic DNA extremely limiting. Indeed, genetic and genomic techniques often cannot be performed at all simply because of insufficient genomic DNA. For this reason, several techniques have been developed to circumvent the problem of minimal

amounts of genetic material by preamplifying as much of the entire genome as possible (whole genome amplification, WGA).

[0004] One of the popular WGA techniques is primer-extension-preamplification-PCR (PEP-PCR) that uses a totally degenerate mixture of 15-mer or 16-mer, which primes frequently throughout the genome. Zhang *et al.,* Whole genome amplification from a single cell: Implications for genetic analysis. Proc. Natl. Acad. Sci. USA, 89: 5847-5851 (1992). In order to improve the fidelity of the amplified genomic DNAs, there was disclosed an improved PEP-PCR that amplified genomic DNA by using a mixture of at least two DNA polymerases, wherein at least one of which possesses proofreading activity. U.S. Pat. No. 6,365,375. However, a major drawback of PEP-PCR is that since the primer is short and completely degenerate, it can prime to regions internal to products amplified in previous cycles, thereby creating shorter products with each additional cycle.

[0005] Another WGA technique is Tagged PCR (T-PCR). T-PCR is a two-step strategy, which uses for the first few low-stringent cycles a primer with a constant 5' end and a random 3' end to allow priming at many sites. The unincorporated primers are then removed and amplification is carried out with a second primer containing only the constant 5' sequence of the above primer under high-stringent conditions to allow exponential amplification. Grothues *et al.,* PCR amplification of megabase DNA with tagged random primers (T-PCR), Nucleic Acids Res., 21: 1321-1322 (1993). This method is more laborious than the other methods due to removal of unincorporated degenerate primers, which can also cause the loss of sample material.

[0006] Yet another WGA technique is Inter-Alu-PCR that uses primers based on the most conserved regions of the Alu repeat family and allows the amplification of fragments flanked by these sequences. Nelson *et al.,* Alu polymerase chain reaction: A method for rapid isolation of human-specific

sequences from complex DNA sources, Proc. Natl. Acad. Sci. USA, 86:6686-6690 (1989). Consequently, sequence targets in regions that are not flanked by Alu sequences within the amplification distance cannot be preamplified.

[0007] Yet another WGA technique is Ligation-mediated PCR (L-PCR). L-PCR first digests genomic DNA with Mse I, leaving a TA overhang for adaptor annealing and subsequent ligation. After ligation, the genomic DNA was amplified by PCR. Klein *et al.,* Comparative genomic hybridization, loss of heterozygosity, and DNA sequence analysis of single cells, Proc. Natl. Acad. Sci. USA, 96:4494-4499 (1999); and WO 00/17390.

[0008] Yet another WGA technique is multiple displacement amplification (MDA). MDA uses a thiophosphate-modified random hexamer and a ϕ29 DNA polymerase to amplify DNA in a 30^{oC} reaction not requiring thermal cycling. Dean *et al.,* Comprehensive human genome amplification using multiple displacement amplification, Proc. Natl. Acad. Sci. USA, 99: 5261-5266 (2002), and U.S. Pat. No. 6,642,034.

[0009] One more promising WGA technique is degenerate oligonucleotide-primed PCR (DOP-PCR) that was developed by Telenius in 1992. Telenius *et al.,* Degenerate oligonucleotide-primed PCR: General amplification of target DNA by a single degenerate primer, Genomics, 13: 718-725. DOP-PCR is based on the principle of priming from short sequences specified by the 3'-end of partially degenerate oligonucleotides used, during initial low annealing temperature cycles of the PCR protocol. Since these 3' end short sequences occur frequently, amplification of target DNA proceeds at multiple loci simultaneously. After these low-stringent cycles the annealing temperature is increased to allow only amplification of fragments that are tagged with the specific part of the primer sequence.

[0010] A modified version of DOP-PCR, termed LL-DOP-PCR (long products from low DNA quantities), used the common DOP-PCR primer 6-

MW (5'-CCGACTCGAGNNNNNNATGTGG-3'; SEQ ID NO: 281), but substituted Taq polymerase with a proofreading enzyme and increased annealing and extension time during the PCR. Kittler *et al.,* A whole genome amplification method to generate long fragments from low quantities of genomic DNA, Anal. Biochem., 300: 237-244 (2002).

[0011] To be most useful, a whole genome amplification method should have the following features: generation of long DNA fragments, successful amplification of small amounts of starting template DNA (pg to ng scale), high amplification fidelity and yield, and good coverage of the genome. The aforementioned genome amplification techniques have one or another drawbacks.

[0012] While the 6-MW primer has been widely used since its publication, there is no systematic ways to search for optimal primers for being used in the PCR-based whole genome application. Therefore, there is a need for determining optimal primers to amplify genomic DNAs with small amount (even single cells) to generate long DNA fragments with a high fidelity and good coverage.

[0013] This invention satisfies this need by disclosing computer-implemented methods and systems for determining optimal primers for unbiased whole genome amplification. Other advantages of this invention will be apparent with reference to the detailed description.

Summary of the Invention

[0014] The present invention provides systems and methods for determining optimal primers to be used for unbiased genome amplification. The methods for determining the optimal primers employ computer-based algorithms. In one embodiment, a genetic algorithms-based method for

selecting primers for genomic DNA amplification is provided. The method comprises the steps of defining a chromosome, wherein the chromosome comprises a plurality of alleles representing individual features of the primers; creating an initiation population of the chromosomes; defining a fitness function for evaluating the quality of each chromosomes; calculating the fitness function of each chromosomes from the initial population; selecting two chromosomes at each time to generate two new chromosomes by way of crossover, wherein the new chromosomes will constitute a next generation population; selecting a predetermined percentage of the new chromosomes from the crossover operation for single mutation, wherein the mutated chromosomes will be part of the next generation population; calculating the fitness function of each chromosomes from the next generation population, wherein chromosomes failing to meet a predetermined value of the fitness function will be excluded from the next generation population; repeating the steps of crossover, single mutation and calculation of the fitness function to determine whether a chromosome will be excluded from the next generation population; terminating the repeating; and outputting the primer sequences.

[0015] In another embodiment, the present invention provides a system for executing genetic algorithms-based methods for selecting primers for genomic DNA amplification. The system comprises a portal PC for transmitting request from an end user; a head node for receiving the request from the portal PC, wherein the head node includes a job manager module for creating tasks and assigning jobs, and a lookup service module for transmitting the tasks; and a computing nodes for executing the genetic algorithms, receiving and returning tasks from the lookup service module, and registering execution results.

[0016] In yet another embodiment, the present invention provides a number of hexamers that can be used for unbiased genome amplification. The sequences of the hexamers are listed in Table 1.

[0017] Accordingly, one object of the present invention is to provide genomic DNAs with adequate quantity and quality for genetic analyses.

[0018] The objects and advantages of the present invention will become apparent from the following detailed description of preferred embodiments thereof in connection with the accompanying drawings.

Brief Description of the Drawings

[0019] FIG 1 shows the operation of the genetic algorithm.

[0020] FIG 2 shows examples of one-point crossover.

[0021] FIG 3 shows examples of two-point crossover.

[0022] FIG 4 illustrates the calculation of the number of amplicons.

[0023] FIG 5 shows the distribution of amplicons in human genomes for (A) NYP6-1, (B) NYP6-2, (C) Telenius, (D) Fiegler-1, (E) Fiegler-2 and (F) Fiegler-3, where the intensity of the bands represents the quantity of amplicons found.

[0024] FIG 6 summarizes comparative genomic hybridisation of normal versus normal human genomic DNA, where left panels are for unamplified DNA, and right panels for DOP-PCR amplified DNA versus unamplified DNA. Green bars signify amplification, red bars deletions.

[0025] FIG 7 summarizes CGH images obtained after DOPPCR using primers:

A. 5'-ATCTCA-3' (SEQ ID NO: 6) (NYP6-1);

B. 5'-CCGACTCGAGNNNNNNATCTCA-3' (SEQ ID NO: 280) (NYP-22);

C. 5'-CCGACTCGAGNNNNNNATGTGG-3' (SEQ ID NO: 281) (Telenius); and

D. [5'-CCGACTCGAGNNNNNNCTAGAA-3' (SEQ ID NO: 282),

5'-CCGACTCGAGNNNNNNTAGGAG-3'(SEQ ID NO: 283),

5'-CCGACTCGAGNNNNNNTTCTAG-3'(SEQ ID NO: 284)] (Fiegler);

where deletions are indicated by a bar on the left of the schematic map of the chromosome while a bar on the right marks amplifications.

[0026] FIG 8 shows a comparative view of chromosome 1.

[0027] FIG 9 illustrates the measurement of deviation from the Normal Expression Line.

[0028] FIG 10 shows Comparison of Squared error of different primers.

[0029] FIG 11 shows the distribution of amplicons calculated from chromosome one.

[0030] FIG 12 shows Predictive Versus Experimental Datasets.

[0031] FIG 13 shows Scatter Plot For Chromosome 1.

[0032] FIG 14A shows Upper Path Regression Curve.

[0033] FIG 14B shows Lower Path Regression Curve.

[0034] FIG 15 shows Plot of the Normalised versus Experimental Datasets.

[0035] FIG 16 shows Corrected Curve.

[0036] FIG 17 is a block diagram of one embodiment of the system.

Detailed Description of the Invention

[0037] The present invention may be understood more readily by reference to the following detailed description of certain embodiments of the invention.

[0038] Various documents are cited throughout the text of this specification. Each of the documents cited herein (including any manufacturer's specifications, instructions, etc.) are hereby incorporated by reference; however, there is no admission that any document cited is indeed prior art of the present invention.

[0039] Fundamental to most genetic analysis is availability of genomic DNA of adequate quality and quantity. Because DNA yield from human samples is frequently limiting, much effort has been invested in developing methods for whole genome amplification (WGA) by random or degenerate oligonucleotide-primed PCR. Whole genome amplification (WGA) is a technique for greatly increasing minuscule amounts of starting DNA to enable genetic and genomic analyses that would otherwise be impossible. However, existing WGA methods including degenerate oligonucleotide-primed PCR suffer from incomplete coverage and inadequate average DNA size. For many applications, WGA must provide complete coverage of the genome with minimal amplification bias. A certain amount of sequence bias may be anticipated with any amplification method and may result from factors such as priming efficiency, template accessibility, GC content, and proximity to telomeres and centromeres.

[0040] The methods and systems of the present invention provide unbiased whole genome amplification of DNAs from various samples for further techniques and methods, wherein the samples may contain as little as one single cell. The unbiased WGA methods and systems have diverse applications including preimplantation and prenatal diagnosis, "immortalization" of clinical samples, genotyping, molecular tumor

pathology, microbial diversity, forensic genetics, bioterrorism genome detection, and molecular anthropology. The unbiasedly amplified DNAs can be used for the techniques including comparative genomic hybridization (CGH), fluorescence in situ hybridization (FISH), polymerase chain reaction (PCR), single strand conformation polymorphism (SSCP), DNA sequence, "loss of heterozygosity (LOH), fingerprint, and restriction fragment length polymorphism (RFLP).

[0041] As used in accordance with the present invention, the term "DNA fragments of similar length" denotes fragments which, at a statistical level, have a size which is of comparable length. DNA fragments of comparable length are, for example, fragments of 1kbp +/- 0.1kbp or of 3kbp +/- 0.3 kbp. The length range of DNA fragments that is preferably generated is advantageously between about 1kb and about 3 kb. DNA fragments of greater or shorter length may be used as well, although they may be amplified or represented to a lesser extent than the above defined fragments.

[0042] The term "primer" as used herein refers to an oligonucleotide whether occurring naturally as in a purified restriction digest or produced synthetically. The primer is preferably single stranded for a maximum of efficiency in the method of the present invention, and is preferably an oligodeoxyribonucleotide. Purification of said primers is generally envisaged, prior to their use in the method of the present invention. Such purification steps can comprise HPLC (high performance liquid chromatography) or PAGE (polyacrylamide gel-electrophoresis), and are known to the person skilled in the art. Oligonucleotides, representing primers as used in the method of the present invention can be identified, obtained and tested according to the state of the art especially represented by computer based sequence analysis and laboratory manuals, e.g. Sambrook (Molecular

Cloning; A Laboratory Manual, 2nd Edition, Cold Spring Harbour Laboratory Press, Cold Spring Harbour, NY (1989)).

