# 1

# JUST ENOUGH KNOWLEDGE…

Agnieszka Kowalczyk

*Formerly of Roche, Nutley, NJ, USA*



## 1.1 INTRODUCTION

At the heart of DNA-Encoded Library (DEL) technology lies Deoxyribonucleic Acid (DNA), a molecule that encodes genetic information in all living organisms. It provides a set of instructions, a blueprint for the development and functioning of an organism. In scientific literature, DNA is too often treated as a schematic, two-ribbon spiral. This model does not convey sufficient information for chemists involved in the production of a DEL. The intent of this chapter is to provide "just enough knowledge" about DNA structure, composition, characteristics, and chemical as well as enzymatic operations so that practitioners may fully embrace DEL technology. Whereas a highly detailed and referenced discussion of these topics is beyond the scope of this chapter, highlighting a few key, basic concepts should assist newcomers to this field. Those wishing for in-depth discussion are advised to search the plethora of textbooks, handbooks, and online materials. For those readers for whom this chapter may seem simplistic and too basic, they are urged to proceed to Chapter 2.

*A Handbook for DNA-Encoded Chemistry: Theory and Applications for Exploring Chemical Space and Drug Discovery*, First Edition. Edited by Robert A. Goodnow, Jr.

## 1.2 DNA STRUCTURE

DNA was first isolated in 1869 by Friedrich Miescher [1], but its tertiary structure eluded scientists for almost a century. In 1953, James Watson and Francis Crick [2] proposed that DNA exists as a double helix. This discovery led to a period of rapid advances in biology, greatly increasing our knowledge of life processes at the molecular level. The key to understanding how hereditary information is encoded and why DNA is uniquely suited for storage of genetic instructions lies in its composition and structure.

DNA is a linear polymer built from monomers called nucleotides. Synthetic DNA molecules, usually containing fewer than 200 nucleotides, are known as oligonucleotides. Often, oligonucleotides are described in terms of "mers," referring to the number of nucleotides within the oligonucleotide. For example, 12-mer will contain 12 nucleotides in its structure. Nucleotides found in DNA are made of three components: a 2-deoxy-D-ribose unit, a nitrogenous base that is connected to the sugar molecule via a glycosidic bond, and a phosphate group (Fig. 1.1). Whereas a nucleotide is phosphorylated at the 5′-hydroxyl group, a nucleoside has a free 5′-hydroxyl.

All natural DNA nucleosides have β-configuration at the anomeric carbon. The four commonly occurring nitrogenous bases in DNA are adenine, guanine, cytosine, and thymine (Fig. 1.2). One-letter abbreviations A, G, C, and T are commonly used to denote these moieties. To avoid confusion with the numbering of atoms within a nucleotide or a nucleoside, the following convention has been adopted: carbon atoms of a

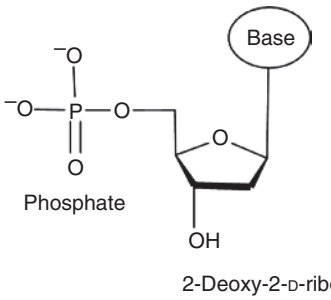


Figure 1.1. Generic structure of a 2-deoxy-2-D-ribose nucleotide.

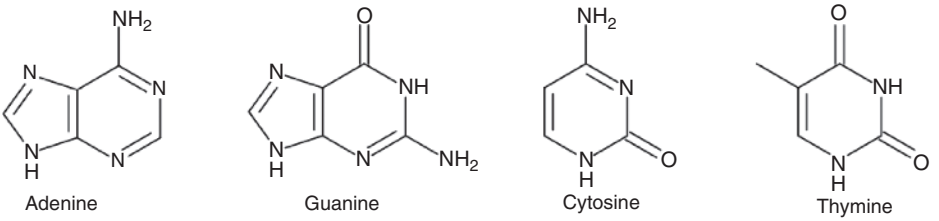


Figure 1.2. DNA bases.

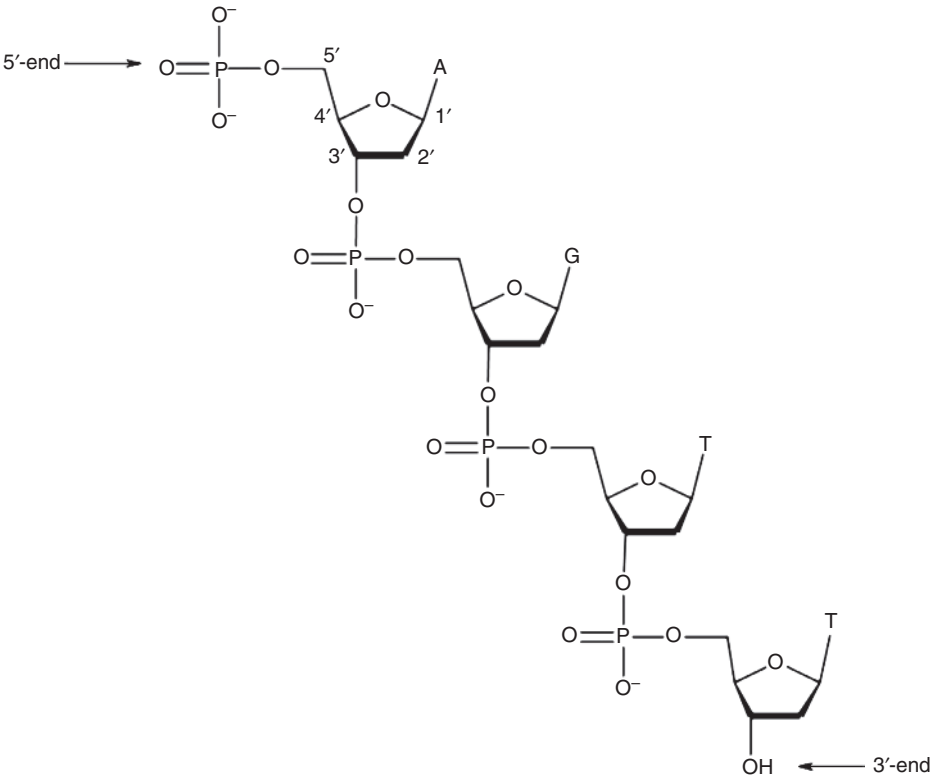


Figure 1.3. Structure of AGTT.

2-deoxy-D-ribose molecule are numbered with prime, 1′, 2′, 3′, etc., while numbers without prime notation are used for atoms in bases. Phosphodiester bonds link 3′- and 5′-hydroxyl groups of neighboring 2-deoxy-D-ribose molecules, forming a DNA backbone. Due to this architecture, a DNA molecule has so-called 5′- and 3′-ends and possesses a defined polarity or direction. The sequence of bases in DNA is conventionally written from left to right starting at the base at the 5′-end and continuing to 3′-end. For simplicity, the prefixes 5′- and 3′- are sometimes dropped. This concept is shown in Figure 1.3. The specific sequence of the bases in DNA, or the order in which they are connected, encodes genetic information.

