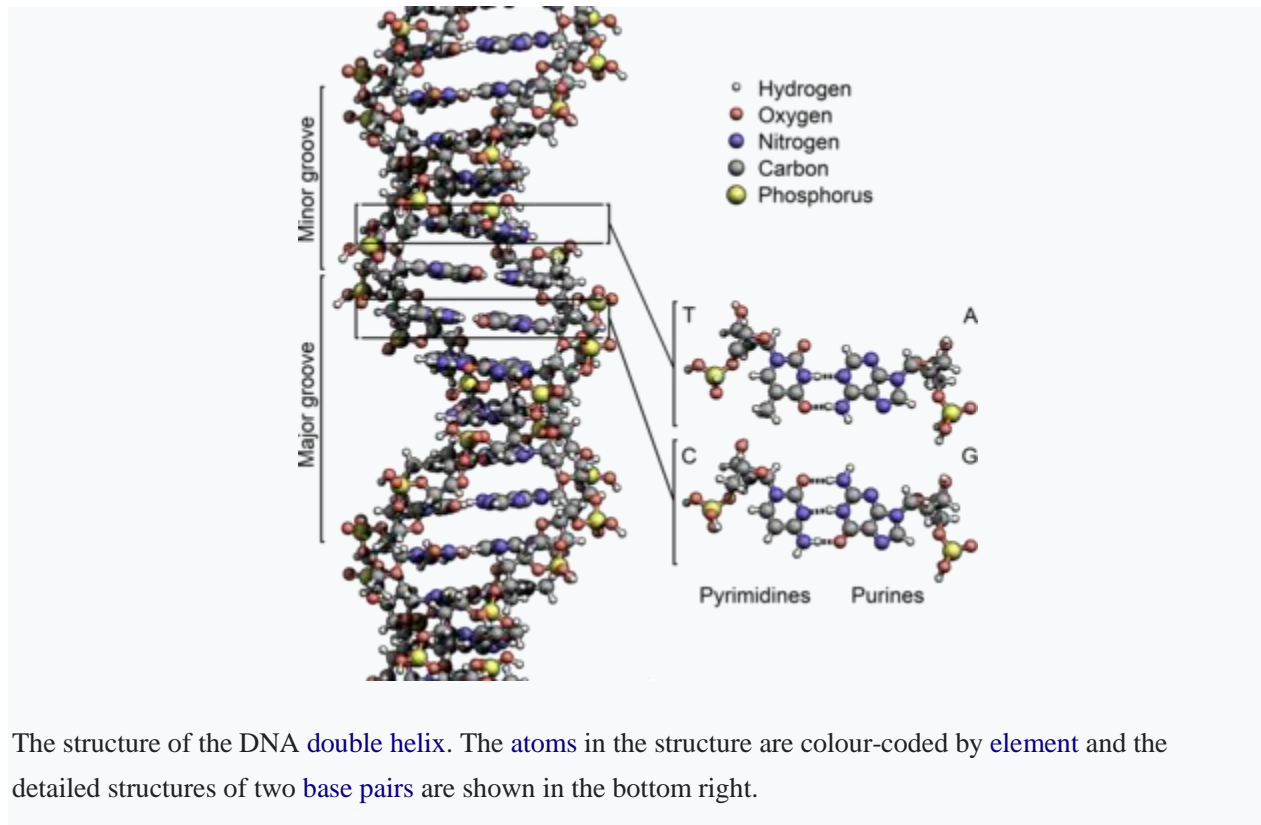
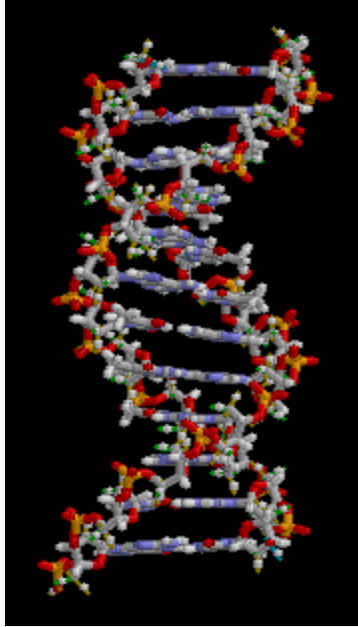


UNIT-I

DNA: Definition, Structure & Discovery

DNA





The structure of part of a DNA **double helix**

Deoxyribonucleic acid is a **molecule** composed of two **polynucleotide** chains that coil around each other to form a **double helix** carrying **genetic** instructions for the development, functioning, growth and **reproduction** of all known **organisms** and many **viruses**. DNA and **ribonucleic acid** (RNA) are **nucleic acids**. Alongside **proteins**, **lipids** and complex carbohydrates (**polysaccharides**), nucleic acids are one of the four major types of **macromolecules** that are essential for all known forms of **life**.