[0043] The most widely practised WGA method is degenerate oligonucleotide-primed polymerase chain reaction (DOP-PCR) using a partially degenerate primer, 5' CCG ACT CGA GNN NNN NAT GTG G 3' (where N may be A, T, C or G) (SEQ ID NO: 281), designed by Telenius. Telenius et al. Genomics 13: 718-725 (1992). The inventors of the present invention discovered by comparative genomic hybridisation (CGH) that whole genome amplification using the Telenius primer design is significantly biased i.e. starting genomic DNA is not uniformly and homogeneously amplified. This is a serious weakness because analytical data from such DNA cannot be considered reliable if substantial bias is known to be already present.

[0044] There was disclosed the design, sequence, and performance of primers for use in DOP-PCR amplification of large insert clones or flow-sorted chromosomes for the construction of DNA microarrays for CGH and other studies. Fiegler *et al.*, DNA microarrays for comparative genomic hygridization based on DOP-PCR amplification of BAC and PAC clones, Genes, Chromosomes & Cancer, 36: 361-374 (2003). The primers were selected by a custom perl script counting the occurrence of all possible 6mers in the E. coli genome and in a number of NCBI human genomic reference sequence contigs. There 6mers were selected that showed the lowest frequency in E. coli DNA with a high frequency in human sequence. There sequences of the three 6mers are:

DOP 1 5'-CCGACTCGAGNNNNNNCTAGAA-3' (SEQ ID NO: 282);

DOP 2 5'-CCGACTCGAGNNNNNNTAGGAG-3' (SEQ ID NO: 283);

DOP 3 5'-CCGACTCGAGNNNNNNTTCTAG-3' (SEQ ID NO: 284).

[0045] However, the disclosed Telenius 6-MW primer and Fiegler primers failed to provide unbiased amplification of human genomic sequences, as shown in the Example section. Therefore, there is a need to obtain optimal primers that can be used in unbiased genomic amplification. It is known that primers are the most crucial elements that will affect the outcomes of PCR reactions. Recent completion of the entire human genome sequence presents an opportunity to redesign primers to achieve unbiased and efficient WGA. The present invention teaches the methods of employing genetic algorithms to search for optimal primers, discloses the obtained optimal primers, and demonstrates that the new primers with only six (6) base pairs meet the goal of unbiased amplification. The present invention provides simple and useful reagents to overcome current bottlenecks in molecular pathology, clinical genomics and forensic pathology.

[0046] A DNA sample to be amplified can be obtained from a single cell or a population of cells isolated from, e.g., bone marrow, peripheral blood, or tumor. In addition, DNA samples may be processed ones including cosmid, P1, P1 artificial chromosome (PAC), and bacterial artificial chromosome (BAC) clones. In the present invention, the description has used human genome as an illustration. It would be appreciated by those skilled in the art that the methods and systems of the present invention are applicable to genomic DNAs from other species such as bacterial, fungi, plants and mammals.

[0047] PCR (polymerase chain reaction) is an extremely powerful in vitro method for the amplification of DNA, which was initially introduced in 1985. Saiki *et al.,* Science 230, 1350-1354 (1985). By repeated thermal denaturation, primer annealing and polymerase extension, PCR can amplify a single target DNA molecule to easily detectable quantities.

[0048] Current methods of PCR amplification involve the use of two primers which hybridize to the regions flanking a nucleic acid sequence of interest such that DNA replication initiated at the primers will replicate the nucleic acid sequence of interest. By separating the replicated strands from the template strand with a denaturation step, another round of replication using the same primers can lead to geometric amplification of the nucleic acid sequence of interest.

[0049] The primers are short oligonucleotide sequences, and can be synthesized using established oligonucleotide synthesis methods. Methods to produce or synthesize oligonucleotides are well known in the art. Such methods can range from standard enzymatic digestion followed by nucleotide fragment isolation (see for example, Sambrook et al., Molecular Cloning: A Laboratory Manual, 2nd Edition (Cold Spring Harbor Laboratory Press, Cold Spring Harbor, N.Y., 1989) Chapters 5, 6) to purely synthetic methods, for example, by the cyanoethyl phosphoramidite method using a Milligen or Beckman System 1Plus DNA synthesizer (for example, Model 8700 automated synthesizer of Milligen-Biosearch, Burlington, Mass. or ABI Model 380B). Synthetic methods useful for making oligonucleotides are also described by Ikuta et al., Ann. Rev. Biochem. 53:323-356 (1984), (phosphotriester and phosphite-triester methods), and Narang et al., Methods Enzymol., 65:610-620 (1980), (phosphotriester method).

[0050] Furthermore, the setting of conditions for the above described steps of PCR is well within the skill of the artisan. Ausubel, "Current Protocols in Molecular Biology", Green Publishing Associates and Wiley Interscience, N.Y. (1989), or Higgins and Hames (Eds) "Nucleic acid hybridization, a practical approach" IRL Press Oxford, Washington DC, (1985). In addition, the GA was used to determine the optimum times

required to complete the denaturation, annealing and extension events of PCR. U.S. Pat. No. 6,493,640.

[0051] There is no disclosure in the prior arts about searching optimal primers for unbiased amplification of genomic DNAs. The present invention discovered that with the assistance of available human genomic sequences, optimal primers could be obtained by computer-implemented methods. The present invention further discloses that genetic algorithms could be employed to obtain optimal primers. While the general concepts of the genetic algorithms will be discussed next, the implementation of genetic algorithms into the searching of optimal primers for human genomic sequences will be detailed in the Example section.

[0052] A general definition of genetic algorithms does not exist. They can be represented by an abstraction of the work of Holland (See "Adaptation in natural and artificial systems". Holland, (1975). The University of Michigan Press; Ann Arbor). Genetic algorithms (GA) are often used for optimizing combinatorial problems. GA takes its roots from evolutionary strategies taken by living organisms – survival of the fittest. Organisms' genetic material randomly undergoes crossovers and mutations during reproduction. Combinations of genes that express superior traits survive, whilst poorer offspring weed off.

[0053] In a basic form of GA, a population of potential solutions to a specified problem are maintained and repeatedly updated according to the principles of evolution, i.e. selection, mutation and/or recombination (crossover). Recombination selects pairs of solutions in the population (parents) and generates a series of new solutions (children) by combining elements from their parents, which are then inserted into a new population of solutions. The principle of selection demands that "good" solutions be preferentially chosen over "bad" solutions. This is achieved by defining a

fitness function that assigns a number to each solution according to how "good" it is. Selection guarantees that only solutions (chromosomes) with best fitness will propagate in future populations. Mutation is used modify solutions in the population without interaction with the rest of the population. Genetic algorithms therefore search for sets of alleles which, together, produce good solutions (co-adapted alleles). The genetic algorithm in its simplest form is described below.

[0054] Initialisation

[0055] The initial population of chromosomes (typically n=25 . . . 100), is created either randomly, or by "shuffling" an input chromosome generated from the artificial neural network element. This reduces the time taken by the algorithm in searching for optima.

[0056] Evaluation

[0057] The fitness of each chromosome is evaluated. In the present invention, the quality of amplification is calculated by using a fitness function as discussed later. These are then assigned fitness functions, which numerically encodes the performance of each chromosome.

[0058] Exploitation

[0059] Chromosomes with the highest fitness score (i.e. the best predicted amplification), are placed one or more times into a mating subset in a semi-random fashion. Two chromosomes are drawn at random from the population. Both chromosomes are returned to the population and the selection process is repeated until the mating subset is full. This ensures that chromosomes with low fitness are removed from the mating population. The chance of selection as a parent is proportionate to chromosomes normalised fitness. This means that better chromosomes will generally produce more children. However, the stochastic nature of the process means that occasionally poor solutions will produce children. The addition of elitism

functions is also beneficial. Here, the single best solution in any parent generation is copied unmodified into the child generation; all other children are generated as normal.

[0060] Exploration

[0061] This uses recombination and mutation operators to modify chromosomes in the next generation of chromosomes (FIG 1). Two chromosomes from the mating subset are randomly selected and mated. The probability of mating is a controllable function and is generally given a high value (.apprxeq.0.90). A recombination operator is used to exchange genes between the two parent chromosomes, to produce two children. In FIG 1, a population of three individuals is shown. Each is assigned a fitness function F. On the basis of these fitnesses, the selection phase assigns the first individual 0 copies, the second 2 copies and the third 1 copy. After selection genetic operators are applied probablistically; the first individual has its first bit mutated from a 1 to a 0, and crossover combines the second two individuals into new ones. (Modified from Forrest (1993). Science, Vol.261, pp.872-878).

[0062] Recombination may involve one-point and/or two-point crossovers (FIG 2 and FIG 3). Where one-point crossover is used, a crossover point is selected along the chromosome and the genes up to that point are swapped between the two parents. Where two-point crossover occurs, two crossover points are selected and the genes between the two points are swapped. Other crossover algorithms such as partially mapped crossover (PMX), order crossover (OX) and cycle crossover (CX) may also be used. All children are produced by crossover of two parents; two children are produced simultaneously. The children then replace the parents in the next generation. Mutation of single genes is another controllable function. Mutation rates are generally assigned a low probability (.apprxeq.0.001), and

can be defined (N*L1/2)-1, where N is the population size and L is the length of chromosome.

[0063] This sequence of events is repeated for a fixed number of generations until convergence occurs. Crowding replacement may be used to reduce the problem of premature convergence. A child is compared with a number of existing parents and replaces a parent most similar to itself. Essentially crowding replacement replaces like with like, this allows sub-populations to explore various parts of the genetic search space. This is particularly useful for multi-modal search spaces where there are a number of scattered fitness maxima.

[0064] The basic structure of the genetic algorithm used can be formalised as:

Initialise population
Evaluate chromosomes
Calculate normalised fitness
For each successive generation
{
Generate and evaluate candidate replacements
Replace numbers of the populations with candidates
Re-evaluate unchanged members
Normalise fitness
Output statistics on population
}

[0065] It is appreciated to those skilled in the art that the present invention is not limited to the employment of the basis operations of GA as discussed above. Other improved operations of GA are also applicable to the present invention. The improvements of GA include U.S. Pat. App.

2003/0050902, U.S. Pat. Nos. 5,970,487 and 6,185,547, and U.S. Pat. App. 2002/0120407.

[0066] FIG 17 illustrates one embodiment of the computing system of the present invention. The main components in the high level architecture **100** comprises a Computing Nodes **110**, a Head Node **120**, and a Portal PC **130**. The Computing Nodes **110** includes a plurality of CPUs suitable for executing genetic algorithms instructions implementing techniques described in the present invention. In one embodiment, 15 CPUs were used to process, send and return tasks in connection with the Head Node **120**. In addition, the CPUs register their results with the Head Node **120.** The Head Node **120** executes two main functions, i.e., Job Manager and Lookup Service. The Job Manager may have two modules for creating tasks and assigning jobs respectively. The Lookup Service serves as a liaison between the Job Manager and the Computing Nodes **110**. The Portal PC **130** sends request from an end user to the Job Manager, and displays the results from the Computing Nodes **110** through the Head Node **120**

[0067] The present invention is not limited to any particular hardware or software configuration. The techniques may be implemented in hardware or software, or a combination of the two. Preferably, the techniques are implemented in computer programs executing on programmable computers that each include a processor, a storage medium readable by the processor (including volatile and non-volatile memory and/or storage elements), at least one input device, and one or more output devices.

[0068] Each program is preferably implemented in high level procedural or object oriented programming language to communicate with a computer system. However, the programs can be implemented in assembly or machine language, if desired. In any case the language may be compiled or interpreted language.

[0069] Each such computer program is preferably stored on a storage medium or device (e.g., CD-ROM, hard disk, or magnetic disk) that is readable by a general or special purpose programmable computer for configuring and operating the computer when the storage medium or device is read by the computer to perform the procedures described herein. The system may also be considered to be implemented as a computer-readable storage medium, configured with a computer program, where the storage medium so configured causes a computer to operate in a specific and predefined manner.