DNA forms a double helix, meaning that two polynucleotide chains are twisted together around an axis, forming a double-helical structure. These two chains, also called strands, run antiparallel to each other: one strand running from 5′- to 3′-end and the other one from 3′- to 5′-end. The bases, due to their hydrophobicity, are stacked inside a helix and perpendicular to its axis, while the backbone that contains alternating 2-deoxy-D-ribose and phosphate moieties is located on the outside of the helix. This spatial arrangement makes the bases hard to access and in this way protects them from undesired interactions that could potentially change the genetic instructions they

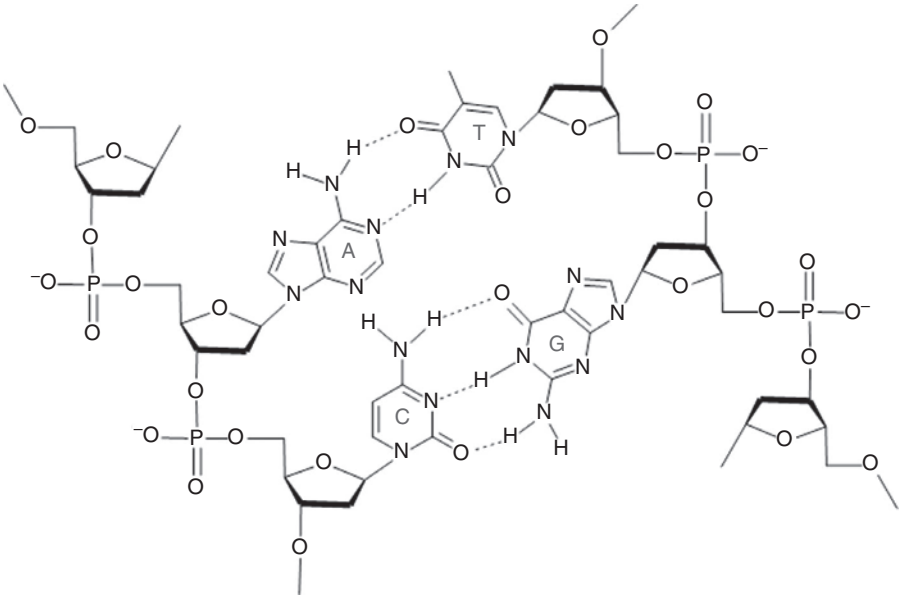


Figure 1.4. Complementary base pairing in DNA.

encode. The formation of a double helix is driven by hydrogen bonding between the bases of opposite strands and van der Waals interactions arising from stacking of the bases. Hydrogen bonding between bases occurs in a specific manner. Adenine forms a base pair with thymine through two H-bonds, while guanine forms a base pair with cytosine through three H-bonds (Fig. 1.4). This means that the opposite strands in a DNA helix are complementary; the sequence of one strand can be used to determine the sequence of the other strand. A DNA molecule that has a double-helical structure is often called a DNA duplex.

Depending on several factors, such as the level of hydration, the salt concentration, and the sequence of bases, the double helix can adopt different conformations while retaining the general spiral-like structure with two antiparallel strands. There are three major conformations of DNA helices: A-, B-, and Z-forms, each form having distinct geometrical features [3]. Both A- and B-forms are right-handed helices, with the anti-conformation around glycosidic bonds. These forms differ in the type of sugar pucker with the A-form adopting a C3′-endo and B-form adopting a C2′-endo conformation (Fig. 1.5). Consequently, the A-form helix is wider and shorter compared to the B-form. The B-form is the most common conformation found under physiological conditions, at low salt concentration and high water content, while the A-form is favored under dehydrated conditions. Z-DNA is strikingly different; it only occurs in DNA with alternating pyrimidine and purine sequences and exists as a left-handed helix. The subtle structural nuances of DNA conformations play an important role in DNA recognition by proteins and other molecules. In DEL chemistry, the most likely form of DNA a chemist will encounter is B-conformation.

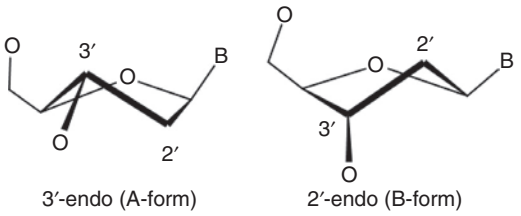


Figure 1.5. Types of sugar pucker found in DNA.

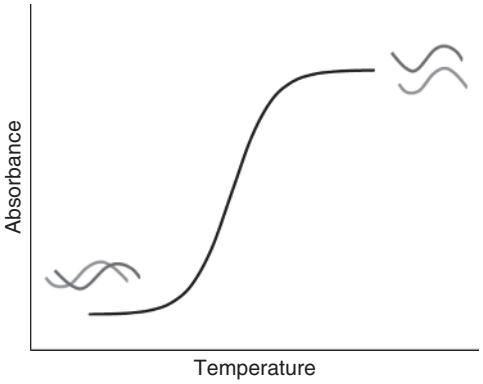


Figure 1.6. DNA melting curve.

## 1.3 DNA DENATURATION

Due to different physical factors such as temperature, salt concentration, the presence of organic solvents, the presence of chaotropic agents, and pH, the base interactions holding the DNA helix together can be destabilized, resulting in a separation of the duplex into single strands in a process called denaturation. Thermal denaturation, or melting, can be easily followed by measuring absorbance at 260 nm while slowly increasing the temperature of a DNA solution (Fig. 1.6). The absorbance of denatured DNA is higher than that of the corresponding duplex. This phenomenon is known as a hyperchromic shift. The temperature at which half of the DNA molecules exist as a duplex and half in a single-stranded form is referred to as the melting temperature (Tm). The Tm is a measure of duplex stability and depends on both the composition and sequence of the DNA. For example, the Tm of a 10-mer DNA duplex can range from 20°C to 40°C. High GC content stabilizes the duplex, thus increasing the Tm. The sequence also plays a role because the base stacking interactions depend on the neighboring base pairs and some combinations are more energetically favorable than others. Base mismatches, such as those caused by errors during the DNA replication process, destabilize the duplex and lead to local melting, providing a recognition mechanism for DNA repair enzymes. Under appropriate conditions, the single strands of denatured DNA will hybridize with their complementary strands to recreate the double helix in a process called hybridization.

