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

(19) World Intellectual Property Organization
International Bureau

(43) International Publication Date
10 May 2013 (10.05.2013)

WIPO | PCT

(10) International Publication Number
WO 2013/064896 A1

(51) International Patent Classification:
C12Q 1/68 (2006.01)

(21) International Application Number:
PCT/IB2012/002423

(22) International Filing Date:
30 October 2012 (30.10.2012)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:
61/553,889 31 October 2011 (31.10.2011) US

(71) Applicant: **GENOMIC VISION** [FR/FR]; 80-84 rue des Meuniers, F-92220 Bagneux (FR).

(72) Inventors: **KOMATSU, Jun**; 4, place de la république, F-92220 Bagneux (FR). **WALRAFEN, Pierre**; 4 avenue Jean Jaurès, F-92120 Montrouge (FR). **CEPPI, Maurizio**; 2bis, Henri Tariel, F-92130 Issy Les Moulineaux (FR). **CONSEILLER, Emmanuel**; 10, rue de Plélo, F-75015 Paris (FR).

(74) Agent: **GUTMANN, Ernest**; 3, rue Auber, F-75009 Paris (FR).

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

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

Published:

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

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

— *with sequence listing part of description (Rule 5.2(a))*

WO 2013/064896 A1

(54) Title: METHOD FOR IDENTIFYING OR DETECTING GENOMIC REARRANGEMENTS IN A BIOLOGICAL SAMPLE

(57) Abstract: A method for detection, visualization and/or comparison of polynucleotide sequences of interest using specially designed sets of long and short probes that enhance resolution and simplify visualization and detection. Probe compositions useful for practicing this method and procedures for identifying useful probes and probe combinations. These methods are useful for the detection of genomic rearrangements, especially those associated with various diseases, disorders and conditions including cancer or for assessment of genomic rearrangements associated with therapy. The probe compositions may be used in kits for detection of genetic rearrangements or in companion diagnostic products or kits, such as kits for the diagnosis or assessment of predisposition to cancer such as colorectal cancer.

TITLE

METHOD FOR IDENTIFYING OR DETECTING GENOMIC REARRANGEMENTS IN A BIOLOGICAL SAMPLE.

BACKGROUND OF THE INVENTION

Field of the Invention

The invention relates to high-resolution, precise method for detecting genomic rearrangements *in vitro* using specially designed combinations of polynucleotide probes. The invention concerns accurate methods of detection and diagnosis of conditions, disorders and diseases associated with rearrangement of genomic DNA.

Description of the Related Art

The multigenic paradigm of human diseases

Advances in genetic analysis of human diseases have provided better insights into the molecular mechanisms contributing to disease initiation and progression. Previous associations were made between particular diseases and association and/or linkage disequilibrium to single base mutations in somatic genetic sequences or with particular single nucleotide polymorphisms ("SNPs") in genomic DNA. Newer technologies have provided evidence that larger genetic alterations and rearrangements are associated with, or can constitute major causes of diseases, disorders or conditions having a genetic origin or basis. Disease associations have now moved from a monogenic to a multigenic paradigm where a disease's origins and progression is mainly linked to more than one single genetic mutation or origin. While these new insights provide better avenues for disease detection and treatments, they also highlight the need for combinatorial genetic analysis that goes beyond detection of single mutational events or SNPs by assessing disease associations with larger genomic rearrangements. Such combinatorial genetic analysis would provide a better, more precise and accurate diagnosis of a particular condition, disorder, disease or pathology, but would also help establishing a more appropriate medical

survey, more accurate therapeutic decisions and interventions, as well as help in assessing the efficacy of such therapies and interventions.

Multigenic causes of genetic disease

Genetic disorders manifesting the same or similar clinical signs and consequences can arise from both single and exclusive, or combined, mutations in various genes. Such mutations can fall within either the single base alteration and/or the class of large genetic rearrangements. A few examples of such genetic disorders are Fragile X syndrome (mutations and expansions in the FMR1 gene), Ataxia Telangectasia (single base pair mutations in either intronic and exonic sequences as well as deletions and translocations of the ATM gene), Seckel syndrome (mutations as well as large rearrangements in SCKL1, SCKL2, SCKL3, PCTN and ATR), autism (mutations as well as large rearrangements in GLO1, MTF1 and SLC11A3), Spinal Muscular Atrophy (mutations, deletions, transconversions as well as cis-duplications involving the SMN1 and SMN2 genes) and myotonic dystrophy (trinucleotide/tetranucleotide expansions in DM1 and DM2).

Multigenic causes of cancer predisposition

In the case of cancer predisposition, there are several examples of familial cancer predisposition syndromes for which one can nominate several causative genes for which both single base alterations and/or large rearrangements were identified.

Breast and Ovary Cancer. Causative genes: BRCA1, BRCA2, ATM…

mutation type: higher proportion of point mutations identified so far.

Hereditary nonpolyposis colorectal cancer (Lynch syndroma). Causative genes: MSH2, MLH1, MSH6, EPCAM,…mutation type: equivalent proportion of point mutations has also been identified.

Multigenic causes of cancer progression

Cancer progression is surely the human disease domain where the monogenic causative hypothesis was definitely ruled out since several years. First, the disease's initiation is strictly dependent of two molecular events (immortalizing and transforming) due to genetic alterations in at least two independent genes classified at either oncogene or tumor suppressor genes. Second, the disease's progression is linked to additional genetic alterations independent from the causative ones. Not only do these additional alterations play a role in cancer progression, they also were demonstrated to be the basis for appearance of resistance to therapy during treatments.

Strikingly, in the list of cancer related genes, if extremely rare examples are only subject to discrete single base mutations (*e.g.*, KRas or BRaf), the large majority is either subject to only large rearrangements (*e.g.*, HER2, ALK…) or to both single base mutations and large rearrangements (p53, c-myc, c-Met, EGFR…).

The identification and characterization of multigenic conditions, disorders and diseases, including cancer, cardiovascular disease, diabetes and other heritable genetic conditions has been made difficult in part due to the imprecision of existing methods of molecular diagnosis. Molecular Combing is probably the sole approach allowing detecting all type of large genetic rearrangements (deletion, amplification, expansions, inversions, translocations…) even in a complex and heterogeneous population (such as tumors).

High resolution barcodes allowing multiplex analysis of patients could help diagnostic at different level such as for patient stratification/classification and/or prognosis.

Multiplex high resolution barcodes for identifying the right genetic alterations as a key driver for therapeutic intervention

The example of myotonic dystrophy

Myotonic Dystrophy (DM1) and Myotonic Dystrophy 2 (DM2) are two muscular dystrophies characterized by trinucleotide/tetranucleotide expansions in two different genes. If severe forms of DM1 can be clinically differentiated from DM2, milder DM1 forms are displayed extremely similar clinical signs than DM2. There is currently no cure for or treatment specific to myotonic dystrophy. However, DM1 patients exhibit Complications of the disease (heart problems, cataracts …) not existing in DM2 that could can be treated but not cured. Differentiating DM1 and DM2 by the use of a multiplex assay of high resolution barcodes could thus help preventing and treating secondary effects

The example of hereditary breast and ovary cancer

In certain countries (U.S.) detecting constitutional alterations in BRCA1/2 drives to therapeutic intervention (surgery/reconstitution). Thus, there is a clear need for an accurate diagnostic comprising all the potentially involved genes. Such a test could be made on the basis of a multiplex assay of high resolution barcodes comprising large chromosomal regions around genes known to be involved in this syndrome; BRCA1, BRCA2, ATM, ATR…

DNA Damage and Response inhibitors example

Synthetic lethality became a strong reality for therapeutic decision to include Cancer patients in specific protocols/regimens. One of the first examples was given with the demonstration that Breast cancer patients with BRCA deficiency exhibit a higher sensitivity to PARP inhibitors, a new category of drug acting on DNA Damage and Response pathway. More recently, this was extended to other type of inhibitors in this category such as ATM inhibitors but also to more traditional anti-cancer drugs including all types of DNA polymerase and replication inhibitors.

Not only does this concept extended to other inhibitors, but it was also demonstrated that it could be extended to other types of cancers such as lung and metastatic melanoma.

Here, a multiplex high resolution barcode will allow detection of genetic alteration in genes involved in DNA damage and response that could help predicting sensitivity to this class of inhibitors. A list of such genes could include BRCA1, BRCA2, ATM, ATR, MSH2, MLH1, MSH6, EPCAM…

The Lung cancer example

Numerous alterations involved in lung cancer could be multiplexed for a better patient classification such as:

- LOH/Deletion (P53, STK11, LKB1, BRG1, KLF6);
- Amplification (FGFR1, MET, EGFR, HER2…);
- Translocation: (ALK);

All these genetic alteration are associated to therapeutic treatments:

- P53: Nutlin (low doses Actinomycin D produce similar effects)
- FGFR1: Masitinib, PD173074, SU5402 TK1258 AZD4547…
- MET: GSK1363089, ARQ197, SGX523, XL184…
- EGFR: Tarceva, Erbitux, Vectibix…
- HER2: Herceptin, Lapatinib…
- ALK: Crizotinib

As at least 30% of NSCLCs were demonstrated to be dependent on at least one of these mutations, defining the genetic profile of the tumor could help driving therapeutic options. This could be made possible by designing multiplex assays combining high resolution barcodes covering this major genetic loci.

Localization of (genetic) sequences of interest

Genetic sequence is the most fundamental information to synthesize functional protein. Alteration of genetic sequence sometimes results in loss of functional protein synthesis. In addition to alteration of genetic sequence, loss or gain of genetic sequence (copy number variation, CNV) also can be problematic for homeostasis of cellular activity. For example, loss of (functional) anti-tumor protein (p53) or gain of proto-oncogene (c-myc) results in cancer-prone cell. When such mutation happens (or exists) in germ cell, this mutation spreads whole cell in an individual who is either carrier or patient of genetic disease, or has a predisposition to cancer. The germline mutation can be heritable. These days CNV becomes more and more important to understand in the field of genetics (*ref 1*). However, copy number count alone is not always sufficient and it is often critical to establish the actual location of sequence elements. This is strikingly the case for e.g. balanced translocations. DNA sequencing and CNV detection methods such as array-based comparative genomic hybridization (aCGH) and quantitative PCR generally cannot detect these balanced mutations because these methods assess whether the sequence and the copy number are correct or not. FISH and its extended forms such as fiber-FISH or molecular combing can address these balanced mutations with different resolutions and precisions depending on methods.

Resolution and precision

The use of BAC/PAC/cosmid probes on targeted regions was successfully conducted to detect large (a few kb to tens of kb) genomic rearrangements (ref 2). In these approaches, the minimum size of detectable events (*e.g.*, the size of the deleted or amplified sequence), hereafter designated as the "resolution" of such an assay, is limited due to the large standard deviation involved in measuring probes or gaps of tens of kilobases. Indeed, in such assays the standard deviation of measurements increases with the length of the measured element. For example, a 40 kb-probe is measured with a standard deviation of ~5 kb. Thus, if 16 measurements of a given probe are made on a slide, the precision on the size of the probe obtained as the mean value of measurements is in the order of magnitude of 2.5 kb (Considering the distribution is gaussian, and the precision is the half-width of the confidence interval, i.e. $2.sd/\sqrt{n}$ where sd = standard deviation and n = number of measurements). For a 10 kb-probe, where the standard deviation is ~2 kb, the precision would be ~1 kb. This illustrates the fact that shorter probes allow for better (lower) resolution.

Besides, the location of such an event (the position of the extremities of the event) may be defined with a precision (hereafter the location precision) limited by the size of the probe or gap within which it occurs: e.g. if a 40 kb probe is estimated to measure 39 kb in a sample, one can conclude that a 1 kb deletion occurred somewhere within the probe, with no further precision – thus, somewhere in a 40 kb genomic region. If the same 1 kb deletion had occurred within a 10 kb probe, the location of that deletion would be known with a better precision, as the range would be reduced to a 10 kb genomic region. Therefore, the smaller the probes and gaps, the better the location precision.

There are limits to small probes: (i) below a certain size, they become difficult to detect; (ii) they involve more complex color schemes (as there are relatively more probes); (iii) there are more distinct probes to cover a given region, and the experiments are therefore more expensive and time-consuming; (iv) most importantly, fast and reliable identification of probes, whether by a human operator or a piece of software, is easier with longer probes, as they are more readily distinguished from background. Indeed, background is mainly constituted of roughly circular fluorescent spots. When large enough, the shape of these spots allows to one to easily distinguish them from probes. However, when their size is small enough, they appear difficult to distinguish from small probes.

In operating conditions according to the invention, probes shorter than ~3 kb are detected with a diminished efficiency. Within the 3-10 kb range, the standard deviation of measurements varies little, and there is therefore little benefit in resolution with the shorter probes within this range. Therefore, this range is usually considered to be a good compromise for probe size. However, in cases where probes are close enough (less than 10 kb gaps), smaller probes (within the 500 – 3000 bp range) are still useful, as they will be detected in at least a fraction of signals and the presence of the corresponding sequences may therefore be established with certainty. It was also found that detection of isolated probes longer than 12 kb (preferably longer than 14 kb) is more reliable, whether for a human operator or for automatic detection software.

Exclusion of repeats

Eukaryotic genomic DNA contains various repetitive sequences, *i.e.*, sequences that appear more than once (and more than statistically predicted based on their length and base content) in a normal haploid genome. Among these, some appear with very high frequency (tens of thousands to millions of copies). In human genomic DNA, the most abundant of these is the

Alu family, which has ~1,000,000 copies constituting ~10 % of the genome. In any hybridization procedure involving human genomic DNA, it is expected that probes carrying such repeats would hybridize on numerous targets, generating non-specific signal from regions throughout the genome. Other types of repetitive sequences exist, with lower frequency, and often more specific localization. The number of copies and repeat sequence length may vary widely, as well as the degree of homology. Beta-satellite sequences, for example, are present in multiple copies (hundreds to thousands), usually as tandem repeat arrays comprising hundreds of copies of the same 50-100 bp long sequence, specifically localized in a limited number of loci.

Strategies to get rid of the non-specific signals depend on the type of procedure and probe. Schematically, when probes are very short sequences of DNA (oligonucleotides, typically less than 100 bp), as in aCGH procedures, the sequence of the oligonucleotides is chosen to be free of repetitive sequences, by comparison with repetitive sequences found in databases. This strategy is only practical for very short probes, as short sequences free of repetitive sequences are relatively abundant, but unpractical for longer probes, as long stretches completely devoid of repetitive elements are rare (although this has been adapted to longer FISH probes, in an approach that suffers multiple drawbacks, see below). Besides, even for short probes, it constrains the design of probes heavily and some genomic regions, rich in repetitive sequences, have lower density of coverage (and thus lower resolution of events) due to this constraint.

When probes are longer (typically PCR products or cloned DNA inserts – 1 to 150 kb), in Southern Blot or in FISH procedures, non-labeled competitive DNA, enriched in repetitive elements such as Alu repeats (usually Cot-1 DNA), is added in large excess along with the labeled probe. Competition of unlabelled probes on the repetitive sequences minimizes the hybridization of labeled probes. This strategy is expensive and since the competitor DNA is not purely made of repetitive sequences, competition also occurs on the unique sequences for which the probes were designed, thus limiting the amount of competitor DNA that may be used. Therefore, the efficiency of this approach is limited.

An alternative approach for longer probes has been proposed by Knoll and collaborators (U.S. Patent 7,014,997), resembling the strategy usually adopted for oligonucleotides: probes are chosen within sequence intervals devoid from repetitive elements. This strategy is based on bioinformatics analysis of the regions of interest and exclusion of known repetitive sequences by comparison with sequence databases. However, this approach has several limitations: prior

knowledge of the repetitive sequences is required, which can be a problem e.g. in species where such knowledge is unavailable. More importantly, intervals longer than 2 kb devoid of repetitive sequences appear only once in 20-30 kb on average and are unevenly distributed(Considering the distribution is gaussian, and the precision is the half-width of the confidence interval, i.e. $2.sd/\sqrt{n}$ where sd = standard deviation and n = number o) so the design of probes would be highly constrained, impairing the possibility to design a high-resolution code. This would prove especially difficult in repeat-rich regions, and/or regions where pseudogenes are located next to homologous genes of interest – such low-copy repetitive sequences being also excluded with the strategy from Knoll and co (ref. 3). Since regions targeted in rearrangement tests, *e.g.*, for diagnostics purposes, often display these features, this approach is not suitable for the design of high-resolution barcodes and especially not if such a code is to be used for diagnostics purposes. Distinctions between this approach and the invention are disclosed in more detail below.

BRIEF SUMMARY OF THE INVENTION

The present invention concerns the field of the *in vitro* diagnosis and detection of genetic rearrangements and is related to a method to identify or detect genetic rearrangements in a biological sample to be tested which are already known or which are new and provide markers for example of diseases as cancers or metabolic or foetal genetic diseases. The invention is characterized by using compositions containing purified or synthesized nucleic acid molecules (polynucleotides) having nucleotide sequences selected as short sequences with a length of less than 10 Kb and associated in the said method with other different nucleic acid molecules (polynucleotides) having nucleotide sequences non-overlapping with the former ones and having a size longer than 12 Kb. The selected nucleotide sequences (polynucleotides) used as probes are partly deleted of their natural frequently repeated sequences. The present invention concerns also improvements brought to the design of set of probe sequences for the detection of genetic rearrangements by hybridization as with fiber-FISH-like technologies such as Molecular Combing. The improvements described herein allow for high precision / high-resolution detection of rearrangements in time- and cost-efficient assays. This invention also relates to the use of probe sequences for diagnostics applications and companion diagnostics tests, to a method of detection of presence or absence of alterations in sequences and to a kit for the above uses.

This is illustrated hereinafter with sets of nucleotide sequences corresponding to parts of at least two genes: MSH2 and MLH1 or to the regions of MSH2 and MLH1, whose mutations increase the risk of occurrence of human colorectal cancer .