The two DNA strands are known as **polynucleotides** as they are composed of simpler **monomeric** units called **nucleotides**. Each nucleotide is composed of one of four **nitrogen-containing nucleobases** (**cytosine** [C], **guanine** [G], **adenine** [A] or **thymine** [T]), a **sugar** called **deoxyribose**, and a **phosphate group**. The nucleotides are joined to one another in a chain by **covalent bonds** (known as the phospho-diester linkage) between the sugar of one nucleotide and the phosphate of the next, resulting in an alternating **sugar-phosphate backbone**.

The nitrogenous bases of the two separate polynucleotide strands are bound together, according to **base pairing** rules (A with T and C with G), with **hydrogen bonds** to make double-stranded DNA. The complementary nitrogenous bases are divided into two groups, **pyrimidines** and **purines**. In DNA, the pyrimidines are thymine and cytosine; the purines are adenine and guanine.

Both strands of double-stranded DNA store the same **biological information**. This information is **replicated** as and when the two strands separate. A large part of DNA (more than 98% for humans) is **non-coding**, meaning that these sections do not serve as patterns for **protein sequences**.

The two strands of DNA run in opposite directions to each other and are thus **antiparallel**. Attached to each sugar is one of four types of nucleobases (informally, *bases*). It is the **sequence** of these four nucleobases along the backbone that encodes genetic information. **RNA** strands are created using DNA strands as a template in a process

called **transcription**, where DNA bases are exchanged for their corresponding bases except in the case of thymine (T), for which RNA substitutes **uracil** (U). Under the **genetic code**, these RNA strands specify the sequence of **amino acids** within proteins in a process called **translation**.

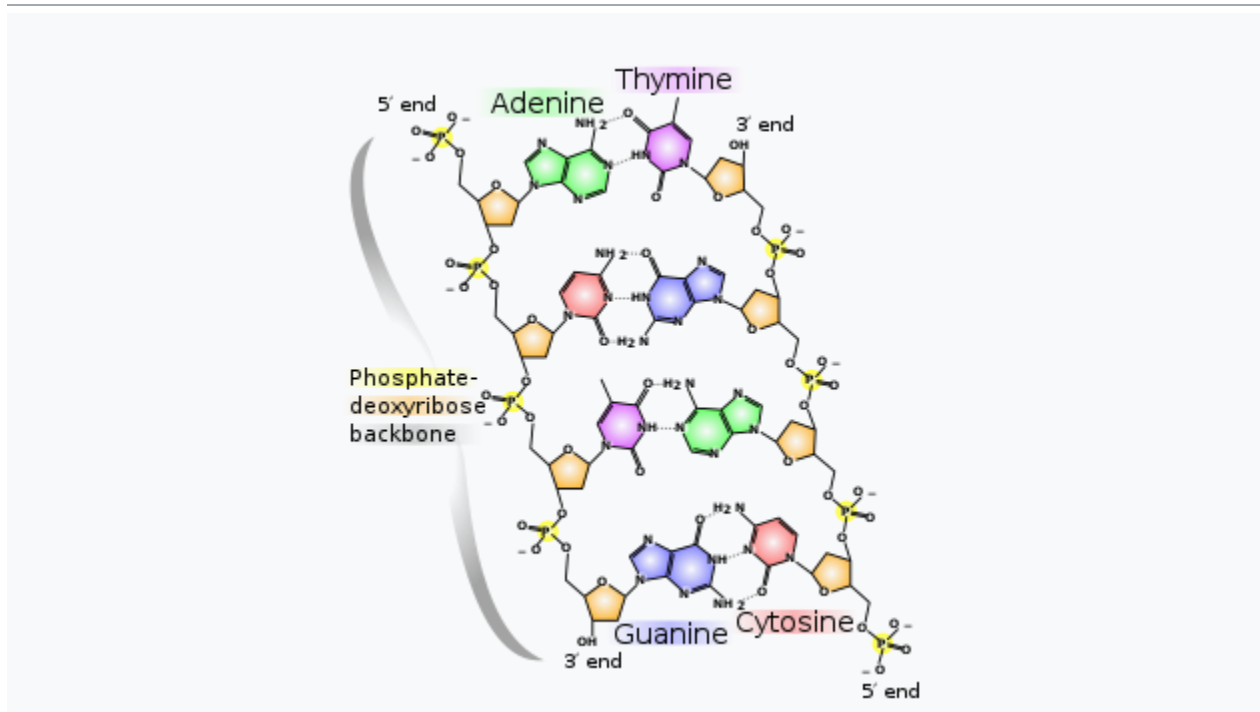
Within eukaryotic cells, DNA is organized into long structures called **chromosomes**. Before typical **cell division**, these chromosomes are duplicated in the process of **DNA replication**, providing a complete set of chromosomes for each daughter cell. **Eukaryotic organisms** (animals, plants, fungi and protists) store most of their DNA inside the **cell nucleus** as **nuclear DNA**, and some in the **mitochondria** as **mitochondrial DNA** or in **chloroplasts** as **chloroplast DNA**.

