Office de la Propriété Intellectuelle du Canada
Un organisme d'Industrie Canada

Canadian Intellectual Property Office
An agency of Industry Canada

CA 2854558 A1 2013/05/16

(21) 2 854 558

(12) DEMANDE DE BREVET CANADIEN
CANADIAN PATENT APPLICATION

(13) A1

(86) Date de dépôt PCT/PCT Filing Date: 2012/11/09
(87) Date publication PCT/PCT Publication Date: 2013/05/16
(85) Entrée phase nationale/National Entry: 2014/05/05
(86) N° demande PCT/PCT Application No.: EP 2012/072258
(87) N° publication PCT/PCT Publication No.: 2013/068528
(30) Priorité/Priority: 2011/11/10 (GB1119390.1)

(51) Cl.Int./Int.Cl. *C12Q 1/68* (2006.01)

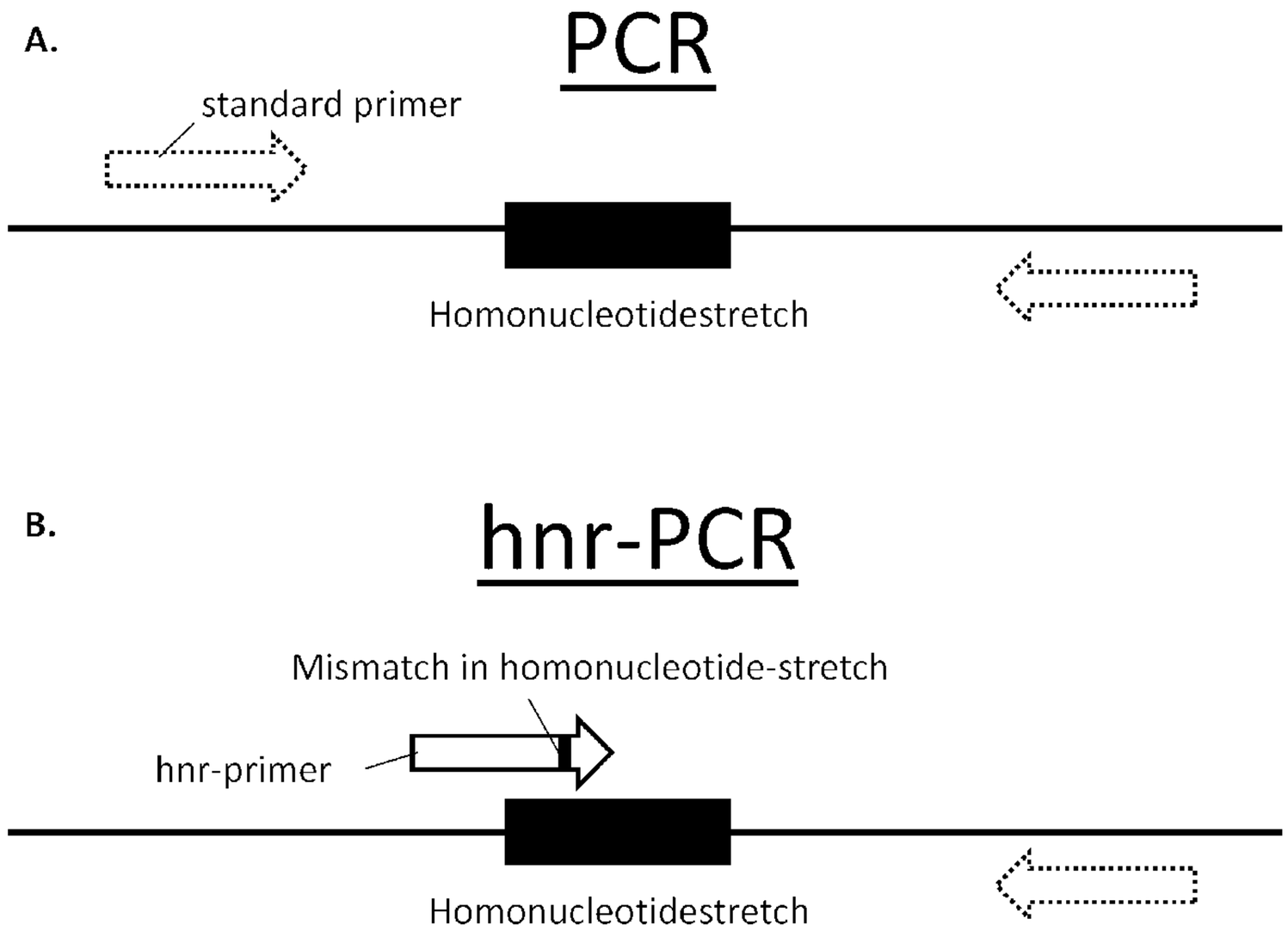
(71) Demandeur/Applicant:
CUPPENS, HARRY, BE

(72) Inventeur/Inventor:
CUPPENS, HARRY, BE

(74) Agent: SMART & BIGGAR

(54) Titre : PROCEDES DE DETERMINATION DE REPETITIONS DE SEQUENCES NUCLEOTIDIQUES
(54) Title: METHODS FOR DETERMINING NUCLEOTIDE SEQUENCE REPEATS

Figure 3



(57) Abrégé/Abstract:
The invention relates to methods of generating one or more copies of a target polynucleotide molecule, that contains a repeat of identical nucleotides. These methods comprise the steps of : using a primer which hybridizes with a part of a homonucleotide stretch in the target molecule, but is not completely complementary with the homonucleotide stretch in the target molecule. Within the portion of the oligonucleotide primer hybridizing to the homonucleotide stretch in the target molecule, at least one mismatch is incorporated, compared to the homonucleotide stretch in the target molecule. The amplified molecule is more easily sequencable than the sequence which contains the entire repeat of identical nucleotides.

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

(19) World Intellectual Property Organization
International Bureau

(43) International Publication Date
16 May 2013 (16.05.2013)

WIPO | PCT

(10) International Publication Number
WO 2013/068528 A1

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

(21) International Application Number:
PCT/EP2012/072258

(22) International Filing Date:
9 November 2012 (09.11.2012)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:
1119390.1 10 November 2011 (10.11.2011) GB

(72) Inventor; and
(71) Applicant : **CUPPENS, Harry** [BE/BE]; Violetstraat 35 bus 11, B-1000 Brussel (BE).

(74) Agent: **DE BAERE, Ivo**; IPLodge bvba, Bondgenotenlaan 93/5, B-3000 Leuven (BE).

(81) Designated States *(unless otherwise indicated, for every kind of national protection available)*: AE, AG, AL, AM, AO, AT, AU, AZ, BA, BB, BG, BH, BN, BR, BW, BY, BZ, CA, CH, CL, CN, CO, CR, CU, CZ, DE, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, 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).

Declarations under Rule 4.17:

— *of inventorship (Rule 4.17(iv))*

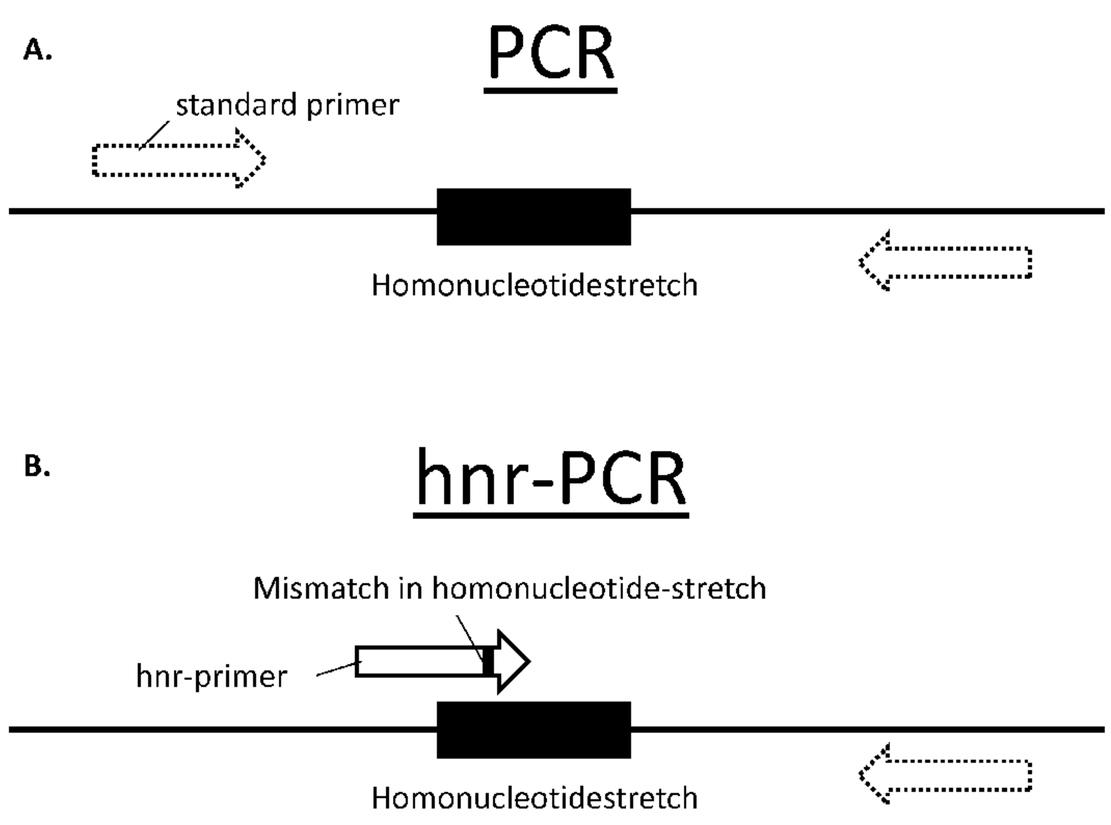
Published:

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

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

(54) **Title**: METHODS FOR DETERMINING NUCLEOTIDE SEQUENCE REPEATS

Figure 3



(57) **Abstract**: The invention relates to methods of generating one or more copies of a target polynucleotide molecule, that contains a repeat of identical nucleotides. These methods comprise the steps of : using a primer which hybridizes with a part of a homonucleotide stretch in the target molecule, but is not completely complementary with the homonucleotide stretch in the target molecule. Within the portion of the oligonucleotide primer hybridizing to the homonucleotide stretch in the target molecule, at least one mismatch is incorporated, compared to the homonucleotide stretch in the target molecule. The amplified molecule is more easily sequencable than the sequence which contains the entire repeat of identical nucleotides.