[0070] The optimal primers obtained from the GA operations will be used in PCR-based genomic amplification. The quality of the PCR amplification can be investigated in many ways. The simple way is to check the size ranges of the PCR products. The ultimate test for the PCR products is their application in genetic and genomic techniques, e.g., CGH, which will be discussed later.

[0071] To aid in detection and quantitation of nucleic acids amplified using the disclosed method, detection labels can be directly incorporated into amplified nucleic acids or can be coupled to detection molecules. As used herein, a detection label is any molecule that can be associated with amplified nucleic acid, directly or indirectly, and which results in a measurable, detectable signal, either directly or indirectly. Many such labels for incorporation into nucleic acids or coupling to nucleic acid probes are known to those of skill in the art. Examples of detection labels suitable for use in the disclosed method are radioactive isotopes, fluorescent molecules, phosphorescent molecules, enzymes, antibodies, and ligands.

[0072] Examples of suitable fluorescent labels include fluorescein (FITC), 5,6-carboxymethyl fluorescein, Texas red, nitrobenz-2-oxa-1,3-diazol4-yl (NBD), coumarin, dansyl chloride, rhodamine, 4'-6-diamidino-2-phenylinodole (DAPI), and the cyanine dyes Cy3, Cy3.5, Cy5, Cy5.5 and

Cy7. Preferred fluorescent labels are fluorescein (5-carboxyfluorescein-N-hydroxysuccirtimide ester) and rhodamine (5,6-etramethyl rhodamine). Preferred fluorescent labels are FITC and the cyanine dyes Cy3, Cy3.5, Cy5, Cy5.5 and Cy7. The absorption and emission maxima, respectively, for these fluors are: FITC (490 nm; 520 nm), Cy3 (554 nm; 568 nm), Cy3.5 (581 nm; 588 nm), Cy5 (652 nm: 672 nm), Cy5.5 (682 nm; 703 nm) and Cy7 (755 nm; 778 nm), thus allowing their simultaneous detection. The fluorescent labels can be obtained from a variety of commercial sources, including Molecular Probes, Eugene, Oreg. and Research Organics, Cleveland, Ohio.

[0073] Labeled nucleotides are a preferred form of detection label since they can be directly incorporated into the amplification products during synthesis. Examples of detection labels that can be incorporated into amplified DNA or RNA include nucleotide analogs such as BrdUrd (Hoy and Schimke, Mutation Research 290:217-230 (1993)), BrUTP (Wansick et al., J. Cell Biology 122:283-293 (1993)) and nucleotides modified with biotin (Langer et al., Proc. Natl. Acad. Sci. USA 78:6633 (1981)) or with suitable haptens such as digoxygenin (Kerkhof, Anal. Biochem. 205:359-364 (1992)). Suitable fluorescence-labeled nucleotides are Fluorescein-isothiocyanate-dUTP, Cyanine-3-dUTP and Cyanine-5-dUTP (Yu et al., Nucleic Acids Res., 22:3226-3232 (1994)). A preferred nucleotide analog detection label for DNA is BrdUrd (BUDR triphosphate, Sigma), and a preferred nucleotide analog detection label for RNA is Biotin-16-uridine-5'-triphosphate (Biotin-16-dUTP, Boehringher Mannheim). Fluorescein, Cy3, and Cy5 can be linked to dUTP for direct labelling. Cy3.5 and Cy7 are available as avidin or anti-digoxygenin conjugates for secondary detection of biotin- or digoxygenin-labelled probes.

[0074] Detection labels that are incorporated into amplified nucleic acid, such as biotin, can be subsequently detected using sensitive methods well-

known in the art. For example, biotin can be detected using streptavidin-alkaline phosphatase conjugate (Tropix, Inc.), which is bound to the biotin and subsequently detected by chemiluminescence of suitable substrates (for example, chemiluminescent substrate CSPD: disodium, 3-(4-methoxyspiro-[1,2,-ioxetane-3-2'-(51'-chloro)tricyclo [3.3.1.13.7]decane]-4-yl) phenyl phosphate; Tropix, Inc.).

[0075] Molecules that combine two or more of these detection labels are also considered detection labels. Any of the known detection labels can be used with the disclosed probes, tags, and method to label and detect nucleic acid amplified using the disclosed method. Methods for detecting and measuring signals generated by detection labels are also known to those of skill in the art. For example, radioactive isotopes can be detected by scintillation counting or direct visualization; fluorescent molecules can be detected with fluorescent spectrophotometers; phosphorescent molecules can be detected with a spectrophotometer or directly visualized with a camera; enzymes can be detected by detection or visualization of the product of a reaction catalyzed by the enzyme; antibodies can be detected by detecting a secondary detection label coupled to the antibody. As used herein, detection molecules are molecules which interact with amplified nucleic acid and to which one or more detection labels are coupled.

[0076] The DNA products from the optimal primers have many applications. In addition to the traditional genetic and genomic techniques, the demand for custom-made cDNA chips/arrays as well as whole genome chips is increasing rapidly. It is apparent to those skilled in the art that the GA-based determination of optimal primers disclosed in the present invention can be applicable to search for optimal primers for specific genes, cosmids, P1, P1 artificial chromosomes, YACs and BACs.

[0077] A variety of techniques are now available for genome-wide screening of alterations in copy-number, structure and expression of genes and DNA sequences. These include molecular cytogenetic techniques such as comparative genomic hybridization (CGH) and multicolor fluorescence in situ hybridization (M-FISH), as well as molecular genetic techniques such as representational difference analysis (RDA), differential display, serial analysis of gene expression (SAGE) and microarray techniques.

[0078] While the present invention will use CGH as a technique to illustrate the verification of the optimal primers obtained from the execution of genetic algorithms in accordance with the present invention, it shall be noted that the optimal primers can be used in other suitable DNA amplification operations. Moreover, the amplified genomic DNAs from the WGA of the present invention can be employed in any molecular cytogentic and genetic techniques.

[0079] CGH was the first molecular cytogenetic tool that allowed comprehensive analysis of the entire genome. Kallioniemi *et al.*, Comparative genomic hybridisation for molecular cytogenetic analysis of solid tumors, Science 258, 818-821 (1992). CGH is an *in situ* hybridization technique using fluorescence that allows a comparative analysis of genetic differences, usually between normal and tumor tissues. In the present invention, CGH was used to compare the human genome against the WGA product. Therefore, in order to screen individual tumor cells by methods like CGH, it is desirable to uniformly and accurately amplify the whole genome of such a cell.

[0080] CGH allows for the screening for DNA sequence copy-number changes and provides a map of those chromosomal regions that are gained or lost in a DNA specimen. Because DNA copy-number alterations are of pathogenetic importance in cancer, most of the applications of CGH are in

cancer research. An important contribution of CGH to cancer research has been in pinpointing putative locations of cancer genes, especially at chromosomal sites undergoing DNA amplification. A large number of subregional chromosomal gains and DNA amplifications have been discovered by CGH in some cancers. Because oncogenes and drug-resistance genes are known to be upregulated by DNA amplifications, it has been speculated that DNA amplification sites in cancer could pinpoint locations of novel genes with important roles in cancer progression. Tumor progression implies the gradual transition of a localized, slow growing tumor to an invasive, metastatic and treatment refractory cancer. This progression is thought to be caused by a stepwise accumulation of genetic changes affecting critical genes.

[0081] By providing genome-scale information of clonal genetic alternations, CGH is extremely useful in the analysis of the biological basis of the tumor progression process. Two cancer specimen taken from the same patient at different stages of progression can be analyzed. For example, genetic changes that are not found in primary tumors, but do occur in their metastases could be informative in pinpointing genetic changes and genes with important roles in the metastatic progression.

[0082] Accordingly, the above described method is particularly useful for screening individual tumor cells by CGH and therefore allowing early diagnosis of e.g. neoplastic disorders or patients susceptible to such disorders. The term "neoplastic disorders" is intended to mean the whole spectrum from initiation of malignant transformation in a single cell to advanced cancer disease, including distant solid metastasis.

[0083] CGH is based on a modified in situ hybridisation. There are two basic approaches. One is that genomic DNA from a subject cell or cell population of cells is isolated, labelled and hybridised to reference

chromosomes. Another is that genomic DNAs from two or more subject cells or cell populations are isolated and differentially labelled, where differentially labeled test (green) and reference (red) DNAs are co-hybridized to normal metaphase spreads. Copy-number differences between test and reference genomes are seen as green:red fluorescence intensity differences on the metaphase chromosomes. DNA gains and amplifications in the test DNA are seen as chromosomal regions with an increased fluorescence ratio, while losses and deletions result in a reduced ratio. For conventional CGH, chromosomal regions were considered to be lost or gained if the hybridization ratio of tumor cell line to normal was less than 0.85:1 or greater than 1.15:1, respectively.

[0084] CGH is a well-established technique. For example, U.S. Pat. No. 6,335,167 discloses a complete protocol for performing CGH on many different tumor cells. Therefore, it is incorporated herein in its entirety.

Examples

[0085] The following examples are presented by way of illustration of the above-described invention and is not limiting to that description.

[0086] The system setup included a 16-node Intel® Xeon™ 2.4 GHz cluster, each having a memory size of 2 gigabytes (GB). Java™ RMI was used to write the application. The search tasks were distributed over 15 nodes, with each node pre-assigned chromosomes based on their memory requirements. Generally, each node did not carry more than two chromosomes. The remaining node was used to schedule and monitor jobs, collect completed jobs and summarize the results.

[0087] The GA search engine was developed using JDEAL (http://www.laseeb.org/sw/jdeal/). The architecture of the GA is described in detail below.

[0088] Human genome sequences (Build 34) were obtained from the University of California, Santa Cruz (http://genome.ucsc.edu/) (Karolchik *et al.*, Nucl. Acids Res, 31: 51-54 (2003)) via download. The datasets were sorted by chromosomes and stored in FASTA format. The human genome consists of 24 pairs of chromosomes, with the longest chromosome reaching length of approximately 245 Mbp. To rapidly search these sequences, the dataset files were pre-loaded into the computer system's primary storage, e.g., RAM.

[0089] Now there is a description of one embodiment employing the GA to obtain optimal primers for unbiased amplification of whole human genome. For the practical purpose, a simple GA architecture was utilized as discussed below.

[0090] Objective

[0091] The objective of performing GA is to minimize the value of the fitness function. In this example, the ideal solution would be one that provides PCR products that have significant uniform distribution over and sufficiently cover the whole genome sequence.

[0092] Solution Representation

[0093] GA solutions are generally described by using a string of bits that can either be integers or real numbers. Integers were used to represent (i) nucleotides and (ii) the maximum range of extension. Each solution consisted of nine (9) alleles encoded as:

A	A	A	A	A	A	B	B	C

where the first 6 alleles represented by **A** were nucleotides of A, C, G, or T, identified by integer values 1 to 4 respectively, the next 2 alleles represented by **B** were nucleotides of None, A, C, G or T, identified by integer values 0 to

4 respectively, and the last allele of C represented extension length (Extension length is shown in FIG 4) with an integer ranged between 1000 and 3000.

[0094] Therefore, each solution encodes primers with a varying length from 6 to 8 nucleotides. Each solution defines a candidate sequence used as a single primer in an *in silico* PCR simulation.

[0095] Now with reference to FIG 4, methods for calculating the number of amplicons are described. Amplicons, as used herein, refer to fragments that can be possibly amplified from a genomic DNA sequence by using a PCR method with a single primer defined in one solution. To identify amplicons, first was to search for forward primers in the genomic DNA sequence, where the forward primer had the same sequence as defined in a solution. Then, within the extension length of the identified forward primer, the genome was searched downstream of the forward primer for the number of reverse sequences, where the reverse sequences were complementary to the forward primer. For the convenience of execution of the algorithms, amplicons were then counted within 1 Mbp segments of the chromosomes.

[0096] Fitness function

[0097] In practice, the fitness of a GA solution is expressed and quantified by using mathematical equations. In the present invention, three components were used to define the fitness of a candidate primer: (i) variance, (ii) average and (iii) coverage. The fitness function is defined as:

$$f(S_c) = \frac{1}{\frac{2\sigma_c}{\mu_c} + \frac{1}{2}(1 - \frac{C_c}{C_{total}})} \qquad (1)$$

where σ_c = variance of the number of amplicons in each 1 Mbp segments, μ_c = average number of amplicons in each 1 Mbp segments, C_c = coverage defined by the number of nucleotides reproduced (amplified), and C_{total} = total number of nucleotides found in Human Genome Build 34.