The invention is related to the sets of polynucleotides or probes labeled or not which are specific of said genes. Presently, the detection of genetic rearrangements using current technologies is often insufficiently reliable for diagnostics use. Unlike most technologies used to detect genetic alterations, which suffer strong intrinsic limitations towards some types of rearrangements, direct technologies such as FISH or Fiber-FISH can intrinsically detect any type of rearrangements. Their use is mainly limited by their resolution. Molecular Combing, on the other hand, may reach sufficient resolution, but probe designs currently used fail to allow cost- and time-efficient high resolution analysis of rearrangements.

These improvements involve the combination within the same sets of probes of -typically shorter – probes designed to optimize the sensitive detection and precise measurement of rearrangements and – typically longer – probes to allow for fast and reliable detection of signals of interest when analyzing results. Alternative designs where the longer probes are replace with a combination of shorter probes having equivalent functions and effects are also disclosed.

Specific aspects of the invention based on the concept of combining small probes for resolution and long probes for ease of detection for the detection on one or more genomic region(s) of interest as disclosed in more detail below.

The invention thus concerns a method for detecting mutated or rearranged genomic polynucleotide (target) sequence comprising:

(a1) hybridizing a target genomic polynucleotide comprising one or more genomic region(s) of interest, where mutations or rearrangements are sought, to a set of short probes that bind to each region of interest without long gaps between the portions of the target sequence bound by the set of short probes, where on each genomic region a subset of short probes are selected so that when taken together they form a long contiguous stretch inside or outside the region of interest, and wherein the probes may optionally have frequent repetitive sequences removed and thus more generally are optionally devoid of such repetitive sequences; or

(a2) hybridizing a target genomic polynucleotide comprising one or more genomic region(s) of interest, where mutations or rearrangements are sought, to a set of short probes that bind to each region of interest without long gaps between the portions of the target sequence

bound by the set of short probes and to one or more long (docking) probe(s) that bind to sequences near but outside of the region(s) of interest; wherein the sequence(s) of the long probe(s) does not overlap that of the short probes and wherein the short and / or long probes may optionally have frequent repetitive sequences removed and thus more generally are optionally devoid of such repetitive sequences;

(b) detecting the locations of hybridized probes on the genomic region(s) of interest; optionally,

(c) comparing the location of the hybridized probes on the target genomic polynucleotide sequence with one or more motifs based on the hybridization of said probes to a reference, control, normal, not mutated, or not rearranged genomic polynucleotide sequence; and optionally,

(d) correlating the presence of a mutated or rearranged genomic polynucleotide with a specific phenotype, disease, disorder, or condition.

The mutated or arranged genomic polynucleotide sequence can be obtained from a subject who has cancer or who is suspected to having cancer, for example, from a subject who has colorectal cancer or who is suspected of having colorectal cancer. In such a case, the short and long probes identify mutations or genomic rearrangements associated with colorectal cancer and a control or reference sample would not contain these mutations or rearrangements. The presence or risk of developing colorectal cancer is assessed by comparing a target genomic polynucleotide sequence with the reference and determining whether a mutation or rearrangement associated with colorectal cancer is present. This method can be practiced with specific probes corresponding to or derived from Probe sets 1, 2, 3 and 4. For colorectal cancer, a genomic region of interest can be selected from genes associated with this disease, such as MSH2, MLH1, MSH6, PMS2 or EPCAM.

Similarly, the method may be applied to samples obtained from subjects having or at risk of developing other kinds of cancer, such as breast cancer, ovary cancer, or lung cancer. The method may also be applied to samples obtained from subjects having or at risk of other kinds of diseases, disorders, or conditions, including cardiovascular disease, diabetes, neuromuscular disorders; such as myotonic dystrophy or spinal muscular atrophy or samples obtained from a subject who has, is suspected of having, or is suspected of being a carrier for a genetic or hereditary disease, disorder or condition, including known or unknown foetal genetic alterations. The sample can be obtained from a subject having a multigenic genetic or hereditary disease,

disorder or condition or for a genetic or hereditary disease, disorder or condition associated with rearrangement of genomic DNA.

In some aspects of the invention, the sample will be obtained from a subject undergoing treatment for a disease, disorder or condition associated with a genomic or somatic genetic rearrangement and the results obtained are compared to results obtained at other time points before, during or after the termination of treatment. A companion test for evaluating the efficiency of a therapeutic drug on the mutated or rearranged nucleotide sequences of the gene or the region of the gene of interest can be performed using the short and long probes according to the invention.

Preferably, in the method described above, the hybridizing with the short and long probes in step a) will be performed simultaneously.

Preferably, the short probes range in length from 0.5 kb to 10 kb and the maximum size of the gaps between the short probes when they are bound to the target is 15 kb, preferably 12 kb and more preferably 10 kb.

The number of short probes employed in the method described above can range from 1, 2, 3 to 10, 15 or more.

The maximum size for the long probes is 150 kb and these probes preferably range from 12 kb to 40 kb in length. Preferably, in order to have "*long probe(s) that bind to sequences near but outside of the region of interest*", distance between the long probes and the region of interest is no longer than 150 kb, and more preferably no longer than 75 kb and even more preferably no longer than 25 kb from the region of interest. The minimum size for a genomic region to be tested or targeted is 50 kb. The minimum number of regions of interest is one for a singleplex test and two or more for a multiplex test. Examples of combinations of short and/or long probes include at least one short (less than 10 kb) sequence and at least one non-overlapping long sequence (more than 15 kb), or at least one group of at least two short sequences, less than 10 kb each, which total group length is longer than 14 kb and less than 150 kb, hybridizing contiguously on the mutated or rearranged polynucleotide sequence. The short probes can comprise a set of contiguous probes that span a stretch of the genomic polynucleotide sequences inside or outside the region of interest that is at least 15 kb.

The long probes may have repetitive DNA sequences excluded. These repetitive sequences to be excluded would ordinarily appear more than once and more often than

statistically predicted based on their length and base content, for example, repetitive sequences between 50 and 400 bp can be excluded, though shorter or longer repetitive sequences that decrease sensitivity or specificity of the method can be identified and excluded. An example of such a sequence is the repetitive Alu family DNA sequences.

According to an embodiment of the invention, in order for the probes, either short probes or long probes, to have repetitive sequences excluded, these probes are designed to hybridize in regions of the genome which are free of such repetitive sequences, i.e. which have less than 10% preferably less than 2% of the selected type(s) of repetititve sequences to be excluded.

In the method described above, the short and long probes are preferably fluorescently tagged and different components of the probe sets may be tagged with different labels, such as labels with different colors. Tagging provides one means to identify motifs or submotifs characteristic of a mutated or rearranged sequence.

Compositions or kits comprising a set of short probes or a combination of short and long probes as described herein and optionally one or more components for binding said probes to a polynucleotide, for performing molecular combing, and/or for detecting whether hybridization has occurred are also contemplated. For example, a composition containing the short and long probe(s) described above, wherein at least two of said probe sequences detect a genetic rearrangement by using Molecular Combing, said composition comprising either at least one short (< 12 kb) sequence and at least one non-overlapping long sequence (> 14 kb), or at least one group of at least two short sequences, less than 10 kb each, which total length is longer than 14 kb and less than 150 kb, hybridizing contiguously on the genetic target. The short probe(s) in such a composition may preferably range from 0.5 kb to 12 kb and the long probe(s) range from 14 kb to 40 kb. Frequent repetitive sequences described above may be removed from the probes. Examples of probe sequences are those that hybridize specifically on the MSH2 gene or in the region of the MSH2 gene or on the MLH1 gene or in the region of the MLH1 gene. Specific kinds of short probe sequence(s) where repetitive sequences have been removed include those selected from the group consisting of or comprising the sequences obtained by PCR amplification on human genomic DNA using the primer pairs described in Table 1 in the lines:

MSH2-v1

P3 (primer pairs P3a_MSH2-v1 to P3c_MSH2-v1, SEQ ID NO:21-26)

P4 (primer pairs P4a_MSH2-v1 to P4b_MSH2-v1, SEQ ID NO:27-30)

P5 (primer pairs P5a_MSH2-v1 to P5c_MSH2-v1, SEQ ID NO:31-36)P6 (primer pairs P6a_MSH2-v1 to P6b_MSH2-v1, SEQ ID NO:37-40)

P7 (primer pairs P7a_MSH2-v1 to P7c_MSH2-v1, SEQ ID NO:41-46)

P8 (primer pairs P8a_MSH2-v1 to P8b_MSH2-v1, SEQ ID NO:47-50)

P9 (primer pairs P9a_MSH2-v1 to P9c_MSH2-v1, SEQ ID NO:51-56)

P10 (primer pairs P10a_MSH2-v1 to P10b_MSH2-v1, SEQ ID NO:57-60)

MLH1-v1

P3 (primer pairs P3a_MLH1-v1 to P3d_MLH1-v1, SEQ ID NO:95-102)

P4 (primer pairs P4a_MLH1-v1 to P4b_MLH1-v1, SEQ ID NO:103-106)

P5 (primer pairs P5a_MLH1-v1 to P5b_MLH1-v1, SEQ ID NO:107-110)

P6 (primer pair P6a_MLH1-v1, SEQ ID NO:111-112)

P7 (primer pair P7a_MLH1-v1, SEQ ID NO:113-114

P8 (primer pairs P8a_MLH1-v1 to P8d_MLH1-v1, SEQ ID NO:115-122)

and the short probes may be used in combination with the long probe sequence(s) selected from the group consisting of or comprising the sequences obtained by PCR amplification on human genomic DNA using the primer pairs described in Table 1 in the lines

MSH2-v1

P11 (primer pairs P11a_MSH2-v1 to P11c_MSH2-v1, SEQ ID NO:61-66)

P12 (primer pairs P12a_MSH2-v1 to P12e_MSH2-v1, SEQ ID NO:67-76)

MLH1-v1

P9 (primer pairs P9a_MLH1-v1 to P9c_MLH1-v1, SEQ ID NO:123-128)

P10 (primer pairs P10a_MLH1-v1 to P10e_MLH1-v1, SEQ ID NO:129-138).

Specific kinds of contiguous short probe sequence(s) forming long stretches include those selected from the group consisting of or comprising the sequences obtained by PCR amplification on human genomic DNA using the primer pairs described in Table 1 in the lines:

MSH2-v2

PE1-2 (primer pairs PE1_MSH2-v2 to PE2_MSH2-v2, SEQ ID NO:163-166) and

PE3-6 (primer pairs PE3_MSH2-v2 to PE5-6_MSH2-v2, SEQ ID NO:167-172), together forming one stretch;

PE9 (primer pairs E9_MSH2-v2 and I9-10_MSH2-v2, SEQ ID NO:185-188),

PE10 (primer pair E10_MSH2-v2, SEQ ID NO:189-190),

PE11 (primer pairs E11_MSH2-v2 and I11-12_MSH2-v2, SEQ ID NO:191-194),

PE12-14 (primer pairs E12_MSH2-v2 and E13-14_MSH2-v2, SEQ ID NO:195-198) and

PE15-16 (primer pairs E15_MSH2-v2 and E16_MSH2-v2, SEQ ID NO:199-202),

together forming one stretch;

MLH1-v2

PE1-2 (primer pairs E1_MLH1-v2 and E2_MLH1-v2, SEQ ID NO:227-230),

PE3-4 (primer pairs I23_MLH1-v2, E3_MLH1-v2 and E4_MLH1-v2, SEQ ID NO:231-236),

PE5-6 (primer pairs E5_MLH1-v2 and E6_MLH1-v2, SEQ ID NO:237-240),

PE7-9 (primer pairs E7-8_MLH1-v2 and E9_MLH1-v2, SEQ ID NO:241-244) and

PE10-11 (primer pairs E10_MLH1-v2 and E11_MLH1-v2, SEQ ID NO:245-248),

together forming one stretch;

The primers designed for the purpose of preparing short probes of the invention may have a sequence of 20 to 40 nucleotides and comprise in their 3' end a sequence of at least 20 contiguous nucleotides that base pairs with the target. The primer sequence thus may also comprise additional nucleotides that do not base pair with the target in its 5' end. The nucleotides which do not base pair may be useful for the construction of the primers or for the cloning of the amplified sequence resulting from polymerization starting from the primers. In a particular embodiment the sequence of the primer that hybridizes to the target is longer than 20 nucleotides.

Molecular Combing is a powerful FISH-based technique for direct visualization of single DNA molecules that are attached, uniformly and irreversibly, to specially treated glass surfaces (Herrick and Bensimon, 2009); (Schurra and Bensimon, 2009). This technology considerably improves the structural and functional analysis of DNA across the genome and is capable of visualizing the entire genome at high resolution (in the kb range) in a single analysis.

Another embodiment of the invention is a method for designing a set of short probes or set of short and long probes as described above comprising:

identifying a polynucleotide containing a genomic region of interest,

selecting long probe sequences outside of the genomic region of interest but within 100 kb of the closest probe in the region of interest, and preferably within 30 kb of the closest probe in the region of interest and optionally removing frequently repeated sequences from said long probe sequences,

selecting a short probe sequences from within the genomic region of interest so that no gaps longer than 20 kb, and preferably no gaps longer than 12 kb appear between the short probes; or selecting a series of short probes that together form a long continuous stretch that covers the genomic region of interest;

hybridizing the probes to a genomic polynucleotide comprising the genomic region of interest,

detecting the hybridized probes, and

determining which sets of probes form motifs that specifically identify the genomic sequence of interest from a reference genomic sequence.

The comparison of the location of the hybridized probes on the target genomic polynucleotide sequence with one or more motifs based on the hybridization of said probes to a reference, control, normal, not mutated, or not rearranged genomic polynucleotide sequence, as disclosed in the databanks or experimentally obtained on samples.

The techniques disclosed herein may be applied to diagnosis of disease as well as for the identification of genetic rearrangements associated with a disease, disorder or condition. They may also be used as companion diagnostics to study the responses of a subject or group of subjects who have particular rearrangements to therapy, responses to environmental agents, or the effects of lifestyle choices. Specifically, the diagnostic products and methods of the invention are useful for diagnosis and assessments for subjects having or at risk of developing colorectal cancer. High resolution barcodes allow multiplex analysis of patients for extended or expanded diagnosis at the levels of patient stratification/classification and prognosis. Thus, the techniques disclosed herein can also be used to predict the course and probably outcome of a disease, disorder or condition as well as the likelihood of progression, stability, or recovery. Multiplex high resolution barcodes also permit the identification of key genetic alterations in a subject that would benefit from a particular kind of therapy as well as a way to assess the reaction of a subject to a particular kind of therapy or therapeutic intervention.

Specific embodiments of the invention include the following, which embodiments are especially carried out *in vitro*.

A method for detecting mutated or rearranged genomic polynucleotide sequence comprising: (a1) hybridizing a target genomic polynucleotide comprising one or more genomic region(s) of interest, where mutations or rearrangements are sought, to a set of short probes that bind to each region of interest without long gaps between the portions of the target sequence bound by the set of short probes said set of short probes optionally including or being in combination with a (sub)set of short probes selected so that on each genomic region some of the short probes when taken together form a long contiguous stretch inside or outside the region of interest and where the short probes may optionally have frequent repetitive sequences removed; or (a2) hybridizing a target genomic polynucleotide comprising one or more genomic region(s) of interest, where mutations or rearrangements are sought, to a set of short probes that bind to each region of interest without long gaps between the portions of the target sequence bound by the set of short probes and to one or more long (docking) probe(s) that bind to sequences near but outside of the region(s) of interest; wherein the sequence(s) of the long probe(s) does not overlap that of the short probes and wherein the short and/or long probes may optionally have some or all of the frequently repeating sequences removed; (b) detecting the locations of hybridized probes on the genomic region(s) of interest; optionally, (c) comparing the location of the hybridized probes on the target genomic polynucleotide sequence with one or more motifs based on the hybridization of said probes to a reference, control, normal, not mutated, or not rearranged genomic polynucleotide)sequence; and optionally, and/or (d) correlating the presence of a mutated or rearranged genomic polynucleotide with a specific phenotype, disease, disorder, or condition.

The invention relates in particular to the method herein described wherein the mutated or rearranged genomic polynucleotide sequence is obtained from a subject who has cancer or who is suspected of having cancer or who is susceptible to have a genetic predisposition to cancer.

The invention also relates in a particular embodiment to a method wherein the mutated or rearranged genomic polynucleotide sequence is obtained from a subject who has colorectal cancer or who is suspected of having colorectal cancer or who is susceptible to have a genetic predisposition to colorectal cancer, wherein said short and long probes identify mutations or genomic rearrangements associated with colorectal cancer, wherein said control, not mutated or

normal genomic sequence is obtained from a subject not at risk for colorectal cancer and wherein the detection of a genomic rearrangement; and assessing presence of or risk of developing colorectal cancer when said genomic rearrangement is detected. In this method the probes can hybridize specifically on the MSH2 gene, in the region of the MSH2 gene, on the MLH1 gene, or in the region of the MLH1 gene.

The invention also relates in a particular embodiment to a method wherein the mutated or rearranged genomic polynucleotide sequence is obtained from a subject who has breast cancer or who is suspected to having breast cancer or who is susceptible to have a genetic predisposition to breast cancer.

The invention also relates in a particular embodiment to a method wherein the mutated or rearranged genomic polynucleotide sequence is obtained from a subject who has ovarian cancer or who is suspected to having ovarian cancer or who is susceptible to have a genetic predisposition to ovarian cancer.

The invention also relates in a particular embodiment to a method wherein the mutated or rearranged genomic polynucleotide sequence is obtained from a subject who has lung cancer or who is suspected to having lung cancer or who is susceptible to have a genetic predisposition to lung cancer.