WO 2013/068528 A1

METHODS FOR DETERMINING NUCLEOTIDE SEQUENCE REPEATS

Field of the invention

The present invention relates to methods for determining the number of nucleotides in homonucleotide stretches.

The present invention relates to the field of so-called next generation sequencing methods.

Background of the invention

DNA (deoxyribonucleic acid) is the universal carrier of genetic information. DNA is an intertwined helix of two polymeric strands, each strand build up of nucleotide units attached to a backbone of deoxyribose sugars and phosphate groups joined by ester bonds. These two strands run in opposite, anti-parallel directions. Each DNA strand is build up of 4 nucleotides A, C, G and T, in a specific order for that DNA molecule. It is the sequence of these four bases along the backbone that encodes genetic information.

The two DNA strands have a complementary nature. An A nucleotide forms a base pair with a T nucleotide in the opposite strand, and vice versa; a G nucleotide forms a base pair with a C nucleotide in the opposite strand, and vice versa.

In eukaryotic cells, DNA is transcribed to RNA (ribonucleic acid). RNA molecules are rather similar to DNA molecules, the single chains of nucleotides are attached to a backbone of ribose sugars and phosphate. Depending on their function, there are different types of RNA molecules. mRNA molecules are used by cellular organisms to carry the genetic information encoded in DNA to direct synthesis of proteins. In some viruses, RNA is even used as the genetic code instead of DNA.

DNA can be replicated by DNA polymerases. A DNA polymerase can only extend an existing DNA strand paired with a template strand. It cannot begin the synthesis of a new strand as such. To begin synthesis, a short fragment of DNA, an oligonucleotide or RNA molecule, called a primer, must be created and paired with the template DNA strand.

DNA polymerase synthesizes a new strand of DNA by extending the 3' end of an existing nucleotide chain, adding new nucleotides to the template strand one at a time through the creation of phosphodiester bonds. The incoming building blocks are the nucleoside triphosphates (dNTPs: dATP, dCTP, dGTP, dTTP). The oxygen of the 3'-hydroxyl end of the growing DNA strand makes a nucleophilic attack on the alpha phosphate (the one closest to the sugar) of the dNTP. The result is that the dNMP (deoxyribonucleoside monophosphate, or nucleotide) becomes covalently bound to the 3' carbon of the sugar at the end of the DNA strand, thus lengthening the strand by one nucleotide. Moreover, pyrophosphate and a proton are released (Figure 1). Then the process repeats.

There is a huge interest in determining genetic DNA information, such as the nucleotide found at a given position, the sequence order found at given locus/loci in the genome, or even the complete genome. Even RNA can be sequenced when it is first converted to cDNA. Genetic information is determined by sequencing technologies, such as Maxim-Gilbert sequencing, Sanger sequencing and derivatives thereof, parallel pyrosequencing (Roche 454 Life Sciences), reversible terminator-based sequencing by synthesis (Illumina), Sequencing by Oligonucleotide Ligation and Detection (SOLiD) (Life Technologies), Ion Semiconductor Sequencing (Ion Torrent, Life Technologies), Single Molecule Real Time sequencing (SMRT) based on zero-mode waveguides properties (Pacific Biosciences), nanopore sensing (Oxford Nanopore Technologies), etc. Depending on the sensitivity of many sequencing technologies, pools (clones) of identical DNA molecules are sequenced in parallel. Sequencing of single DNA strands in parallel is only possible by Single Molecule Real Time sequencing and nanopore sensing. In most sequencing technologies, a double-stranded DNA molecule is denatured, and one of these single-stranded DNA molecules is then sequenced. In essence, this single DNA strand is used as a template in a sequencing reaction for the synthesis of a second complementary DNA strand, based on the complementary nature of DNA. A new DNA strand can be synthesized when a small DNA fragment, an oligonucleotide, binds to the DNA template. This oligonucleotide is a primer for

further extension of a new growing DNA strand by incorporation of nucleotides on the complementary principle. Such oligonucleotides are typically about 10-25 nucleotides long and can be easily synthesized. By monitoring the synthesis of this new DNA strand, i.e. the order in which nucleotides are incorporated in the new DNA strand, the DNA sequence of that strand can be determined. Given the complementary nature of the two DNA strands, the sequence of the other original DNA strand is then also known.

Despite the progress in sequencing techniques, the sequence of the number of nucleotides in homonucleotide stretches cannot be accurately determined with certain newer generation sequencing technologies. For example pyrosequencing was invented in the early nineties, highly parallel sequencing was introduced in 2005, while Ion Semiconductor sequencing was only introduced in 2009. The pitfall of inaccurate calling of homonucleotide stretches is already known for almost two decades.

Summary of the invention

In this invention, the length of longer homonucleotide stretches is reduced to a series of shorter nucleotide parts; either both parts are shorter nucleotide stretches, or one part is a shorter nucleotide stretch and the second part is only 1 nucleotide long. For example when a stretch of 8 T residues is modified at the 7th position, the above indicated shorter nucleotide stretch contains 6 T nucleotides and the T at the 8th position represents the above mentioned “ second part of only 1 nucleotide long”. Both shorter nucleotide parts can be more accurately determined, and the combined accurate analysis of the smaller parts allows more accurate determination of the length of the original longer homonucleotide stretch.

A first aspect of the invention relates to methods of generating one or more copies of a target polynucleotide molecule, or part thereof, that contains a repeat of identical nucleotides. These methods comprise the steps of :

- altering one nucleotide in this repeat of identical nucleotides, or
- altering different single nucleotides separated at intervals in said repeat of

identical nucleotides into another nucleotide,

in order to divide said repeat of identical nucleotides into two or more smaller altered parts of identical nucleotides in the copied molecules, wherein an oligonucleotide is used for this purpose which extends in the unaltered repeat of identical nucleotides, either until the end of the unaltered repeat of identical nucleotides or not, and wherein said oligonucleotide primer is not 100% complementary within the sequence complementary to the repeat in the target sequence.

In other words as illustrated in figure 4, the primer which is used which hybridizes with a part of the homonucleotide stretch in the target molecule, but does not hybridize completely with the homonucleotide stretch in the target molecule. Within the portion of the oligonucleotide primer hybridizing to the homonucleotide stretch in the target molecule, at least one mismatch is incorporated, compared to the homonucleotide stretch in the target molecule.

Typically, the polynucleotide is DNA.

In certain embodiments of these methods, the smaller altered parts of identical nucleotides are generated through an enzymatic or chemical reaction, such as DNA synthesis, ligation and/or amplification.

In certain embodiments of these methods, one nucleotide in the repeat of complementary identical nucleotides of the primer is replaced by another nucleotide type so that the unaltered repeat of consecutive number of identical nucleotides is split in two shorter altered parts of these identical nucleotides interrupted by the replaced nucleotide type.

In certain embodiments of these methods the part of identical nucleotides at the 3' of the primer is shorter than the part of identical nucleotides at the 5' end of the primer.

In certain embodiments of these methods, several single nucleotides in the primer in the repeat of complementary identical nucleotides are replaced at regular intervals so that all shorter parts of identical nucleotides do not exceed a given length for example not longer than 5, 6 or 7 nucleotides.

In other embodiments of these methods the obtained shorter part of identical nucleotides is no longer than 4-6 nucleotides.

In particular embodiments, the primer contains in addition at the 5' end one or more adapter nucleotides which are not complementary to the target sequence. These adapter nucleotides can be used for discriminating them from other fragment types and/or for further processing such as amplification, sequencing, inclusion of bar code sequences.

In certain embodiments of these methods different DNA synthesis /ligation/ amplification reactions of the same type, or of different types, are combined in a multiplex(-like) format.

A second aspect of the invention relates to a method for determining the number of nucleotides in a nucleotide repeat of a template DNA molecule in which the sequence of a copied DNA molecule is determined by sequencing, and which uses the prior knowledge of the altered nucleotides and positions generated in the copied molecules as described in the above cited methods of the first aspect, comprising the step of counting the number of identical nucleotides in the generated smaller parts and substituted nucleotides for generating the smaller parts of identical nucleotides.

Counting is most important for the last shorter part, not necessary for the other part were it can be assumed because they are provided by the primer for which you already known the exact count, even if it is much longer than 7 nucleotides.

Optionally, these methods further comprise the step of performing a method of determining the qualitative nature of the stretch of identical nucleotides and its downstream and upstream DNA regions, wherein the fragments obtained from each method are discriminated and separately analysed using said adapter sequences.

A further aspect of the invention relates to a data carrier comprising program instructions for analysing and providing the results of the above described method, when executed on a computer.

A further aspect of the invention relates to the use of one or more oligonucleotides in a DNA synthesis, ligation, or amplification reaction to reduce larger repeats of identical nucleotides in two or more smaller parts of identical nucleotides in a method as described above.

Figure legends:

Figure 1: Synthesis of a new DNA strand. A pyrophosphate is released from the incoming new nucleotide, and a proton is released from the extending DNA strand.

Figure 2: Absolute and relative signal differences obtained in assays such a pyrosequencing and Ion Semiconductor Sequencing of different homonucleotide stretches.