[0098] The constants in the fitness function are used to control the weight of variables. In the fitness function of the present invention, the variance was doubled to highlight its importance over the average, and the significance of coverage was reduced by halving it.

[0099] Crossover & Mutation

[0100] As discussed above, the crossover could be conducted in one or more points simultaneously. In this example, single-point crossover operations were performed. A pair of solutions were picked from the mating pool and subjected to a crossover operation with a defined probability p_c. The operator chose a point for crossover, restricted to the length of the solution. Each of the two offspring was accorded the proportioned amount of genetic information from each of the parents. This action was repeated until all members of the mating pool were accessed.

[0101] As discussed above, mutations can be done in one or more alleles of the solution. In this example, a single flip mutation operation was performed on the offspring with a defined probability p_m following the crossover operations. The operator selects a bit for mutation and changes it to a new value within its eligible range.

[0102] Selection

[0103] Except for the first generation, which was populated randomly, each generation was a population of entirely new solutions. A rank selection was used to determine subsequent generations. Solutions were selected based solely on their fitness value. Chromosomes with the highest fitness score were ranked first.

[0104] Termination

[0105] The GA was terminated when it has passed the defined maximum number of generations. In this example it was set to 200.

[0106] Identification of optimal primers

[0107] The GA was allowed to evolve over 200 generations. The representative optimal primers are listed in Table 1. The list was generated upon pure fitness calculation of all possible combination of 6-mer with a fixed extension length of 3000, because it is not possible to give a ranking if a specific extension length was not chosen.

[0108] Table 1 The listing of the identified optimal primers

Primer Sequence	σ_c	μ_c	C_c	Fitness Score	SEQ ID NO
TTTCCT	448.9	1302.1	69.1	1.185	1
AGGAAA	442.3	1282.7	69.0	1.184	2
TTTTGT	802.0	2184.9	74.1	1.158	3
ACAAAA	805.1	2190.5	74.0	1.156	4
TGAGAT	195.7	593.7	55.6	1.134	5
ATCTCA	191.7	582.2	54.9	1.131	6
TTTCTG	431.2	1184.3	68.3	1.128	7
AACAGA	206.2	626.2	54.0	1.126	8
CAGAAA	422.0	1156.3	68.1	1.124	9
TTCCTT	370.2	1039.8	63.9	1.121	10
AAAAAG	848.8	2223.5	74.2	1.120	11
AAGGAA	372.3	1044.9	63.9	1.120	12
TTCTGT	278.4	803.9	59.6	1.118	13
ATCTCT	174.0	528.9	52.6	1.117	14
AGAGAT	173.9	527.3	52.4	1.114	15
TTCTCT	412.8	1138.7	65.4	1.114	16
ACAGAA	263.9	764.6	57.9	1.110	17
TCTGTT	202.0	607.4	52.7	1.109	18
CTTTTT	832.5	2157.6	74.0	1.109	19
AAAAGA	850.5	2205.3	73.9	1.109	20
TTTTTG	1032.0	2630.7	74.9	1.099	21
CACTTT	266.7	746.0	60.8	1.098	22
TCTTTT	848.2	2171.4	73.8	1.096	23
AGAGAA	413.1	1117.7	65.4	1.096	24
CTGTTT	226.0	651.1	55.7	1.092	25

AAACAG	221.7	639.6	55.5	1.092	26
ATTCTC	173.2	516.3	50.6	1.090	27
CAAAAA	1072.7	2706.7	74.9	1.089	28
CATGTT	172.9	511.3	51.4	1.088	29
TTTCCA	324.0	885.1	62.5	1.088	30
AACATG	169.9	503.7	50.9	1.087	31
TCACTT	183.2	539.4	51.8	1.087	32
AAAGTG	282.2	773.5	61.4	1.084	33
AAGTGA	183.1	538.5	51.4	1.083	34
CTGAAG	125.1	388.0	44.3	1.083	35
CAGAAG	183.5	538.4	51.5	1.082	36
TTGCTT	210.4	604.6	54.3	1.082	37
AAGAAA	1273.5	3111.5	78.8	1.082	38
GGAAAA	377.0	1013.0	63.9	1.081	39
CTTCTG	183.9	539.4	51.1	1.080	40
TTTTCC	366.4	993.8	62.2	1.079	41
CTTCAG	125.1	386.7	44.1	1.079	42
TTTCTC	437.6	1151.8	66.4	1.078	43
GGGAAA	176.0	521.7	49.3	1.077	44
GAGAAA	441.3	1159.5	66.5	1.077	45
TTTCTT	1270.5	3079.5	78.8	1.074	46
GTGAAA	193.0	567.2	49.7	1.073	47
CTTTTC	269.7	750.8	57.2	1.072	48
TTTTGA	504.2	1298.9	68.7	1.072	49
TTTCCC	176.5	519.3	49.2	1.071	50
GAGAAT	192.3	549.4	53.3	1.071	51
GAAAAG	278.6	767.9	58.3	1.071	52
AAAACA	869.9	2170.0	73.4	1.070	53
AAGAGA	295.4	811.3	58.7	1.070	54
TCAAAA	500.5	1292.5	67.3	1.066	55
TTCTCC	279.7	765.6	58.5	1.066	56
ATTTTT	2998.8	7000.8	83.7	1.066	57
TGTTTT	863.2	2141.4	73.5	1.065	58
CAAAAC	151.8	455.5	45.5	1.065	59
AAAGGA	312.8	846.1	60.0	1.065	60
TCTCTT	303.9	827.3	59.0	1.064	61
AAAAAT	3043.7	7096.1	83.5	1.064	62
TCCTTT	305.5	826.0	59.7	1.063	63
CCATTT	230.8	641.6	55.6	1.062	64
TGGAAA	308.4	832.6	59.6	1.061	65
TGAACT	109.1	334.4	41.8	1.060	66
AGTTCA	104.9	321.6	41.2	1.057	67
AAAGAG	276.6	756.2	57.0	1.057	68
CCAAAA	226.6	636.2	53.1	1.056	69
AGAAGA	285.4	774.8	57.9	1.056	70

TGCCAT	96.2	299.9	38.9	1.055	71
GGTTTT	156.9	463.5	45.8	1.055	72
AAAACC	154.6	456.0	45.8	1.054	73
TTTTGG	225.6	630.8	53.1	1.053	74
GTTTTG	145.0	432.2	44.2	1.052	75
TACAAA	391.9	1023.2	63.1	1.052	76
TCAAGA	121.3	364.3	43.1	1.052	77
TCTTGA	123.7	370.7	43.3	1.052	78
TTTCAG	263.2	710.1	58.0	1.051	79
CTGAAA	259.5	700.5	57.9	1.051	80
AAATGG	245.4	670.3	56.1	1.051	81
CATTTC	233.9	640.0	55.8	1.051	82
CTCTTT	280.6	761.2	57.0	1.050	83
AAGCAA	202.1	569.9	51.3	1.050	84
TGGTTT	184.4	526.3	49.5	1.049	85
TTCTGC	138.3	409.7	44.3	1.049	86
GCAGAA	128.4	382.7	43.3	1.048	87
GGAGAA	287.3	769.0	58.6	1.048	88
CCTGAA	107.5	327.2	40.5	1.048	89
AAAGCA	269.1	725.7	57.4	1.048	90
AAACCA	185.4	527.7	49.6	1.047	91
AAACTC	126.4	375.7	43.5	1.047	92
AAGGCA	130.3	387.1	43.5	1.047	93
TTTTTA	1957.2	4557.7	80.6	1.046	94
GAAATG	232.6	633.7	55.6	1.046	95
TTTGTA	386.4	1002.5	62.9	1.046	96
TTTCAC	190.5	545.7	48.3	1.046	97
AAAAAA	2235.4	5247.2	78.8	1.044	98
TTCAGG	107.9	326.5	40.4	1.042	99
TAAAAA	2006.7	4654.4	80.6	1.042	100
TGGGAA	199.2	556.9	51.2	1.042	101
AAAAGG	234.5	643.3	53.8	1.042	102
AGATGA	152.1	436.8	47.2	1.041	103
TTTTTT	1998.0	4729.7	76.7	1.040	104
CAAACT	125.6	370.1	43.3	1.040	105
TTCACT	148.6	426.7	46.8	1.039	106
AAAGAA	1155.4	2725.8	77.0	1.039	107
ATGCTG	109.1	326.3	41.0	1.038	108
CCTTTT	235.0	641.6	53.8	1.038	109
TCATCT	153.2	437.9	47.2	1.037	110
CATTCT	220.8	597.3	55.1	1.037	111
TTTTAG	367.1	943.8	62.7	1.037	112
TTGCTG	123.4	366.5	41.8	1.037	113
AAGAAG	232.6	636.8	53.1	1.036	114
AGTTTG	126.3	370.6	43.2	1.036	115

CAGCAT	109.6	326.9	41.0	1.036	116
TTCCCA	197.9	550.2	50.8	1.036	117
CTGCTT	143.8	418.5	44.4	1.036	118
TGTTTG	185.0	518.7	49.5	1.036	119
TTCAGC	80.5	250.4	35.4	1.035	120
CAGCAA	122.0	361.3	41.7	1.035	121
ATGCCA	87.4	268.3	36.6	1.032	122
GCCATT	69.0	218.3	32.7	1.032	123
AGTGAA	149.8	425.8	46.9	1.032	124
TGCCTT	123.4	363.3	42.0	1.032	125
GCTGAA	80.4	249.0	35.2	1.031	126
TGGCAT	87.8	269.0	36.6	1.031	127
AATGGC	71.3	224.5	33.0	1.031	128
TGTGTT	188.4	524.7	49.5	1.030	129
TCTTCT	272.8	736.3	54.1	1.030	130
ATGGCA	92.6	283.0	36.6	1.030	131
GAGTTT	125.1	367.1	42.0	1.029	132
GAAGAA	298.2	780.5	58.5	1.029	133
AACACA	190.1	527.7	49.7	1.028	134
CTTCTT	238.4	645.5	53.3	1.028	135
CTTCCA	159.5	451.8	46.7	1.028	136
CAAACA	185.2	514.9	49.3	1.028	137
TTGGCT	105.8	317.1	38.8	1.027	138
TTCCAG	194.1	532.7	51.0	1.027	139
AGAAAA	1447.3	3363.5	77.2	1.026	140
AAACAA	642.4	1581.3	67.6	1.026	141
TTCTTC	300.3	782.7	58.6	1.026	142
CAACAG	84.9	263.4	34.0	1.026	143
CATCTT	140.9	402.5	44.9	1.025	144
GTGAGA	117.1	344.7	40.7	1.025	145
TTCTTT	1106.0	2597.4	75.1	1.024	146
TTGTTT	636.8	1561.9	67.7	1.023	147
CTGGAA	195.6	532.2	51.2	1.021	148
AGGATT	96.8	288.4	38.4	1.021	149
GTTTCA	173.5	479.7	48.7	1.021	150
AAGCAG	142.3	407.4	43.7	1.020	151
AAGACT	98.7	293.6	38.3	1.019	152
AAGATG	139.7	396.8	44.6	1.019	153
TGGCTT	132.5	379.7	43.3	1.019	154
AATCTC	104.2	307.0	39.5	1.019	155
TCCAAA	182.6	501.5	49.3	1.018	156
CTGTTG	79.1	245.7	32.4	1.018	157
AGCCAA	106.1	314.1	38.6	1.018	158
GAGATT	107.8	315.7	40.1	1.018	159
TGAGAA	309.5	785.4	61.1	1.018	160