The invention also relates in a particular embodiment to a method wherein the mutated or rearranged genomic polynucleotide sequence is obtained from a subject who has a cardiovascular disease, disorder or condition or who is suspected of having cardiovascular disease, disorder or condition or who is susceptible to have a genetic predisposition to cardiovascular disease, disorder or condition.

The invention also relates in a particular embodiment to a method wherein the mutated or rearranged genomic polynucleotide sequence is obtained from a subject who has a diabetes or who is suspected of having diabetes or who is susceptible to have a genetic predisposition to diabetes.

The invention also relates in a particular embodiment to a method wherein the mutated or rearranged genomic polynucleotide sequence is obtained from a subject who has a neuromuscular disorder or who is suspected of having a neuromuscular disorder.

The invention also relates in a particular embodiment to a method wherein the mutated or rearranged genomic polynucleotide sequence is obtained from a subject who has, is suspected of

having, or is susceptible of being a carrier for a genetic or hereditary disease, disorder or condition.

The invention also relates in a particular embodiment to a method wherein the short and long probe sequences are specific to human genes or to human genomic regions associated with cancer, colorectal cancer or a foetal genetic alteration known or unknown when said region or gene is mutated or genetically rearranged.

The invention also relates in a particular embodiment to a method wherein the mutated or rearranged genomic polynucleotide sequence is obtained from a subject who has, is suspected of having, or is suspected of being a carrier for a multigenic genetic or hereditary disease, disorder or condition or for a genetic or hereditary disease, disorder or condition associated with rearrangement of genomic DNA.

The invention also relates in a particular embodiment to a method wherein the mutated or rearranged genomic polynucleotide sequence is obtained from a subject undergoing treatment for a disease, disorder or condition associated with a genomic inherited or acquired rearrangement and the results obtained are compared to results obtained at other time points before, during or after the termination of treatment.

The invention relates to method of any of the embodiments described herein, characterized by the following features taken individually or in any combination: the hybridizing with the short and long probes in (a2) is performed simultaneously; the short probes are 10 kb or less; and/or the short probe(s) comprise at least one short (less than 10 kb) sequence and at least one non-overlapping long sequence (more than 12 kb), or at least one group of at least two short sequences, less than 5, 6, 7, 8, 9 or 10 kb each, total group length is longer than 12 kb and less than 150 kb, hybridizing contiguously on the mutated or rearranged polynucleotide sequence. In these methods the short probes may comprise a set of contiguous probes that span a stretch of the genomic polynucleotide sequences inside or outside the region of interest that is at least 14 kb; and/or the long probe(s) may comprise one or more docking probes of more than 14 kb and less than 40 kb. The long probe(s) may have a length of at least 14 kb and bind to a polynucleotide sequence outside the region of interest.

Both the long and short probes may be designed to exclude frequently occurring repetitive DNA sequences. These repetitive DNA sequences, which may be excluded from the long and short probes, will generally appear more than once and more often than statistically

predicted based on their length and base content. For example, a repetitive DNA sequence between 50 and 400 contiguous nucleotides in length, which appear more than once and more often than statistically predicted based on their length and base content, can be excluded from the short and /or long probe(s). One example of a repetitive sequence that can be excluded from the short and long probes is or are members of the repetitive Alu family DNA sequences.

In some embodiments of the invention the probes in (b) of the first embodiment are fluorescently tagged so that they can be detected fluorometrically. In other embodiments in b) each probe is tagged with one of two or more fluorescent tags.

According to other embodiments of the methods above, motifs or easily identifiable subsets of the probes are detected and compared instead of every probe sequence.

The methods described above may employ at least 3, 4, 5, 6, 7, 8, 9, 10, 11, 12 or more short probes. These short probes may each have a length of least 500, 600, 700, 800, 900 or more base pairs (bp). In some embodiments of the methods above, the probes will be selected so that the gaps between short probes in the genomic region of interest are no more than 12 kb each. In further embodiments the short probes will bind to a single contiguous genomic region of interest or the short probes can be selected to bind to more than one non-contiguous genomic region of interest. The long probes used in the method above may be selected so as to be no more than 20, 30 or 40 kb. The or each of the genomic region(s) of interest in the methods described above can be selected to be longer than 50 kb.

Another embodiment of the invention is a kit comprising a set of short probes or a set of short and a set of long probe(s); and optionally one or more components for binding said probes to a polynucleotide, for performing molecular combing, and/or for detecting whether hybridization has occurred; (i) wherein the short probes comprise a set of probes that taken together bind to a long continuous stretch of the genomic region of interest; or(ii) wherein the long probes bind to sequences outside the genomic region of interest, do not overlap the short probe sequences; and optionally, where the repetitive sequences have been removed from the long and / or short probes. A kit of the invention is suitable and/or is specific for use in a method of the invention as disclosed herein. In a particular embodiment its short and/or long probes are characterized by the features described herein in relation with the methods. Such a kit may be employed for or contain instructions for the detection of genomic rearrangements associated with colorectal cancer or genetic predisposition to colorectal cancer; for the detection of genomic

rearrangements associated with breast cancer or genetic predisposition to breast cancer; for the detection of genomic rearrangements associated with ovarian cancer or genetic predisposition to ovarian cancer; for the detection of genomic rearrangements associated with lung cancer or genetic predisposition to lung cancer.

Another embodiment of the invention is a composition containing the short, or short and long probe(s) described by the first embodiment above, wherein at least two of said probe sequences detect a genetic rearrangement by using Molecular Combing, said composition comprising either (a) at least one short (less than 10 kb) sequence and at least one non-overlapping long sequence (more than 14 kb), or (b) at least one group of at least two short sequences, less than 10 kb each, which total length is longer than 14 kb and less than 150 kb, hybridizing contiguously on the genetic target. In this composition the short probe(s) can range from 0.5 kb to 9 kb and the long probe(s) can range from 14 kb to 40 kb. The size of the short probes may range from 0.5 to 9 kb and at least 90% of the frequent repetitive sequences can be been removed from the short probe sequences. This composition may contain probes sequences that hybridize specifically on the MSH2 gene or in the region of the MSH2 gene or on the MLH1 gene or in the region of the MLH1 gene.

In yet another embodiment the invention involves a method for designing short and long probes described herein in relation to methods comprising (a) identifying a polynucleotide containing a genomic region of interest, (b) selecting long probe sequences outside of the genomic region of interest but within 100 kb of the closest probe within the region of interest and optionally removing frequently repeated sequences from the long probe sequences, (c) selecting a set of short probe sequences from within the genomic region of interest so that no gaps longer than 15 kb appear between the short probes; or selecting a series of short probes that together form a long continuous stretch that covers the genomic region of interest; (d) hybridizing the probes to a genomic polynucleotide comprising the genomic region of interest, (e) detecting the hybridized probes, and (f) determining which sets of probes form motifs that distinguish the genomic sequence of interest from a reference genomic sequence.

BRIEF DESCRIPTION OF THE DRAWINGS

Figure 1. (A) Dot-plot of MSH2 gene sequence on RP11-1084A21 BAC clone. (B) probe code v1 (without repetitive element) on RP11-1084A21. (C) probe code-v2 on RP11-1084A21. Diagonal lines are perfectly matched region of DNA between two sequences. Dots are representatives of repetitive elements. Higher density of dots (or grey band) are higher density of repetitive element.

Figure 2. Dot plot analysis of MLH1 region. (A) Dot-plot of MLH1 gene sequence on RP11-426N19 BAC clone. (B) probe code v1 (without repetitive element) on RP11-426N19 . (C) probe code-v2 on RP11-426N19.

Figure 3. Designed probe set for MSH2 by exclusion of repetitive element. A) theoretical probe set (labeled in red and green in microscopy experiments represented here in grey and black, respectively), and position of exon (small numbered dots). (B) actual hybridization image corresponding to MSH2-v1 probe set. Original microscopy images consist of three channel images where each channel is the signal from a given fluorophore – these are acquired separately in the microscopy procedure. These channels are represented here as different shades on a grayscale: green probes are shown in black and red probes in gray, while the background (absence of signal) is white. The aspect ratio was not preserved, signals have been "widened" (i.e. stretched perpendicularly to the direction of the DNA fiber) in order to improve the visibility of the probes.

Figure 4. Designed probe set for MLH1 by exclusion of repetitive element. A) theoretical probe set (red and green), and position of exon (purple dot). (B) actual hybridization image corresponding to MLH1-v1 probe set. The same color conventions are used for diagrams and microscopy images as in panels A and B of figure 3.

Figure 5. Designed probe set for MSH2 with docking probes (v2). (A) theoretical probe set). B) actual hybridization image corresponding to MSH2-v2 probe set. The color conventions in this and the other 3-color microscopy images (and corresponding diagrams) is as follows: blue probes are represented in black, green probes in dark gray, red probes in light gray and the background is white.

Figure 6. Designed probe set for with docking probes (v2). (A) theoretical probe set). (B) actual hybridization image corresponding to MLH1-v1 probe set. The same color conventions are used for diagrams and microscopy images as in figure 5.

Figure 7. Validation of genomic rearrangement in MSH2 in LoVo cell line with v2 probe set. Sketches of both theoretical probe set (top) and validated rearrangement (middle) by molecular combing. The photo (bottom) is the recurrent abnormal signal set which corresponding to deletion from exon 3 to exon 8 of MSH2 (as in middle). The same color conventions are used for diagrams and microscopy images as in figure 5

Figure 8. Validation of genomic rearrangement in MLH1 in SK-OV-3 cell line with v2 probe set. Sketches of both theoretical probe set (top) and validated rearrangement (middle) by molecular combing. The photo (bottom) is the representative (but few cases) signal set corresponding to the upper stream of MLH1 probe set (left side of theoretical probe set). The difference of observation number between MSH2 probe signal (normal) and MLH1 (a part of left side) clearly demonstrates that deletion of exon 4 to 19 in MLH1 is homozygous, (consistent with reference 7). Molecular combing test also revealed that the breakpoint of deletion is larger than previously reported (downstream probes from exon 19 are all deleted). The same color conventions are used for diagrams and microscopy images as in figure 5

Table 1. describes primer sequences and coordinates on human genomic DNA used for hybridization fragment synthesis to design the probes of the invention. These primers or variant therefore obtained by adding nucleotides in the ends of the described sequences and having up to 40 nucleotides, are part of the invention..

Table 2. Analysis of sequence of probe sets and their covering region. These sequences and the sets of probes that are disclosed in particular, are part of the invention.

Sequence of each of probe sets or region was subjected to RepeatMasker test and some of representative values are shown in the table. Sum length: sum up of sequence of all probes in each set. For MLH1 and MSH2 regions, this is the total length of each region. Repeat length: sum of sequences recognized as sorts of repeat in human genome. This includes sequences other than SINE. Total repeat. % of repeat length in sum length. SINE: % of sequences categorized as SINE in sum length. ALUs: % of sequences categorized as Alu family sequences in sum length.

DETAILED DESCRIPTION OF THE INVENTION

The above described strategies, for the reasons mentioned, are unsuitable to design a high-resolution code for diagnostics applications using technologies such as molecular combing.

In the present invention, the probes are defined as follows: a short probe is a nucleic acid sequence complementary to a genomic sequence, which probe can be detected with a given marker (such as a fluorochrome) once hybridized on the genomic sequence. One probe may be either made of (i) one single fragment covering the whole sequence, or of (ii) several exactly contiguous fragments, and/or (iii) slightly overlapping fragments (with an overlap less than 250 bp) and/or (iv) fragments separated by a very short gap (less than 1000 bp). With such short overlaps or gaps, using Molecular combing in our current setup, the fragments appears almost contiguous. The distance may be adjusted depending on the specific technique and experimental conditions. For example, with less resolutive conditions, longer gaps (less than 2 kb) or overlaps may be tolerated, provided fragments separated by such a gap still appear contiguous. Under more resolutive conditions, gaps should be shorter (less than 200 bp) in order for the fragments to appear contiguous. Short probes range in size from 500 bp to 10 kb.

A long probe is a nucleic acid sequence complementary to a genomic sequence, which probe can be detected with a given marker (such as a fluorochrome) once hybridized on the genomic sequence. One probe may be either made of (i) one single fragment covering the whole sequence, or of (ii) several exactly contiguous fragments, and / or (iii) slightly overlapping fragments (with an overlap less than 250 bp) and / or (iv) fragments separated by a gap (less than 3.5 kb), provided that more than 70 % of the target sequence stretch is covered by probes (i.e. provided the gaps represent less than 30 % of the target sequence). With such overlaps or gaps, using Molecular combing in our current setup, the fragments are efficiently detected. The distance may be adjusted depending on the specific technique and experimental conditions. For example, with less resolutive conditions, longer gaps (less than 5 kb each, representing in total less than 50 % of the sequence) or overlaps may be tolerated, provided fragments separated by such gaps are still detected efficiently. Also, under such conditions, longer probes should be used (more than 20 kb) to allow for efficient detection. Under more resolutive conditions, gaps should be shorter (less than 2 kb) in order for the fragments to be efficiently detected, and probes may still be efficiently detected with shorter size (more than 10 kb). Long probes range in size from 12 kb to 150 kb.

In the present invention, the size of probes reflects the length of the genomic sequence where the probe hybridizes, independently of the number of strands in the DNA molecules. Therefore, a probe may be described as 1 kb (1 kilobase = 1000 bases) or, indifferently, as 1000

bp (base pairs): in both cases, the probe hybridizes over 1000 bases of one of the strands of the target DNA molecule (and, if the probe is double stranded, also on the 1000 complementary bases of the other strand of the target molecule).

In the present invention, a "barcode" designates a specific motif formed by a set of probes labeled with different markers, where the motif characteristics are the lengths of the probes in the set, the lengths of the gaps separating successive probes and the colors in which the probes are detected (or, more generally, the markers with which the probes are labeled).

If a high coverage barcode is to be designed for *high resolution*, probe and space lengths need to be roughly in the 0.5 kb to 10 kb range (see above). This makes it unpractical to design probes that completely exclude rearrangements, and yet are spaced closely enough for the code to allow high *location precision*. On the other hand, some *non-specific hybridization* (i.e. hybridization of [parts of] a probe on genomic regions that are not the designed target of that probe) of a probe is acceptable when using a code strategy for the reading of signals. Indeed, in applications such as Southern blot where the hybridization of a single probe is assessed or aCGH where hybridization of every probe is considered separately, the *non-specific hybridization* of probes on even a very limited number of regions may lead to completely unusable results. To a lesser extent, this is also the case with multiple-probe applications such as FISH, since the resolution of FISH is insufficient to distinguish genomic regions as far apart as several tens of megabases: a single non-specific hybridization would lead to unusable results if it were located close enough to the targeted region.

In molecular combing and other similar applications using a code strategy, the quantity of non-specifically hybridized probes is not in issue per se. If a probe (or fragments of a probe) hybridizes even multiple times outside the region of interest, it is unlikely it will recreate a motif sufficiently similar to the code to be confusing. Also, *non-specific hybridization* over short sequences (<<1 kb), even within the region of interest, would most likely not be detected, unless they are sufficiently clustered to generate a long (> 1 kb) stretch of *non-specific hybridization*. For the above reasons, the inventors have developed an alternative approach for the design of probes when the main issue is the design of a (several) *high resolution code(s)* in a (several) given genomic region(s). The main step of this approach relies only on the knowledge of the sequence of the region(s) themselves. When designing such a code, the major issue is to avoid significant non-specific hybridization within the regions of interest(s). *Non-specific hybridization*

becomes an issue only if several probes display *non-specific hybridization* on neighboring sequences outside the region of interest. In the latter case, there is a risk that the pattern of probes resembles the original code, or a rearranged version of it, and this would likely lead to false conclusions. Although the invention described herein does not allow excluding such occurrences, this is relatively easily done once the method described herein has been used to exclude other *non-specific hybridizations* (see below).

The basis for this approach is the detection and exclusion of sequences that are repetitive within the region(s) of interest. For this, only the corresponding sequence(s) (the *target sequence(s)*) have to be known. One easy way to detect such repeats is the search for local sequence alignments within the *target sequence(s)*, which can be done with e.g. a dot-plot comparison of each *target sequence* with itself and the other *target sequences*. A dot-plot is a graph with the two (sets of) sequences that are being compared forming the two axis, while dots are printed at every point where the coordinates correspond to a local homology. For example, if nucleotide x from sequence A (horizontal axis) matches nucleotide y from sequence B (vertical axis), then a dot will appear at the point with (x; y) coordinates. Graphically, local alignments appear as diagonal lines. Some more elaborate tools inspired from dot-plots are available, that compare short sequences ("words", typically a few nucleotides / tens of nucleotides long) rather than single nucleotides, and display dots in various shades of gray depending on the extent of homology, thus allowing a direct visual reading of relaxed homologies (*non-specific hybridization* may well appear with incomplete homology). The comparison may also be done directly on both strands for one of the sequences, so homologies appear for both sense and reverse complement orientations. An example of such a tool is "Dotter" (ref. 4).

With these tools, very frequent repetitive sequences, such as Alu sequences in the Human genome, appear quite clearly, as they have local homologies with numerous other sequences within the target regions. Therefore, stretches with a high frequency of these sequences appear as a gray band (horizontal or vertical depending on whether the stretch is located on the vertical or horizontal axis). The exact appearance of these stretches with dot-plot display tools will depend on settings, and possibly word size. Settings were selected such that sequence stretches longer than 200 bp with more than 80 % homology appear clearly and can be located with a roughly 10 bp precision.

A sequence of 200 bp or more that contains more than 10 significant homologous sequences (less than 1, 2, 3, 4, 5, 10, 15 or 20 % nucleotide mismatch or insertion/deletion) within the regions of interest is a frequent repetitive sequence, prone to generate significant non-specific hybridization. It is generally possible to design probes in such a way that they are void of these frequent repetitive sequences, thus increasing the specificity and the high resolution of the present technology compared to the published previous methods.