Figure 3: A. Standard PCR in which the primers, indicated by arrows, flank the region of interest, in this case a region which contains a homonucleotide stretch. [prior art] B. hnr-PCR, in accordance with an embodiment of methods of the present invention, in which one of the primers extends in the homonucleotide stretch of interest. This hnr-primer contains one or more mismatches at the (complementary site of the) homonucleotide stretch.

Figure 4.

A. hnr-PCR, in accordance with an embodiment of methods of the present invention, in which one of the primers extends in the homonucleotide stretch of interest.

B. Example of hnr-PCR, in accordance with an embodiment of methods of the present invention, in which the homonucleotide stretch contains a stretch of 9 T-residues. The hnr-PCR-primer extends 5 nucleotides in the homonucleotide stretch, and contains a mismatch at position 4 at the site of the homonucleotide stretch. After synthesis of a new DNA strand, a homonucleotide stretch of 3 T-residues and 5 T-residues, separated by one A-nucleotide will be obtained. When after sequencing 5-T nucleotides are found in the last 3' stretch, the original stretch contained 9 T-nucleotides; if 4-T nucleotides are found in the last stretch, the original stretch contained 8 T-residues, etc.

Figure 5: Example, in accordance with an embodiment of methods of the present invention as shown in figure 4, wherein a sequence tag (a given nucleotide sequence) is added to the primers as adapters. When both standard and hnr-PCR is used in a multiplex(-like) PCR format, standard PCR primers (one or both) may contain one type of adapter, while the primers used in hnr-PCR (one or both) may contain another type of tag. If more than one amplicon

is generated in a multiplex(-like) PCR, the tags of the primers used for generation of the 'standard'-amplicons may be identical or not. If more than one hnr-amplicon is generated in a multiplex(-like) format, the tags of the hnr-primers may be identical or not. In this way all amplicons can be combined and separated for analysis.

Figure 6: Example in accordance with an embodiment of methods of the present invention as shown in figure 5, wherein one or more adapter sequences or attached to the primers for further processing, such as a priming site for a second DNA synthesis or PCR, sequencing, barcode, etc., or a combination thereof

Detailed description of the invention

Different sequencing technologies use different technologies for determining the nucleotide sequence of a DNA strand. A common feature of current high throughput sequencing technologies is that many DNA strands are sequenced in parallel, mostly on the surface of a small plate (picotiterplate, flow cell), yielding up to more than 600 Gb (giga bases) sequence information. Typically, these DNA fragments carry identical sequence ends, e.g. through the linking of small nucleotide adapters during the DNA preparation phase for sequencing, so that the same primer can be used for the sequencing of all DNA fragments in parallel.

Current high throughput sequencing technologies can be broadly divided into two categories on the basis by which sequencing is performed; sequencing by synthesis (Illumina, Roche, Ion Torrent, Helicos and Pacific Biosciences) and sequencing by ligation (SOLiD, Life Technologies).

Sequencing by synthesis can be further divided into two distinct categories. In the first subcategory, each incorporated nucleotide is detected as such, such as in the reversible terminator approach in which each nucleotide is labelled with a different fluorophore and a single base elongation with any base is assayed

concurrently. Only one single base incorporation is possible in each cycle/step of the sequencing reaction. Indeed, the nucleotides are protected, such that they can be incorporated in a growing DNA strand, but once incorporated, another nucleotide cannot be incorporated in the same cycle/step. In a new cycle, the incorporated nucleotide of the previous cycle is first deprotected, so that one new protected nucleotide can be incorporated in the new cycle. In the second subcategory, a side product of nucleotide incorporation is monitored, such as pyrophosphate in pyrosequencing or a proton in Ion Semiconductor Sequencing. In both the latter strategies, individual native nucleotides are added sequentially. Pyrosequencing or Ion Semiconductor Sequencing follows DNA polymerase progression along a DNA strand by allowing only a single dNTP to be available for incorporation in a given cycle, and then takes advantage of the chemical reaction that occurs when the dNTP is incorporated by the polymerase. This reaction is detected either by inducing a bioluminescence cascade starting from pyrophosphate and detecting the emitted light (pyrosequencing), or by directly detecting protons released during incorporation as a change in pH (Ion Semiconductor Sequencing). A new cycle is then started for another dNTP. The four different dNTPs are thus separately added in four different cycles, and this is repeated for a given number of rounds, in which each round is built up of the same order of 4 cycles of administration of the 4 different dNTPs.

More specifically for pyrosequencing, a sequencing primer is hybridized to a single stranded DNA template and incubated with the enzymes DNA polymerase, ATP sulfurylase, luciferase, the substrates adenosine 5' phosphosulfate (APS) and luciferin. The addition of one of the four deoxynucleotide triphosphates (dNTPs) (dATP is not used, but rather dATPαS which is not a substrate for luciferase) initiates the second step. DNA polymerase incorporates only the given dNTP if it is complementary onto the template. This incorporation releases pyrophosphate (PPi) stoichiometrically. ATP sulfurylase then quantitatively converts PPi to ATP in the presence of adenosine 5' phosphosulfate. This ATP acts as fuel to the luciferase-mediated

conversion of luciferin to oxyluciferin that generates visible light in amounts that is proportional to the amount of ATP. The light produced in the luciferase-catalysed reaction is detected by a camera and analysed in a program. Unincorporated nucleotides and ATP are then degraded by apyrase, and the reaction can restart with another nucleotide. The four DNA nucleotides are added sequentially in a fixed order across the picotiterplate. During the nucleotide flow, thousands of copies of DNA bound to each of the beads are sequenced in parallel. When a nucleotide complementary to the template strand is added into a well, the polymerase thus extends the existing DNA strand by adding nucleotide(s). In case that the template carries a stretch of identical nucleotides, a number of nucleotides equal to the length of the stretch can be incorporated in the same cycle/step. Indeed, native unprotected dNTPs are used, so that more than one nucleotide can be incorporated in a given cycle. The signal strength is proportional to the number of nucleotides. The signal strength for homopolymer stretches, however, is only linear up to seven consecutive nucleotides. The sequence of solutions which produce chemiluminescent signals allows the determination of the sequence of the template in the end.

More specifically for Ion Semiconductor Sequencing, a microwell containing clonal template DNA strands on a bead is flooded with a single species of deoxyribonucleotide (dNTP). If the introduced dNTP is complementary to the leading template, the nucleotide is incorporated into the growing complementary strand. This causes the release of a hydrogen ion that triggers a hypersensitive ion sensor, which indicates that a reaction has occurred. If the introduced dNTP is not complementary to the leading template nucleotide, no nucleotide is incorporated and no hydrogen ions are released. The chip is sequentially flooded with one nucleotide after another. Again here, since native unprotected dNTPs are used, multiple dNTP molecules will be incorporated in a single cycle at homonucleotide stretches. This leads to a corresponding number of released hydrogens and a proportionally higher electronic signal. Again here, the

accuracy of correct calling of the number of nucleotides in homonucleotide stretches decreases for longer homonucleotide stretches.

The Ion Semiconductor Sequencing technology differs from other sequencing technologies in that no optics is used. Ion semiconductor sequencing may also be referred to as Ion Torrent sequencing, pH-mediated sequencing, silicon sequencing, or semiconductor sequencing. Ion Semiconductor Sequencing creates a direct connection between the chemical and the digital worlds, enabling fast, simple, massively scalable sequencing. Ion Semiconductor Sequencing Chips are designed, manufactured and packaged like any other semiconductor chips. Wafers are cut from a silicon boule. The transistors and circuits are then pattern-transferred and subsequently etched onto the wafers using photolithography. This process is repeated 20 times or more, creating a multi-layer system of circuits. Ion Semiconductor sequencing benefits from four decades of exponential improvement in semiconductor technology, also known as Moore's Law.

Pyrosequencing and Ion Semiconductor Sequencing allows sequencing to proceed at a much faster rate than e.g. reversible terminator sequencing, because fewer steps are required to detect a base and to continue the extension of a template. As such, it is common to achieve currently 100-base-pair reads in fewer than 3 h with pyrosequencing or Ion Semiconductor Sequencing. Moreover, longer sequence reads can be obtained in pyrosequencing and Ion Semiconductor sequencing, of which the templates can be generated by classical PCR amplification. In diagnostics this is highly desirable, since PCR is much less labour-intensive then other template preparation technologies such as the construction of DNA libraries. Another plus of the preparation of sequencing template by PCR is that the standard length of obtained amplicons, as well as the read lengths obtained with these sequencing technologies, are in the range of the size, or over the size, of the average length of exons. Therefore the number of amplicons needed for the sequencing of a given gene is of the same order of the exons in most genes, so

that most amplicons cover single exons. When performed in a multiplex format, all exons of interest can be enriched in one or a limited number of steps. This can also achieved in a multiplex-like format, such as highly parallel amplification on integrated fluidic circuits (Fluidigm). Indeed, although each PCR reactor of the integrated fluidic circuit generates only one type of amplicon, the parallel nature of the integrated fluidic circuit achieves and assures a multiplex-like format.