AAGCCA	131.6	376.5	43.1	1.017	161
AAATAC	372.8	938.6	62.0	1.016	162
CTGAGA	231.5	613.0	54.1	1.016	163
CTAAAA	340.8	869.1	59.9	1.015	164
CAACAT	158.1	440.3	46.6	1.015	165
TTCCTG	244.1	645.0	54.3	1.015	166
TTGCCA	82.1	248.5	35.0	1.014	167
TTCTCA	327.1	823.2	61.7	1.014	168
TTTTCT	1469.0	3359.5	77.6	1.014	169
AACAAA	663.8	1610.2	67.5	1.013	170
AGTCTT	97.9	289.3	37.9	1.013	171
AGCCAT	85.6	258.2	35.2	1.013	172
GTATTT	371.8	932.2	62.0	1.012	173
CAGGAA	243.5	642.2	54.1	1.012	174
TTGCCT	98.8	292.5	37.6	1.012	175
TGCTGA	122.6	352.4	41.5	1.012	176
TTTGGA	184.6	501.9	49.4	1.012	177
AAGACA	207.6	559.9	50.6	1.012	178
TGCTTT	263.2	700.0	52.7	1.012	179
ATGGCT	84.0	253.3	34.9	1.011	180
TGGCAA	83.0	249.5	35.0	1.010	181
AGGCAA	98.3	290.5	37.4	1.010	182
AGGAAT	154.6	425.3	47.2	1.009	183
TTTGTT	665.2	1604.7	67.6	1.009	184
ACAAGA	92.3	275.1	35.9	1.009	185
CTTTCC	169.8	469.5	46.3	1.008	186
GGCAAA	83.1	250.4	34.3	1.008	187
TTTGCC	84.0	252.8	34.5	1.008	188
TGAAAC	165.4	456.1	46.5	1.007	189
AGACTG	90.2	267.7	36.2	1.007	190
ATCTTG	93.6	276.2	36.9	1.007	191
ATCTGT	115.1	330.1	40.9	1.007	192
TCTCAG	222.0	587.6	52.4	1.007	193
ACAGAT	115.3	330.8	40.7	1.007	194
ATTCCT	157.8	432.1	47.3	1.006	195
TTCAGA	196.6	524.9	51.0	1.006	196
CAAGAT	92.7	273.4	36.7	1.006	197
GAAGCA	98.1	288.4	37.1	1.006	198
GAAACA	194.4	522.9	49.8	1.005	199
GGAAAG	168.1	463.9	45.9	1.005	200
AATCCT	104.4	302.5	39.0	1.005	201
TGTTTC	194.5	523.1	49.7	1.005	202
TCAGCA	121.8	346.7	41.3	1.004	203
GATTCT	107.5	311.3	39.0	1.004	204
GCTTTT	173.3	472.3	47.6	1.004	205

CTTGTT	94.7	279.5	36.4	1.004	206
TCAGTG	102.6	298.4	38.4	1.004	207
AAAAGC	175.8	478.3	47.7	1.004	208
GAAAGA	318.5	818.8	56.3	1.004	209
TCTGTA	122.5	347.8	41.6	1.004	210
ACTTTG	236.7	612.0	55.4	1.003	211
TGGAAG	146.0	412.6	42.1	1.003	212
AGGTTT	109.7	316.2	39.2	1.002	213
TACAGA	119.4	339.1	41.3	1.002	214
TCTTGT	98.6	288.6	37.0	1.002	215
TTCCTC	193.3	520.5	48.9	1.002	216
TCTCTG	297.9	763.4	56.4	1.002	217
CAGTCT	84.1	252.2	33.6	1.002	218
CACTGA	102.8	297.8	38.4	1.001	219
TCTCCA	193.7	517.9	49.7	1.000	220
AGAATC	109.8	315.3	39.3	1.000	221
GAAAAA	990.3	2269.7	74.4	1.000	222
TCAAAC	80.4	240.0	33.8	1.000	223
ATGTTG	162.0	441.6	46.5	0.999	224
AACAAG	94.4	276.7	36.3	0.999	225
CCACTT	78.3	235.0	33.1	0.999	226
GAGGAA	192.9	517.4	48.9	0.998	227
GTTTGA	79.6	237.4	33.7	0.998	228
TTCCCT	193.2	519.3	48.4	0.998	229
TGCTTC	98.8	287.9	36.9	0.998	230
AGGGAA	195.7	524.9	48.6	0.997	231
AAGTGG	79.0	236.2	33.2	0.997	232
CCAGAA	169.3	455.7	47.8	0.996	233
ATTGCT	123.9	346.0	42.5	0.996	234
TCTGAA	189.5	503.9	49.5	0.996	235
TGTCTT	200.3	536.7	48.5	0.996	236
CAGAAT	188.6	499.9	50.0	0.996	237
TTGTGG	74.5	224.0	32.1	0.995	238
AAACCT	112.6	320.4	39.5	0.995	239
TTCTGG	170.0	456.4	47.8	0.994	240
GTTTCT	202.7	533.9	50.7	0.994	241
TGTATT	368.5	905.1	61.6	0.994	242
GTGATT	79.8	236.3	33.7	0.994	243
TCCAGT	73.8	221.6	31.8	0.993	244
AATACA	366.9	902.5	61.3	0.993	245
CTCATT	168.8	451.1	48.4	0.993	246
GGAAGA	190.0	510.7	47.3	0.993	247
CGACTA	0.3	1.0	0.2	0.992	248
AAGGGA	135.1	377.9	41.4	0.992	249
TTTTTC	991.0	2251.9	74.4	0.992	250

TTTGCT	242.0	624.2	53.4	0.991	251
TGGAGA	198.9	525.0	49.8	0.991	252
TGAAGA	199.5	524.4	50.4	0.991	253
AATGAG	170.2	453.2	48.3	0.991	254
ACTGGA	74.1	221.5	31.8	0.990	255
TCCCTT	134.3	374.6	41.4	0.990	256
CTCTCA	113.5	321.5	39.1	0.990	257
AGAAGT	138.7	379.0	44.4	0.990	258
TCTTCA	210.7	546.8	52.0	0.989	259
AGCAAA	241.5	621.5	53.3	0.989	260
TCTCAC	108.0	313.9	35.4	0.989	261
CCACAT	103.0	295.8	37.1	0.989	262
CATGAG	112.1	317.2	39.0	0.988	263
AGAACA	204.1	529.9	51.7	0.988	264
AATCAC	81.9	239.9	34.1	0.988	265
ACTTCT	141.2	384.3	44.5	0.988	266
CCTTTC	133.7	373.8	40.7	0.988	267
AGATCA	115.9	326.8	39.4	0.988	268
AGAAAC	201.5	526.9	50.5	0.988	269
CCTTCA	100.7	289.1	36.8	0.988	270
AGAATG	243.3	615.3	55.6	0.987	271
CATTTG	201.2	523.1	51.3	0.987	272
GTGGAA	83.2	244.7	33.4	0.987	273
CTCATG	115.1	325.3	38.9	0.987	274
CAAAGA	246.1	628.6	53.9	0.987	275
TTCCAA	189.4	496.0	50.0	0.987	276
CCACAA	80.5	236.9	33.1	0.986	277
TGTGGT	131.3	363.8	41.5	0.986	278
TGAAGG	105.0	300.8	36.8	0.986	279

[0109] Primers are evaluated by analyzing the distribution of corresponding PCR products. For example, the 6-MW was evaluated by scanning the whole genome sequence for the appearance of the sequence ATGTGG, the3' end hexamer of the primer with SEQ ID NO: 281. Whenever one sequence was found, its complementary reverse sequence (CCACAT) was searched and counted within 2000 bp downstream of the sequence, which was the estimated maximum extension length given the conditions used for PCR. For the convenience of operation, the chromosomes were divided into 1 Mbp segments. The numbers of PCR products derived

each segment were then plotted against the chromosomal segments to view the distribution pattern. A Java™ Swing-based application was developed to visualize the distribution.

[0110] The primers underwent this computational analysis included ATCTCA (NYP6-1, SEQ ID NO: 6), TGAGAT (NYP6-2, SEQ ID NO: 5), ATGTGG (Telenius, 3' end hexamer of SEQ ID NO: 281), CTAGAA (Fiegler-1, 3' end hexamer of SEQ ID NO: 282), TAGGAG (Fiegler-2, 3' end hexamer of SEQ ID NO: 283), and TTCTAG (Fiegler-3, 3' end hexamer of SEQ ID NO: 284). Table 2 below shows their performances evaluated using the method described herein.

[0111] Table 2. Summary of fitness scores of six PCR primers

	Sequence	Ext. Length	δc	μc	C_c (%)	Fitness Score	Reference
NYP6-1	ATCTCA	2537	162.52	493.13	45.62	1.07	
NYP6-2	TGAGAT	2926	191.07	579.39	54.35	1.13	
Telenius	ATGTGG	2000	74.42	188.48	20.93	0.84	Telenius *et al* ., 1992
Fiegler-1	CTAGAA	2000	89.75	193.86	21.47	0.76	Fiegler *et al* ., 2003
Fiegler-2	TAGGAG	2000	34.64	96.21	13.05	0.87	
Fiegler-3	TTCTAG	2000	88.15	191.67	21.43	0.76	

[0112] A Graphical User Interface (GUI) was developed to provide an overview of the coverage for all six primers. FIG 5 shows the distribution of amplicons for these primers. The intensity of the bands represents the quantity of amplicons.

[0113] The DOP-PCR protocol designed by Telenius and co-workers (Telenius et al., 1992) was used for WGA. DOP-PCR is a specialized polymerase chain reaction that uses a single primer, starting with a fix lead sequence, followed by a region of six degenerate nucleotides. At the 3' end is the hexameric sequence ATGTGG, predicted to occur approximately every 4630 bp (Vivian *et al.*, 1996). In addition, WGA was performed using only

the hexameric sequence as a primer. The hexameric sequences obtained from the GA were tested, along with those originally used by Telenius *et al.* (1992) and the 3-primer set designed by Fiegler *et al.* (2003). When hexamers, e.g., SEQ ID NOs: 5 and 6 were used for PCR, the cycles were slightly modified. For instance, the PCR was performed with the following steps: 94°C for 9 minutes for initial denaturation, followed by 40 rounds of amplification with 94 °C for 1 minutes, 20 °C for 1 minutes and 72 °C for 3 minutes, and finished with a final extension at 72°C for 8 minutes.

[0114] The amplified genomic DNAs from above PCR reactions were used for the CGH experiments. As discussed above, the CGH is a well-known technique. Thus, there is no detail description about the performance of the CGH in the present invention. Briefly, the PCR-amplified DNAs were purified before labeling by nick translation. The labeled probes having a fragment size in the range of 500-2000 bp were purified by micro-columns and precipitated for hybridization. The target metaphase chromosomes were prepared from PHA stimulated peripheral blood lymphocyte cultures following conventional methods. Then, the hybridization was done by conventional methods.

[0115] The data from two different comparative genomic hybridisation (CGH) experiments demonstrated biased amplification when whole genome amplification (WGA) was performed using the Telenius primers. FIG 6 (left panel) shows normal human genomic DNA hybridised to itself without prior amplification, i.e. unamplified normal vs. unamplified normal. The corresponding CGH profiles confirm no copy number abnormalities under these conditions, as predicted. FIG 6 (right panel) shows normal human genomic DNA that has been DOP-PCR amplified with the Telenius primers hybridised against the same DNA but without amplification i.e. amplified normal vs. unamplified normal. The corresponding CGH profiles now show

multiple spurious amplifications. A bar on the left-hand-side of the ideogram indicates deletion, and a bar on the right-hand-side of the ideogram indicates amplification.

[0116] Four CGH experiments were conducted to compare their performance respectively. Normal unamplified human genomic DNA was used to compare with amplified normal human genomic DNA using 4 different primers. The results of the experiments are shown in FIG 7. In one of the experiment, NYP-22 primer was used. However, from the CGH result, it is obvious that NYP-6 primer produced the most even coverage. FIGs 7 and 8 summarizes the CGH images obtained from PCR amplification using the below primers:

A. 5'-ATCTCA-3' (SEQ ID NO: 6) (NYP6-1);

B. 5'-CCGACTCGAGNNNNNNATCTCA-3' (SEQ ID NO: 280) (NYP-22);

C. 5'-CCGACTCGAGNNNNNNATGTGG-3' (SEQ ID NO: 281) (Telenius); and

D. [5'-CCGACTCGAGNNNNNNCTAGAA-3' (SEQ ID NO: 282),

5'-CCGACTCGAGNNNNNNTAGGAG-3' (SEQ ID NO: 283),

5'-CCGACTCGAGNNNNNNTTCTAG-3' (SEQ ID NO: 284)] (Fiegler).