In contrast, **prokaryotes** (bacteria and archaea) store their DNA only in the **cytoplasm**, in **circular chromosomes**. Within eukaryotic chromosomes, **chromatin** proteins, such as **histones**, compact and organize DNA. These compacting structures guide the interactions between DNA and other proteins, helping control which parts of the DNA are transcribed.

DNA was first isolated by **Friedrich Miescher** in 1869. Its molecular structure was first identified by **Francis Crick** and **James Watson** at the **Cavendish Laboratory** within the **University of Cambridge** in 1953, whose model-building efforts were guided by **X-ray diffraction** data acquired by **Raymond Gosling**, who was a post-graduate student of **Rosalind Franklin** at **King's College London**. DNA is used by researchers as a **molecular tool** to explore physical laws and theories, such as the **ergodic theorem** and the theory of **elasticity**

. The unique material properties of DNA have made it an attractive molecule for material scientists and engineers interested in micro- and nano-fabrication. Among notable advances in this field are **DNA origami** and DNA-based hybrid materials

Properties



DNA is a long **polymer** made from repeating units called **nucleotides**, each of which is usually symbolized by a single letter: either A, T, C, or G. The structure of DNA is dynamic along its length, being capable of coiling into tight loops and other shapes.

In all species it is composed of two helical chains, bound to each other by **hydrogen bonds**. Both chains are coiled around the same axis, and have the same pitch of 34 **angstroms** (Å) (3.4 **nanometres**). The pair of chains has a radius of 10 angstroms (1.0 nanometre). According to another study, when measured in a different solution, the DNA chain measured 22 to 26 angstroms wide (2.2 to 2.6 nanometres), and one nucleotide unit measured 3.3 Å (0.33 nm) long.

Although each individual nucleotide is very small, a DNA polymer can be very large and contain hundreds of millions, such as in **chromosome 1**. Chromosome 1 is the largest human **chromosome** with approximately 220 million **base pairs**, and would be 85 mm long if straightened.

DNA does not usually exist as a single strand, but instead as a pair of strands that are held tightly together.¹ These two long strands coil around each other, in the shape of a **double helix**. The nucleotide contains both a segment of the **backbone** of the molecule (which holds the chain together) and a **nucleobase** (which interacts with the other DNA strand in the helix).

A nucleobase linked to a sugar is called a **nucleoside**, and a base linked to a sugar and to one or more phosphate groups is called a **nucleotide**. A **biopolymer** comprising multiple linked nucleotides (as in DNA) is called a **polynucleotide**.

The backbone of the DNA strand is made from alternating **phosphate** and **sugar** groups. The sugar in DNA is **2-deoxyribose**, which is a **pentose** (five-carbon) sugar. The sugars are joined together by phosphate groups that form **phosphodiester bonds** between the third and fifth carbon **atoms** of adjacent sugar rings. These are known as the **3'-end** (three prime end), and **5'-end** (five prime end) carbons, the prime symbol being used to distinguish these carbon atoms from those of the base to which the deoxyribose forms a **glycosidic bond**. Therefore, any DNA strand normally has one end at which there is a phosphate group attached to the 5' carbon of a ribose (the 5' phosphoryl) and another end at which there is a free hydroxyl group attached to the 3' carbon of a ribose (the 3' hydroxyl).