These sequencing formats, especially Ion Semiconductor Sequencing, has the highest potential to become the most important sequencing format in future routine genetics tests, especially tests in which only one or a few genes are analysed. The inaccurate calling of (longer) homonucleotide stretches, however, is a serious pitfall. Especially homonucleotide stretches of 7 nucleotides long, or longer, are not always correctly called. Most genes carry at least 1 or a few homonucleotide-stretches in this size range. Moreover, homonucleotide stretches are prone to mutations, e.g. because of slippage of DNA polymerase. A correct characterization of homonucleotide stretches is therefore a prerequisite for a diagnostic test that aims to analyse complete regions of a genome, such as (a) gene(s). For example, the coding region and exon/intron junctions of the *CFTR* gene, in which mutations cause cystic fibrosis, contains three homonucleotide stretches of at least 7 nucleotides. Most commercial *CFTR* tests, the most common genetic tests performed in the Caucasian population and which are so far not based on sequencing, only screen for about the 30 most common *CFTR* mutations, which include three mutant loci located in these homonucleotide stretches. The pitfall of inaccurate characterization of homonucleotide stretches thus prohibits the implementation of pyrosequencing and Ion Semiconductor Sequencing in routine genetic diagnostics. Indeed, either no accurate typing can be performed at these homonucleotide-stretches, or a second assay such as classical Sanger is performed across these problematic homonucleotide-stretches. This is not desirable in routine genetic testing, and a ‘one-stop’ or ‘one-shop’ test is preferred.

For most routine diagnostic genetic tests, there is only an interest in the analysis of one or a few genes. In fact, only the exons and exon/intron junctions of these genes are characterized. These DNA regions of interest are therefore first enriched from the total genome in order to reduce the 'background signal' of DNA regions not of interest, and to increase the sensitivity of detection.

Detailed description of the invention

Neighbouring nucleotide positions in a DNA molecule may harbour a(n) identical nucleotide(s). In the context of the present invention a repeat of identical nucleotides is also called a “homonucleotide stretch”. A repeat can have as little as 2 or 3 identical nucleotides, and can extend up to more than 15 to 25 identical nucleotides. Repeats of 7, 8, 9 or 10 are more common, repeats of more than 15 are less common. Because of unpredictable structural conformations, some repeats of 5 or 6 repeats may be difficult to type. Particular embodiments include ranges having as lower and upper limit any of the above cited values, For example, 7 T-nucleotides in a row would thus be called a homonucleotide stretch of 7 nucleotides.

In pyrosequencing and Ion Semiconductor sequencing, native (unprotected) nucleotides are added to the sequencing reaction in each cycle. At a given homonucleotide-stretch, more than 1 nucleotide will thus be incorporated in a sequencing reaction. The number of complementary nucleotides that will be incorporated equals the number of nucleotides in the homonucleotide stretch, so that the signal strength is proportional to the number of nucleotides. The signal strength for homopolymer stretches, however, is only linear up to seven consecutive nucleotides. Indeed, the longer the homonucleotide stretch, the smaller the difference in signal intensity for 1 less or 1 more nucleotide in the homonucleotide stretch (Figure 2). For example, a 3-homonucleotide stretch results in a 50% higher signal then a 2-homonucleotide stretch, while a 8-homonucleotide stretch only results in a 14 % higher signal then a 7-homonucleotide stretch. At a given moment, the signal difference that needs to be detected in order to determine the exact length of a homonucleotide-stretch

is of the same order as the variability in background signal, so that the system reaches its limit of detection of signal differences. Some shorter homonucleotide stretches than 7 nucleotides might even incorrectly called since sequence-context specific secondary structures may hamper DNA polymerization.

In this invention, the longer homonucleotide stretch which cannot be accurately determined is reduced to two, or more (e.g. 3, 4, 5 or 6), smaller parts, in which at least the 3' part can each be accurately determined in sequencing assays, such as pyrosequencing and Ion Semiconductor sequencing. Only the last one of the repeats needs to be in this range of less than 5 to -7 identical nucleotides, the first one may still be more than 5 to 7 nt, since the first one is completely located in the oligonucleotide and should have the expected number when well synthesized and purified. However, to control for lower quality oligo's, a very long stretch might be reduced to more than 2 parts so that they are all less than 5 to7 nt long so that they can be counted as an extra control. Their combined length then allows an accurate determination of the original longer homonucleotide stretch. Typically only the last smaller homonucleotide-stretch thus has a length that can be typed at the highest accuracy.

The reduction in homonucleotide stretch can be achieved in a DNA synthesis reaction, which is defined in this invention as 'homonucleotide-stretch-reduction (hnr) DNA synthesis'. When a PCR reaction is used for DNA synthesis/amplification, the technique is defined in this invention as "homonucleotide-stretch-reduction PCR" (hnr-PCR).

In fact, in order to obtain sufficient strong sequencing signals and/or to sequence specific regions of the total genome only, the DNA molecules to be sequenced have to be enriched by copying or even amplification. hnr-PCR can be a means of copying/amplification so that the homonucleotide stretch reduction and amplification of DNA target can be performed at the same time in a single step.

For amplification, primers are designed so that they bind to their complementary target DNA through hybridization. Primers are normally designed to be 100% complementary to their target DNA. However, even a primer that is not 100% complementary may bind to its target DNA region, especially when DNA synthesis is performed in less stringent conditions. A new DNA strand is then synthesized through extension of the 3' end of the primer. Although a primer needs not to be 100% complementary in order to prime DNA synthesis, the 3' nucleotide needs to be 100% complementary and be hydrogen-bounded with its DNA target in order to initiate the synthesis of a new DNA strand. For hnr-PCR, an hnr-primer can be designed at a homo-nucleotide stretch which extends in the homonucleotide-stretch (Figures 3 and 4), either partly in the homonucleotide stretch or until its complete end. Of course, the primer preferably needs to contain more unique complementary sequences 5' preceding the targeted homo-nucleotide-stretch in order to increase the specificity of binding. In practice, additional adapter sequences may be added 5', to allow further processing of the newly synthesized DNA strands such as adapter sequences for primer needed in sequencing, second PCR steps, emulsion-PCR, barcode sequences, etc. The 3' end of an hnr-primer will thus harbour a homonucleotide stretch, or even 1 nucleotide, in which the nucleotide type is complementary to the nucleotide type in the target homonucleotide stretch. Since a primer needs not to be necessarily 100% complementary to its target region, non-complementary nucleotides can be incorporated at certain positions to obtain an hnr-primer, e.g. the 2nd last 3' position of the hnr-primer, the 3rd last 3' position of the hnr-primer, etc. … (in this context the nucleotide at the 3' end is at the "first" 3' position"). When one nucleotide is substituted for a non-complementary nucleotide in the region of the homonucleotide-stretch, the homonucleotide-stretch gets disrupted and split in two smaller parts so that the original homo-nucleotide-stretch is reduced to two smaller parts. The two parts might be either two smaller homonucleotide stretches of the same length, or not. The 3' part might even become one nucleotide long, and may thus even not be a homonucleotide stretch anymore. An hnr-primer which does not extend until the end of the homonucleotide stretch will detect deletions and insertions in

the homonucleotide stretch. An hnr-primer which does extend until the end of the homonucleotide stretch will detect insertions, but not deletions in the homonucleotide stretch.

For example, for a homonucleotide stretch of 8 nucleotides, an hnr-primer which extends 6 nucleotides in the homonucleotide-stretch and in which the second last nucleotide of the hnr-PCR is substituted by a non-complementary nucleotide, a newly synthesized DNA strand will contain a 4-homonucleotide stretch, followed by the non-complementary nucleotide, and followed by a 3-homonucleotide stretch. If a sequencing reaction detects at this last position a homonucleotide-stretch of 3 nucleotides long, the DNA fragment under investigation then harbours 8 nucleotides in the original homonucleotide stretch. If a sequencing reaction detects at this position a homonucleotide stretch of 2 nucleotides long, the DNA fragment under investigation then harbours 7 nucleotides in the homonucleotide stretch. And if a sequencing reaction detects at this position a homonucleotide stretch of 4 nucleotides long, the DNA fragment under investigation then harbours 9 nucleotides in the homonucleotide stretch. In this way, and in this example, the inaccurate calling of 7, 8 or 9 nucleotides in a homonucleotide stretch is transformed to an accurate calling of 2, 3 or 4 nucleotides respectively.

In another example, for a homonucleotide stretch of 8 nucleotides, an hnr-primer which extends until the of the homonucleotide-stretch and in which the second last nucleotide of the hnr-PCR is substituted by a non-complementary nucleotide, a newly synthesized DNA strand will contain a 6-homonucleotide stretch, followed by the non-complementary nucleotide, and 1 nucleotide of the same type as the 6-homonucleotide stretch.

In another example, for a homonucleotide stretch of 8 nucleotides, an hnr-primer which extends 7 nucleotides in the homonucleotide-stretch and in which the third last nucleotide of the hnr-PCR is substituted by a non-complementary nucleotide, a newly synthesized DNA strand will contain a 4-homonucleotide stretch, followed by the non-complementary nucleotide, and followed by a 3-homonucleotide stretch.

More than 1 nucleotide mismatch can be incorporated at regularly intervals in a hnr-primer, so that different smaller homonucleotide stretches are obtained as a further control. For a homonucleotide stretch of 15 nucleotides in which the 6th and 12th nucleotide is mismatched in the hnr-primer, the 15-homonucleotide stretch will be reduced to two 5-homonucleotide stretches and a third 3-homonucleotide stretch.

Of course, sequencing reactions of such hnr-amplicons provide no information about potential mutations at the site of the primer, since the observed sequenced sequence at that site is derived from the primer instead of the target DNA under investigation. In the example above in which an 8 nucleotide long homonucleotide stretch is reduced by hnr-PCR, in which the hnr-primer extends 6 nucleotides in the homonucleotide-stretch and in which the second last nucleotide of the hnr-PCR is substituted by a non-complementary nucleotide, the 4 nucleotides of the first smaller homonucleotide stretch, the mismatched nucleotide, and the first nucleotide of the second 3-homonucleotide stretch, does not provide sequence information of the DNA fragment under investigation since it is derived from primer sequence. The second and third nucleotide of the second 3-homonucleotide stretch, however, does provide sequence information of the DNA under investigation. Also for this reason it may be preferable that hnr-primers do not extend until the complete end of the homonucleotide stretch, so that sequence information is obtained of the junction of the homonucleotide stretch.