Deletions are indicated by a bar on the left of the schematic map of the chromosome while a bar on the right-side marks amplifications.

[0117] The CGH results show that current invention produces better WGA primers that are validated. To further quantify the result. The deviation of the intensity curve were measured from the ratio=1 line. FIG 9 depicts the method used to calculate the deviation. The deviation was considered as error generated due to un-even amplification. Comparisons of the squared error for different primers are shown in FIG 10.

[0118] Despite the effort put in to sequence the complete human genome, there are still loci where the DNA sequences are not available. In this region, the coverage of the primers is not predictable. There are also certain loci where the coverage is either below or above the mean values. It is almost impossible to obtain primer that produces the ideal coverage, knowing the complexity of the human genome. Based on the performance of the primers measured using CGH, normalization method is developed to correct the minor biasness in those loci.

[0119] Since the error is quantifiable, based on the predictive data, CGH result can be actually normalized to further improve the performance of the WGA primers. FIG 11 shows the typical distribution of amplicon calculated based on the method described earlier. FIG 11 shows distribution of amplicons calculated from chromosome one.

[0120] The deviation is obtained from the CGH images where WGA primer is applied onto normal genomic DNA. With the two sets of numerical values available, the correlation between them was found. From now on, it is described the in-silico predicted PCR result dataset as predictive dataset and the deviation result as experimental dataset.

[0121] The mean value was first deducted from the predictive dataset.

This is to obtain the deviation of the dataset from the mean value to make it comparable to the experimental dataset. Then, the moving average was applied to the predictive dataset to smoothen the curve. A window of 20 samples before and after the current data point was used. The filter weight coefficients used are calculated using Normal (Gaussian) function. FIG 12 shows the smoothened predictive dataset together with the experimental dataset. FIG 12 shows predictive versus experimental datasets.

[0000] From the dataset, a scatter plot was plotted to find correlation between the two datasets. FIG 13 shows a sample scatter plot for chromosome 1 datasets. Scatter Plot For Chromosome 1. It was found that there are two correlation paths. Then, polynomial regression for each of them was performed. FIGs 14A and 14B show the 3rd order polynomial regression function applied to each of the path. The polynomial function and its squared Pearson coefficient are shown in the figure itself. FIG 14A shows Upper Path Regression Curve, and FIG 14B shows Lower Path Regression Curve.

[0000] With the known polynomial regression functions, they were applied onto the predictive dataset and obtained the normalised dataset. FIG 15 shows the plots of the 3 datasets for chromosome 1. With these normalised dataset, it can be used to correct the experimental curve. The eventual corrected result is shown in FIG 16. The deviation from the normal expression curve is significantly reduced. FIG 16 shows the Corrected Curve.

[0000] From the foregoing it will be appreciated that, although specific embodiments of the invention have been described herein for purposes of illustration, various modifications may be made without deviating from the spirit and scope of the invention. Accordingly, the invention is not limited except as by the appended claims.

CLAIMS

What is claimed is:

1. A genetic algorithms-based method for selecting primers for genomic DNA amplification, comprising the steps of:

defining a chromosome, wherein the chromosome comprises a plurality of alleles representing individual features of the primers;

creating an initiation population of the chromosomes;

defining a fitness function for evaluating the quality of each chromosomes;

calculating the fitness function of each chromosomes from the initial population;

selecting two chromosomes at each time to generate two new chromosomes by way of crossover, wherein the new chromosomes will constitute a next generation population;

selecting a predetermined percentage of the new chromosomes from the crossover operation for single mutation, wherein the mutated chromosomes will be part of the next generation population;

calculating the fitness function of each chromosomes from the next generation population, wherein chromosomes failing to meet a predetermined value of the fitness function will be excluded from the next generation population;

repeating the steps of crossover, single mutation and calculation of the fitness function to determine whether a chromosome will be excluded from the next generation population;

terminating the repeating; and

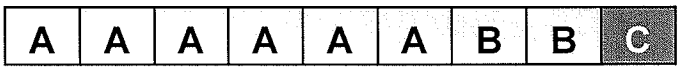
outputting the primer sequences.

2. The method of claim 1, wherein the genomic DNA is in any form selected from the group consisting of plasmid, cosmid, P1, P1 artificial chromosome, YAC, BAC, isolated chromosomes, and isolated whole genome DNA.

3. The method of claim 2, wherein the genomic DNA is from any species selected from the group consisting of bacteria, fungi, plants, and mammals.

4. The method of claim 3, wherein the genomic DNA is from human.

5. The method of claim 1, wherein, in the step of defining a chromosome, the chromosome is represented by the following structure:



wherein the first 6 alleles represented by **A** were nucleotides of A, C, G, or T, identified by integer values 1 to 4 respectively, the next 2 alleles represented by **B** were nucleotides of None, A, C, G or T, identified by integer values 0 to 4 respectively, and the last allele of **C** represented extension length with an integer ranged between 1000 and 3000.

6. The method of claim 1, wherein the initial population of the chromosomes is created randomly.

7. The method of claim 1, wherein the fitness function is defined by the following equation:

$$f(S_c) = \frac{1}{\frac{2\sigma_c}{\mu_c} + \frac{1}{2}(1 - \frac{C_c}{C_{total}})}$$

wherein σ_c = variance of the number of amplicons in each 1 Mbp segments, μ_c = average number of amplicons in each 1 Mbp segments, C_c = coverage defined as the number of nucleotides reproduced (amplified), and C_{total} = total number of nucleotides.

8. The method of claim 1, wherein the crossover is conducted in a single point.

9. The method of claim 1, wherein the crossover is conducted in two points.

10. The method of claim 1, wherein the repeating is terminated at a preset number of generations.

11. The method of claim 1, wherein the repeating is terminated at outputting an optimal primer.

12. A method for unbiased amplification of genomic DNA, comprising steps of:

determining optimal primers by the methods of any one of claims 1-11; and

amplifying the genomic DNA by using the determined optimal primers.

13. The unbiased amplification method of claim 12, wherein the genomic DNA is amplified by PCR

14. The unbiased amplification method of any of claims 12-13, wherein the genomic DNA is from human.

15. The unbiased amplification method of claim 14, wherein the primers are oligonucletides comprising at 3' ends of the oligonucleotides a hexamer selected from the group consisting of SEQ ID NOs: 1-279.

16. The unbiased amplification method of claim 14, wherein the primers are oligonucleotides selected from the group consisting of: SEQ ID NOs: 1-279.

17. A method for DNA analysis comprising using the amplified DNA fragments obtained by the method of any one of claims 12 to 16.

18. The DNA analysis method of claim 17, wherein the methods for DNA analysis is selected from the group consisting of comparative genomic hybridization (CGH), representational difference analysis (RDA)-, analytical PCR-, restriction fragment length polymorphism analysis (RFLP)-, single strand conformation polymorphism (SSCP)-, DNA sequence-, "loss of heterozygosity" (LOH), fingerprint- and/or FISH analysis.

19. A method of identifying genome abnormalities, comprising the steps of:

determining optimal primers by the methods of any one of claims 1-11;

amplifying the genomic DNA by using the determined optimal primers; and

identifying genome abnormalities by hybridizing the amplified genomic DNA to reference chromosomes.

20. The genome abnormality identification method of claim 19, wherein the genomic DNA is amplified by PCR.

21. The genome abnormality identification method of claim 19, wherein the genome is from human.

22. The genome abnormality identification method of claim 21, wherein the primers are oligonucletides comprising at 3' ends of the oligonucleotides a hexamer selected from the group consisting of SEQ ID NOs: 1-279.

23. The genome abonormality identification method of claim 21, wherein the primers are oligonucleotides selected from the group consisting of SEQ ID NOs: 1-279.

24. The genome abnormality identification method of claim 18, wherein the hybridisation technique is selected from the group consisting of comparative genomic hybridization (CGH), representational difference analysis (RDA)-, analytical PCR-, restriction fragment length polymorphism analysis (RFLP)-, single strand conformation polymorphism (SSCP)-, DNA sequence-, "loss of heterozygosity" (LOH), fingerprint- and/or FISH analysis..

25. A system for executing genetic algorithms-based programs for selecting primers for genomic DNA amplification, comprising:

a portal PC for transmitting request from an end user;

a head node for receiving the request from the portal PC, wherein the head node includes a job manager module for creating tasks and assigning jobs, and a lookup service module for transmitting the tasks; and

a computing nodes for executing the genetic algorithms, receiving and returning tasks from the lookup service module, and registering execution results.

26. An oligonucleotide for being used as a primer for unbiased amplification of human genomic DNA comprising at 3' ends of the oligonucleotides a hexamer selected from the group consisting of SEQ ID NOs: 1-279.

27. An oligonucleotide for being used as a primer for unbiased amplification of human genomic DNA is consisting of a hexamer selected from the group consisting of SEQ ID NOs: 1-279.

28. A kit for genomic DNA amplification, comprising a primer as defined in any one of claims 1-11.

29. A kit for human genomic DNA amplification, comprising a primer, wherein the primer comprising at 3' ends of the oligonucleotides a hexamer selected from the group consisting of SEQ ID NOs: 1-279.

30. A kit for human genomic DNA amplification, comprising a primer, wherein the primer is consisting of a hexamer selected from the group consisting of SEQ ID NOs: 1-279.

31. A method for normalizing the intensity of hybridization from a comparative genomic hybridization experiment, wherein the CGH experiment uses amplified genomic DNAs as probes, and the amplified genomic DNAs are obtained by using a selected primer, the method comprising steps of:

determining the number of amplicons of a target genomic DNA when the selected primer is used;

determining the deviation from the CGH images where the selected primer is applied onto normal genomic DNA;

determining a correlation between a predictive dataset and an experimental dataset; and

applying the polynomial regression to the experimental dataset to obtain a normalized CGH intensity.