## 1.4 DNA REPLICATION

For genetic material to be passed on to the next generation, DNA must be copied during cell division. This process is known as DNA replication and involves the complex interplay of several enzymes. Some of these enzymes are utilized by DEL technology; we will look at them and the processes they catalyze in more detail. During replication, the parent DNA molecule is unwound into two single strands that serve as templates and guide the synthesis of two brand-new strands, resulting in the formation of two daughter molecules identical to the parent molecule. Each daughter molecule contains one original parent strand and one newly synthesized strand. This type of replication is known as semiconservative.

The replication is initiated by helicase that unzips the two strands of the double helix creating the replication fork. Proteins known as single-stranded DNA-binding proteins bind to freshly unwound single strands, preventing them from annealing and protecting them from digestion by nucleases. DNA polymerases polymerize nucleotide triphosphates complementary to the template strands. When the polymerase "sees" guanine on the parent strand, it will add cytosine nucleotide to the new strand in the complementary position, and in the case of adenine being on the parent strand, it will add a thymine nucleotide. Different template-dependent DNA polymerases exhibit a range of accuracies with the highest fidelity demonstrating error rates as low as 1 in 10 million [4]. Template-dependent DNA polymerases can only generate DNA chains by adding complementary nucleotides to the free 3′-hydroxyl end of an annealed complementary strand—the primer. Consequently, a new strand is synthesized in the 5′ to 3′ direction. This poses a problem with replication of a duplex because two unwound template strands from the duplex are running antiparallel to each other but both of them can only be synthesized in the 5′ to 3′ direction. One strand, called a leading strand, can be synthesized continuously because its polarity is consistent with the direction of duplex unwinding and polymerase action. It only needs one primer, a short piece of RNA to form an initial duplex that will be continuously extended. On the other hand, the second strand, known as a lagging strand, is synthesized in multiple fragments, called Okazaki fragments, which are later conjoined by a ligase. The lagging strand requires many primers, one for each Okazaki fragment. This process is shown schematically in Figure 1.7.

Two of the classes of enzymes involved in DNA replication, DNA polymerase [5, 6] and DNA ligase [7, 8], play important roles in DEL technology. A DNA polymerase is utilized in the amplification of selected sequences, and a ligase is often used to join encoding tags.

There are several types of DNA polymerases involved in the complex biological processes of DNA replication and repair. The structure of the catalytic unit is highly conserved between different polymerases, indicating that the process they carry out is extremely ancient. All known polymerases catalyze the same reaction—elongation of the DNA chain by addition of a nucleotide to the free 3′-hydroxyl end of the existing chain. Template-dependent DNA polymerases cannot synthesize a new chain de novo; they require both a template strand and a primer for their function.

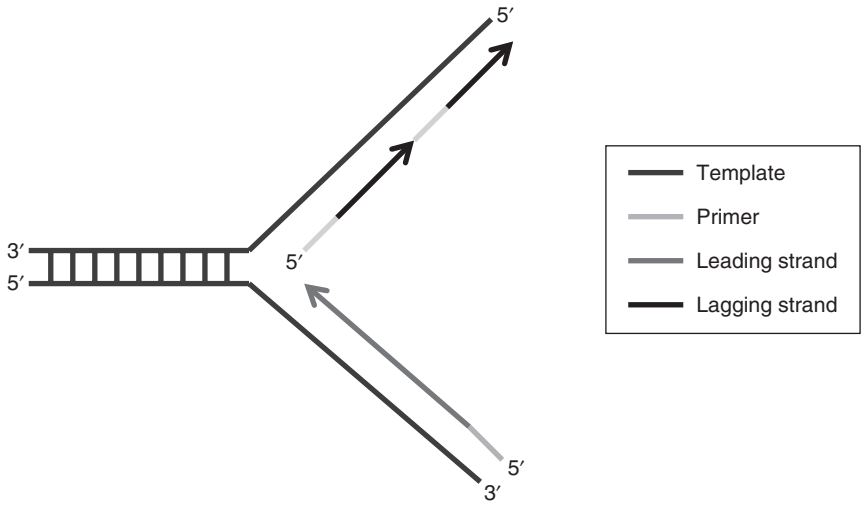


Figure 1.7. DNA replication—synthesis of a leading and a lagging strand.

Some polymerases have a proofreading ability, meaning they are able to detect the incorrectly added nucleotide and replace it with a correct one. A mismatched base pair destabilizes the duplex, causing local melting, and consequently provides a mechanism for its detection. When such a mismatched base pair is found, the polymerase reverses its slide along a template strand and excises the incorrect nucleotide—then, it adds the proper nucleotide as directed by the template strand and resumes its action of chain elongation. This function of a polymerase is also known as 3′ to 5′ exonuclease activity, and it explains the high fidelity that may be achieved by DNA replication.

For the purpose of *in vitro* DNA amplification, a variety of thermostable template-dependent polymerases are utilized. The most well-known example is Taq polymerase [9], isolated from the bacterium *Thermus aquaticus* found in thermal springs. Taq polymerase has optimal activity between 75°C and 80°C [10]. Due to their heat resistance, Taq polymerase and other thermostable DNA polymerases are widely used in the Polymerase Chain Reaction, abbreviated PCR, a molecular biology technique employed in DNA amplification. PCR methodology allows DNA to be rapidly copied *in vitro* from a single or few DNA fragments many million times. PCR utilizes the enzymatic replication of DNA and therefore requires a polymerase to assemble the new strands, primers to initiate the replication, and free nucleotides to serve as building blocks. PCR is performed in a thermocycler, a programmable apparatus capable of incubating at defined temperatures between 4°C and 100°C and of rapidly transitioning between these temperatures at defined rates. At high temperature, the DNA is denatured. Subsequent cooling allows primers to anneal to the freshly separated DNA strands. Primers are used in excess compared to the sequence being amplified so that the DNA strands will anneal with primers rather

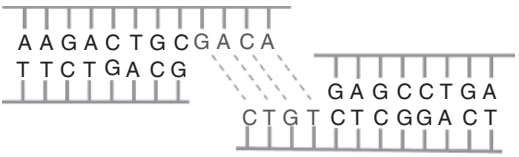


Figure 1.8. Cohesive ends.

than with themselves. The next step is the synthesis of new strands by the thermostable polymerase. This sequence of events is repeated a defined number of times. Since newly copied DNA fragments formed in one cycle serve as templates in the next cycle, the amplification process is exponential. It is understandable now why the use of thermostable polymerases, such as Taq polymerase, is very advantageous in the PCR process. Thermostable enzymes are able to survive repeated exposure to elevated temperatures (typically 94°C) and consequently do not require new addition of polymerase with each cycle.