In many instances, the sequence nevertheless needs to be determined at the site of the hnr-primer or even upstream. Indeed, in many instances, a homonucleotide stretch is located somewhere in an exon, and the complete sequence of the exon and exon/intron junctions needs to be determined. Therefore, a classical PCR amplification may also be performed using primers flanking the complete region of interest, including the homonucleotide stretch

(Figure 3). hnr-PCR might thus be used as such, or in combination with classical PCR.

The combined analysis of a standard amplicon and hnr-amplicon against a given region will then conclude the actual sequence, in which the standard amplicon provides the qualitative information of the sequence, except the exact length of the homonucleotide stretch, while the hnr-amplicon will determine the exact length of the homonucleotide stretch.

One single type of hnr-fragment, or homonucleotide stretch, might be generated and analysed, and therefore a single pair of primers is used. On the other hand, a combination of different types of hnr-fragments, directed against different homonucleotide stretches may be generated and analysed in a multiplex(-like) format in which multiple pairs of primers are used. In such a multiplex-format, even the respective standard amplicons might be included, such as amplicons containing the complete homonucleotide stretches under investigation

When multiplex-(like) PCR formats are used in which standard amplicons and hnr-amplicons are amplified together, the non-hnr-primer used for generation of the hnr-amplicon might be either the same, or not the same, as the one used for the amplification of the respective standard amplicon.

In new generation sequencing technologies many sequence reads need to be analysed at any position and need to be aligned to the reference sequence. Indeed, the analysis has a probabilistic nature and these sequencing technologies have a relatively high sequencing error per sequencing read. Typically, every nucleotide in a sequence under investigation needs to be detected in 20 to 30 sequence reads to deduce a diploid sequence with a high accuracy. Moreover, all sequences of multiple amplicons or DNA targets are generated in a single experiment and a mixture of sequence reads of all these amplicons will be obtained. Typically, a mixture of reads derived from a mixture of amplicons derived from different targets, such as all exons of a gene, are obtained. When standard amplicons and hnr-amplicons to the same target are combined, the obtained mixture will be even more complex and will contain for a

given homonucleotide stretch both sequence reads obtained from standard PCR amplicons and reads from hnr-PCR amplicons. Apart from their difference in length (the hnr-PCR amplicon only starts or ends around the homonucleotide stretch sequence), the hnr-amplicon contains one or more intently introduced mismatched nucleotide(s) which behave as (a) point mutation(s). When standard amplicons and hnr-amplicons are aligned together, both types of amplicons will correctly align to the reference sequence, and the mismatched nucleotide will be wrongly called as a mutation. The standard amplicons and hnr-amplicons thus need to be separated for analysis and separately aligned so that their alignment and analysis does not interfere. This can be easily realized when different 5' adapter sequences are introduced in the primer used for standard amplification and hnr-amplification (Figure 5). Although the adapter sequences may differ for each amplicon, all standard amplicons preferably contain the same adapter sequence, and if more than one homonucleotide stretch needs to be analysed by hnr-PCR, all hnr-amplicons preferably contain a same different adapter sequence. The pool of obtained sequence reads may then be separated in two pools based on the adapter sequence, so that a pool of standard amplicons and a pool of hnr-amplicons is obtained. Each pool will then be separately aligned to the reference sequence so that their alignment and analysis does not interfere. Preferably the reference sequence for the homonucleotide stretch may be adapted so that the mismatched nucleotide is taken up in the reference sequence. To all primers, both standard primers and hnr-primers, additional adapter sequences may be added for further processing, such as sequences for oligonucleotide hybridization for another amplification, sequencing, and barcode sequences (Figure 6).

In the context of the present application, PCR, and therefore hnr-PCR is typically the method of hnr-DNA-synthesis. However the methods of the present invention can be performed with any technique that syntheses one or more new DNA strands from an original DNA strand.. Suitable methods known in the art are for example, isothermal amplification (rolling circle amplification), single strand displacement amplification, nucleic acid sequence-based amplification

(NASBA), solid phase PCR (on beads or arrays) (Raindance), padlock probes, selector probes, collector probes, Haloplex PCR (Halogenomics), ligase chain reaction and amplification of 'extension-ligation of bound oligos'-products (TSCA; TruSeq Custom Amplicon, Illumina), or a combination thereof.

Claims

1. A method of generating one or more copies of a target polynucleotide molecule, or part thereof, that contains a repeat of identical nucleotides, comprising the steps of :
- altering one nucleotide in this repeat of identical nucleotides, or
- altering different single nucleotides separated at intervals in said repeat of identical nucleotides into another nucleotide, in order to reduce said repeat of identical nucleotides into two or more smaller altered parts of said identical nucleotides in the copied molecules,
- wherein this alternation is performed by using an oligonucleotide primer which extends in the unaltered repeat of identical nucleotides of the target polynucleotide, and wherein said oligonucleotide primer is not 100% complementary within the sequence complementary to the repeat in the target sequence.

2. The method according to claim 1, wherein said oligonucleotide primer does not extend until the end of the unaltered repeat of identical nucleotides of the target polynucleotide

3. The method according to claim 1 or 2, in which the polynucleotide is DNA.

4. The method according to any one of claims 1 to 3, wherein the smaller altered parts of identical nucleotides are generated through an enzymatic or chemical reaction, such as DNA synthesis, ligation and/or amplification, or a combination thereof.

5. The method according to any one of claim 1 to 4, in which one nucleotide in the repeat of complementary identical nucleotides of the primer is replaced by another nucleotide type so that the unaltered repeat of consecutive number of identical nucleotides is reduced to two shorter

altered parts of these identical nucleotides interrupted by the replaced nucleotide type.

6. The method according to any one of claims 1 to 5, wherein the repeat of identical nucleotides at the 3' of the primer is shorter than the repeat of identical nucleotides at the 5' end of the primer.

7. The method according to any one of claims 1 to 6, wherein the nucleotide in the repeat of complementary identical nucleotides of the primer which is replaced by another nucleotide type is at the second, third, fourth or fifth position counted from the 3' end of the oligonucleotide primer.

8. The method according to any one of claims 1 to 7, wherein several single nucleotides in the primer in the repeat of complementary identical nucleotides are replaced at regular intervals so that all shorter parts of identical nucleotides do not exceed a given length.

9. The method according to claim any one of claim 1 to 8, wherein the obtained 3' part is no longer than 4, 5 or 6 nucleotides.

10. The method according to any one of claims 1 to 9, wherein the primer contains in addition at the 5' end one or more adapter nucleotides which are not complementary to the target sequence.

11. The method according to any one of claims 1 to 10, wherein different DNA synthesis/ligation/amplification reactions of the same type, or of different types, are combined in a multiplex(-like) format.

12. The method according to any one of claims 1 to 11 further comprising the step of performing a method of determining the qualitative nature of the stretch of identical nucleotides and its downstream and upstream DNA

regions, wherein the fragments obtained from each method are discriminated and separately analysed using said adapter sequences.

13. A method for determining the number of nucleotides in a nucleotide repeat of a template DNA molecule in which the sequence of a copied DNA molecule is determined by sequencing, and which uses the prior knowledge of the altered nucleotides and positions generated in the copied molecules according to any one of claims 1 to 12, comprising the step of counting the number of identical nucleotides in the generated smaller parts and substituted nucleotides for generating the smaller parts of identical nucleotides.

14. A data carrier comprising program instructions for analysing and providing the results of a method according to claim 12 or 13, when executed on a computer.

15. Use of one or more oligonucleotides in a DNA synthesis, ligation, or amplification reaction, or combinations thereof, to reduce larger repeats of identical nucleotides in two or more smaller parts of identical nucleotides in a method according to any one of claims 1 to 12.