The orientation of the 3' and 5' carbons along the sugar-phosphate backbone confers **directionality** (sometimes called polarity) to each DNA strand. In a **nucleic acid double helix**, the direction of the nucleotides in one strand is opposite to their direction in the other strand: the strands are **antiparallel**.

The asymmetric ends of DNA strands are said to have a directionality of five prime end (5'), and three prime end (3'), with the 5' end having a terminal phosphate group and the 3' end a terminal hydroxyl group. One major difference between DNA and **RNA** is the sugar, with the 2-deoxyribose in DNA being replaced by the alternative pentose sugar **ribose** in RNA.



A section of DNA. The bases lie horizontally between the two spiraling strands ([animated version](#)).

The DNA double helix is stabilized primarily by two forces: [hydrogen bonds](#) between nucleotides and [base-stacking](#) interactions among [aromatic](#) nucleobases. The four bases found in DNA are [adenine](#) (A), [cytosine](#) (C), [guanine](#) (G) and [thymine](#) (T).

These four bases are attached to the sugar-phosphate to form the complete nucleotide, as shown for [adenosine monophosphate](#). Adenine pairs with thymine and guanine pairs with cytosine, forming A-T and G-C base pairs.

Nucleobase classification

The nucleobases are classified into two types: the [purines](#), A and G, which are fused five- and six-membered [heterocyclic compounds](#), and the [pyrimidines](#), the six-membered rings C and T. A fifth pyrimidine nucleobase, [uracil](#) (U), usually takes the place of thymine in RNA and differs from thymine by lacking a [methyl group](#) on its ring. In addition to RNA and DNA, many artificial [nucleic acid analogues](#) have been created to study the properties of nucleic acids, or for use in biotechnology.

Non-canonical bases

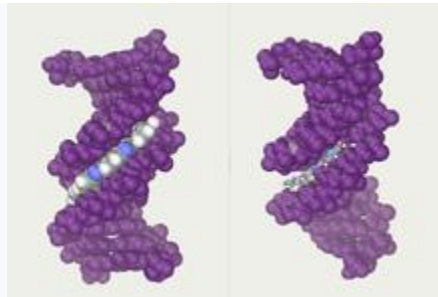
Modified bases occur in DNA. The first of these recognised was [5-methylcytosine](#), which was found in the [genome](#) of *Mycobacterium tuberculosis* in 1925. The reason for the presence of these noncanonical bases in bacterial viruses ([bacteriophages](#)) is to avoid the [restriction enzymes](#) present in bacteria. This enzyme system acts at least in part as a molecular immune system protecting bacteria from infection by viruses. Modifications of the bases cytosine and adenine the more common and modified DNA bases plays vital roles in the [epigenetic](#) control of gene expression in plants and animals.

Listing of non-canonical bases found in DNA

A number of non canonical bases are known to occur in DNA. Most of these are modifications of the canonical bases plus uracil.

- Modified **Adenosine**
 - N6-carbamoyl-methyladenine

- N6-methyladenine
- **Modified Guanine**
 - 7-Deazaguanine
 - 7-Methylguanine
- **Modified Cytosine**
 - N4-Methylcytosine
 - 5-Carboxylcytosine
 - 5-Formylcytosine
 - 5-Glycosylhydroxymethylcytosine
 - 5-Hydroxycytosine
 - 5-Methylcytosine
- **Modified Thymidine**
 - α -Glutamylthymidine
 - α -Putrescinythymine
- **Uracil** and modifications
 - Base J
 - Uracil
 - 5-Dihydropentauracil
 - 5-Hydroxymethyldeoxyuracil
- **Others**
 - Deoxyarchaeosine
 - 2,6-Diaminopurine



DNA major and minor grooves. The latter is a binding site for the [Hoechst stain dye 33258](#).

Grooves

Twin helical strands form the DNA backbone. Another double helix may be found tracing the spaces, or grooves, between the strands. These voids are adjacent to the base pairs and may provide a [binding site](#). As the strands are not symmetrically located with respect to each other, the grooves are unequally sized. One groove, the major groove, is 22 [angstroms](#) (Å) wide and the other, the minor groove, is 12 Å wide.