Ligases are a class of enzymes that covalently join, or ligate, two DNA strands together. For example, T4 DNA ligase catalyzes the formation of a phosphodiester bond between a 5′-phosphate and an adjacent 3′-hydroxyl group of a nicked strand in a duplex. Ligases are involved in both DNA replication and DNA repair processes. To be competent for a ligation, the oligonucleotides must be monophosphorylated at their respective 5′-ends and have free 3′-hydroxyl ends. Double-stranded oligonucleotides with either cohesive or blunt ends can be ligated; however, the latter usually requires much higher ligase concentration. Cohesive ends are overhangs on each oligonucleotide made of unpaired nucleotides that are complementary to each other; thus, they can anneal and hold the two DNA fragments to be ligated together (Fig. 1.8). The term blunt ends means that there are no overhangs and the duplex ends in a complementary base pair. The use of cohesive, or "sticky," ends is preferred because it is more efficient and ensures that the ligation proceeds only in one orientation determined by complementarity of the overhangs. There are two major ligase families, NAD+ and ATP dependent, indicating the cofactor needed for their action. NAD+-dependent ligases are found only in bacteria, while eukaryotes and bacteriophages require ATP.

T4 DNA ligase, isolated from T4 bacteriophage, has been extensively used in many molecular biology applications such as cloning and DEL technology. This ligase operates best in the pH range from 7.5 to 8. Cohesive-end ligations are typically performed at room temperature or below to stabilize the transiently annealed oligonucleotide junctions, although the optimum temperature for the ligase itself is higher. T4 DNA ligase utilizes ATP as a cofactor. The first step in the ligation process involves adenylation of the amino group of a lysine residue at the active site of the ligase with concomitant release of pyrophosphate. Next, AMP is transferred from the ligase to DNA, specifically to the 5′-phosphate group of one of the strands to be ligated. The resulting pyrophosphate is attacked by the 3′-OH group of the other strand, creating a phosphodiester bond and linking two strands covalently together. These steps are shown in Figure 1.9.

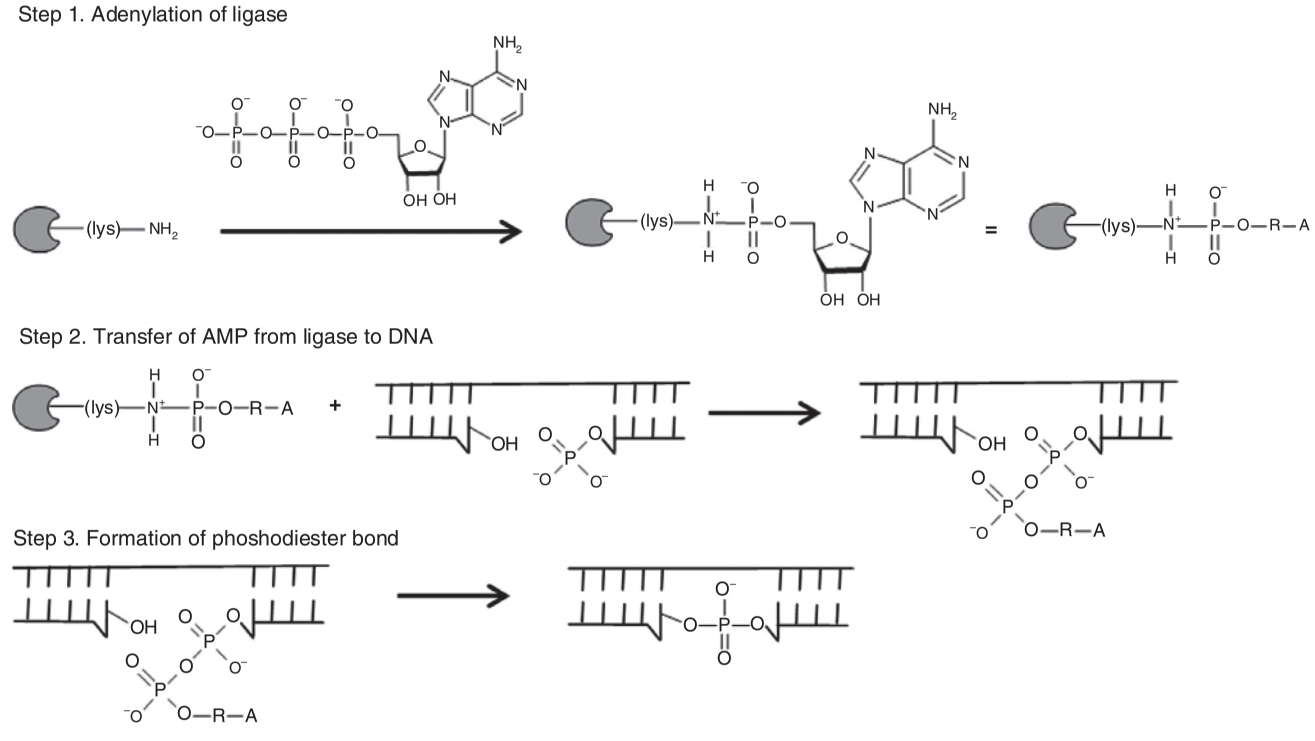


Figure 1.9. Mechanism of action of the ATP-dependent ligase.

## 1.5 CHEMICAL SYNTHESIS OF DNA

DNA oligonucleotides can be routinely synthesized chemically using solid-phase methodology [11]. Usually, the encoding tags used in DEL technology are prepared in this way. DNA synthesis proceeds by sequential addition of nucleoside building blocks to a growing oligonucleotide covalently attached to solid support. This process is fully automated in commercially available apparatuses called DNA synthesizers that have now been available for more than 30 years.

The coupling of nucleoside building blocks is based on phosphoramidite chemistry. Briefly, nucleoside phosphoramidites, which are used as nucleotide equivalents, have the following features: the 5′-hydroxyl moiety is protected with a dimethoxytrityl (DMT) group and the 3′-OH is derivatized as a 2-cyanoethyl *N*, *N*-diisopropyl phosphoramidite (Fig. 1.10). In addition, the protection of the exocyclic amino functionalities of guanine, adenine, and cytosine is required to avoid side reactions and to improve the solubility of the nucleoside building blocks. For this purpose, an acylation reaction is commonly employed; the N-6 amino group of adenine and the N-4 amino group of cytosine are usually blocked with benzoyl groups, while the N-2 position of guanine is protected with an isobutyryl group. The choice of protecting groups is dictated by the chemistry employed in the synthetic process. Different functionalities must be temporarily blocked and the blocking groups later easily removed at various stages of the oligonucleotide synthesis. This orthogonal protection must be compatible with all reagents and conditions used in the process.