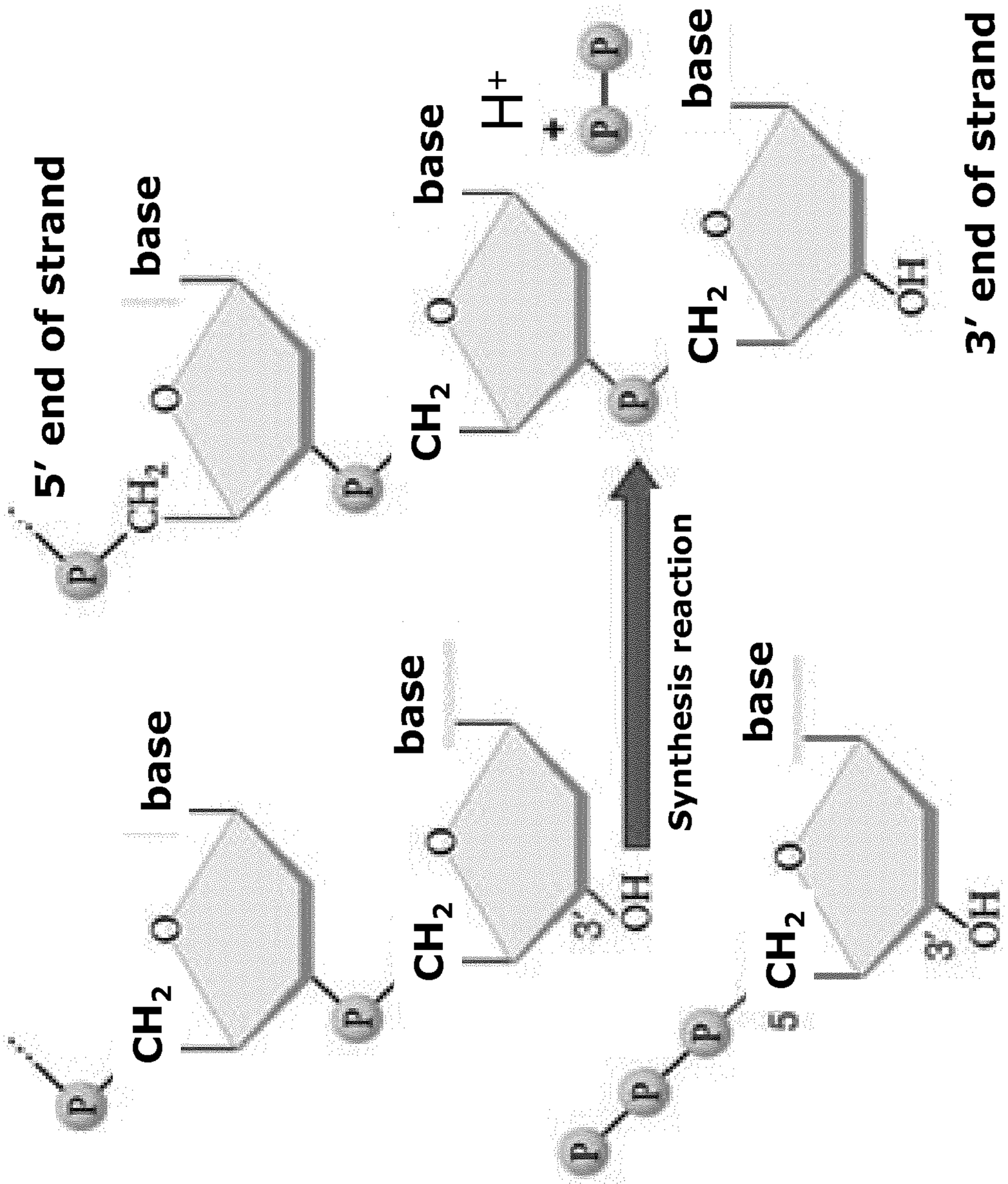

5′ end of strand
base
O
CH2
P
base
O
CH2
P
H+
+
P
P
base
O
CH2
OH
3′ end of strand
Synthesis reaction
P
P
P
5
base
O
CH2
3′
OH
base
O
CH2
3′
OH

Figure 1

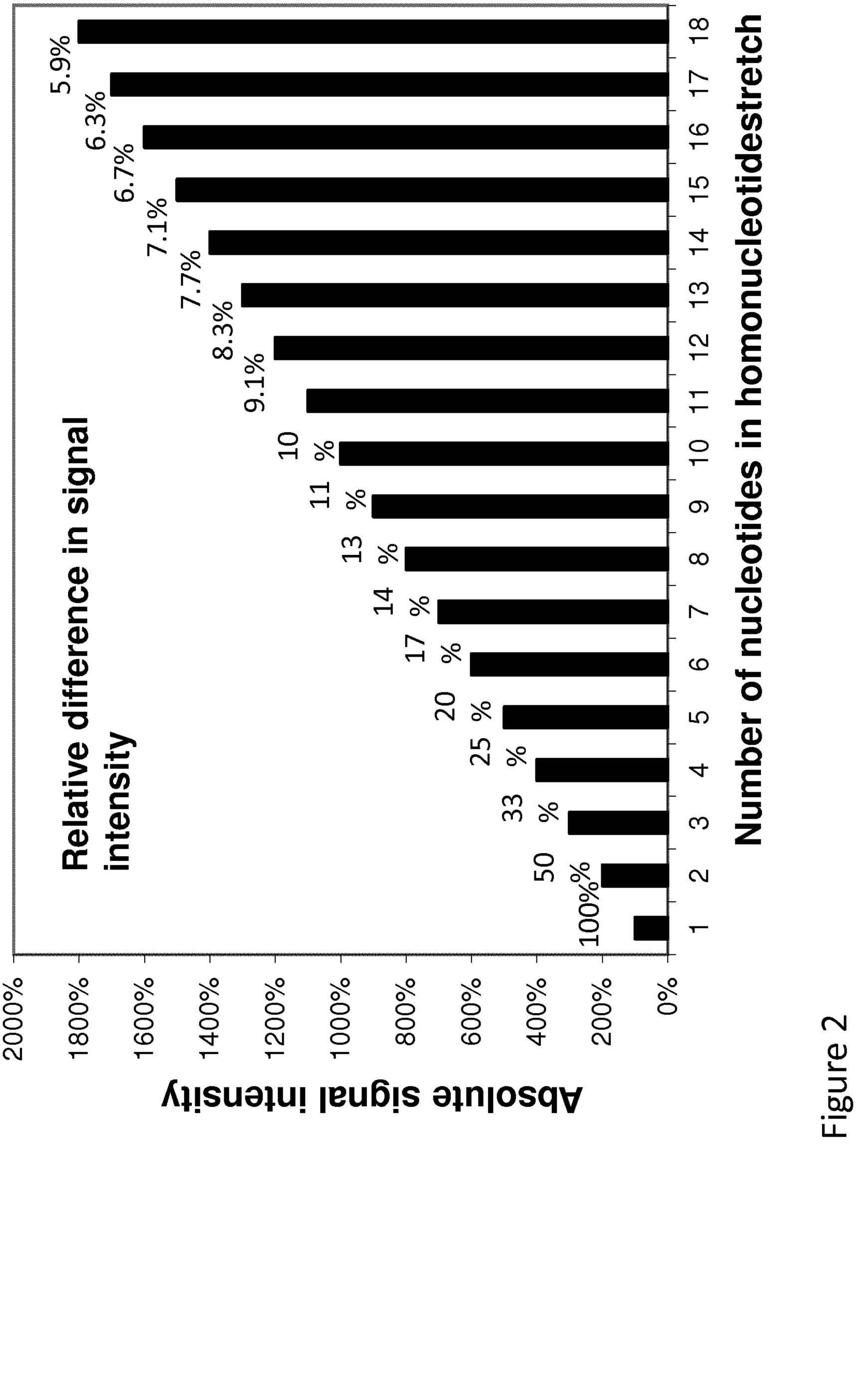

Absolute signal intensity
2000%
1800%
1600%
1400%
1200%
1000%
800%
600%
400%
200%
0%
Relative difference in signal intensity
100%
50 %
33 %
25 %
20 %
17 %
14 %
13 %
11 %
10 %
9.1%
8.3%
7.7%
7.1%
6.7%
6.3%
5.9%
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18
Number of nucleotides in homonucleotidestretch

Figure 2

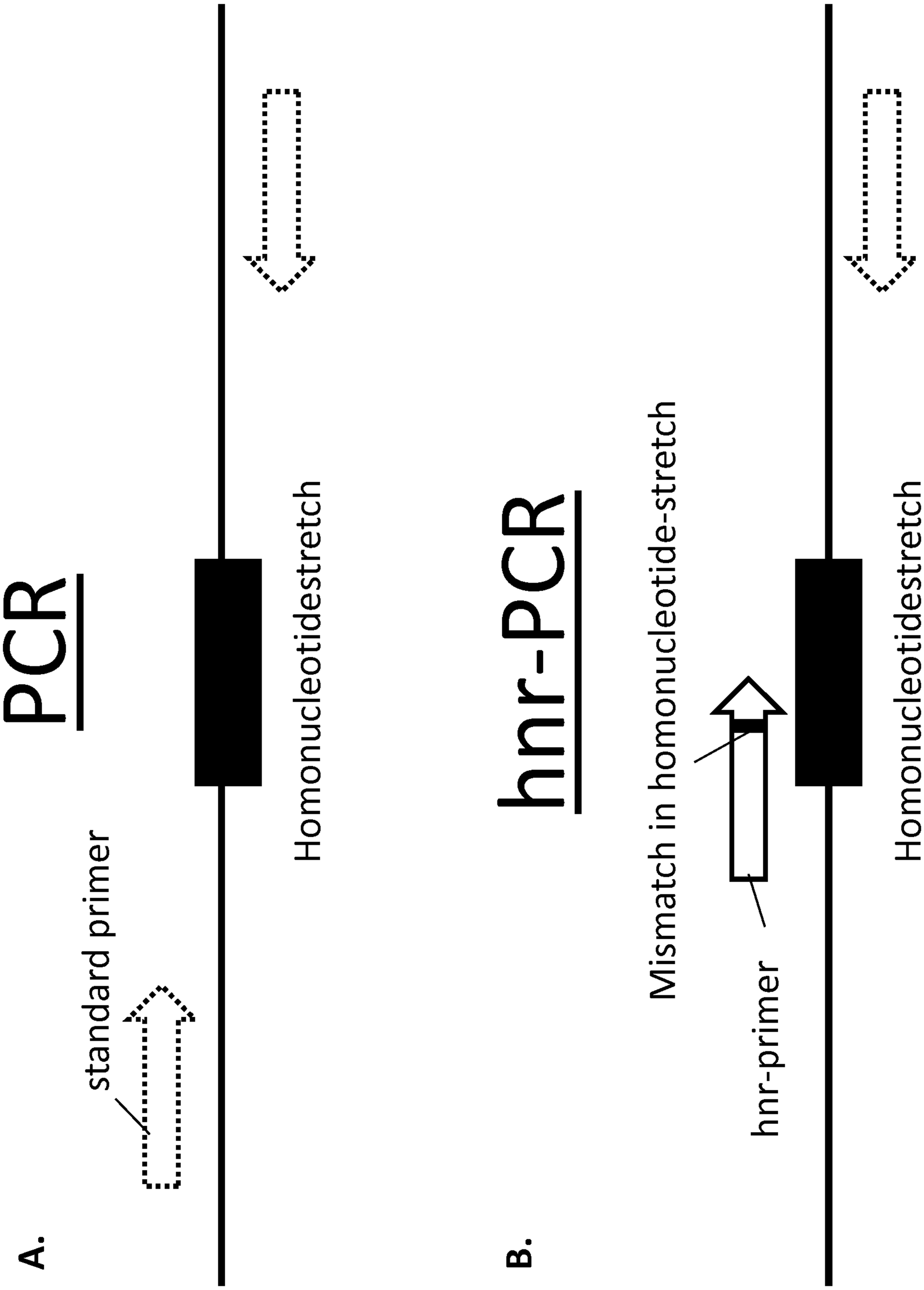
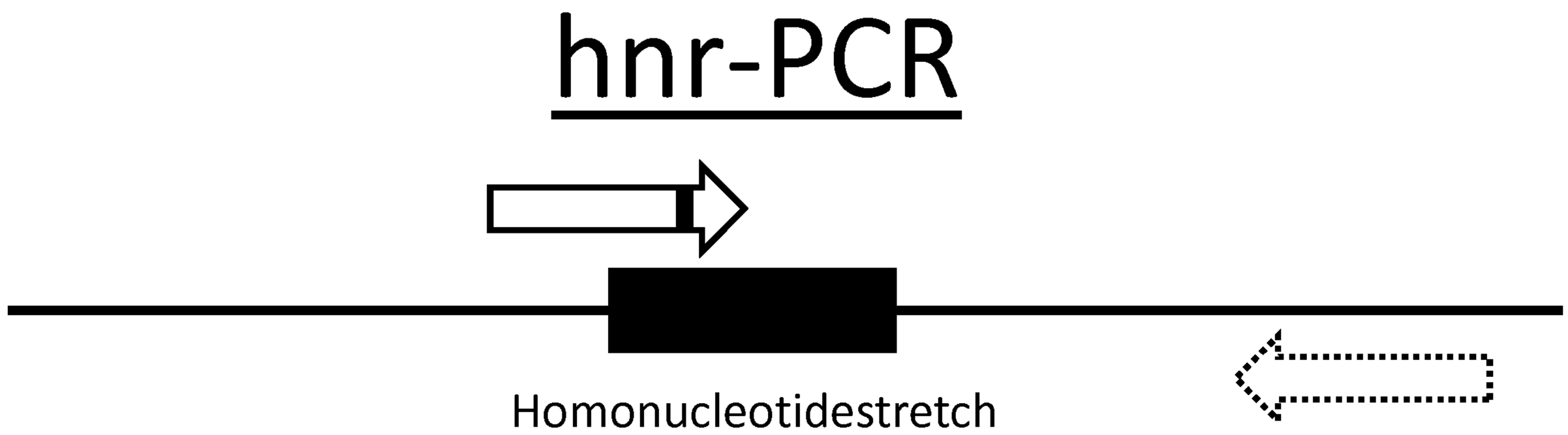