The width of the major groove means that the edges of the bases are more accessible in the major groove than in the minor groove. As a result, proteins such as [transcription factors](#) that can bind to specific sequences in double-stranded DNA usually make contact with the sides of the bases exposed in the major groove.

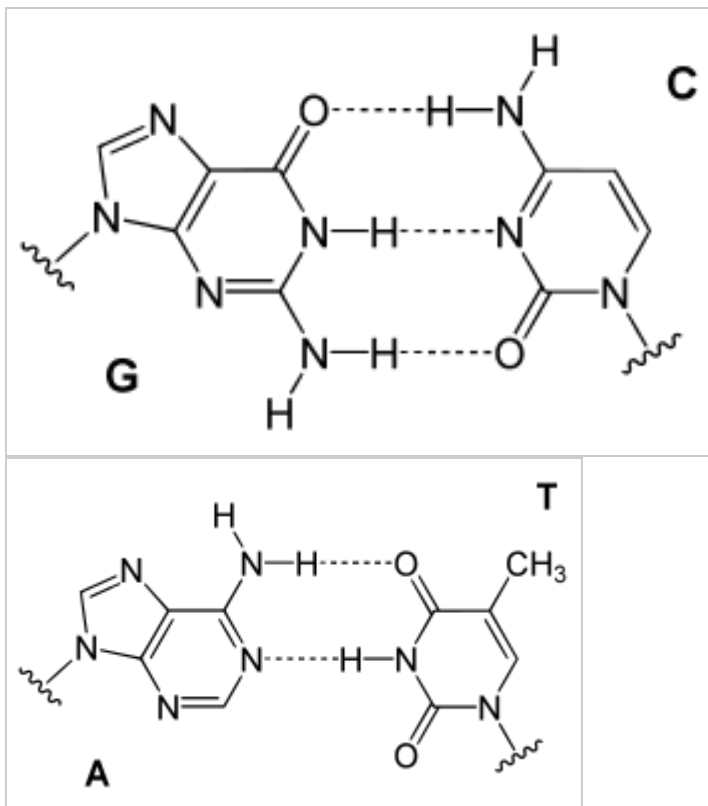
This situation varies in unusual conformations of DNA within the cell, but the major and minor grooves are always named to reflect the differences in size that would be seen if the DNA is twisted back into the ordinary B form.

Base pairing

In a DNA double helix, each type of nucleobase on one strand bonds with just one type of nucleobase on the other strand. This is called **complementary base pairing**. Purines form **hydrogen bonds** to pyrimidines, with adenine bonding only to thymine in two hydrogen bonds, and cytosine bonding only to guanine in three hydrogen bonds.

This arrangement of two nucleotides binding together across the double helix is called a Watson-Crick base pair. DNA with high **GC-content** is more stable than DNA with low GC-content. A Hoogsteen base pair is a rare variation of base-pairing. As hydrogen bonds are not **covalent**, they can be broken and rejoined relatively easily. The two strands of DNA in a double helix can thus be pulled apart like a zipper, either by a mechanical force or high **temperature**.

As a result of this base pair complementarity, all the information in the double-stranded sequence of a DNA helix is duplicated on each strand, which is vital in DNA replication. This reversible and specific interaction between complementary base pairs is critical for all the functions of DNA in organisms.



Top, a **GC** base pair with three **hydrogen bonds**. Bottom, an **AT** base pair with two hydrogen bonds. Non-covalent hydrogen bonds between the pairs are shown as dashed lines.

As noted above, most DNA molecules are actually two polymer strands, bound together in a helical fashion by noncovalent bonds; this double-stranded (**dsDNA**) structure is maintained largely by the intrastrand base stacking interactions, which are strongest for G,C stacks. The two strands can come apart—a process known as melting—to form two single-stranded DNA (**ssDNA**) molecules. Melting occurs at high temperature, low salt and high **pH** (low pH also melts DNA, but since DNA is unstable due to acid depurination, low pH is rarely used).

The stability of the dsDNA form depends not only on the GC-content (% G,C basepairs) but also on sequence (since stacking is sequence specific) and also length (longer molecules are more stable). The stability

can be measured in various ways; a common way is the "melting temperature", which is the temperature at which 50% of the ds molecules are converted to ss molecules; melting temperature is dependent on ionic strength and the concentration of DNA.