The entire synthesis is implemented as a series of repeated cycles; in each cycle, one nucleoside is added to the oligonucleotide growing on a solid support. The cycle consists of four distinct steps: (i) detritylation, (ii) coupling, (iii) capping, and (iv) oxidation

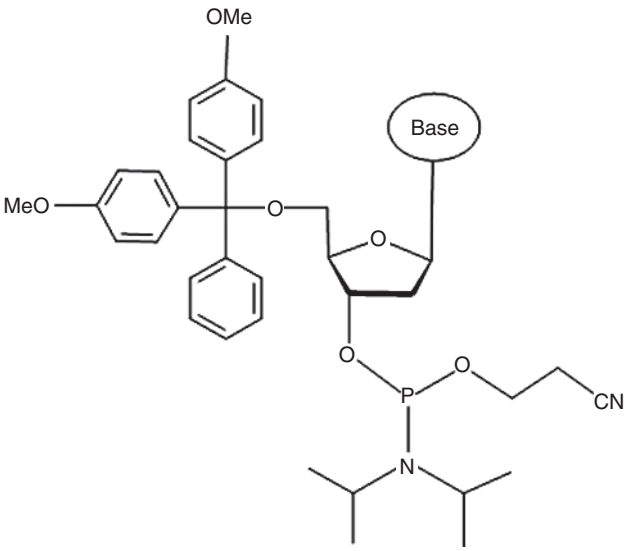


Figure 1.10. Building block for DNA synthesis.

(Fig. 1.11). The very first nucleoside is already attached to the solid support via its 3′-OH group, and after detritylation, its 5′-hydroxyl group will be able to react with the next nucleoside. This design implies that the oligonucleotide is synthesized from the 3′- to the 5′-end, which is opposite to the direction of DNA assembly by a polymerase. The synthesis cycle starts with the removal of the acid-labile DMT group to expose a 5′-hydroxyl group for coupling with the next building block. The DMT cation has a bright orange color in solution. The measurement of its absorbance serves as an indicator of coupling efficiency because the DMT group must be removed from the last nucleoside incorporated into the oligonucleotide prior to coupling with the next nucleoside. The incoming phosphoramidite nucleoside is activated by tetrazole, enabling the formation of an unstable phosphite bond between two nucleosides. Even though the coupling efficiency is usually very high (over 98%), a small amount of uncoupled material may persist. This uncoupled material could potentially react further in subsequent steps forming a sequence that lacks one base and is difficult to separate from the full-length product. To avoid this undesired process, the capping step is implemented immediately after coupling. During capping, all remaining free 5′-hydroxyl groups are acetylated, thus preventing them from

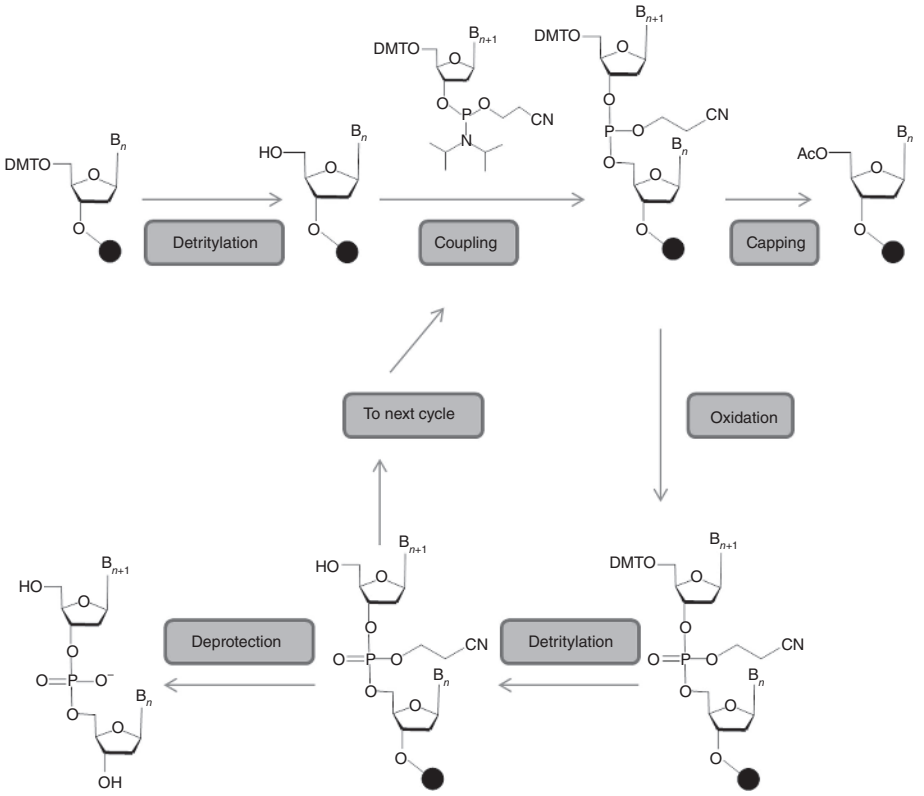


Figure 1.11. Solid-phase oligonucleotide synthesis by phosphoramidite method.

any further reaction. This results in failure, or truncated, sequences that are sufficiently different from the full-length sequence to be easily removed after the synthesis. The phosphite triester that links two newly coupled nucleosides is unstable and following the capping step is oxidized to a more stable phosphate triester. A commonly used oxidizing system is a mixture of iodine and pyridine in water and THF. After all nucleosides have been added, the product needs to be cleaved from the solid support and further processed to remove both the protecting groups from the bases and 2-cyanoethyl groups from the phosphate triesters to yield a functional oligonucleotide. This is achieved by heating a solid-support-bound product in concentrated ammonium hydroxide at 55°C for 1 h. The crude product is then desalted and, if needed, purified further using reversed phase or ion exchange HPLC.

## 1.6 OLIGONUCLEOTIDE CHARACTERIZATION

Making a DEL involves steps that require assessing the identity and/or purity of oligonucleotides. Quality control must be performed after ligation and addition of a chemical building block. Two analytical tools routinely used for these purposes are electrophoresis and Mass Spectrometry (MS).

Electrophoresis [12, 13] allows for the separation of DNA fragments based on their size. DNA molecules have a net negative charge due to the negatively charged phosphate groups of the backbone. This charge enables them to migrate through an inert gel matrix when an electric field is applied across the gel. The rate of migration of charged molecules in electrophoresis is known as electrophoretic mobility. Molecules differing in size move through the pores of the gel with different rates. Smaller DNA fragments migrate faster than larger ones forming a distinct band pattern on the gel. These bands correspond to different lengths of DNA fragments. The gel can be "calibrated" by running a mixture of molecular weight size markers (DNA fragments of known lengths) along with a sample of unknown DNA to estimate its size. Two matrix materials commonly employed in gel electrophoresis are polyacrylamide and agarose. Polyacrylamide is a synthetic polymer prepared from acrylamide monomer and the cross-linking agent *N*,*N*′-methylenebisacrylamide. The relative amounts of these two reagents determine the porosity of the gel and can be optimized to obtain the best conditions for a specific separation. When high concentrations of a chaotropic agent such as urea are present in a polyacrylamide gel, hydrogen bonds are destabilized and single-stranded oligonucleotides may be characterized at high resolution; oligonucleotides differing in size by only one base pair can be resolved. Such gels are referred to as denaturing. This high resolving power is offset by a low range of polyacrylamide separations, up to a couple of 1000 bp long. Gels made from agarose, a natural polysaccharide isolated from seaweed, have a large pore size and thus are well suited for the separation of much larger DNA fragments than polyacrylamide gels, but their resolution is limited. Agarose gels cannot be made denaturing and are generally only used with double-stranded DNA. Agarose gels can be prepared and poured in the lab prior to use. Alternatively, precast gels, both denaturing and native polyacrylamide and agarose, can be purchased from commercial vendors.