"Docking" probes

Although, as shown above, shorter probes make for more precise localization of breakpoints and measurement of deleted or amplified sequences, they are, generally speaking, more difficult to detect with fiber-fish techniques and molecular combing, as they appear as shorter stretches of signal, *i.e.*, they are both smaller and less easy to distinguish from noise (fluorescent spots either unrelated to probes or to hybridization of probes). This is particularly true when considering automatic (computer-based) detection of signals.

It is therefore desirable to include longer probes in the code (for example, more than 12 kb and less than 150 kb, preferably more than 14 kb and less than 40 kb, in particular for the detection of genetic rearrangements in the regions of MSH2 or MLH1 genes). These probes would appear as actual lines (rather than spots), readily distinguishable from noise and easily detectable due to their size. Once the signals of interest are detected, the detection of other probes located on the same DNA fiber is easier.

This is especially true using technologies such as Molecular Combing where the linearity of the fibers implies the other probes, if any, are located in the alignment of the first probe. Therefore, the invention provides that the inclusion of longer (>12 kb, preferably > 14 kb) probes in the set of probes is a step towards easier detection of signals of interest. Not all probes in the set need to be that long: in a fast and "rough" detection step, the long probes are sought, which allows the localization of signals of interest. These probes are called "docking probes" as they allow to "land" on the regions of interest efficiently. In a second step, the shorter probes are sought in the neighborhood of the docking probes (and more specifically in the case of Molecular Combing or related technologies, in the alignment of these probes). Although when performed by a human operator these steps can hardly be formally executed consecutively, if an operator may limit his search to longer probes, he can browse through images more rapidly, which would only allow him to detect these probes and spend more time on images where a

docking probe is seen in order to look for other shorter probes. As the longer docking probes would locally diminish the location precision and the resolution of the code, it is preferable for them not to be located in the region where rearrangements are sought. This is possible if the probes are located near, but not in, the region of interest, e.g. at either end of this region.

If it is desirable to only consider complete signals in the analysis of a given region (i.e. signals covering the entire contiguous region), these longer probes may also be used to assess the integrity of the region: if there is a probe located at each end and both probes are present, no breakage of the fiber has occurred during the DNA preparation or stretching step.
In cases where several non contiguous regions are analyzed in a single test, obviously each region has to have its "docking" probes in order to be correctly detected.

Continuous stretch of short probes

An alternative to the "docking probes" approach above is to design the set of probes in such a way that at least some groups of shorter probes form a continuous stretch of signal. This is possible if probe sequences are adjacent. In that case, several probes, although short enough (less than 10 kb) to provide for sufficient resolution, may well combine to form a long enough (more than 14 kb) signal for fast and reliable detection. Indeed, if the operator may combine color channels to view images, this stretch would still appear as a long line rather than a spot, allowing its distinction from background noise. This is possible by using either common optical setups such as tri-color filters in fluorescence microscopy, or by using common image viewing software. In the case of automatic detection, it is also possible to use combined color information and therefore to make use of the very characteristic aspect of a multicolor line relatively to background spot-like noise.

Measurements

The probe designs described above likely lead to a large number of probes to be measured in a test. The usual approach for probe measurement is to measure all of the probes constituting a signal, as well as the gaps separating them. In a test with a large number of probes, the amount of work required for analyzing results is increased. In order to balance this, the invention relates to a more efficient designed approach for signal measurement. This approach consists in the measurement of subgroups of probes constituting easily recognizable motifs. The subgroups are two or several consecutive probes and the gaps between them, and possibly gaps

at either end, chosen in order for their total length to remain within reasonably precise measurement range (10-30 kb).

There is likely to be a systematic bias in the measurement of digitalized images of fluorescent segments. Indeed, at the extremity of such a fragment, the intensity of the signal decreases gradually when moving away from the center, to reach the level of the background. Depending on where the operator / the software sets the threshold for the determination of the actual end there may be a systematic over- or under-estimation of the lengths. This bias is compensated for if the measured motifs have a probe at one end and a gap at the other. Therefore, it is preferable to design motifs in this way.

If a motif is found to have an abnormal length (different from the expected theoretical length) in a given sample, it remains possible to measure the probes and gaps within this motif in order to further precise the location of the rearrangement. With this approach, it is possible to measure in a fast and efficient way all of the signals for initial screening, while keeping the location precision allowed by small probes. The somewhat lower precision on measurements due to the larger size of the subgroups compared to the probes is essentially compensated for by the higher number of signals that can be measured within the same operator time.

Application to HNPCC - rationale

Colorectal cancer is the 4th most frequent form of cancer in human and around 5% of the cancer is considered as a hereditary form. The most frequent form of hereditary colorectal cancer is known as Lynch syndrome, or HNPCC (hereditary non-polyposis colorectal cancer). HNPCC increases a lifetime risk of cancer development in up to 80% (lifetime risk is around 7% in normal population US). HNPCC also increases other cancers (endometrial, ovarian, stomach).

Genetic aspect of HNPCC is known as a result of mutation in some of Mismatch Repair (MMR) genes such as MSH2, MLH1, MSH6, PMS2, etc. MSH2 and MLH1 mutation accounts for more than 80 % of all mutation of MMR genes in HNPCC. Both point mutation and large rearrangements are reported in mutation of those genes, and especially high % of large mutation in MSH2 is observed because of high level of small repetitive element in its genetic sequence. Today the molecular diagnosis is done after studies of familial cancer history, tumor characterization by microsatellite instability test.

Normally mutation one alleles of one of MMR genes is sufficient for molecular diagnosis of HNPCC. All HNPCC individuals have both wild and mutated genes. Point mutation of targeted MMR genes can be detected by sequencing of genes and current sequencing test investigates only the sequence of exons. In case of large rearrangements such as deletion and amplification (loss and gain of genetic elements, respectively), sequencing does not detect them because altered sequences do not exist, and frequently primer binding regions for sequencing are deleted. As a result, sequence information comes from only wild allele and gives false negative. Indeed, MSH2 and MLH1 genes are higher percentage of repetitive elements of SINE in their genetic sequence. To address this large rearrangement, the test should detect presence of deletion or amplification in the MMR genes. One approach is cartography of MMR genes with designed probes of hybridization. Causal large rearrangement has a wide range from sub-kb to loss of total gene (up to 100kb). A given cartography has to be sensitive to this wide dynamic range of mutation. To cope with it specific probe design was done for MSH2 and MLH1 loci.

The present invention is also related to the detection of known or unknown genomic rearrangements. It is also related to kits containing probes according to the invention, for the detection of known or unknown genomic rearrangements and the associated pathologies, or associated predispositions to pathologies such as cancers or cardiovascular diseases for example.

EXAMPLES

Application to HNPCC – Materials and method

probe design v1

Each probe (probe means continuous hybridization signal, can consist of multiple cloned DNA fragments, *e.g.*, probe 1 of MSH2-v2 covers a 15 kb stretch and consists of five cloned DNA fragments of 3kb. Since gap or overlap of each junction of these five fragments are smaller than resolution (< 50 bp), they are considered and indeed look like continuous single probe of 15kb) on a region of gene sequence itself has a length between 3-6kb. In case of larger rearrangement than probe or gap size, obvious change of color pattern of designed probe will be observed. As well as large rearrangement in probe region, such rearrangement is also detectable in gap region, meaning any rearrangement larger than 1 kb at any position in the target genes are detectable. This is a uniqueness of cartography method with high resolution probe hybridization.

Other techniques (MLPA, aCGH) can detect only such rearrangement involving probe sequence. For genes with high frequency of large rearrangement such as MSH2 and MLH1, presence of repetitive element in their genetic sequence limits a freedom of probe design for the other technology. Inclusion of repetitive element sequence in their probe design increases false detection a lot, their probe designing has to be free of repetitive element in principle.

Probe sequence was chosen by a dot plot analysis. BAC clone sequence of each gene (RP11-1084A21 (Ch2:47,574,044-47,785,729 for MSH2, RP11-426N19 (Ch3: 36,992,516-37,161,490) for MLH1 was self-plotted and all grey bands region were excluded from the target region of PCR primer design. PCR primer set was designed in the target regions by Primer3plus PCR primer design tool (ref 6). A list of the primers' sequence is shown in table 1A and B. Exclusion of *Alu* repeat was verified by both dot-plot analysis and RepeatMasker (http://www._repeatmasker.org). Fig. 1B and Fig. 2B show a lot less grey band on dot-plot of probe fragment sequence on BAC clone than dot-plot of gene (containing Alu repeat) on BAC clone. This indicates that sequence of designed probes does not include recurrent repetitive sequence in this target regions. RepeatMasker analysis (with default setting of web server) also clearly shows a dramatic reduction of % of *Alu* sequence in designed probe sequence.(table 2).

probe design v2

To facilitate "recognition" of barcodes on hybridization images, an alternative design of probe set (called v2) was done as said in "Docking" probe section. Design process is same as v1 except no exclusion of repetitive elements based on dot-plot. For v2 probe design, each probe was designed to have more than 3 kb length, close to limit to be recognized as "line", and all exon sequences are covered by a probe stretch (no exons fall in gaps) .Docking probes were designed on both extremities of each gene with 15-20 kb length. For MSH2-v2 code, specific probes covering EPCAM gene (see rationale part) was also included between two docking probes. DNA sequence of designed code v2 was subjected to dot-plot analysis to make sure that there is no segmental repeats inside of designed region (Fig.1C and 2C).

Cloning of probe fragments and labeling for hybridization probe

Each fragment of probes was amplified by PCR, then the fragment was ligated into plasmid vector (pNEB193, pCR2.1-TOPO, pCRXL-TOPO). The ligation product was transformed into *E. coli* competent cells and end-sequences of cloned fragment were verified. Purified plasmid DNA set of each gene was separated into two (v1) or three (v2) gropes

according to colors corresponding to theoretical barcodes (Fig. 3A and Fig. 4A for v1, Fig. 5 and Fig. 6 for v2 probe sets). Each group of plasmid DNA was labeled by random priming method. Either whole plasmids containing probe fragments' sequence or PCR amplified probe fragments were used as a template for random priming. There are three haptens to be used for three color detection, biotin (Biot), digoxigenin (Dig) and Alexa Fluor 488 (A488). Biot-labeling was done by BioPrime DNA labeling system (Invitrogen) with manufacture's instruction. For Dig and A488 labeling, dNTP mixture in the kit was replaced with home-blend dNTP mixtures (either 0.1mM Digoxigenin -11-dUTP (Roche applied science) for Dig labeling or 0.1mM ChromaTide® Alexa Fluor® 488-7-OBEA-dCTP (Invitrogen) for A488 labeling, 0.1mM unmodified equivalent (dTTP or dCTP) and 0.2mM each of other three deoxynucleotides in final labeling reaction solution.).

Sample DNA preparation

3 cell human cell lines were used for validation for large rearrangement detection in either MSH2 or MLH1. Cell line GM17939 was used as non-mutated sample. Cell line LoVo was used for MSH2 rearrangement validation, which is homozygous for deletion of exon 3-exon8 in MSH2. Another cell line SK-OV-3 was used for rearrangement validation of MLH1, which was reported as homozygous deletion of exon 4-exon 19 in MLH1. For each cell line, cell culture was prepared according to cell bank's instruction. Cultured cells were harvested (for LoVo and SK-OV-3 when 50-70% confluency) or collected by centrifuge (for GM17939 when between 300,000-400,000cells /ml of medium. Cell pellet was resuspended in 1xPBS/Trypsin mixture to have 1,000,000 cells in 45µl the cell suspension was mixed with an equal volume of 1.2%(w/v) NuSieve GTG agarose solution in 1xPBS (melted and equilibrated at 50°C in advance). The cell/agarose mixture as poured into a well of gel plug mold, followed by gelification at 4°C for 30 min. the gelified agarose plug was immersed in a mixture of 2 mg/ml of Proteinase K, 1%(w/v) of sarcosyl in 0.5M EDTA (pH8.0, 250µl for each plug). The agarose plug was incubated at 50°C overnight.

Next day the incubated plug was washed in 1xTE (10mM Tris-HCl, 1mM EDTA, pH8.0) 3 times for 1 hour each. The DNA plug can be stored in 0.5mEDTA at 4°C. The washed plug was stained in 100µl of 33µM YOYO-1 (Invitrogen) in TE40.2 (40mM Tris-HCl, 2mM EDTA pH8.0) for 1 hour in the dark. The stained plug was heated at 68°C in 1 ml of combing buffer

(0.5M MES pH5.5) for 20min, then cooled at 42°C 10min prior to add 1.5 unit of beta agarase I (NEB). Beta agarase treatment was carried overnight at 42°C in the dark.

The following day the treated DNA solution was poured into a combing reservoir and a level of the solution in the reservoir was adjusted with additional combing buffer.

Molecular combing

The DNA solution was set on a Molecular Combing Machine (MCS, Genomic Vision). Molecular combing was performed on a silanized coverslips (Combicoverslips, Genomic Vision). The combed coverslips was fixed at 68°C for 4 hours, then used for hybridization (or stored at -20°C until use).

Hybridization and detection of probe

For one hybridization, 5µl of each of labeled probe solutions (of both MSH2 and MLH1) was combined together and with 10µg of sonicated herring or salmon sperm DNA and 10µg of human Cot1-DNA (only for V2 probe sets), then purified by standard ethanol precipitation. The precipitate was resuspended with 20µl of hybridization buffer (50% formamide, 2x SSC, 1% SDS and BlockAid blocking solution (Invitrogen)). The resuspended probe solution was set on a clean glass slide and covered with a DNA combed coverslip. The slide was heated at 90°C for 5 min for co-denaturation of both probe and combed DNA then incubated at 37°C overnight with an humidity for hybridization between labeled probes and combed DNA.

The hybridized coverslips was washed in 50% Formamid/2xSSC solution 3 times for 5 min each, followed by another 3 times washing with 2xSSC for 5 min each. The washed coveslips was then developed with two or three layers of fluorescently labeled antibodies or streptavidin. For each layer, antibodies for all haptens were diluted 25 times in BlockAid blocking solution (20µl in final volume) and incubated for 20min at 37°C. For Biot, Streptavidin Alexa Fluor 594 (Invitrogen) was used for the 1st and the 3rd layer, biotin conjugated-goat anti-streptavidin antibody was used for the 2nd layer. Fr Dig, mouse anti-Digoxin AMCA conjugated (Jackson immunoresearch) was for the 1st layer, rat anti-mouse AMCA conjugated (Jackson immunoresearch) conjugated was for the 2nd, the goat anti-rat Alexa Fluore 350 conjugated (Invitrogen) was used for the 3rd layer. For A488, rabbit anti-Alexa Fluor 488 (Invitrogen) was used for the 1st layer, goat anti-rabbit Alexa Fluor 488 conjugated was used for the 2nd layer (no third antibody for A488). After 20 min incubation of each layer of antibody, the coverslip was washed in 2x SSC/1% Tween 20 washing solution 3 times for 5 min each at room temperature.

After the washing of 3rd layer, the coverslip was rinsed in 1xPBS, followed by successive bath of 70, 90 and 100 % ethanol for 1 min each. The coverslip was dried at room temperature prior to microscopy.

Signal acquisition and measurement

Fluorescent signal of developed antibody on the coverslip was obtained by standard epi-fluorescent microscope system or automated fluorescent microscope system (Image Xpress Micro, Molecular Devices) with custom scanning configuration for molecular combing signal. Every set of linearly aligned fluorescent signals and gaps was measured by ImageJ. Each measured set of signals (with color information) was subjected to pattern matching to determine position (if the set is a part of one of probe set) and orientation by comparison with the theoretical probe sets. All unclassified sets (did not match with any positions and orientations of theoretical probe sets) were subjected to similarity check between them to find whether recurrent abnormal pattern appears or not.

Application to HNPCC – results

Figure 3B and 4B are representative images of signal from hybridized DNA. Some of probes look like "dot" rather than "line" as expected from their length. There are some "random" spots on images of hybridization, but these spots do not interfere recognition of designed code. Although signals of some small probes (arrowed in Fig. 3B, for example) is not evident to measure "length" of probe signals for size evaluation, measurement of "distance" between probe signals is possible and equivalent to measurement of the length of probe and gaps in normal probe set hybridization

Figure 5B and 6B are the representative image of hybridization signal of barcodes-v2. Fluorescent signals are more continuous than the signals of barcodes-v1, and easier to find docking probes and measure the length of each probe and gap. These barcodes-v2 were used to visualize large genomic rearrangements of characterized cancer cell lines, LoVo and SK-OV-3 (ref. 5).

Figure 7 is a result of hybridization of barcodes v2 on combed DNA from LoVo cell line; LoVo cell line is homozygous for deletion in MSH2 (from exon 3 to 8). Hybridization slide had many normal (identical to theoretical code) signal of MLH1 gene but none of normal MSH2 signals. Instead, there was a recurrent signal of truncated form of the normal MSH2 signal (fig

7B). By deduction from the truncated signals, this truncation results from loss of probes and gaps corresponding to ex3 to8 of MSH2 gene.

Figure 8 is a result of barcodes-v2 on SK-OV-3 cell line DNA, homozygous for deletion in MLH1 (from ex4 to 19). Among many normal MSH2 signals, only a few signals of part of MLH1 (from probe 1 to probe 3) were observed. This means a lack of following sequence of MLH1, which is consistent with reference. Moreover, a lack of the right (downstream of MLH1) docking probe indicates that this deletion affects beyond exon 19 of MLH1.