As a result, it is both the percentage of GC base pairs and the overall length of a DNA double helix that determines the strength of the association between the two strands of DNA. Long DNA helices with a high GC-content have stronger-interacting strands, while short helices with high AT content have weaker-interacting strands. In biology, parts of the DNA double helix that need to separate easily, such as the TATAAT **Pribnow box** in some **promoters**, tend to have a high AT content, making the strands easier to pull apart.

In the laboratory, the strength of this interaction can be measured by finding the temperature necessary to break half of the hydrogen bonds, their **melting temperature** (also called T_m value). When all the base pairs in a DNA double helix melt, the strands separate and exist in solution as two entirely independent molecules. These single-stranded DNA molecules have no single common shape, but some conformations are more stable than others.

Sense and antisense

A **DNA sequence** is called a "sense" sequence if it is the same as that of a **messenger RNA** copy that is translated into protein. The sequence on the opposite strand is called the "antisense" sequence. Both sense and antisense sequences can exist on different parts of the same strand of DNA (i.e. both strands can contain both sense and antisense sequences). In both prokaryotes and eukaryotes, antisense RNA sequences are produced, but the functions of these RNAs are not entirely clear. One proposal is that antisense RNAs are involved in regulating **gene expression** through RNA-RNA base pairing.

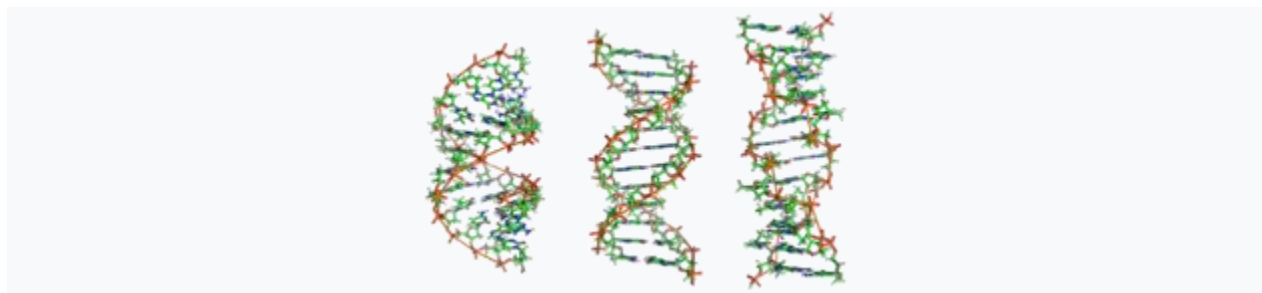
A few DNA sequences in prokaryotes and eukaryotes, and more in **plasmids** and **viruses**, blur the distinction between sense and antisense strands by having **overlapping genes**. In these cases, some DNA sequences do double duty, encoding one protein when read along one strand, and a second protein when read in the opposite direction along the other strand. In **bacteria**, this overlap may be involved in the regulation of gene transcription, while in viruses, overlapping genes increase the amount of information that can be encoded within the small viral genome.

Supercoiling

DNA can be twisted like a rope in a process called **DNA supercoiling**. With DNA in its "relaxed" state, a strand usually circles the axis of the double helix once every 10.4 base pairs, but if the DNA is twisted the strands become more tightly or more loosely wound. If the DNA is twisted in the direction of the helix, this is positive supercoiling, and the bases are held more tightly together.

If they are twisted in the opposite direction, this is negative supercoiling, and the bases come apart more easily. In nature, most DNA has slight negative supercoiling that is introduced by **enzymes** called **topoisomerases**.

These enzymes are also needed to relieve the twisting stresses introduced into DNA strands during processes such as **transcription** and **DNA replication**.



From left to right, the structures of A, B and Z DNA

Alternative DNA structures

DNA exists in many possible **conformations** that include **A-DNA**, **B-DNA**, and **Z-DNA** forms, although, only B-DNA and Z-DNA have been directly observed in functional organisms. The conformation that DNA adopts depends on the hydration level, DNA sequence, the amount and direction of supercoiling, chemical modifications of the bases, the type and concentration of metal **ions**, and the presence of **polyamines** in solution.

The first published reports of A-DNA **X-ray diffraction patterns**—and also B-DNA—used analyses based on **Patterson transforms** that provided only a limited amount of structural information for oriented fibers of DNA. An alternative analysis was then proposed by Wilkins *et al.*, in 1953, for the *in vivo* B-DNA X-ray diffraction-scattering patterns of highly hydrated DNA fibers in terms of squares of **Bessel functions**.