DNA on a gel is not visible to the human eye, and thus, it needs to be stained or otherwise rendered visible in order to be detected. Ethidium bromide is widely used for the visualization of double-stranded DNA. It works by intercalating into the DNA duplex and fluorescing under UV light, revealing the localization of the DNA bands on a gel. Similar fluorescent dyes are also available for single-stranded nucleic acids. In general, dyes can be incorporated into either gel matrix prior to the separation or alternatively a gel can be stained with a dye after the separation. The latter method is employed in case of polyacrylamide gels since dyes can interfere with the polymerization reaction.

In the DEL process with double-stranded DNA, agarose gel electrophoresis is used to monitor the efficiency of ligation of the tags. Tag ligation increases the molecular weight of the DNA fragment, and therefore, observed retardation of the ligated product on the gel indicates that the ligation was successful.

The second indispensable analytical tool for a DEL chemist is an MS system suited for oligonucleotide analysis. In most cases, the mass spectrum of an oligonucleotide provides sufficient information about its identity. Two MS methods have been widely used in oligonucleotide analysis: Electrospray Ionization (ESI) [14–16] and Matrix-Assisted Laser Desorption/Ionization–Time Of Flight (MALDI-TOF) [17–19]. MALDI-TOF is slightly more sensitive than ESI for shorter oligos (<50 bp). It produces singly charged species, simplifying the interpretation of the spectrum. Ionization efficiency drops significantly for oligonucleotides longer than 50 bases in MALDI, limiting its use to the shorter sequences. On the other hand, the ESI technique delivers good accuracy and sensitivity over a wide range of oligonucleotide length/mass. In the process of ESI, multiply charged species of a parent molecule are formed, necessitating the use of special algorithms to deconvolute the spectra. Deconvolution allows for the reconstruction of the molecular weight of a parent molecule from the $m/z$ values of its charged ions. It is especially useful for analyzing mixtures of oligonucleotides. Automated deconvolution software packages for processing ESI data are commercially available. The ESI system can be easily coupled with Liquid Chromatography (LC) to provide both sample separation and sample identification.

An LC/MS system optimized for oligonucleotide analysis is an absolutely necessary tool for a DEL chemist. It allows for monitoring chemical reactions on DNA during library development and synthesis. A more detailed discussion of DNA analytical methods is found in Chapter 8.

## 1.7 DNA SEQUENCING

DNA sequencing makes it possible to determine the exact order of nucleotides within a fragment of DNA. The development of DNA sequencing methodologies was spurred by the need to decipher the functions of single genes and, ultimately, of the entire genome. Eventually, DNA sequencing enabled new applications in diagnostics and forensics as well as various novel biotechnologies such as DEL.

One of the first DNA sequencing methods was developed by Maxam and Gilbert in 1977 [20]. It relies on a series of four carefully selected and performed chemical

processes that cleave DNA at specific bases. One of the 5′-ends of the DNA is labeled, usually with radioactive phosphorus ($^{32}$P). Reaction conditions are optimized in such a way that on average one cleavage occurs per one DNA molecule. It ensures that all possible fragments that include a 5′-labeled end will be represented. Each reaction is run in a separate vessel and generates DNA fragments of varying length. These four sets of fragments are then separated side by side by polyacrylamide gel electrophoresis with a resolution of a single nucleotide, and the sequence is deduced from the pattern on the gel. Maxam–Gilbert methodology is labor-intensive and complex. It was soon replaced by a conceptually very different method established by Sanger [21]. Sanger sequencing, also called the chain-termination method, utilizes an *in vitro* DNA replication process with a small but profound modification. Along with natural 2′-deoxyribonucleotides, their 2′,3′-dideoxy analogs are used. These dideoxy analogs are devoid of a 3′-hydroxyl group, and thus, they can be added to the growing DNA chain via their 5′-ends but cannot form a phosphodiester bond with an incoming nucleotide. Consequently, when a dideoxy analog is incorporated, chain elongation stops generating a terminated DNA fragment. This process is carried out separately for each of the chain-terminating nucleotides. The new chain starts growing from the labeled primer, and therefore, all terminated fragments will be also labeled, enabling their detection. The DNA sequence can be then easily read from a gel electropherogram of the terminated DNA fragments.

Sanger methodology quickly became the method of choice for DNA sequencing. It was simple and amenable to automation. The use of dye terminators, dideoxy analogs labeled with different fluorescent dyes, further simplified the process. With this improvement, the sequencing process could be carried out in one vessel instead of four, since the last nucleotide of the terminated fragment is easily identified by its unique wavelength after excitation. Gel electrophoresis was replaced with capillary gel electrophoresis. Eventually, the DNA sequencing process became fully automated with first-generation sequencers based on the Sanger chain-termination method.

More recently, these sequencing processes have been adapted such that *in vitro* DNA cloning and amplification can be performed relatively simply on a massive scale. Such technologies are known as “next-generation” or “deep” sequencing [22–24]. This allows for parallel sequencing of DNA libraries at high speed and low cost. There are several sequencing platforms commercially available. Despite employing diverse processes and techniques, most next-generation methodologies share a common strategy. DNA fragments are amplified into clusters with each cluster arising from a single DNA molecule. These clusters, made of identical DNA fragments, are sequenced by synthesis. Sequencing by synthesis relies on detecting either labeled nucleotides being incorporated into a new DNA strand or by-products of the synthesis such as pyrophosphate. After each nucleotide incorporation cycle, a synthesis is temporarily halted, and a snapshot image that captures all clusters simultaneously is acquired, revealing the sequence as it is being assembled in each cluster.

DEL technology has been greatly facilitated by the adoption of deep sequencing technologies because it permits the detection of relatively weak signals such as are necessarily found in selection experiments with large libraries. The two platforms commonly used in DEL technology are the Illumina/Solexa and 454/Roche systems.