The sequences selected to detect predisposition to colorectal cancer linked to rearrangements in the MSH2 genomic region or the MLH1 genomic region are preferably chosen among the following nucleotide sequences and their corresponding complementary sequences and are described as:

The short probes covering the MSH2 gene region and constituting contiguous stretches (PE1-2 and PE3-6 (SEQ ID NO:354-358); PE9 to PE15-16 (SEQ ID NO:365-373) in table 1 under the header MSH2-v2) and the other short probes covering MSH2 gene region (PE7 and PE8, SEQ ID NO:359-364 in table 1 under the header MSH2-v2); the long probes neighboring the MSH2 gene (tPP1, EPCAM5', EPCAM3' (SEQ ID NO:342-353) and cPP1 (SEQ ID NO:374-378) in table 1 under the header MSH2-v2); the short probes covering the MLH1 gene region and constituting a contiguous stretch (PE1-2 to PE10-11, SEQ ID NO:386-396, in table 1 under the header MLH1-v2) and the other short probes covering MLH1 gene region (PE12-13, PE14-15 and PE16-19, SEQ ID NO:397-401, in table 1 under the header MLH1-v2); the long probes neighboring the MLH1 gene (tPP1 (SEQ ID NO:379-385) and cPP1 (SEQ ID NO:402-408) in table 1 under the header MLH1-v2). For example, these probes may be obtained by amplification of the fragments using the primers listed in Table 1 under the headers MSH2-v2 (SEQ ID NO:139-212) and MLH1-v2 (SEQ ID NO:213-272).

Incorporation by Reference

Each document, patent, patent application or patent publication cited by or referred to in this disclosure is incorporated by reference in its entirety, especially with respect to the specific subject matter surrounding the citation of the reference in the text. However, no admission is made that any such reference constitutes background art and the right to challenge the accuracy and pertinence of the cited documents is reserved.

Table 1

MSH2-v1							
Name of probe	Name of fragment	SEQ ID NO (fragment)	For / Rev	SEQ ID NO (primer)	Sequence (5'-3')	start	end
P1	P1a_MSH2-v1	273	forward	1	TTCTTCCCAAGAGAGCCAAG	47595911	47595930
			reverse	2	CTGTTTTGGAACCCCAAGTC	47597074	47597093
	P1b_MSH2-v1	274	forward	3	GGCTTCAATCTGGGACTACG	47598716	47598735
			reverse	4	GCTGTCACCGCCTCTTTTAC	47599478	47599497
	P1c_MSH2-v1	275	forward	5	GCCAGGCACTTAGGCAGTAG	47600433	47600452
			reverse	6	TTGGTCCTGACATCCTTTCC	47601671	47601690
	P1d_MSH2-v1	276	forward	7	TTAGTTGAACAGGGCATGACAC	47602097	47602118
			reverse	8	GGTAAAGGGGCCTGATGTC	47602743	47602761
	P1e_MSH2-v1	277	forward	9	GAGCCTTGATGTTCCCTCTTAAC	47603695	47603717
			reverse	10	ACCCAGATCCGAAACTGTTG	47604324	47604343
	P1f_MSH2-v1	278	forward	11	CCGGCCTTACCTTTCATTTC	47605735	47605754
			reverse	12	CCAGGATCCAGATCCAGTTG	47606965	47606984
P2	P2a_MSH2-v1	279	forward	13	GAGTTCCATGGCAGATCACC	47612521	47612540
			reverse	14	GCAGCTTTCAATCACAAATCAG	47614067	47614088
	P2b_MSH2-v1	280	forward	15	GAAGGGTTGGTCTTGCTGTC	47615115	47615134
			reverse	16	ACCCTTTGCACCTCTCTGTG	47615632	47615651
	P2c_MSH2-v1	281	forward	17	CCCGGTGTTGAATCATTTG	47616079	47616097
			reverse	18	TTCAGCCCTGAAGGTAGAGG	47617513	47617532
	P2d_MSH2-v1	282	forward	19	CTGGCCACTTTTTGGAAGAG	47618884	47618903
			reverse	20	TGGGACGCAGAGTGATACAG	47619394	47619413
P3	P3a_MSH2-v1	283	forward	21	TTACTGGCGATCCTCAGAGC	47629651	47629670
			reverse	22	AACGCCTCTTCCGTTGTATG	47631623	47631642
	P3b_MSH2-v1	284	forward	23	GAAAGGACAGACCAAGTGCAG	47632605	47632625
			reverse	24	AGCCTGTGCAGGGAAACTC	47633083	47633101
	P3c_MSH2-v1	285	forward	25	AGTGGGATGCAGCTGAAAAG	47633591	47633610

			reverse	26	CAACAGCATGGGAAAGATCC	47635238	47635257
P4	P4a_MSH2-v1	286	forward	27	TTGAAAGTTGGTCTTAGGAAGAGG	47643286	47643309
			reverse	28	CCCAACAAACCTGGCTTTAG	47644179	47644198
	P4b_MSH2-v1	287	forward	29	AGACGCCCAAAATCAACAAC	47645155	47645174
			reverse	30	CCGCTTGCTGCTAAAAATTG	47646042	47646061
P5	P5a_MSH2-v1	288	forward	31	TGATTGCCAAGGAAGATTCAC	47657647	47657667
			reverse	32	TGGAAGTAAATGCAGGTGCTC	47658763	47658783
	P5b_MSH2-v1	289	forward	33	TCATTCTTGGGTGTTTCTCG	47659578	47659597
			reverse	34	ATGGCGGTTTTGTGGAATAG	47660015	47660034
	P5c_MSH2-v1	290	forward	35	GAGGGAGAGGGAACCTTTTG	47661699	47661718
			reverse	36	GGGGACTATACCGCATTCAC	47662243	47662262
P6	P6a_MSH2-v1	291	forward	37	TGTTGATTCATGGGCATTTG	47669651	47669670
			reverse	38	GCTGGGGAATCATGTATGAAG	47671879	47671899
	P6b_MSH2-v1	292	forward	39	CATCAAGCACAGTTCCATTG	47672243	47672262
			reverse	40	TTCTCTTTCCGTTTCCAGTG	47673113	47673132
P7	P7a_MSH2-v1	293	forward	41	GGAGCTTGGGAATTCAACTG	47678126	47678145
			reverse	42	AGAAACGGGCATGTCATAGG	47679330	47679349
	P7b_MSH2-v1	294	forward	43	CAGCCTACGTGCCCATTTC	47679649	47679667
			reverse	44	TCAAAAGATGGCCAAAATGC	47681179	47681198
	P7c_MSH2-v1	295	forward	45	GTGTTGCACCCATTAACTCG	47681915	47681934
			reverse	46	AGCCTGGTGAGAGGTGACTG	47684723	47684742
P8	P8a_MSH2-v1	296	forward	47	CACGATGCCAGTCCAATTC	47689478	47689496
			reverse	48	AAGGTGGACTTTAATGCAAAGG	47690835	47690856
	P8b_MSH2-v1	297	forward	49	GGAGTGAGAGCGACACCTTG	47691634	47691653
			reverse	50	CGACAGCTGACTGCTCTATGG	47694068	47694088
P9	P9a_MSH2-v1	298	forward	51	CACAATGGGAAAGGATGTAGC	47701939	47701959
			reverse	52	CAGAGAAAAACACCCATGACC	47704112	47704132
	P9b_MSH2-v1	299	forward	53	CACCGTGATCCTCCTTATTTC	47704395	47704415
			reverse	54	GAACAAACAACGGATGAAAGG	47704945	47704965
	P9c_MSH2-v1	300	forward	55	GTGGCATATCCTTCCCAATG	47705311	47705330

			reverse	56	CCCCCAGACTGTGAATTAAGG	47705787	47705807
P10	P10a_MSH2-v1	301	forward	57	GATGCAGATCAGGGAAATGC	47711630	47711649
			reverse	58	ATCTTGCTGGATGGACAAGG	47715272	47715291
	P10b_MSH2-v1	302	forward	59	CTTAATCCTGAAAGGCAGGTG	47715788	47715808
			reverse	60	TGTTTCTCAGGCAACCACAG	47717266	47717285
P11	P11a_MSH2-v1	303	forward	61	GAAACCACAGAATCGCCTTC	47731087	47731106
			reverse	62	ACCTGGACAGTCCCACAGAC	47733482	47733501
	P11b_MSH2-v1	304	forward	63	CAGTGCTTTTGCATCCTTCC	47734903	47734922
			reverse	64	ATTTAATCCCCTGGCCAATC	47741649	47741668
	P11c_MSH2-v1	305	forward	65	CACCTGTGCCCATCACATAG	47742239	47742258
			reverse	66	GAGTCCCCTCTTGGAGAACC	47747829	47747848
P12	P12a_MSH2-v1	306	forward	67	AAAGCCATTTCCAGTGTCG	47753989	47754007
			reverse	68	ATTGTGCAGCCAGAATTGAG	47758158	47758177
	P12b_MSH2-v1	307	forward	69	TTCACAGCAAAGTGGCTCAG	47760593	47760612
			reverse	70	GCTATTATGGGCTGCAAAGC	47764302	47764321
	P12c_MSH2-v1	308	forward	71	TTCACTCCCAACAAGCACTG	47764863	47764882
			reverse	72	TGCCCAGTCCTTTTTCACT	47765618	47765636
	P12d_MSH2-v1	309	forward	73	AATCCCTCCTGCACACTTTC	47765925	47765944
			reverse	74	AATGGATGCTTCCACTGTCC	47767687	47767706
	P12e_MSH2-v1	310	forward	75	CCATCTGTGCAATTCCTTCC	47768105	47768124
			reverse	76	GTTCAAAGGCAGAAGCCATC	47769886	47769905

MLH1-v1							
Name of probe	**Name of fragment**	**SEQ ID NO (fragment)**	**For / Rev**	**SEQ ID NO (primer)**	**sequence (5'-3')**	**start**	**end**
P1	P1a_MLH1-v1	311	forward	77	GTCTGGATTCTTTCACAATGTAGC	37005551	37005576
			reverse	78	TGCCAATCTTCTCCTCTGTTC	37006562	37006582
	P1b_MLH1-v1	312	forward	79	AACCACCCAATGTGTTCACC	37006815	37006836
			reverse	80	GTTCATTCCTGCGAGTAGGC	37007422	37007441
	P1c_MLH1-v1	313	forward	81	GCCAAAGGTGGAAAATGTTG	37008987	37009008
			reverse	82	GCCTTCTTCATGAAAGCACTG	37009873	37009893
	P1d_MLH1-v1	314	forward	83	CCAGAAGGTGGAAGCTACAG	37011079	37011100
			reverse	84	TGGGGTCAATGAAGCAAG	37011830	37011847
	P1e_MLH1-v1	315	forward	85	ACATCGACCCAGAAAGTTCC	37012314	37012335

			reverse	86	AATGTGCTTCGTACCACTGC	37012867	37012886
	P1f_MLH1-v1	316	forward	87	AGCGTGCCATTGTACTCTCC	37013822	37013843
			reverse	88	TTTCTGAGCCCATGATTTCC	37015267	37015286
P2	P2a_MLH1-v1	317	forward	89	GTGCCCAGCTAGTTCCATTC	37023623	37023644
			reverse	90	TCAAGAGCGCTAATCCCATC	37025002	37025021
	P2b_MLH1-v1	318	forward	91	TGCACATGCTCACTGAAAGAC	37026505	37026527
			reverse	92	TTTTGCCTGCAAACTGACC	37027818	37027836
	P2c_MLH1-v1	319	forward	93	CAGCAAGCACCAAATCACTG	37028305	37028326
			reverse	94	AGTACCAGCCGTCCAAACTG	37032621	37032640
P3	P3a_MLH1-v1	320	forward	95	CCTGGCCAGAAAATTCATTG	37037607	37037628
			reverse	96	ACCCTGCATTCCAAACTCAC	37039199	37039218
	P3b_MLH1-v1	321	forward	97	GCAGTCCTTTGAGGATTTAGC	37042493	37042515
			reverse	98	GAAAGATATCCAACAGGAAGTGAG	37043300	37043323
	P3c_MLH1-v1	322	forward	99	TGGCCTTGTTTAAGGTCCTG	37043746	37043767
			reverse	100	ATGGTCCTGCTGCTTCAGAG	37044723	37044742
	P3d_MLH1-v1	323	forward	101	ACCCCGTCATAGCACAGTTC	37045295	37045316
			reverse	102	CAAAGGCCATTCATCAGTTTC	37046439	37046459
P4	P4a_MLH1-v1	324	forward	103	GTGGCGTGATATCCTTGATTC	37053034	37053056
			reverse	104	CTCTGGAATGACTGCTGCTG	37054289	37054308
	P4b_MLH1-v1	325	forward	105	TGTGCTAGATGCCTCACTGG	37055182	37055203
			reverse	106	TTGCCAAGAAGCACAACAAG	37058326	37058345
P5	P5a1_MLH1-v1	326	forward	107	CGGAGGCTCTACTGTTGGAC	37062345	37062366
			reverse	108	TGCTGTCCACTCTGGAACTG	37064753	37064772
	P5b_MLH1-v1	327	forward	109	ACATCAGAAGCCCTGGTTTG	37064571	37064592
			reverse	110	GCTGGGAGTTCAAGCATCTC	37067377	37067396
P6	P6a_MLH1-v1	328	forward	111	TCGGTCTCAGTCACCATTTG	37072097	37072118
			reverse	112	AACGCACCTGGCTGAAATAC	37075920	37075939
P7	P7a_MLH1-v1	329	forward	113	TGAACCTGCAATATCTCAGAGG	37079607	37079630
			reverse	114	CTTACCGATAACCTGAGAACACC	37083805	37083827
P8	P8a_MLH1-v1	330	forward	115	CCCAGCCCATATATTTTAAAGC	37088387	37088410
			reverse	116	CCAGCCACTCTCTGGACTATC	37089049	37089069
	P8b_MLH1-v1	331	forward	117	GACATGGAGAGCCGAATCC	37089669	37089689
			reverse	118	CCATTAAAATCGGGTCTGAAAG	37091446	37091467
	P8c_MLH1-v1	332	forward	119	TCCAGACCCAGTGCACATC	37091887	37091907
			reverse	120	CATGGTCAGTGCCATCAGAG	37092412	37092431
	P8d_MLH1-v1	333	forward	121	AGCCTCCCAAAGTTAAGTGC	37092788	37092809
			reverse	122	CCCAGCTAAAACCAACACAC	37093346	37093365
P9	P9a_MLH1-v1	334	forward	123	TGCCCTCAGCTACTCACTCC	37103285	37103306
			reverse	124	AGGGCTCAGCCTTTAGGAAC	37105620	37105639
	P9b_MLH1-v1	335	forward	125	GCCAGACTCTCGTTCCATTC	37106390	37106411
			reverse	126	ACTCCCCATTCAGTCCCTTC	37111053	37111072
	P9c_MLH1-v1	336	forward	127	AGGCACAACGTCAGGTTTTC	37114109	37114130
			reverse	128	TTGGAATTTGTCCTGGTGTG	37117519	37117538
P10	P10a_MLH1-v1	337	forward	129	CACCATTGCCAACACTTCTG	37132898	37132919
			reverse	130	GCCATTGGTTTGAAGGTGAC	37134201	37134220
	P10b_MLH1-v1	338	forward	131	CTTAGTCACCGCCTGTCCTC	37134738	37134759
			reverse	132	TAGCTGCATGTGGCTAATCG	37136986	37137005
	P10c_MLH1-v1	339	forward	133	TGTGGCTCGCATTACATTTC	37137579	37137600
			reverse	134	CGCTGTCATTACCTGCTTTG	37139742	37139761
	P10d_MLH1-v1	340	forward	135	TGACCTCCAAAATCATCCAG	37140449	37140470
			reverse	136	TTCTGAGCTAGGAGGTGCTG	37141321	37141340
	P10e_MLH1-v1	341	forward	137	CCAGATTTGTAAATCCCTGTTC	37142008	37142031
			reverse	138	TGTGTGGTTCTTAAGCATTCC	37142420	37142440