In the same journal, **James Watson** and **Francis Crick** presented their **molecular modeling** analysis of the DNA X-ray diffraction patterns to suggest that the structure was a double-helix.

Although the *B-DNA form* is most common under the conditions found in cells, it is not a well-defined conformation but a family of related DNA conformations, that occur at the high hydration levels present in cells. Their corresponding X-ray diffraction and scattering patterns are characteristic of molecular **paracrystals** with a significant degree of disorder.

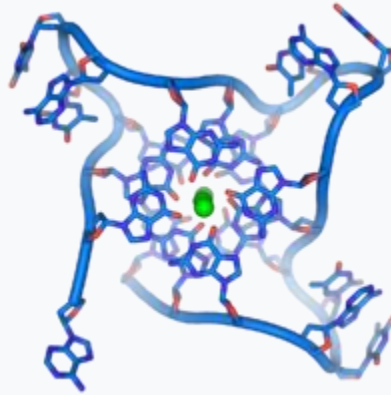
Compared to B-DNA, the A-DNA form is a wider **right-handed** spiral, with a shallow, wide minor groove and a narrower, deeper major groove. The A form occurs under non-physiological conditions in partly dehydrated samples of DNA, while in the cell it may be produced in hybrid pairings of DNA and RNA strands, and in enzyme-DNA complexes. Segments of DNA where the bases have been chemically modified by **methylation** may undergo a larger change in conformation and adopt the **Z form**. Here, the strands turn about the helical axis in a left-handed spiral, the opposite of the more common B form. These unusual structures can be recognized by specific Z-DNA binding proteins and may be involved in the regulation of transcription.

Alternative DNA chemistry

For many years, **exobiologists** have proposed the existence of a **shadow biosphere**, a postulated microbial biosphere of Earth that uses radically different biochemical and molecular processes than currently known life. One of the proposals was the existence of lifeforms that use **arsenic instead of phosphorus in DNA**. A report in 2010 of the possibility in the **bacterium GFAJ-1**, was announced, though the research was disputed, and evidence suggests the bacterium actively prevents the incorporation of arsenic into the DNA backbone and other biomolecules.

Quadruplex structures

At the ends of the linear chromosomes are specialized regions of DNA called **telomeres**. The main function of these regions is to allow the cell to replicate chromosome ends using the enzyme **telomerase**, as the enzymes that normally replicate DNA cannot copy the extreme 3' ends of chromosomes. These specialized chromosome caps also help protect the DNA ends, and stop the **DNA repair** systems in the cell from treating them as damage to be corrected. In **human cells**, telomeres are usually lengths of single-stranded DNA containing several thousand repeats of a simple TTAGGG sequence.

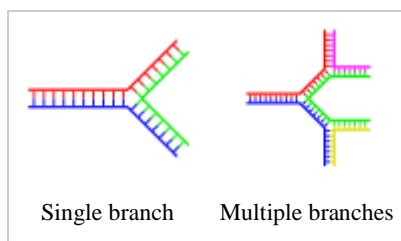


DNA quadruplex formed by [telomere](#) repeats. The looped conformation of the DNA backbone is very different from the typical DNA helix. The green spheres in the center represent potassium ions.

These guanine-rich sequences may stabilize chromosome ends by forming structures of stacked sets of four-base units, rather than the usual base pairs found in other DNA molecules. Here, four guanine bases, known as a [guanine tetrad](#), form a flat plate. These flat four-base units then stack on top of each other to form a stable [G-quadruplex](#) structure.

These structures are stabilized by hydrogen bonding between the edges of the bases and [chelation](#) of a metal ion in the centre of each four-base unit. Other structures can also be formed, with the central set of four bases coming from either a single strand folded around the bases, or several different parallel strands, each contributing one base to the central structure.

In addition to these stacked structures, telomeres also form large loop structures called telomere loops, or T-loops. Here, the single-stranded DNA curls around in a long circle stabilized by telomere-binding proteins. At the very end of the T-loop, the single-stranded telomere DNA is held onto a region of double-stranded DNA by the telomere strand disrupting the double-helical DNA and base pairing to one of the two strands. This [triple-stranded](#) structure is called a displacement loop or [D-loop](#).



[Branched DNA](#) can form networks containing multiple branches.