SEQUENCE LISTING

<110> Nanyang Polytechnic
Kon, Oi Lian

<120> Method and System for Unbiased Genome Amplification

<130> 1138.P043/GDL

<160> 284

<170> PatentIn version 3.2

<210> 1
<211> 6
<212> DNA
<213> human

<400> 1
tttcct 6

<210> 2
<211> 6
<212> DNA
<213> human

<400> 2
aggaaa 6

<210> 3
<211> 6
<212> DNA
<213> human

<400> 3
ttttgt 6

<210> 4
<211> 6
<212> DNA
<213> human

<400> 4
acaaaa 6

<210> 5
<211> 6
<212> DNA
<213> human

<400> 5
tgagat 6

<210> 6
<211> 6
<212> DNA
<213> human

<400> 6
atctca 6

<210> 7
<211> 6
<212> DNA
<213> human

<400> 7
tttctg 6

<210> 8
<211> 6
<212> DNA
<213> human

<400> 8
aacaga 6

<210> 9
<211> 6
<212> DNA
<213> human

<400> 9
cagaaa 6

<210> 10
<211> 6
<212> DNA
<213> human

<400> 10
ttcctt 6

<210> 11
<211> 6
<212> DNA
<213> human

<400> 11
aaaaag 6

<210> 12
<211> 6
<212> DNA
<213> human

<400> 12
aaggaa 6

<210> 13
<211> 6
<212> DNA
<213> human

<400> 13
ttctgt 6

<210> 14
<211> 6
<212> DNA
<213> human

<400> 14
atctct 6

<210> 15
<211> 6
<212> DNA
<213> human

<400> 15
agagat 6

<210> 16
<211> 6
<212> DNA
<213> human

<400> 16
ttctct 6

<210> 17
<211> 6
<212> DNA
<213> human

<400> 17
acagaa 6

<210> 18
<211> 6
<212> DNA
<213> human

<400> 18
tctgtt 6

<210> 19
<211> 6
<212> DNA
<213> human

<400> 19
cttttt 6

<210> 20
<211> 6
<212> DNA
<213> human

<400> 20
aaaaga 6

<210> 21
<211> 6
<212> DNA
<213> human

<400> 21
tttttg 6

<210> 22
<211> 6
<212> DNA
<213> human

<400> 22
cacttt 6

<210> 23
<211> 6
<212> DNA
<213> human

<400> 23
tctttt 6

<210> 24
<211> 6
<212> DNA
<213> human

<400> 24
agagaa 6

<210> 25
<211> 6
<212> DNA
<213> human

<400> 25
ctgttt 6

<210> 26
<211> 6
<212> DNA
<213> human

<400> 26
aaacag 6

<210> 27
<211> 6
<212> DNA
<213> human

<400> 27
attctc 6

<210> 28
<211> 6
<212> DNA
<213> human

<400> 28
caaaaa 6

<210> 29
<211> 6
<212> DNA
<213> human

<400> 29
catgtt 6

<210> 30
<211> 6
<212> DNA
<213> human

<400> 30
tttcca 6

<210> 31
<211> 6
<212> DNA
<213> human

<400> 31
aacatg 6

<210> 32
<211> 6
<212> DNA
<213> human

<400> 32
tcactt 6

<210> 33
<211> 6
<212> DNA
<213> human

<400> 33
aaagtg 6

<210> 34
<211> 6
<212> DNA
<213> human

<400> 34
aagtga 6

<210> 35
<211> 6
<212> DNA
<213> human

<400> 35
ctgaag 6

<210> 36
<211> 6
<212> DNA
<213> human

<400> 36
cagaag 6

<210> 37
<211> 6
<212> DNA
<213> human

<400> 37
ttgctt 6

<210> 38
<211> 6
<212> DNA
<213> human

<400> 38
aagaaa 6

<210> 39
<211> 6
<212> DNA
<213> human

<400> 39
ggaaaa 6

<210> 40
<211> 6
<212> DNA
<213> human

<400> 40
cttctg 6

<210> 41
<211> 6
<212> DNA
<213> human

<400> 41
ttttcc 6

<210> 42
<211> 6
<212> DNA
<213> human

<400> 42
cttcag 6

<210> 43
<211> 6
<212> DNA
<213> human

<400> 43
tttctc 6

<210> 44
<211> 6
<212> DNA
<213> human

<400> 44
gggaaa 6

<210> 45
<211> 6
<212> DNA
<213> human

<400> 45
gagaaa 6

<210> 46
<211> 6
<212> DNA
<213> human

<400> 46
tttctt 6

<210> 47
<211> 6
<212> DNA
<213> human

<400> 47
gtgaaa 6

<210> 48
<211> 6
<212> DNA
<213> human

<400> 48
cttttc 6

<210> 49
<211> 6
<212> DNA
<213> human

<400> 49
ttttga 6

<210> 50
<211> 6
<212> DNA
<213> human

<400> 50
tttccc 6

<210> 51
<211> 6
<212> DNA
<213> human

<400> 51
gagaat 6

<210> 52
<211> 6
<212> DNA
<213> human

<400> 52
gaaaag 6

<210> 53
<211> 6
<212> DNA
<213> human

<400> 53
aaaaca 6

<210> 54
<211> 6
<212> DNA
<213> human

<400> 54
aagaga 6

<210> 55
<211> 6
<212> DNA
<213> human

<400> 55
tcaaaa 6

<210> 56
<211> 6
<212> DNA
<213> human

<400> 56
ttctcc 6

<210> 57
<211> 6
<212> DNA
<213> human

<400> 57
attttt 6

<210> 58
<211> 6
<212> DNA
<213> human

<400> 58
tgtttt 6

<210> 59
<211> 6
<212> DNA
<213> human

<400> 59
caaaac 6

<210> 60
<211> 6
<212> DNA
<213> human

<400> 60
aaagga 6

<210> 61
<211> 6
<212> DNA
<213> human

<400> 61
tctctt 6

<210> 62
<211> 6
<212> DNA
<213> human

<400> 62
aaaaat 6

<210> 63
<211> 6
<212> DNA
<213> human

<400> 63
tccttt 6

<210> 64
<211> 6
<212> DNA
<213> human

<400> 64
ccattt 6

<210> 65
<211> 6
<212> DNA
<213> human

<400> 65
tggaaa 6

<210> 66
<211> 6
<212> DNA
<213> human

<400> 66
tgaact 6

<210> 67
<211> 6
<212> DNA
<213> human

<400> 67
agttca 6

<210> 68
<211> 6
<212> DNA
<213> human

<400> 68
aaagag 6

<210> 69
<211> 6
<212> DNA
<213> human

<400> 69
ccaaaa 6

<210> 70
<211> 6
<212> DNA
<213> human

<400> 70
agaaga 6

<210> 71
<211> 6
<212> DNA
<213> human

<400> 71
tgccat 6

<210> 72
<211> 6
<212> DNA
<213> human

<400> 72
ggtttt 6

<210> 73
<211> 6
<212> DNA
<213> human

<400> 73
aaaacc 6

<210> 74
<211> 6
<212> DNA
<213> human

<400> 74
ttttgg 6

<210> 75
<211> 6
<212> DNA
<213> human

<400> 75
gttttg 6

<210> 76
<211> 6
<212> DNA
<213> human

<400> 76
tacaaa 6

<210> 77
<211> 6
<212> DNA
<213> human

<400> 77
tcaaga 6

<210> 78
<211> 6
<212> DNA
<213> human

<400> 78
tcttga 6

<210> 79
<211> 6
<212> DNA
<213> human

<400> 79
tttcag 6

<210> 80
<211> 6
<212> DNA
<213> human

<400> 80
ctgaaa 6

<210> 81
<211> 6
<212> DNA
<213> human

<400> 81
aaatgg 6

<210> 82
<211> 6
<212> DNA
<213> human

<400> 82
catttc 6

<210> 83
<211> 6
<212> DNA
<213> human

<400> 83
ctcttt 6

<210> 84
<211> 6
<212> DNA
<213> human

<400> 84
aagcaa 6

<210> 85
<211> 6
<212> DNA
<213> human

<400> 85
tggttt 6

<210> 86
<211> 6
<212> DNA
<213> human

<400> 86
ttctgc 6

<210> 87
<211> 6
<212> DNA
<213> human

<400> 87
gcagaa 6

<210> 88
<211> 6
<212> DNA
<213> human

<400> 88
ggagaa 6

<210> 89
<211> 6
<212> DNA
<213> human

<400> 89
cctgaa 6

<210> 90
<211> 6
<212> DNA
<213> human

<400> 90
aaagca 6

<210> 91
<211> 6
<212> DNA
<213> human

<400> 91
aaacca 6

<210> 92
<211> 6
<212> DNA
<213> human

<400> 92
aaactc 6

<210> 93
<211> 6
<212> DNA
<213> human

<400> 93
aaggca 6

<210> 94
<211> 6
<212> DNA
<213> human

<400> 94
ttttta 6

<210> 95
<211> 6
<212> DNA
<213> human

<400> 95
gaaatg 6

<210> 96
<211> 6
<212> DNA
<213> human

<400> 96
tttgta 6

<210> 97
<211> 6
<212> DNA
<213> human

<400> 97
tttcac 6

<210> 98
<211> 6
<212> DNA
<213> human

<400> 98
aaaaaa 6

<210> 99
<211> 6
<212> DNA
<213> human

<400> 99
ttcagg 6

<210> 100
<211> 6
<212> DNA
<213> human

<400> 100
taaaaa 6

<210> 101
<211> 6
<212> DNA
<213> human

<400> 101
tgggaa 6

<210> 102
<211> 6
<212> DNA
<213> human

<400> 102
aaaagg 6

<210> 103
<211> 6
<212> DNA
<213> human

<400> 103
agatga 6

<210> 104
<211> 6
<212> DNA
<213> human

<400> 104
tttttt 6

<210> 105
<211> 6
<212> DNA
<213> human

<400> 105
caaact 6

<210> 106
<211> 6
<212> DNA
<213> human

<400> 106
ttcact 6

<210> 107
<211> 6
<212> DNA
<213> human

<400> 107
aaagaa 6

<210> 108
<211> 6
<212> DNA
<213> human

<400> 108
atgctg 6

<210> 109
<211> 6
<212> DNA
<213> human

<400> 109
cctttt 6

<210> 110
<211> 6
<212> DNA
<213> human

<400> 110
tcatct 6

<210> 111
<211> 6
<212> DNA
<213> human

<400> 111
cattct 6

<210> 112
<211> 6
<212> DNA
<213> human

<400> 112
ttttag 6

<210> 113
<211> 6
<212> DNA
<213> human

<400> 113
ttgctg 6

<210> 114
<211> 6
<212> DNA
<213> human

<400> 114
aagaag 6

<210> 115
<211> 6
<212> DNA
<213> human

<400> 115
agtttg 6

<210> 116
<211> 6
<212> DNA
<213> human

<400> 116
cagcat 6

<210> 117
<211> 6
<212> DNA
<213> human

<400> 117
ttccca 6

<210> 118
<211> 6
<212> DNA
<213> human

<400> 118
ctgctt 6

<210> 119
<211> 6
<212> DNA
<213> human

<400> 119
tgtttg 6

<210> 120
<211> 6
<212> DNA
<213> human

<400> 120
ttcagc 6

<210> 121
<211> 6
<212> DNA
<213> human

<400> 121
cagcaa 6

<210> 122
<211> 6
<212> DNA
<213> human

<400> 122
atgcca 6

<210> 123
<211> 6
<212> DNA
<213> human

<400> 123
gccatt 6

<210> 124
<211> 6
<212> DNA
<213> human

<400> 124
agtgaa 6

<210> 125
<211> 6
<212> DNA
<213> human

<400> 125
tgcctt 6

<210> 126
<211> 6
<212> DNA
<213> human

<400> 126
gctgaa 6

<210> 127
<211> 6
<212> DNA
<213> human

<400> 127
tggcat 6

<210> 128
<211> 6
<212> DNA
<213> human

<400> 128
aatggc 6

<210> 129
<211> 6
<212> DNA
<213> human

<400> 129
tgtgtt 6

<210> 130
<211> 6
<212> DNA
<213> human

<400> 130
tcttct 6

<210> 131
<211> 6
<212> DNA
<213> human

<400> 131
atggca 6

<210> 132
<211> 6
<212> DNA
<213> human

<400> 132
gagttt 6

<210> 133
<211> 6
<212> DNA
<213> human

<400> 133
gaagaa 6

<210> 134
<211> 6
<212> DNA
<213> human

<400> 134
aacaca 6

<210> 135
<211> 6
<212> DNA
<213> human

<400> 135
cttctt 6

<210> 136
<211> 6
<212> DNA
<213> human

<400> 136
cttcca 6

<210> 137
<211> 6
<212> DNA
<213> human

<400> 137
caaaca 6

<210> 138
<211> 6
<212> DNA
<213> human

<400> 138
ttggct 6

<210> 139
<211> 6
<212> DNA
<213> human

<400> 139
ttccag 6

<210> 140
<211> 6
<212> DNA
<213> human

<400> 140
agaaaa 6

<210> 141
<211> 6
<212> DNA
<213> human

<400> 141
aaacaa 6

<210> 142
<211> 6
<212> DNA
<213> human

<400> 142
ttcttc 6

<210> 143
<211> 6
<212> DNA
<213> human

<400> 143
caacag 6

<210> 144
<211> 6
<212> DNA
<213> human

<400> 144
catctt 6

<210> 145
<211> 6
<212> DNA
<213> human

<400> 145
gtgaga 6

<210> 146
<211> 6
<212> DNA
<213> human

<400> 146