Illumina is a popular next-generation DNA sequencing platform in the field [25]. It relies on PCR bridge amplification [26, 27] and reversible dye terminator [28, 29] methodologies. Before actual sequencing, all DNA fragments must be equipped with adaptors, short oligonucleotides that are added by ligation or PCR to both ends of a DNA fragment. These adaptors allow for the hybridization of the DNA fragments with primers that are covalently attached to a flow cell surface. There are only two different types of primers present in multiple copies on the flow cell. One type of primer has a cleavable site engineered in its sequence. A denatured DNA library is loaded on a flow cell, allowing single strands of DNA fragments to hybridize with the primers. This step ensures the spatial separation of different DNA fragments on the surface. Next, the complementary strands are synthesized enzymatically starting from the primers. This process leaves newly assembled strands immobilized on the surface since they originated from the tethered primers. The complementary strands, arising from the original DNA fragments, are not tethered and thus are washed off after denaturation. Then, the strands are amplified in a process called bridge amplification. It occurs when an immobilized DNA strand hybridizes with a proximal primer on the slide surface forming a bridge. DNA polymerase then extends the chain from the primer, resulting in two DNA strands covalently attached to the flow cell after denaturation. This cycle of chain extension and denaturation is repeated many times, leading to a formation of a DNA cluster. Each cluster contains up to 1000 copies of a single original DNA fragment in close proximity immobilized on a surface. Several millions of spatially separated clusters can be formed on a single flow cell. In order to enable sequencing, a DNA cluster must contain only one type of single-stranded DNA; however, after amplification, both the original strand and the reverse strand are tethered to the surface in multiple copies. The reverse strand is attached via a primer with a cleavable site and therefore can be easily removed. Next, all 3′-ends of immobilized strands are blocked with dideoxynucleotide analogs to prevent potential extension of DNA strands on each other. The actual sequencing starts with annealing of the sequencing primer and extension of the chain using four reversible dye terminator nucleotides. These reversible terminators are labeled with different fluorescent dyes and transiently blocked at their 3′-hydroxyl groups. The reversible terminators are incorporated one at a time. After each incorporation, the image of the entire flow cell is captured, enabling the identification of the last added nucleotide in all clusters simultaneously. This cycle of chain extension and imaging resumes after the removal of a fluorescent label and unblocking of 3′-hydroxyl functionality of the last added nucleotide and continues until the sequence of the defined region is obtained in each DNA cluster. Transient blocking of the 3′-hydroxyl group of reversible terminators ensures that the synthesis is temporarily halted after single nucleotide incorporation, allowing for image acquisition.

The 454 sequencing platform [30] is based on a similar strategy: separation of DNA fragments and their amplification, followed by parallel sequencing by synthesis. The DNA fragments are amplified by emulsion PCR [31] and then sequenced by pyrosequencing method [32]. At low template concentration, emulsion PCR is a clonal amplification method carried out in a water-in-oil emulsion. Prior to amplification, DNA fragments are ligated with two different adaptors. These adaptor-flanked DNA fragments are combined with two primers complementary to the adaptors, microbeads

covered with one of the primers and PCR reagents and then emulsified in oil. In this process, DNA fragments and beads get trapped inside tiny aqueous droplets suspended in oil. The low concentration of DNA fragments ensures that on average only one DNA molecule is encapsulated per droplet, creating separate microreactors for the amplification of each DNA fragment. The DNA molecule inside a droplet is amplified by PCR, coating the beads with multiple copies of itself. Once the emulsion is broken, the complementary strands that are not attached to the beads are denatured and washed off and the sequencing primer is annealed. The beads with amplified DNA fragments are then deposited on a plate with picoliter-sized wells for pyrosequencing. Each well can accommodate only a single bead. Much smaller beads with two immobilized enzymes, namely, ATP sulfurylase and luciferase, are added to all wells, surrounding the bigger beads with clonally amplified DNA. In pyrosequencing, only one of the four nucleotides is added at a time. If the added nucleotide is complementary to the one on the template, it gets incorporated into the growing chain with a concomitant release of a pyrophosphate. The pyrophosphate triggers a series of enzymatic reactions involving ATP sulfurylase and luciferase that result in the emission of a bioluminescence signal. The presence or absence of this signal makes it possible to elucidate the sequence. When using pyrosequencing for sequencing of homopolymer segments, some attention is due. The synthesis is only halted after one nucleotide incorporation if the next nucleotide to be incorporated is of different identity. However, in case of stretches containing the same consecutive nucleotides, for example, AAAAA, the synthesis continues for the entire length of homopolymer segment before it stops. For short nucleotide repeats, up to six nucleotides, there is a direct proportionality between the number of nucleotides incorporated and the intensity of the bioluminescence signal. However, longer nucleotide repeats may result in insertion and deletion errors as in such cases the number of nucleotides cannot be easily deduced from the intensity of light.

DNA sequencing technologies are still rapidly expanding and improving, resulting in increased speed and capacity with reduction of cost. They greatly accelerate progress in biomedical research, forensics, agriculture, and diagnostics and enable novel technologies such as DEL. One expects the impact of high-throughput sequencing to continue to have an impact on the practice and applications of DNA-encoded chemistry methods.

## REFERENCES

1. Dahm, R. (2008). Discovering DNA: Friedrich Miescher and the early years of nucleic acid research. *Hum. Genet.*, 122, 565–581.
2. Watson, J. D., Crick, F. H. C. (1953). Molecular structure of nucleic acids: a structure for deoxyribose nucleic acid. *Nature*, 171, 737–738.
3. Dickerson, R. E., Drew, H. R., Conner, B. N., Wing, R. M., Fratini, A. V., Kopka, M. L. (1982). The anatomy of A-, B-, and Z-DNA. *Science*, 216, 475–485.
4. McCulloch, S. D., Kunkel, T. A. (2008). The fidelity of DNA synthesis by eukaryotic replicative and translesion synthesis polymerases. *Cell Res.*, 18, 148–161.
5. Steitz, T. A. (1999). DNA polymerases: structural diversity and common mechanisms. *J. Biol. Chem.*, 274, 17395–17398.