MSH2-v2							
Name of probe	Name of fragment	SEQ ID NO (fragment)	For / Rev	SEQ ID NO (primer)	sequence (5'-3')	start	end
tPP1	tPP1a_MSH2-v2	342	forward	139	CTCAGTCCATCAGCCTCCTC	47574824	47577784
			reverse	140	TGCTGTGCCCTGAGATTAAG	47574823	47577783
	tPP1b_MSH2-v2	343	forward	141	AACTTAATCTCAGGGCACAGC	47577763	47580677
			reverse	142	TGCAGCTTCAGCCTCTTG	47577762	47580676
	tPP1c_MSH2-v2	344	forward	143	GCGTGGTGTTTCGTACCAG	47580604	47583785
			reverse	144	GCTACTGGCCAGAAATCTTCC	47580603	47583784
	tPP1d_MSH2-v2	345	forward	145	GCCCAGCCCTACTAAGGAAG	47583750	47586723
			reverse	146	CTGTGCTCCCCTGCTAGAAC	47583749	47586722
	tPP1e_MSH2-v2	346	forward	147	GTCGTCCTCTTCGACCTAGC	47586769	47589967
			reverse	148	CAGCGCCTATTCTACAGCAG	47586768	47589966
EPCAM5'	EPCa_MSH2-v2	347	forward	149	TTCTTCCCAAGAGAGCCAAG	47595912	47598965
			reverse	150	CCACCTTTAATCTGCCCAAC	47595911	47598964
	EPCb_MSH2-v2	348	forward	151	GTGTTGGGCAGATTAAAGGTG	47598944	47602122
			reverse	152	GCAGTGTCATGCCCTGTTC	47598943	47602121
	EPCc_MSH2-v2	349	forward	153	CTCTTTGTGCCCTTTCTTTTG	47601745	47604931
			reverse	154	AGTTCCTTAAAGCAGAGAAGATGG	47601744	47604930
EPCAM3'	EPCd_MSH2-v2	350	forward	155	AACCTGTCCCTGTGGATGAG	47604796	47607923
			reverse	156	CCGAAGCATCCTTACATTCC	47604795	47607922
	EPCe_MSH2-v2	351	forward	157	AATACCTGAACCCCCAAACC	47607722	47609876
			reverse	158	CTCAGGCTATTTTCCAGATTCAC	47607721	47609875
	EPCf_MSH2-v2	352	forward	159	GCATGCCTGTCATTCTGG	47609695	47612812
			reverse	160	TCCAAGGGACTGAAACACAC	47609694	47612811
	EPCg_MSH2-v2	353	forward	161	TTAGTGTGTTTCAGTCCCTTGG	47612790	47615135
			reverse	162	GACAGCAAGACCAACCCTTC	47612789	47615134
PE1-2	E1_MSH2-v2	354	forward	163	GCACATTACGAGCTCAGTGC	47629942	47633045
			reverse	164	CTACCAGGAGAACAGCACAGG	47629941	47633044
	E2_MSH2-v2	355	forward	165	TGGGTTAGCATTGTGTTAGGTG	47632899	47636029
			reverse	166	CCACAGGTGTGTGCCAATAG	47632898	47636028
PE3-6	E3_MSH2-v2	356	forward	167	AAGTTGCAGTTTGGCTGGTC	47635845	47638929
			reverse	168	TTATCTCCAGCGGTGCTTATG	47635844	47638928
	E4_MSH2-v2	357	forward	169	TACCATAAGCACCGCTGGAG	47638906	47642053
			reverse	170	ACTCCACCAAGCCCAGTCTC	47638905	47642052
	E5-6_MSH2-v2	358	forward	171	TTTAGAGACTGGGCTTGGTG	47642030	47644205
			reverse	172	CTCTTCCCCAACAAACCTG	47642029	47644204
PE7	I6-7_MSH2-v2	359	forward	173	CCCAGTTTCAAGCGATTAAG	47651443	47654570
			reverse	174	AGGAAAAGCATGTTATCTCCAG	47651442	47654569
	E7_MSH2-v2	360	forward	175	TTCCGTAGCAGTAGGCATCC	47654026	47657170
			reverse	176	TCACCACCACCAACTTTATGAG	47654025	47657169
	I7-8_MSH2-v2	361	forward	177	TCCCAGATCTTAACCGACTTG	47656956	47660035
			reverse	178	ATGGCGGTTTTGTGGAATAG	47656955	47660034
PE8	E8_MSH2-v2	362	forward	179	CCCAAACAACAGCATTAGCC	47670887	47673915
			reverse	180	ACATCAGCCTCGGGACAAG	47670886	47673914
	I8-9a_MSH2-v2	363	forward	181	TGAGCCCGTTGAATATAGTGG	47673830	47675514
			reverse	182	AGTTTTCCTAAACGGGATGATG	47673829	47675513
	I8-9b_MSH2-v2	364	forward	183	ATGGGTGTGCACGTGTGTAG	47675368	47678365
			reverse	184	GCCATGTGCAATTGTGAGTC	47675367	47678364
PE9	E9_MSH2-v2	365	forward	185	CCTTGCATAGTTTGCTTCTGG	47688375	47690450
			reverse	186	ATCATACAAGGGCCTGTTGG	47688374	47690449

	I9-10_MSH2-v2	366	forward	187	AAACAGAAATCGCCCAACAG	47690418	47692377
			reverse	188	TAGAGACCCACCCAGAAACG	47690417	47692376
PE10	E10_MSH2-v2	367	forward	189	CAGTCCGATTTCGTTTCTGG	47692347	47695506
			reverse	190	CACACCTAGATTTGGCAATGG	47692346	47695505
PE11	E11_MSH2-v2	368	forward	191	TTCCATTGCCAAATCTAGGTG	47695484	47698468
			reverse	192	GGCCCTAGTGTTTCCTTTCC	47695483	47698467
	I11-12_MSH2-v2	369	forward	193	AAGGAAACACTAGGGCCTACAAC	47698452	47700589
			reverse	194	CCTGGCCTCAGTACACTTTTG	47698451	47700588
PE12-14	E12_MSH2-v2	370	forward	195	AGGGATTCTCCCCACTTAGC	47700228	47702718
			reverse	196	ATTGGAGGACTGGCTCAAAG	47700227	47702718
	E13-14_MSH2-v2	371	forward	197	GCTTACCTTTGAGCCAGTCC	47702694	47705819
			reverse	198	ACATGTTCCTACCCCCAGAC	47702693	47705818
PE15-16	E15_MSH2-v2	372	forward	199	TTTCTGCATCAGTTGGTTGC	47706613	47709532
			reverse	200	GCCAAGTTATTGCTGCTTCAG	47706612	47709531
	E16_MSH2-v2	373	forward	201	AGCCCTGTGAGGTTGGTAAC	47709413	47712504
			reverse	202	TCAACAACAGCTGGAACTGC	47709412	47712503
cPP1	cPP1a_MSH2-v2	374	forward	203	CCTCTCAGGTCAGGCTTCTG	47730898	47733882
			reverse	204	GCTCCCGCTAGAGAAACTCC	47730897	47733881
	cPP1b_MSH2-v2	375	forward	205	GAGCGAAGCACCTAAAGCAC	47733879	47736946
			reverse	206	AATTGGAGGGGGTGGAGTAG	47733878	47736945
	cPP1c_MSH2-v2	376	forward	207	TGTCACCCAGTCAGGTCATC	47736760	47739876
			reverse	208	TTGGAAGGAATCCAACAAGG	47736759	47739875
	cPP1d_MSH2-v2	377	forward	209	TTCCCAGAACTCCTTGTTGG	47739846	47742962
			reverse	210	TGCAAACCCCTTCTTTTCAG	47739845	47742961
	cPP1e_MSH2-v2	378	forward	211	ACCCCATGCAGAAGCAATAG	47743027	47746218
			reverse	212	AAATCCTGAAGGTGGGTTCC	47743026	47746217

MLH1v2							
Name of probe	**Name of fragment**	**SEQ ID NO (fragment)**	**For / Rev**	**SEQ ID NO (primer)**	**sequence (5'-3')**	**start**	**end**
tPP1	tPP1b_MLH1-v2	379	forward	213	AGTTTCAGCCATGTTGCAG	37005587	37005605
			reverse	214	TTGGCAAAATTGTGACTGAG	37007511	37007530
	tPP1c_MLH1-v2	380	forward	215	CAGTCACAATTTTGCCAAGG	37007513	37007532
			reverse	216	AGTTCGTGGCATCTAACTATCG	37009688	37009709
	tPP1d_MLH1-v2	381	forward	217	GGTCCATGTGCTCCAAAAAG	37009460	37009479
			reverse	218	TCCAAAACTGGGAACAAACC	37012624	37012643
	tPP1e_MLH1-v2	382	forward	219	TGGTTTGTTCCCAGTTTTGG	37012623	37012642
			reverse	220	TAGTGCACCACAGCCTCAAG	37015706	37015725
	tPP1f_MLH1-v2	383	forward	221	GGATCACTTGAGGCTGTGGT	37015700	37015719
			reverse	222	TCCAACAACTGCTGTGAAGG	37018677	37018696
	tPP1g_MLH1-v2	384	forward	223	CACCACTGACCTTCCCTTCC	37018492	37018511
			reverse	224	GCACAGAAAGACAAATATCACATGC	37020534	37020558
	tPP1h_MLH1-v2	385	forward	225	CTCTTCCTCGTCTCCTCCTG	37020430	37020449
			reverse	226	CCAATTCAATGCAAAACCTG	37022464	37022483
PE1-2	E1_MLH1-v2	386	forward	227	CGAGCAGCTCTCTCTTCAGG	37034273	37034292
			reverse	228	AGCCTATAAGCACAGACCAACTG	37037250	37037272

	E2_MLH1-v2	387	forward	229	TTCTCTAGCAGTTGGTCTGTGC	37037242	37037263
			reverse	230	ACCCTGCATTCCAAACTCAC	37039199	37039218
PE3-4	I23_MLH1-v2	388	forward	231	GTTCATTTTGGGGCATGTTC	37039148	37039167
			reverse	232	CTGCAACCTCCTTTGAGACAG	37042218	37042238
	E3_MLH1-v2	389	forward	233	TGTCTCAAAGGAGGTTGCAG	37042219	37042238
			reverse	234	CCAAAATGAAACTGCCTTCC	37044367	37044386
	E4_MLH1-v2	390	forward	235	AGTTCCCTGGGTCATTTTCC	37044393	37044412
			reverse	236	TTGTGGGAAGGCAAACTAGC	37046381	37046400
PE5-6	E5_MLH1-v2	391	forward	237	CCTGTGCTAGTTTGCCTTCC	37046376	37046395
			reverse	238	GGTGGTCACCGTGGTAAAAG	37049553	37049572
	E6_MLH1-v2	392	forward	239	GACCACCATGTGATTTCCAAG	37049566	37049586
			reverse	240	TTGGTTGGCGGTTATTTCTC	37052510	37052529
PE7-9	E7-8_MLH1-v2	393	forward	241	TAACCGCCAACCAAGAAAAG	37052516	37052535
			reverse	242	TGTCTGGAGACCTTCCCAAG	37055360	37055379
	E9_MLH1-v2	394	forward	243	TGTGCTAGATGCCTCACTGG	37055182	37055201
			reverse	244	ACTTGCCTACATTGCCCATC	37058175	37058194
PE10-11	E10_MLH1-v2	395	forward	245	ATGGGCAATGTAGGCAAGTC	37058176	37058195
			reverse	246	TCTGCAGCCATGAATAAGTCC	37061070	37061090
	E11_MLH1-v2	396	forward	247	CAGAGCTGAGGCGATAAATTG	37060960	37060980
			reverse	248	TGCTCCTCTCCAATCCATTC	37063973	37063992
PE12-13	E12_MLH1-v2	397	forward	249	ATACTTTCCCAGCCCAAACC	37066434	37066453
			reverse	250	TGATGGGGAAATGAGAGGAG	37069438	37069457
	E13_MLH1-v2	398	forward	251	AGTGGCCTTTGTCCATTGAG	37069405	37069424
			reverse	252	GACAGAGGTGAGAGCCTAGGAG	37071540	37071561
PE14-15	E14-15_MLH1-v2	399	forward	253	AATGTGTTGGGGAAGTGGTC	37081262	37081281
			reverse	254	TTTGGACCACGGCTTTAGAC	37084405	37084424
PE16-19	E16-18_MLH1-v2	400	forward	255	AAGCTGAGGTCACGGATTTG	37087522	37087541
			reverse	256	GATGGGCAAGTTTCATCTCC	37090568	37090587
	E19_MLH1-v2	401	forward	257	TGGGACGAAGAAAAGGAATG	37090401	37090420
			reverse	258	CACCGTGCCTCAGCCTATAC	37093446	37093465
cPP1	cPP1a_MLH1-v2	402	forward	259	GGACTAACCCACCTCCCTTC	37103239	37103258
			reverse	260	GCTATAGGCAGCCCAGAGTG	37106372	37106391
	cPP2a_MLH1-v2	403	forward	261	GCCAGACTCTCGTTCCATTC	37106390	37106409
			reverse	262	AGGATTTGCCGTATGGACTC	37109450	37109469
	cPP3a_MLH1-v2	404	forward	263	TCGCCCAAAGTCACAGTAAG	37109303	37109322
			reverse	264	GATCTGTAGGCCCAGGATTTC	37112356	37112376
	cPP4a_MLH1-v2	405	forward	265	AGGGGTTTCTATGGCTGGTC	37112314	37112333
			reverse	266	CCTCCCTCAAACCTCCTCTC	37114423	37114442
	cPP5a_MLH1-v2	406	forward	267	TTCTCCTGCAGAGGAAGAGG	37114369	37114388
			reverse	268	TTGGAATTTGTCCTGGTGTG	37117519	37117538

	cPP6a_MLH1-v2	407	forward	269	AAAGCCAGGGAGTGAATGG	37117566	37117584
			reverse	270	ATGTGCATCTCCCTGGTGAC	37120703	37120722
	cPP7a_MLH1-v2	408	forward	271	TGTGGGGAAATCAAAACCTG	37120784	37120803
			reverse	272	GGGTAGACTGTGCGTGTGTG	37123930	37123949

Table 2

	MLH1-v2 probe	MLH1-v1 probe	MLH1 region		MSH2-V2 probe	MSH2-V1 probe	MSH2 region	
sum length	86366	55582	121536		106534	73609	171394	bp
repeat length	44684	18525	64712		53243	22133	94584	bp
total repeat	51.74	33.33	53.25		49.98	30.07	55.19	%
SINE	24.93	2.58	23.85		34.68	5.03	35.95	%
ALUs	22.38	0.09	21.85		32.85	0.76	34.15	%

References

1. "Gene copy number variation and common human disease", Fanciulli, *et. al. Clinical Genetics*, 2010 **77**, 201-213
2. " Dynamic molecular combing : stretching the whole human genome for high-resolution studies" Michalet, *et al*., *Science* 1997 **277**, 1518-1523 and "Bar code screening on combed DNA for large rearrangemens of the BRCA1 and BRCA2 gene in French breast cancer families", Gad, *et. al., J. Medical Genetics,*2002, **39**, 817-821
3. "Sequence-based design of single-copy genomic DNA probes for fluorescence in situ hybridization" Rogan, *et. al.,*. Genome Res. 2001 11, 1086-94.
4. "A dot-matrix program with dynamic threshold control suited for genomic DNA and protein sequence analysis". Erik L.L. Sonnhammer and Richard Durbin. Gene 1995, **167**:GC1-10
5. "Microsatellite instability, mismatch repair deficiency and genetic defects in human cancer cel lines", Boyer J.C., *et al. Cancer Research* 1995 **55,** 6063-6070,
6. "Primer3Plus, an enhanced web interface to Primer3", Untergasser A., *et al. Nucleic Acids Research* 2007 **35**, W71-W74

Claims

Claim 1. A method of *in vitro* detecting mutated or rearranged genomic polynucleotide (target) sequence comprising:

(a1) hybridizing a target genomic polynucleotide comprising one or more genomic region(s) of interest, where mutations or rearrangements are sought, to a set of short probes that bind to each region of interest without long gaps between the portions of the target sequence bound by the set of short probes said set of short probes optionally including or being in combination with a (sub)set of short probes selected so that on each genomic region some of the short probes when taken together form a long contiguous stretch inside or outside the region of interest and wherein the short probes may optionally have frequent repetitive sequences removed; or

(a2) hybridizing a target genomic polynucleotide comprising one or more genomic region(s) of interest, where mutations or rearrangements are sought, to a set of short probes that bind to each region of interest without long gaps between the portions of the target sequence bound by the set of short probes and to one or more long (docking) probe(s) that bind to sequences near but outside of the region(s) of interest; wherein the sequence(s) of the long probe(s) does not overlap that of the short probes and wherein the short and / or long probes may optionally have frequent repetitive sequences removed;

(b) detecting the locations of hybridized probes on the genomic region(s) of interest; optionally,

(c) comparing the location of the hybridized probes on the target genomic polynucleotide sequence with one or more motifs based on the hybridization of said probes to a reference, control, normal, not mutated, or not rearranged genomic polynucleotide sequence; and optionally,

(d) correlating the presence of a mutated or rearranged genomic polynucleotide with a specific phenotype, disease, disorder, or condition.

Claim 2. The method of claim 1, wherein the mutated or rearranged genomic polynucleotide sequence is obtained from a subject who has cancer or who is suspected to having cancer or who is susceptible to have a genetic predisposition to cancer.

Claim 3. The method of claim 1, wherein the mutated or rearranged genomic polynucleotide sequence is obtained from a subject who has colorectal cancer or who is suspected of having colorectal cancer or who is susceptible to have a genetic predisposition to colorectal cancer,

wherein said short and long probes identify mutations or genomic rearrangements associated with colorectal cancer,

wherein said control, not mutated or normal genomic sequence is obtained from a subject not at risk for colorectal cancer and wherein the detection of a genomic rearrangement; and

assessing presence of or risk of developing colorectal cancer when said genomic rearrangement is detected.

Claim 4. The method of claim 3, wherein the probes hybridize specifically on the MSH2 gene, in the region of the MSH2 gene, or on the MLH1 gene, or in the region of the MLH1 gene.

Claim 5. The method of claim 1, wherein the mutated or rearranged genomic polynucleotide sequence is obtained from a subject who has breast cancer or who is suspected to having breast cancer or who is susceptible to have a genetic predisposition to breast cancer.

Claim 6. The method of claim 1, wherein the mutated or rearranged genomic polynucleotide sequence is obtained from a subject who has ovarian cancer or who is suspected to having ovarian cancer or who is susceptible to have a genetic predisposition to ovarian cancer.

Claim 7. The method of claim 1, wherein the mutated or rearranged genomic polynucleotide sequence is obtained from a subject who has lung cancer or who is suspected to having lung cancer or who is susceptible to have a genetic predisposition to lung cancer.

Claim 8. The method of claim 1, wherein the mutated or rearranged genomic polynucleotide sequence is obtained from a subject who has a cardiovascular disease, disorder or condition or who is suspected of having cardiovascular disease, disorder or condition or who is susceptible to have a genetic predisposition to cardiovascular disease, disorder or condition.

Claim 9. The method of claim 1, wherein the mutated or rearranged genomic polynucleotide sequence is obtained from a subject who has a diabetes or who is suspected of having diabetes or who is susceptible to have a genetic predisposition to diabetes.

Claim 10. The method of claim 1, wherein the mutated or rearranged genomic polynucleotide sequence is obtained from a subject who has a neuromuscular disorder or who is suspected of having a neuromuscular disorder.

Claim 11. The method of claim 1, wherein the mutated or rearranged genomic polynucleotide sequence is obtained from a subject who has, is suspected of having, or is susceptible of being a carrier for a genetic or hereditary disease, disorder or condition.

Claim 12. The method of any of claims 1 to 4, wherein the short and long probe sequences are specific to human genes or to human genomic regions associated with cancer, colorectal cancer or a foetal genetic alteration known or unknown when said region or gene is mutated or genetically rearranged.