ttcttt 6

<210> 147
<211> 6
<212> DNA
<213> human

<400> 147
ttgttt 6

<210> 148
<211> 6
<212> DNA
<213> human

<400> 148
ctggaa 6

<210> 149
<211> 6
<212> DNA
<213> human

<400> 149
aggatt 6

<210> 150
<211> 6
<212> DNA
<213> human

<400> 150
gtttca 6

<210> 151
<211> 6
<212> DNA
<213> human

<400> 151
aagcag 6

<210> 152
<211> 6
<212> DNA
<213> human

<400> 152
aagact 6

<210> 153
<211> 6
<212> DNA
<213> human

<400> 153
aagatg 6

<210> 154
<211> 6
<212> DNA
<213> human

<400> 154
tggctt 6

<210> 155
<211> 6
<212> DNA
<213> human

<400> 155
aatctc 6

<210> 156
<211> 6
<212> DNA
<213> human

<400> 156
tccaaa 6

<210> 157
<211> 6
<212> DNA
<213> human

<400> 157
ctgttg 6

<210> 158
<211> 6
<212> DNA
<213> human

<400> 158
agccaa 6

<210> 159
<211> 6
<212> DNA
<213> human

<400> 159
gagatt 6

<210> 160
<211> 6
<212> DNA
<213> human

<400> 160
tgagaa 6

<210> 161
<211> 6
<212> DNA
<213> human

<400> 161
aagcca 6

<210> 162
<211> 6
<212> DNA
<213> human

<400> 162
aaatac 6

<210> 163
<211> 6
<212> DNA
<213> human

<400> 163
ctgaga 6

<210> 164
<211> 6
<212> DNA
<213> human

<400> 164
ctaaaa 6

<210> 165
<211> 6
<212> DNA
<213> human

<400> 165
caacat 6

<210> 166
<211> 6
<212> DNA
<213> human

<400> 166
ttcctg 6

<210> 167
<211> 6
<212> DNA
<213> human

<400> 167
ttgcca 6

<210> 168
<211> 6
<212> DNA
<213> human

<400> 168
ttctca 6

<210> 169
<211> 6
<212> DNA
<213> human

<400> 169
ttttct 6

<210> 170
<211> 6
<212> DNA
<213> human

<400> 170
aacaaa 6

<210> 171
<211> 6
<212> DNA
<213> human

<400> 171
agtctt 6

<210> 172
<211> 6
<212> DNA
<213> human

<400> 172
agccat 6

<210> 173
<211> 6
<212> DNA
<213> human

<400> 173
gtattt 6

<210> 174
<211> 6
<212> DNA
<213> human

<400> 174
caggaa 6

<210> 175
<211> 6
<212> DNA
<213> human

<400> 175
ttgcct 6

<210> 176
<211> 6
<212> DNA
<213> human

<400> 176
tgctga 6

<210> 177
<211> 6
<212> DNA
<213> human

<400> 177
tttgga 6

<210> 178
<211> 6
<212> DNA
<213> human

<400> 178
aagaca 6

<210> 179
<211> 6
<212> DNA
<213> human

<400> 179
tgcttt 6

<210> 180
<211> 6
<212> DNA
<213> human

<400> 180
atggct 6

<210> 181
<211> 6
<212> DNA
<213> human

<400> 181
tggcaa 6

<210> 182
<211> 6
<212> DNA
<213> human

<400> 182
aggcaa 6

<210> 183
<211> 6
<212> DNA
<213> human

<400> 183
aggaat 6

<210> 184
<211> 6
<212> DNA
<213> human

<400> 184
tttgtt 6

<210> 185
<211> 6
<212> DNA
<213> human

<400> 185
acaaga 6

<210> 186
<211> 6
<212> DNA
<213> human

<400> 186
ctttcc 6

<210> 187
<211> 6
<212> DNA
<213> human

<400> 187
ggcaaa 6

<210> 188
<211> 6
<212> DNA
<213> human

<400> 188
tttgcc 6

<210> 189
<211> 6
<212> DNA
<213> human

<400> 189
tgaaac 6

<210> 190
<211> 6
<212> DNA
<213> human

<400> 190
agactg 6

<210> 191
<211> 6
<212> DNA
<213> human

<400> 191
atcttg 6

<210> 192
<211> 6
<212> DNA
<213> human

<400> 192
atctgt 6

<210> 193
<211> 6
<212> DNA
<213> human

<400> 193
tctcag 6

<210> 194
<211> 6
<212> DNA
<213> human

<400> 194
acagat 6

<210> 195
<211> 6
<212> DNA
<213> human

<400> 195
attcct 6

<210> 196
<211> 6
<212> DNA
<213> human

<400> 196
ttcaga 6

<210> 197
<211> 6
<212> DNA
<213> human

<400> 197
caagat 6

<210> 198
<211> 6
<212> DNA
<213> human

<400> 198
gaagca 6

<210> 199
<211> 6
<212> DNA
<213> human

<400> 199
gaaaca 6

<210> 200
<211> 6
<212> DNA
<213> human

<400> 200
ggaaag 6

<210> 201
<211> 6
<212> DNA
<213> human

<400> 201
aatcct 6

<210> 202
<211> 6
<212> DNA
<213> human

<400> 202
tgtttc 6

<210> 203
<211> 6
<212> DNA
<213> human

<400> 203
tcagca 6

<210> 204
<211> 6
<212> DNA
<213> human

<400> 204
gattct 6

<210> 205
<211> 6
<212> DNA
<213> human

<400> 205
gctttt 6

<210> 206
<211> 6
<212> DNA
<213> human

<400> 206
cttgtt 6

<210> 207
<211> 6
<212> DNA
<213> human

<400> 207
tcagtg 6

<210> 208
<211> 6
<212> DNA
<213> human

<400> 208
aaaagc 6

<210> 209
<211> 6
<212> DNA
<213> human

<400> 209
gaaaga 6

<210> 210
<211> 6
<212> DNA
<213> human

<400> 210
tctgta 6

<210> 211
<211> 6
<212> DNA
<213> human

<400> 211
actttg 6

<210> 212
<211> 6
<212> DNA
<213> human

<400> 212
tggaag 6

<210> 213
<211> 6
<212> DNA
<213> human

<400> 213
aggttt 6

<210> 214
<211> 6
<212> DNA
<213> human

<400> 214
tacaga 6

<210> 215
<211> 6
<212> DNA
<213> human

<400> 215
tcttgt 6

<210> 216
<211> 6
<212> DNA
<213> human

<400> 216
ttcctc 6

<210> 217
<211> 6
<212> DNA
<213> human

<400> 217
tctctg 6

<210> 218
<211> 6
<212> DNA
<213> human

<400> 218
cagtct 6

<210> 219
<211> 6
<212> DNA
<213> human

<400> 219
cactga 6

<210> 220
<211> 6
<212> DNA
<213> human

<400> 220
tctcca 6

<210> 221
<211> 6
<212> DNA
<213> human

<400> 221
agaatc 6

<210> 222
<211> 6
<212> DNA
<213> human

<400> 222
gaaaaa 6

<210> 223
<211> 6
<212> DNA
<213> human

<400> 223
tcaaac 6

<210> 224
<211> 6
<212> DNA
<213> human

<400> 224
atgttg 6

<210> 225
<211> 6
<212> DNA
<213> human

<400> 225
aacaag 6

<210> 226
<211> 6
<212> DNA
<213> human

<400> 226
ccactt 6

<210> 227
<211> 6
<212> DNA
<213> human

<400> 227
gaggaa 6

<210> 228
<211> 6
<212> DNA
<213> human

<400> 228
gtttga 6

<210> 229
<211> 6
<212> DNA
<213> human

<400> 229
ttccct 6

<210> 230
<211> 6
<212> DNA
<213> human

<400> 230
tgcttc 6

<210> 231
<211> 6
<212> DNA
<213> human

<400> 231
agggaa 6

<210> 232
<211> 6
<212> DNA
<213> human

<400> 232
aagtgg 6

<210> 233
<211> 6
<212> DNA
<213> human

<400> 233
ccagaa 6

<210> 234
<211> 6
<212> DNA
<213> human

<400> 234
attgct 6

<210> 235
<211> 6
<212> DNA
<213> human

<400> 235
tctgaa 6

<210> 236
<211> 6
<212> DNA
<213> human

<400> 236
tgtctt 6

<210> 237
<211> 6
<212> DNA
<213> human

<400> 237
cagaat 6

<210> 238
<211> 6
<212> DNA
<213> human

<400> 238
ttgtgg 6

<210> 239
<211> 6
<212> DNA
<213> human

<400> 239
aaacct 6

<210> 240
<211> 6
<212> DNA
<213> human

<400> 240
ttctgg 6

<210> 241
<211> 6
<212> DNA
<213> human

<400> 241
gtttct 6

<210> 242
<211> 6
<212> DNA
<213> human

<400> 242
tgtatt 6

<210> 243
<211> 6
<212> DNA
<213> human

<400> 243
gtgatt 6

<210> 244
<211> 6
<212> DNA
<213> human

<400> 244
tccagt 6

<210> 245
<211> 6
<212> DNA
<213> human

<400> 245
aataca 6

<210> 246
<211> 6
<212> DNA
<213> human

<400> 246
ctcatt 6

<210> 247
<211> 6
<212> DNA
<213> human

<400> 247
ggaaga 6

<210> 248
<211> 6
<212> DNA
<213> human

<400> 248
cgacta 6

<210> 249
<211> 6
<212> DNA
<213> human

<400> 249
aaggga 6

<210> 250
<211> 6
<212> DNA
<213> human

<400> 250
ttttttc 6

<210> 251
<211> 6
<212> DNA
<213> human

<400> 251
tttgct 6

<210> 252
<211> 6
<212> DNA
<213> human

<400> 252
tggaga 6

<210> 253
<211> 6
<212> DNA
<213> human

<400> 253
tgaaga 6

<210> 254
<211> 6
<212> DNA
<213> human

<400> 254
aatgag 6

<210> 255
<211> 6
<212> DNA
<213> human

<400> 255
actgga 6

<210> 256
<211> 6
<212> DNA
<213> human

<400> 256
tccctt 6

<210> 257
<211> 6
<212> DNA
<213> human

<400> 257
ctctca 6

<210> 258
<211> 6
<212> DNA
<213> human

<400> 258
agaagt 6

<210> 259
<211> 6
<212> DNA
<213> human

<400> 259
tcttca 6

<210> 260
<211> 6
<212> DNA
<213> human

<400> 260
agcaaa 6

<210> 261
<211> 6
<212> DNA
<213> human

<400> 261
tctcac 6

<210> 262
<211> 6
<212> DNA
<213> human

<400> 262
ccacat 6

<210> 263
<211> 6
<212> DNA
<213> human

<400> 263
catgag 6

<210> 264
<211> 6
<212> DNA
<213> human

<400> 264
agaaca 6

<210> 265
<211> 6
<212> DNA
<213> human

<400> 265
aatcac 6

<210> 266
<211> 6
<212> DNA
<213> human

<400> 266
acttct 6

<210> 267
<211> 6
<212> DNA
<213> human

<400> 267
cctttc 6

<210> 268
<211> 6
<212> DNA
<213> human

<400> 268
agatca 6

<210> 269
<211> 6
<212> DNA
<213> human

<400> 269
agaaac 6

<210> 270
<211> 6
<212> DNA
<213> human

<400> 270
ccttca 6

<210> 271
<211> 6
<212> DNA
<213> human

<400> 271
agaatg 6

<210> 272
<211> 6
<212> DNA
<213> human

<400> 272
catttg 6

<210> 273
<211> 6
<212> DNA
<213> human

<400> 273
gtggaa 6

<210> 274
<211> 6
<212> DNA
<213> human

<400> 274
cacatg 6

<210> 275
<211> 6
<212> DNA
<213> human

<400> 275
caaaga 6

<210> 276
<211> 6
<212> DNA
<213> human

<400> 276
ttccaa 6

<210> 277
<211> 6
<212> DNA
<213> human

<400> 277
ccacaa 6

<210> 278
<211> 6
<212> DNA
<213> human

<400> 278
tgtggt 6

<210> 279
<211> 6
<212> DNA
<213> human

<400> 279
tgaagg 6

<210> 280
<211> 22
<212> DNA
<213> human

<220>
<221> misc_feature
<222> (11)..(16)
<223> n is a, c, g, or t

<400> 280
ccgactcgag nnnnnnatct ca 22

<210> 281
<211> 22
<212> DNA
<213> human

<220>
<221> misc_feature
<222> (11)..(16)
<223> n is a, c, g, or t

<400> 281
ccgactcgag nnnnnnatgt gg 22

<210> 282
<211> 22
<212> DNA
<213> human

<220>
<221> misc_feature
<222> (11)..(16)
<223> n is a, c, g, or t

<400> 282
ccgactcgag nnnnnnctag aa 22

<210> 283
<211> 22
<212> DNA
<213> human

<220>
<221> misc_feature
<222> (11)..(16)
<223> n is a, c, g, or t

<400> 283
ccgactcgag nnnnnntagg ag 22

<210> 284
<211> 22
<212> DNA
<213> human

<220>
<221> misc_feature
<222> (11)..(16)
<223> n is a, c, g, or t

<400> 284
ccgactcgag nnnnnnttct ag 22