6. Hubscher, U., Spadari, S., Villani, G., Maga, G. (2010). *DNA Polymerases: Discovery, Characterization and Functions in Cellular DNA Transactions*. World Scientific Publishing Company, Hackensack.
7. Weiss, B., Richardson, C. C. (1967). Enzymatic breakage and joining of deoxyribonucleic acid, I. Repair of single-strand breaks in DNA by an enzyme system from Escherichia coli infected with T4 bacteriophage. *Proc. Natl. Acad. Sci. U.S.A.*, 57, 1021–1028.
8. Lehman, I. R. (1974). DNA ligase: structure, mechanism, and function. *Science*, 186, 790–797.
9. Chien, A., Edgar, D. B., Trela, J. M. (1976). Deoxyribonucleic acid polymerase from the extreme thermophile Thermus aquaticus. *J. Bacteriol.*, 127, 1550–1557.
10. Lawyer, F. C., Stoffel, S., Saiki, R. K., Chang, S. Y., Landre, P. A., Abramson, R. D., Gelfand, D. H. (1993). High-level expression, purification, and enzymatic characterization of full-length Thermus aquaticus DNA polymerase and a truncated form deficient in 5′ to 3′ exonuclease activity. *Genome Res.*, 2, 275–287.
11. Pon, R. T. (2003). Chemical synthesis of oligonucleotides: from dream to automation. In Khudyakov, Y. E., Fields, H. A., Eds. *Artificial DNA: Methods and Applications*. CRC Press LLC, Boca Raton, pp. 1–70.
12. Martin, R. (1996). *Gel Electrophoresis: Nucleic Acids* (*Introduction to Biotechniques*). BIOS Scientific Publishers Ltd., Oxford.
13. Sambrook, J. F., Russell, D. W. (2001). *Molecular Cloning: A Laboratory Manual*, 3rd edition. Cold Spring Harbor Laboratory Press, Cold Spring Harbor.
14. Fenn, J. B., Mann, M., Meng, C. K., Wong, S. F., Whitehouse, C. M. (1989). Electrospray ionization for mass spectrometry of large biomolecules. *Science*, 246, 64–71.
15. Apffel, A., Chakel, J. A., Fischer, S., Lichtenwalter, K., Hancock, W. S. (1997). Analysis of oligonucleotides by HPLC-electrospray ionization mass spectrometry. *Anal. Chem.*, 69, 1320–1325.
16. Potier, N., Van Dorsselaer, A., Cordier, Y., Roch, O., Bischoff, R. (1994). Negative electrospray ionization mass spectrometry of synthetic and chemically modified oligonucleotides. *Nucleic Acids Res.*, 22, 3895–3903.
17. Karas, M., Bahr, U., Giesmann, U. (1991). Matrix-assisted laser desorption ionization mass spectrometry. *Mass Spectrom. Rev.*, 10, 335–357.
18. Wu, K. J., Steding, A., Becker, C. H. (1993). Matrix-assisted laser desorption time-of-flight mass spectrometry of oligonucleotides using 3-hydroxypicolinic acid as an ultraviolet-sensitive matrix. *Rapid Commun. Mass Spectrom.*, 7, 142–146.
19. Stults, J. T. (1995). Matrix-assisted laser desorption/ionization mass spectrometry (MALDI-MS). *Curr. Opin. Struct. Biol.*, 5, 691–698.
20. Maxam, A. M., Gilbert, W. (1977). A new method for sequencing DNA. *Proc. Natl. Acad. Sci. U.S.A.*, 74, 560–564.
21. Sanger, F., Nicklen, S., Coulson, A. R. (1977). DNA sequencing with chain-terminating inhibitors. *Proc. Natl. Acad. Sci. U.S.A.*, 74, 5463–5467.
22. Shendure, J., Ji, H. (2008). Next-generation DNA sequencing. *Nat. Biotechnol.*, 26, 1135–1145.
23. Fuller, C. W., Middendorf, L. R., Benner, S. A., Church, G. M., Harris, T., Huang, X., Jovanovich, S. B., Nelson, J. R., Schloss, J. A., Schwartz, D. C., Vezenov, D. V. (2009). The challenges of sequencing by synthesis. *Nat. Biotechnol.*, 27, 1013–1023.
24. Metzker, M. L. (2010). Sequencing technologies—the next generation. *Nat. Rev. Genet.*, 11, 31–46.

25. Mohamed, S., Syed, B. A. (2013). Commercial prospects for genomic sequencing technologies. *Nat. Rev. Drug Discov.*, 12, 341–342.
26. Adessi, C., Matton, G., Ayala, G., Turcatti, G., Mermod, J. -J., Mayer, P., Kawashima, E. (2000). Solid phase DNA amplification: characterization of primer attachment and amplification mechanism. *Nucleic Acids Res.*, 28, e87.
27. Fedurco, M., Romieu, A., Williams, S., Lawrence, I., Turcatti, G. (2006). BTA, a novel reagent for DNA attachment on glass and efficient generation of solid-phase amplified DNA colonies. *Nucleic Acids Res.*, 34, e22.
28. Ju, J., Kim, D. H., Bi, L., Meng, Q., Bai, X., Li, Z., Li, X., Marma, M. S., Shi, S., Wu, J., Edwards, J. R., Romu, A., Turro, N. J. (2006). Four-color DNA sequencing by synthesis using cleavable fluorescent nucleotide reversible terminators. *Proc. Natl. Acad. Sci. U. S. A.*, 103, 19635–19640.
29. Turcatti, G., Romieu, A., Fedurco, M., Tairi, A. P. (2008). A new class of cleavable fluorescent nucleotides: synthesis and optimization as reversible terminators for DNA sequencing by synthesis. *Nucleic Acids Res.*, 36, e25.
30. Margulies, M., Egholm, M., Altman W. E., Attiya, S., Bader, J. S., Bemben, L. A., Berka, J., Braverman, M. S., Chen, Y. -J., Chen, Z., Dewell, S. B., Du, L., Fierro, J. M., Gomes, X. V., Godwin, B. C., He, W., Helgesen, S., Ho, C. H., Irzyk, G. P., Jando, S. C., Alenquer, M. L. I., Jarvie, T. P., Jirage, K. B., Kim, J.-B., Knight, J. R., Lanza, J. R., Leamon, J. H., Lefkowitz, S. M., Lei, M., Li, J., Lohman, K. L., Lu, H., Makhijani, V. B., McDade, K. E., McKenna, M. P., Myers, E. W., Nickerson, E., Nobile, J. R., Plant, R., Puc, B. P., Ronan, M. T., Roth, G. T., Sarkis, G. J., Simons, J. F., Simpson, J. W., Srinivasan, M., Tartaro, K. R., Tomasz, A., Vogt, K. A., Volkmer, G. A., Wang, S. H., Wang, Y., Weiner, M. P., Yu, P., Begley, R. F., Rothberg, J. M. (2005). Genome sequencing in microfabricated high-density picolitre reactors. *Nature*, 437, 376–380.
31. Dressman, D., Yan, H., Traverso, G., Kinzler, K. W., Vogelstein, B. (2003). Transforming single DNA molecules into fluorescent magnetic particles for detection and enumeration of genetic variations. *Proc. Natl. Acad. Sci. U.S.A.*, 100, 8817–8822.
32. Ronaghi, M., Uhlen, M., Nyren, P. (1998). A sequencing method based on real-time pyrophosphate. *Science*, 281, 363–365.