Claim 13. The method of claim 1, wherein the mutated or rearranged genomic polynucleotide sequence is obtained from a subject who has, is suspected of having, or is suspected of being a carrier for a multigenic genetic or hereditary disease, disorder or condition or for a genetic or hereditary disease, disorder or condition associated with rearrangement of genomic DNA.

Claim 14. The method of claim 1, wherein the mutated or rearranged genomic polynucleotide sequence is obtained from a subject undergoing treatment for a disease, disorder or condition associated with a genomic inherited or acquired rearrangement and the results obtained are compared to results obtained at other time points before, during or after the termination of treatment.

Claim 15. The method of any one of claims 1 to 14, wherein the hybridizing with the short and long probes in (a2) is performed simultaneously.

Claim 16. The method of any one of claims 1 to 14 or 15, wherein the short probes are 10 kb or less.

Claim 17. The method of any one of claims 1 to 14 or 15, wherein the short probe(s) comprise

at least one short (less than 10 kb) sequence and at least one non-overlapping long sequence (more than 12 kb), or

at least one group of at least two short sequences, less than 10 kb each, which total group length is longer than 12 kb and less than 150 kb, hybridizing contiguously on the mutated or rearranged polynucleotide sequence.

Claim 18. The method of any one of claims 1-14 or 15 to 17, wherein the short probes comprise a set of contiguous probes that span a stretch of the genomic polynucleotide sequences inside or outside the region of interest that is at least 14 kb.

Claim 19. The method of any one of claims 1 to 14 or 15 to 18, wherein the long probe(s) comprise one or more docking probes of more than 14 kb and less than 40 kb.

Claim 20. The method of any one of claims 1 to 14 or 15 to 19, wherein the long probe(s) is at least 14 kb and binds to a polynucleotide sequence outside the region of interest.

Claim 21. The method of any one of claims 1 to 14 or 15 to 20, wherein frequently occurring repetitive DNA sequences have been excluded from the short and/or long probes.

Claim 22. The method of any one of claims 1 to 14 or 15 to 21, wherein repetitive DNA sequences, which appear more than once and more often than statistically predicted based on their length and base content, have been excluded from the short and/or long probes.

Claim 23. The method of any one of claims 1 to 14 or 15 to 20, wherein repetitive DNA sequences between 50 and 400 contiguous nucleotides in length, which appear more than once and more often than statistically predicted based on their length and base content, have been excluded from the short and/or long probe(s).

Claim 24. The method of any one of claims 1 to 14 or 15 to 23, wherein most of repetitive Alu family DNA sequences, have been excluded from the short and / or long probes.

Claim 25. The method of any one of claims 1 to 14 or 15 to 24, wherein in b) the probes are fluorescently tagged are detected fluorometrically.

Claim 26. The method of any one of claims 1 to 14 or 15 to 24, wherein in b) each probe is tagged with one of two or more fluorescent tags.

Claim 27. The method of any one of claims 1 to 14 or 15 to 26, wherein motifs or easily identifiable subsets of the probes are detected and compared instead of every probe sequence.

Claim 28. The method of any one of claims 1 to 14 or 15 to 27, wherein at least 3 short probes are employed.

Claim 29. The method of any one of claims 1 to 14 or 15 to 27, wherein at least 10 short probes are employed.

Claim 30. The method of any one of claims 1 to 14 or 15 to 29, wherein the short probes are at least 500 bp each.

Claim 31. The method of any one of claims 1 to 14 or 15 to 30, wherein the gaps between short probes in the genomic region of interest are no more than 12 kb each.

Claim 32. The method of any one of claims 1 to 14 or 15 to 31, wherein the long probes are no more than 40 kb each.

Claim 33. The method of any one of claims 1 to 14 or 15 to 32, wherein each of the genomic region(s) of interest is (are) longer than 50 kb.

Claim 34. The method of any one of claims 1 to 14 or 15 to 33, wherein the short probes bind to a single contiguous genomic region of interest.

Claim 35. The method of any one of claims 1 to 14 or 15 to 33, wherein the short probes bind to more than one non-contiguous genomic region of interest.

Claim 36. A kit comprising a set of short probes or a set of short and a set of long probe(s); and optionally one or more components for binding said probes to a polynucleotide, for performing molecular combing, and/or for detecting whether hybridization has occurred;

(i) wherein the short probes comprise a set of probes that taken together bind to a long continuous stretch of the genomic region of interest; or

(ii) wherein the long probes bind to sequences outside the genomic region of interest, do not overlap the short probe sequences;

and optionally, where the repetitive sequences have been removed from the long and / or short probes.

Claim 37 a kit according to Claim 36 for the detection of genomic rearrangements associated with colorectal cancer or genetic predisposition to colorectal cancer.

Claim 38 a kit according to Claim 36 for the detection of genomic rearrangements associated with breast cancer or genetic predisposition to breast cancer.

Claim 39 a kit according to Claim 36 for the detection of genomic rearrangements associated with ovarian cancer or genetic predisposition to ovarian cancer.

Claim 40 a kit according to Claim 36 for the detection of genomic rearrangements associated with lung cancer or genetic predisposition to lung cancer.

Claim 41. A composition containing the short, or short and long probe(s) described by claim 1, wherein at least two of said probe sequences detect a genetic rearrangement by using Molecular Combing, said composition comprising either

at least one short (less than 10 kb) sequence and at least one non-overlapping long sequence (more than 14 kb), or

at least one group of at least two short sequences, less than 10 kb each, which total length is longer than 14 kb and less than 150 kb, hybridizing contiguously on the genetic target.

Claim 42. The composition of claim 41, wherein the short probe(s) range from 0.5 kb to 9 kb.

Claim 43. The composition according to claim 41 or 42, wherein the long probe(s) range from 14 kb to 40 kb.

Claim 44. The composition according to claim 41, wherein the size of the short probes range from 0.5 to 9 kb and wherein at least 90 % of the frequent repetitive sequences have been removed from the short sequences.

Claim 45. The composition of any of claims 41 to 43, wherein the probe sequences hybridize specifically on the MSH2 gene or in the region of the MSH2 gene or on the MLH1 gene or in the region of the MLH1 gene.

Claim 46. The composition of claim 41, wherein said short probe sequence(s) are selected from the group consisting of the group of short probes obtained by amplification using the primer pairs disclosed as SEQ ID NO: 21-60, SEQ ID NO:95-122; SEQ ID NO:163-172; SEQ ID NO:185-202 and SEQ ID NO:227-248 or the long probe sequence(s) are selected from the group consisting of the group of long probe obtained by amplification using the primer pairs disclosed as SEQ ID NO: 61-76 and SEQ ID NO:123-138.

Claim 47. A method for designing short and long probes described by claim 1 comprising:

identifying a polynucleotide containing a genomic region of interest,

selecting long probe sequences outside of the genomic region of interest but within 100 kb of the closest probe within the region of interest and optionally removing frequently repeated sequences from the long probe sequences,

selecting a short probe sequences from within the genomic region of interest so that no gaps longer than 15 kb appear between the short probes; or selecting a series of short probes that together form a long continuous stretch that covers the genomic region of interest;

hybridizing the probes to a genomic polynucleotide comprising the genomic region of interest,

detecting the hybridized probes, and

determining which sets of probes form motifs that distinguish the genomic sequence of interest from a reference genomic sequence.

Claim 48. A method of claim 47, wherein the short probe(s) and/or long probe(s) are respectively as defined in any of claims 16 to 24 or 41 to 46.

Figure 1

Dot plot analysis of of MSH2 region. A) Dot-plot of MSH2 gene sequence on RP11-1084A21 BAC clone. B) probe code v1 (without repetitive element) on RP11-1084A21. C) probe code-v2 on RP11-1084A21. Diagonal lines are perfectly matched region of DNA between two sequences. Dots are representatives of repetitive elements . Higer density of dots (or grey band) are higher density of repetitive element.

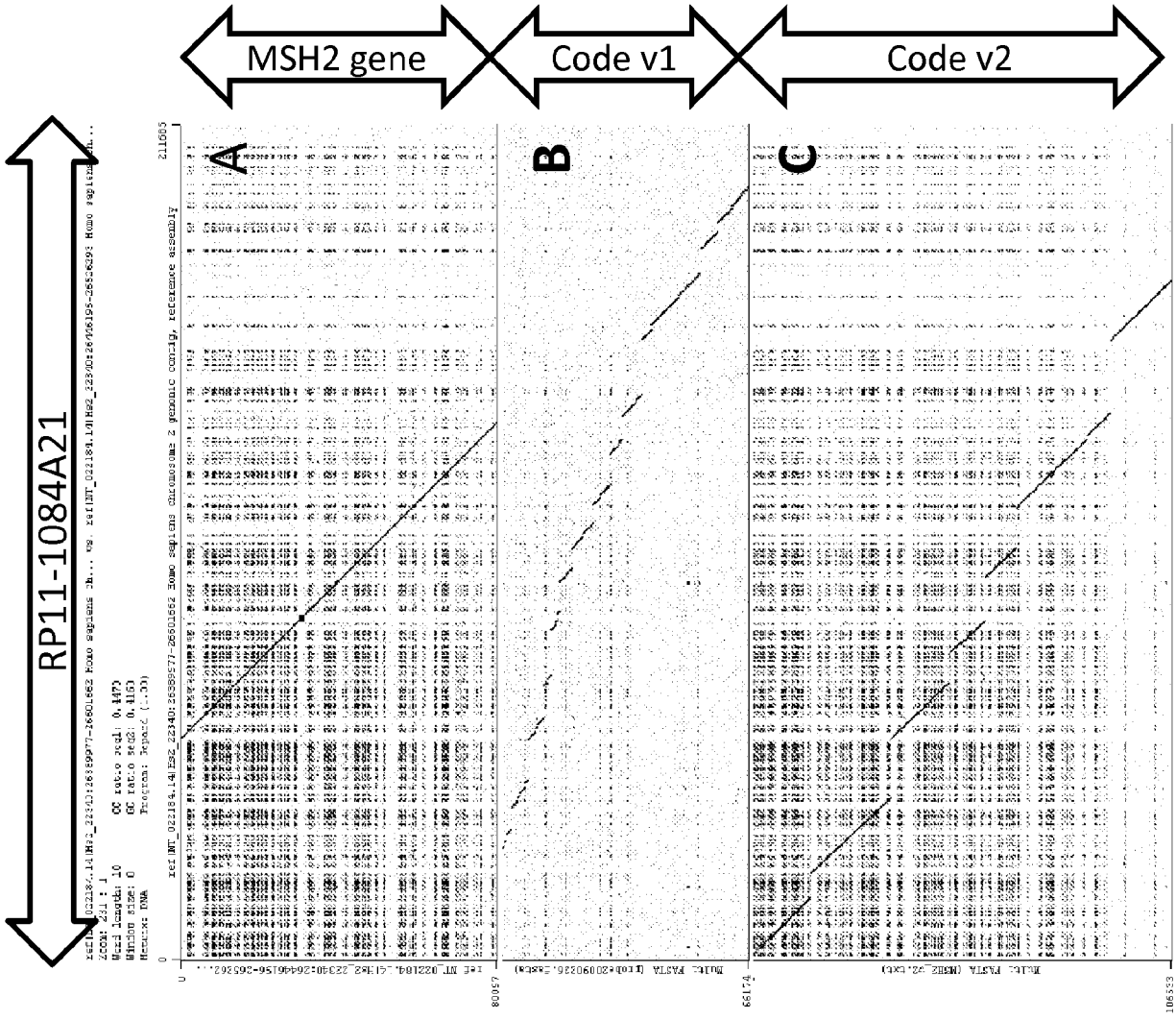


Figure 2

Dot plot analysis of of MLH1 region. A) Dot-plot of MLH1 gene sequence on RP11-426N19 BAC clone. B) probe code v1 (without repetitive element) on RP11-426N19 . C) probe code-v2 on RP11-426N19.

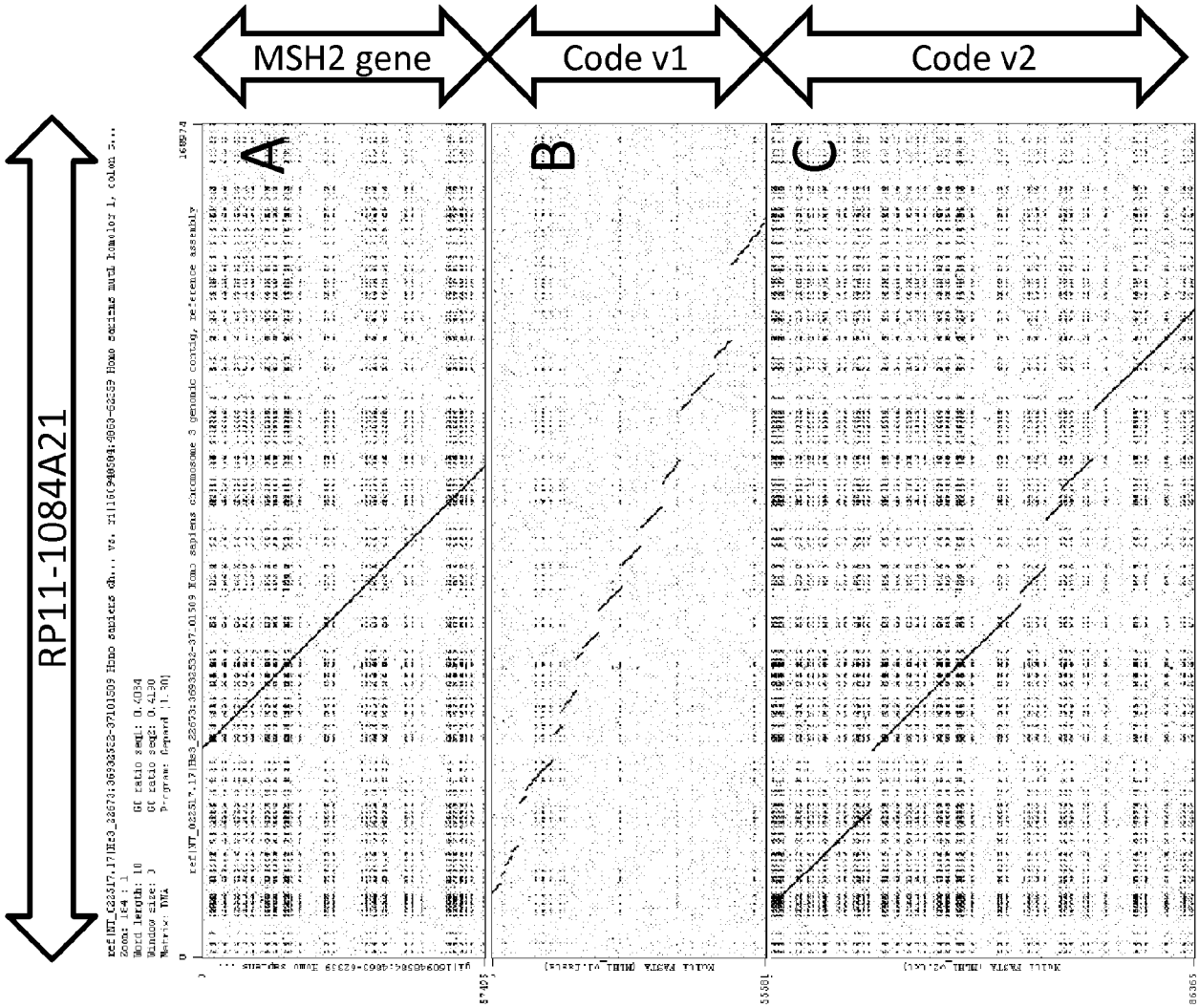
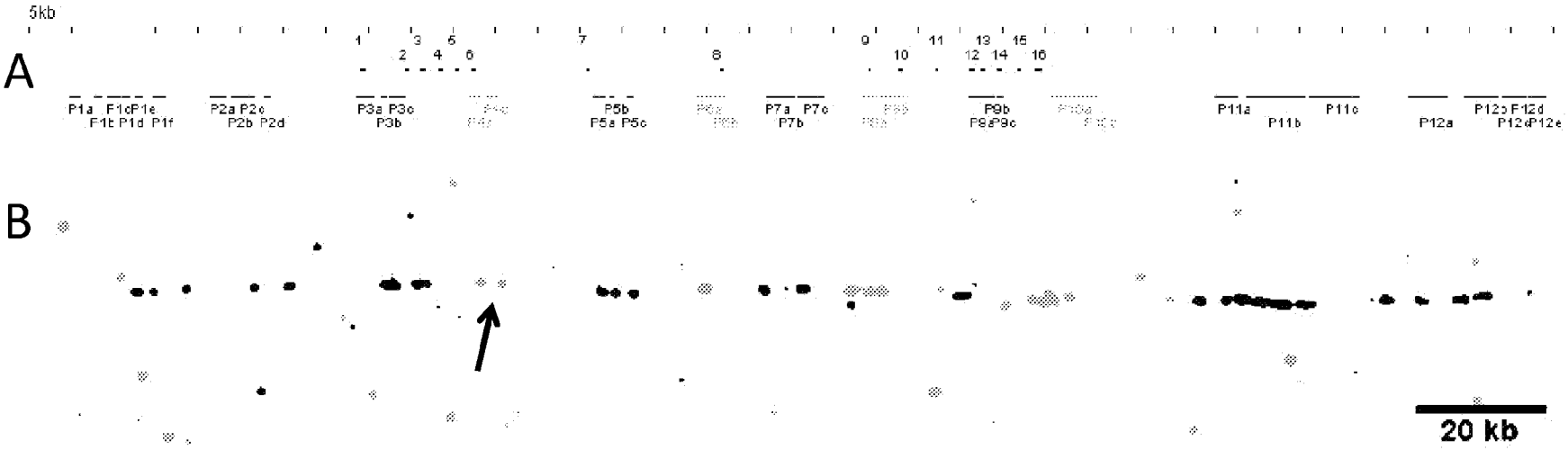


Figure 3

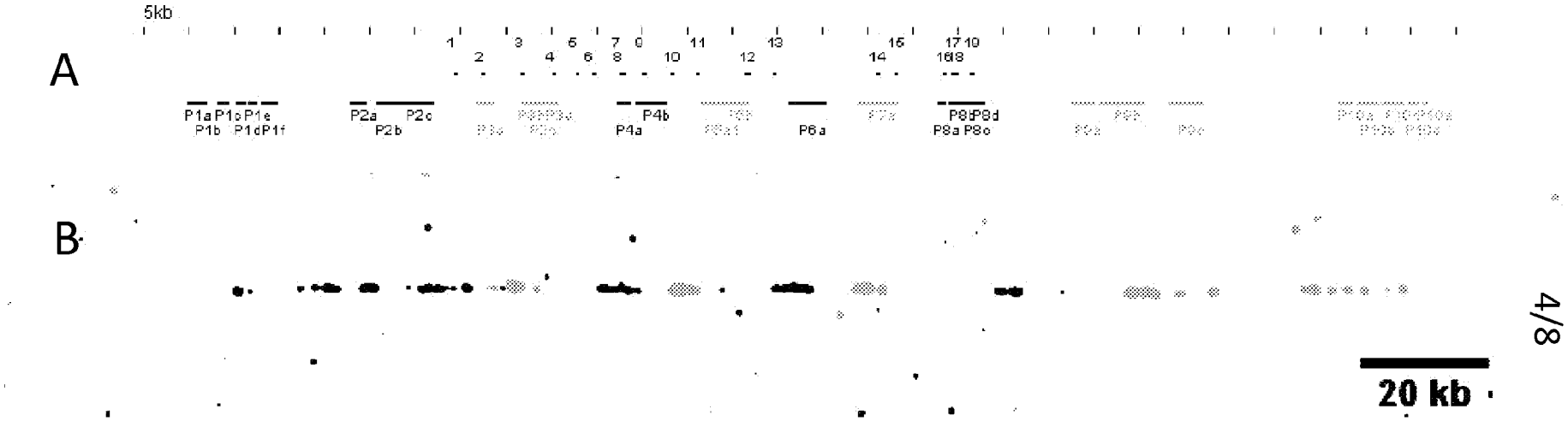
MSH2-v1



Designed code for MSH2 by exclusion of repetitive element. A) theoretical code (black and grey at bottom), and position of exon (black short segments with number of exon). B) actual hybridization image corresponding to MSH2-v1 code.