A.
PCR
standard primer
Homonucleotidestretch
B.
hnr-PCR
Mismatch in homonucleotide-stretch
hnr-primer
Homonucleotidestretch

Figure 3

A.


hnr-PCR
Homonucleotidestretch

B.

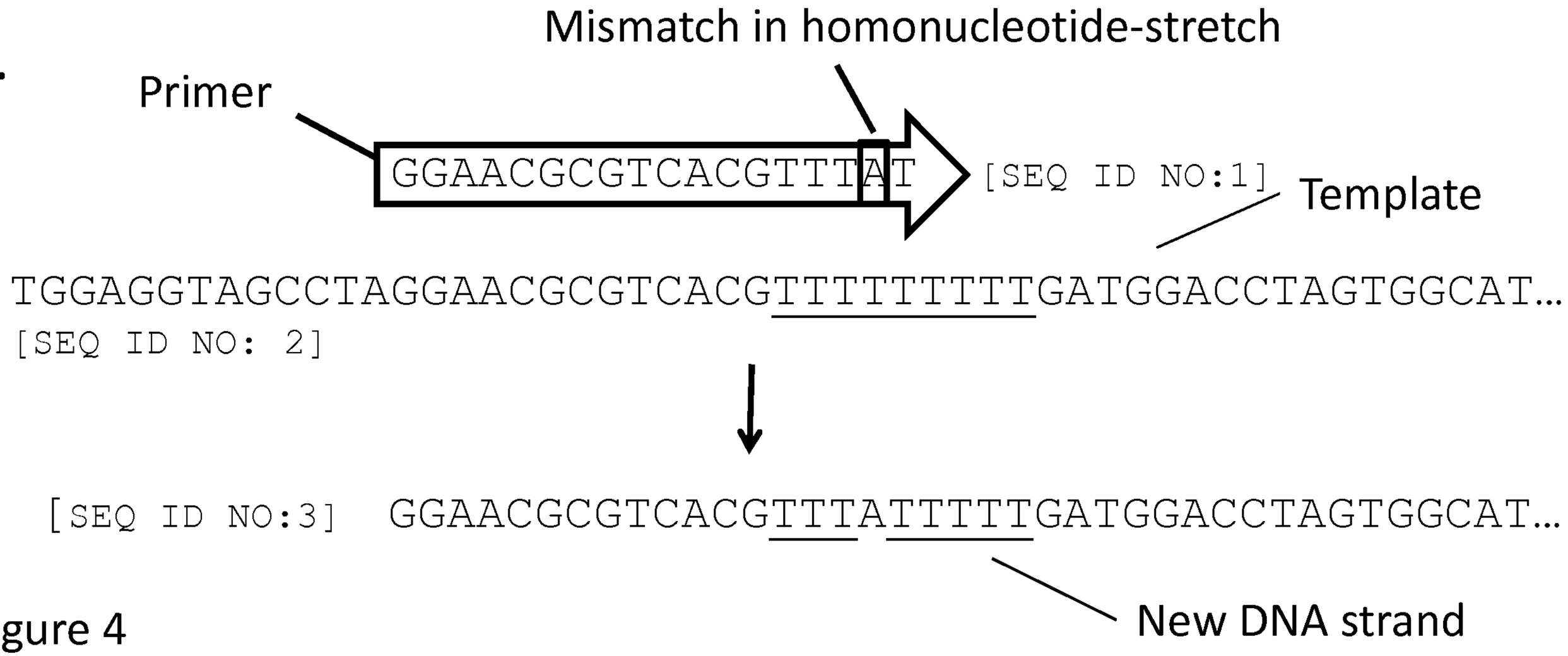

Mismatch in homonucleotide-stretch
Primer
GGAACGCGTCACGTTTAT
[SEQ ID NO:1]
Template
TGGAGGTAGCCTAGGAACGCGTCACGTTTTTTTTTGATGGACCTAGTGGCAT…
[SEQ ID NO: 2]
[SEQ ID NO:3] GGAACGCGTCACGTTTATTTTTGATGGACCTAGTGGCAT…
New DNA strand

Figure 4

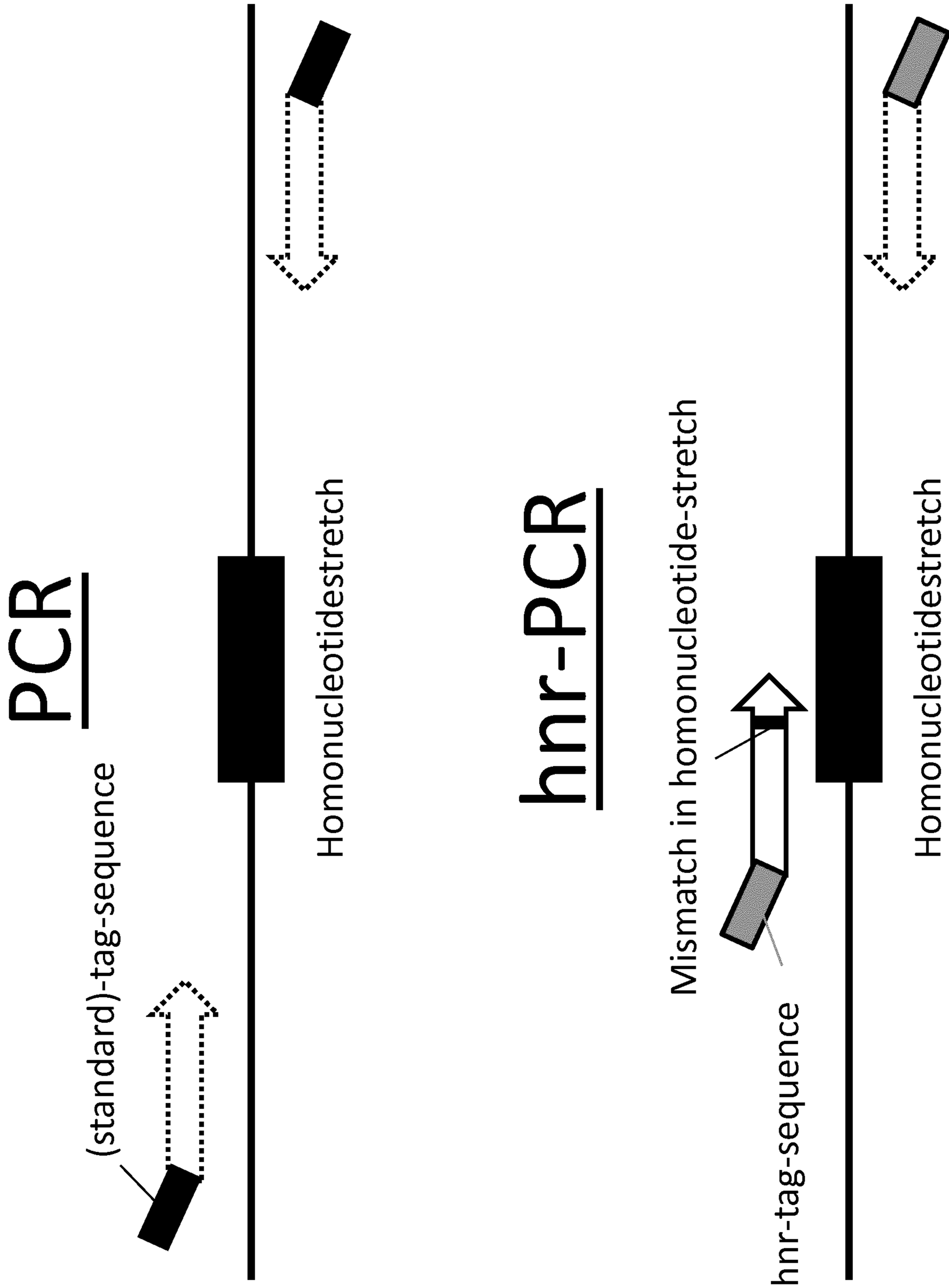

PCR
Homonucleotidestretch
(standard)-tag-sequence
hnr-PCR
Homonucleotide-stretch
Mismatch in homonucleotide-stretch
hnr-tag-sequence