Figure 4

MLH1-v1



Designed code for MLH1 by exclusion of repetitive element. A) theoretical code (red and green), and position of exon (purple dot). B) actual hybridization image corresponding to MLH1-v1 code.

Figure 5

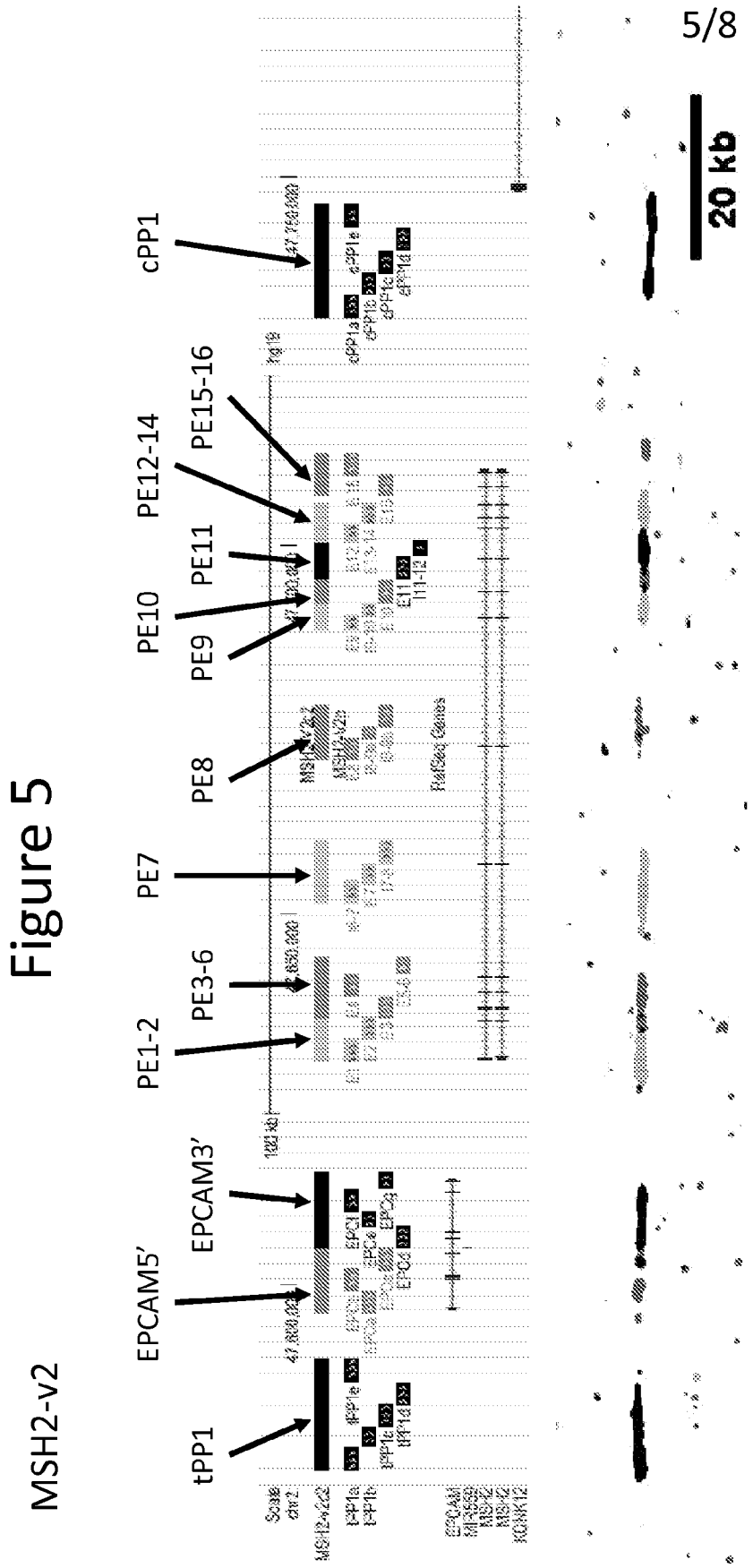

MSH2-v2
tPP1
EPCAM5'
EPCAM3'
PE1-2
PE3-6
PE7
PE8
PE9
PE10
PE11
PE12-14
PE15-16
cPP1
20 kb

Figure 6


MLH1-v2
MLH1
tPP1
PE1-2
PE3-4
PE5-6
PE7-9
PE10-11
PE12-13
PE14-15
PE16-19
cPP1
20 kb

Figure 7

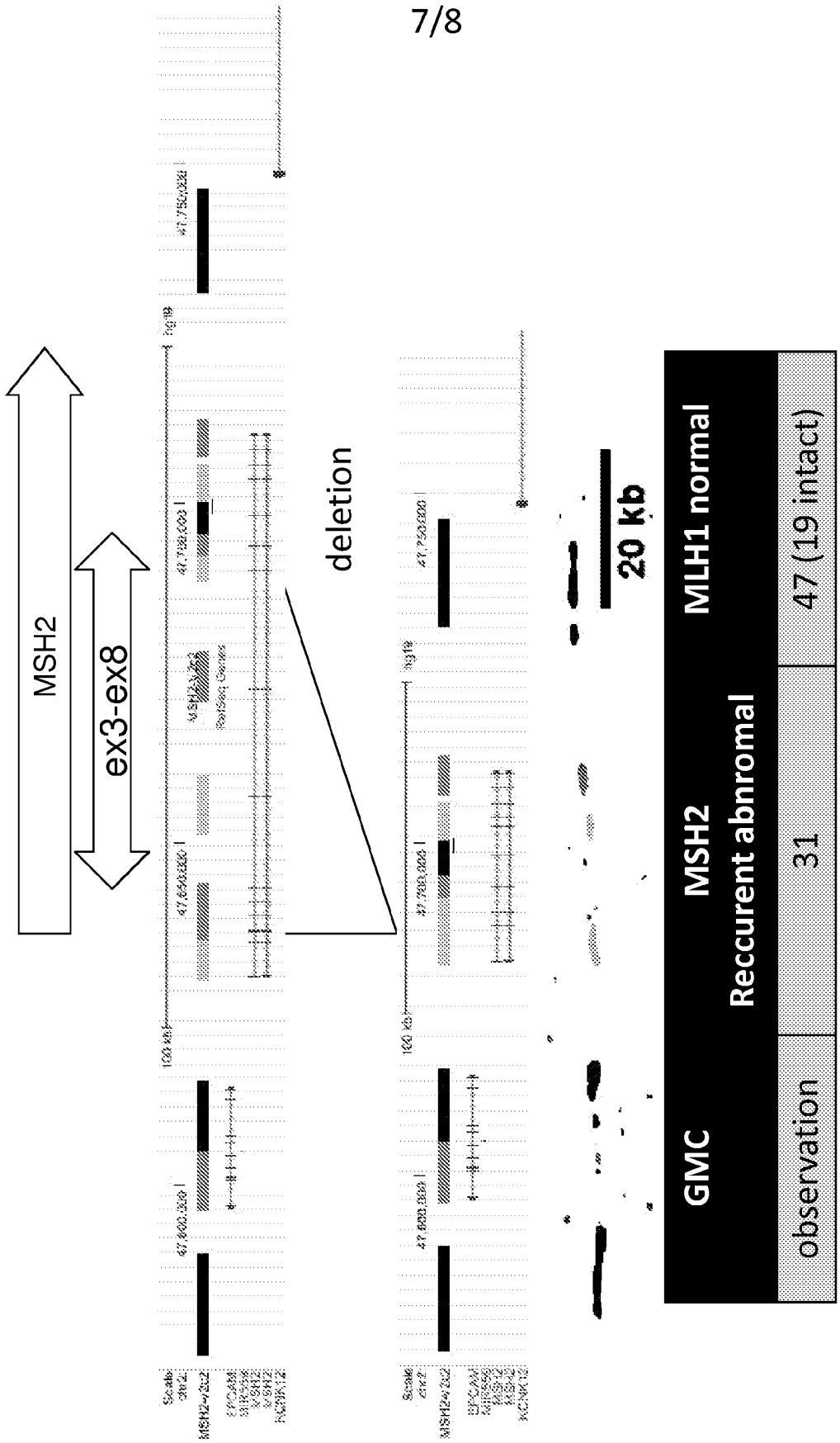
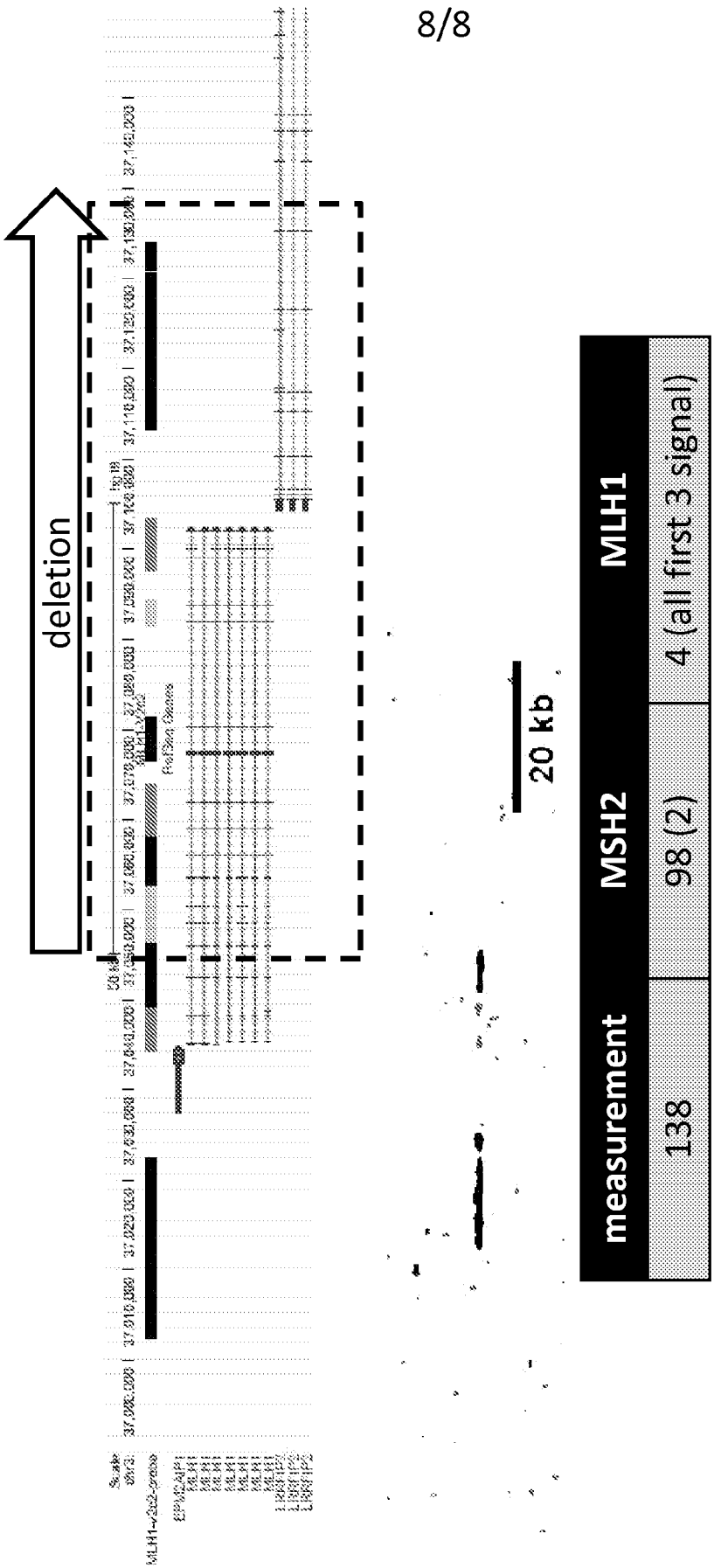

MSH2
ex3-ex8
deletion
20 kb
GMC
MSH2
Reccurent abnromal
MLH1 normal
observation
31
47 (19 intact)

Figure 8


deletion
20 kb
measurement
138
MSH2
98 (2)
MLH1
4 (all first 3 signal)

INTERNATIONAL SEARCH REPORT

International application No
PCT/IB2012/002423

A. CLASSIFICATION OF SUBJECT MATTER
INV. C12Q1/68
ADD.

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

B. FIELDS SEARCHED

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

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

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)
EPO-Internal, BIOSIS, Sequence Search, EMBASE, WPI Data

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	WO 98/18959 A1 (PASTEUR INSTITUT [FR]; CENTRE NAT RECH SCIENT [FR]; BENSIMON AARON [FR) 7 May 1998 (1998-05-07) the whole document -----	1-47
X	WO 2008/028931 A1 (PASTEUR INSTITUT [FR]; GENOMIC VISION [FR]; LEBOFSKY RONALD [CA]; BENS) 13 March 2008 (2008-03-13) the whole document -----	1-47

-/--

[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 March 2013	18/03/2013

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

2

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

C(Continuation). DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	GAD SOPHIE ET AL: "Color bar coding the BRCA1 gene on combed DNA: A useful strategy for detecting large gene rearrangements", GENES CHROMOSOMES & CANCER, JOHN WILEY & SONS, INC, US, vol. 31, no. 1, 1 May 2001 (2001-05-01), pages 75-84, XP002512886, ISSN: 1045-2257, DOI: 10.1002/GCC.1120 the whole document	1-47
X	S GAD: "bar code screening on combed DNA for large rearrangements of the BRCA1 and BRCA2 genes in French breast cancer families", J MED GENET, vol. 39, 1 January 2002 (2002-01-01), pages 817-821, XP55054670, cited in the application the whole document	1-47
X,P	KEVIN CHEESEMAN ET AL: "A diagnostic genetic test for the physical mapping of germline rearrangements in the susceptibility breast cancer genes BRCA1 and BRCA2", HUMAN MUTATION, vol. 33, no. 6, 4 April 2012 (2012-04-04), pages 998-1009, XP55054668, ISSN: 1059-7794, DOI: 10.1002/humu.22060 the whole document	1-47

2

INTERNATIONAL SEARCH REPORT

Information on patent family members

International application No
PCT/IB2012/002423

Patent document cited in search report	Publication date	Patent family member(s)	Publication date
WO 9818959 A1	07-05-1998	AT 301724 T	15-08-2005
		AU 4953397 A	22-05-1998
		DE 69733958 D1	15-09-2005
		DE 69733958 T2	18-05-2006
		DK 0935674 T3	05-12-2005
		EP 0935674 A1	18-08-1999
		ES 2247623 T3	01-03-2006
		FR 2755149 A1	30-04-1998
		JP 4223538 B2	12-02-2009
		JP 4223539 B2	12-02-2009
		JP 4226078 B2	18-02-2009
		JP 2001507929 A	19-06-2001
		JP 2008061651 A	21-03-2008
		JP 2008067716 A	27-03-2008
		US 6344319 B1	05-02-2002
		US 2002048767 A1	25-04-2002
		US 2004033510 A1	19-02-2004
		US 2009123926 A1	14-05-2009
		WO 9818959 A1	07-05-1998
WO 2008028931 A1	13-03-2008	EP 2059609 A1	20-05-2009
		IL 197392 A	30-04-2012
		JP 2010502206 A	28-01-2010
		US 2008064114 A1	13-03-2008
		US 2010041036 A1	18-02-2010
		WO 2008028931 A1	13-03-2008