Figure 5

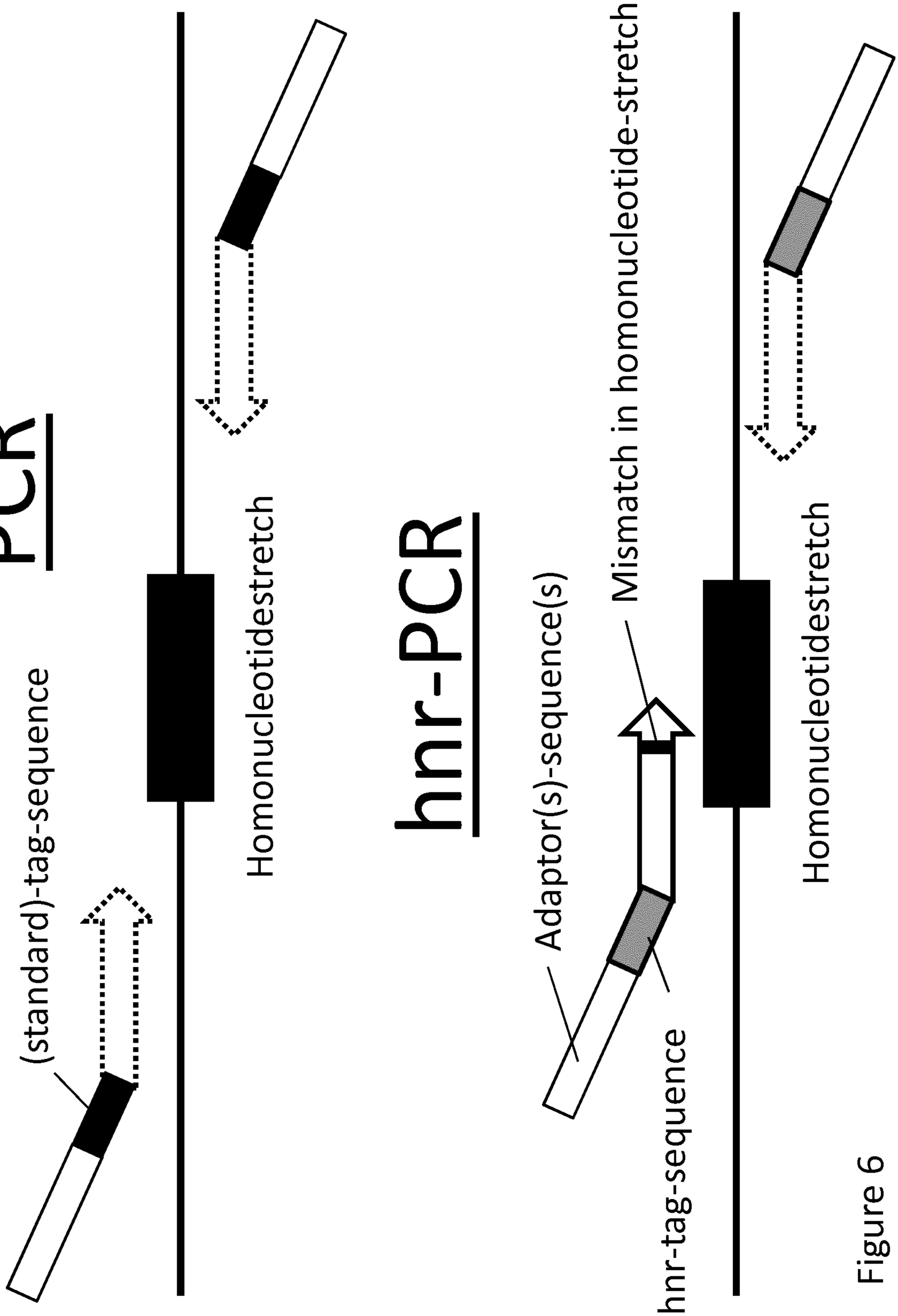
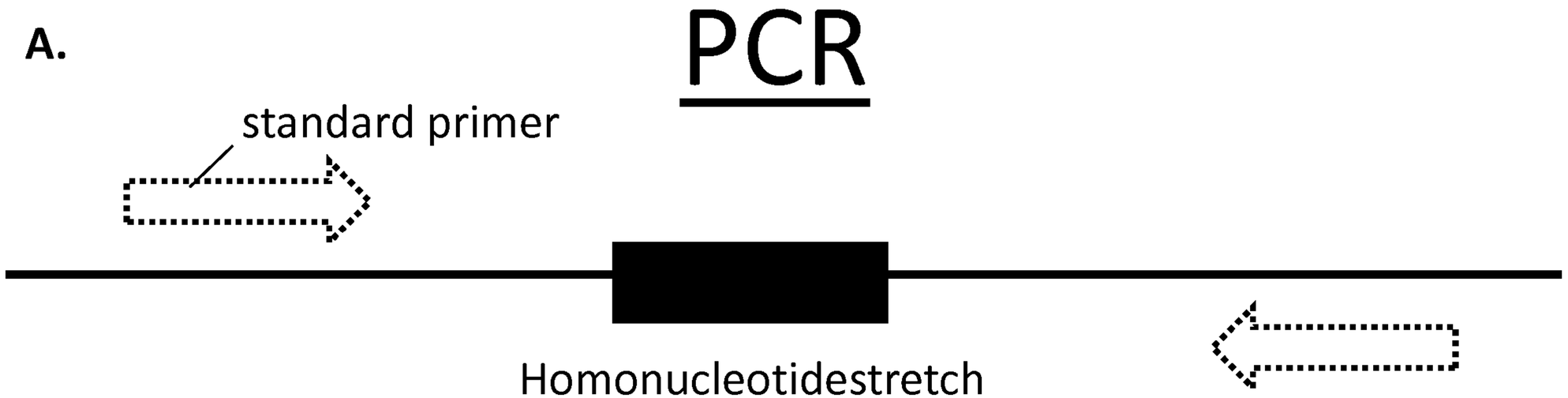
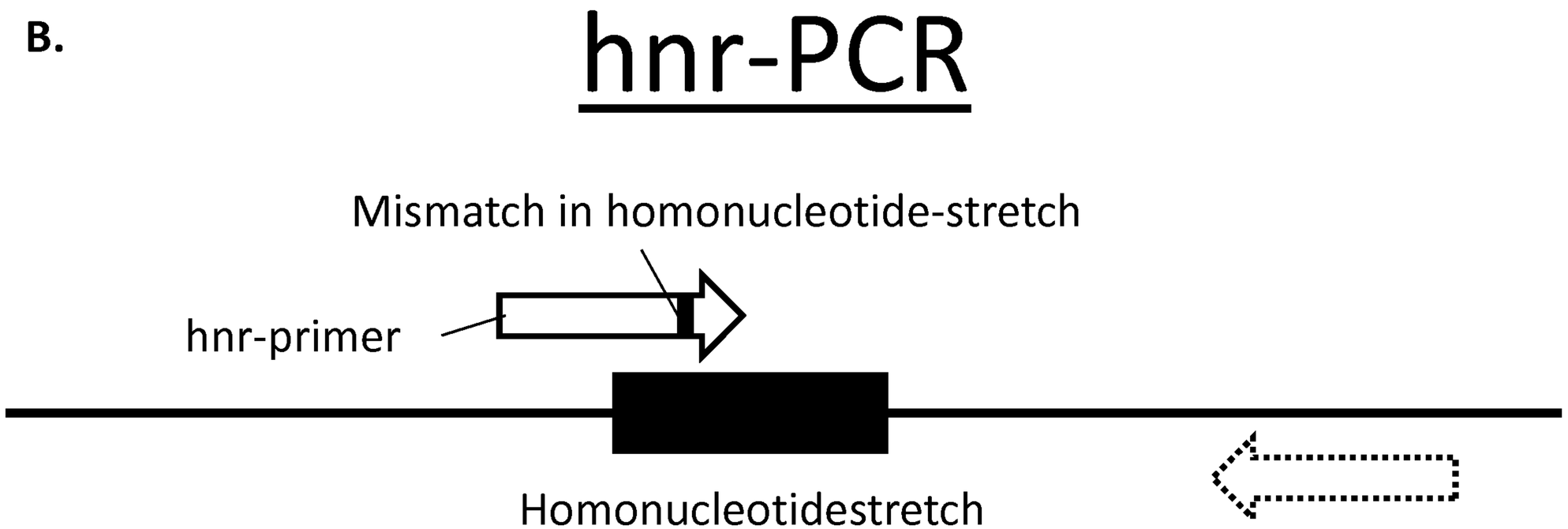

PCR
(standard)-tag-sequence
Homonucleotidestretch
hnr-PCR
Adaptor(s)-sequence(s)
Mismatch in homonucleotidestretch
hnr-tag-sequence
Homonucleotidestretch

Figure 6

Figure 3


A.
PCR
standard primer
Homonucleotidestretch


B.
hnr-PCR
Mismatch in homonucleotide-stretch
hnr-primer
Homonucleotidestretch