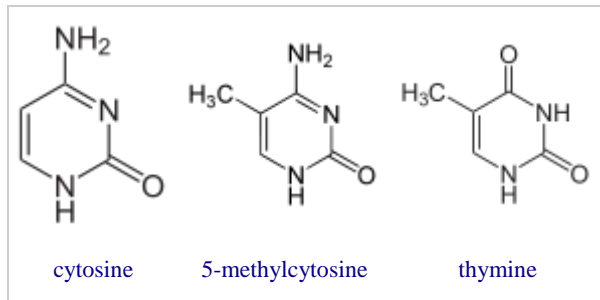
Branched DNA

A, [fraying](#) occurs when non-complementary regions exist at the end of an otherwise complementary double-strand of DNA. However, branched DNA can occur if a third strand of DNA is introduced and contains adjoining regions able to hybridize with the frayed regions of the pre-existing double-strand. Although the simplest example of branched DNA involves only three strands of DNA, complexes involving additional strands and multiple branches are also possible. Branched DNA can be used in [nanotechnology](#) to construct geometric shapes, see the section on [uses in technology](#) below.

Artificial bases

Several artificial nucleobases have been synthesized, and successfully incorporated in the eight-base DNA analogue named **Hachimoji DNA**. Dubbed S, B, P, and Z, these artificial bases are capable of bonding with each other in a predictable way (S–B and P–Z), maintain the double helix structure of DNA, and be transcribed to RNA. Their existence implies that there is nothing special about the four natural nucleobases that evolved on Earth.

Chemical modifications and altered DNA packaging



Structure of cytosine with and without the 5-methyl group. **Deamination** converts 5-methylcytosine into thymine.

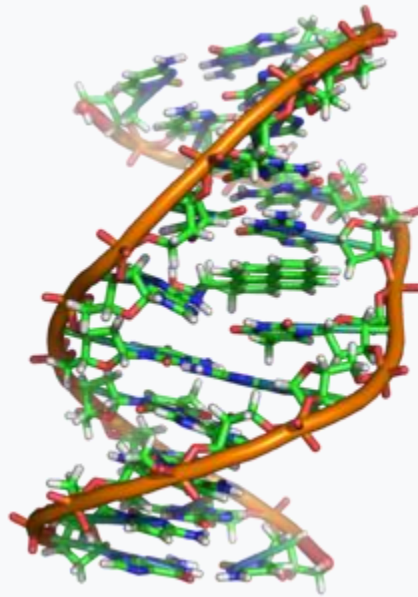
Base modifications and DNA packaging

The expression of genes is influenced by how the DNA is packaged in chromosomes, in a structure called **chromatin**. Base modifications can be involved in packaging, with regions that have low or no gene expression usually containing high levels of **methylation** of **cytosine** bases.

DNA packaging and its influence on gene expression can also occur by covalent modifications of the **histone** protein core around which DNA is wrapped in the chromatin structure or else by remodeling carried out by chromatin remodeling complexes (see **Chromatin remodeling**). There is, further, **crosstalk** between DNA methylation and histone modification, so they can coordinately affect chromatin and gene expression.

For one example, cytosine methylation produces **5-methylcytosine**, which is important for **X-inactivation** of chromosomes. The average level of methylation varies between organisms—the worm *Caenorhabditis elegans* lacks cytosine methylation, while **vertebrates** have higher levels, with up to 1% of their DNA containing 5-methylcytosine. Despite the importance of 5-methylcytosine, it can **deaminate** to leave a thymine base, so methylated cytosines are particularly prone to **mutations**. Other base modifications include adenine methylation in bacteria, the presence of **5-hydroxymethylcytosine** in the **brain**, and the **glycosylation** of uracil to produce the "J-base" in **kinetoplastids**.

Damage



A covalent adduct between a metabolically activated form of benzo[*a*]pyrene, the major mutagen in tobacco smoke, and DNA

DNA can be damaged by many sorts of **mutagens**, which change the **DNA sequence**. Mutagens include **oxidizing agents**, **alkylating agents** and also high-energy **electromagnetic radiation** such as **ultraviolet light** and **X-rays**. The type of DNA damage produced depends on the type of mutagen. For example, UV light can damage DNA by producing **thymine dimers**, which are cross-links between pyrimidine bases.

On the other hand, oxidants such as **free radicals** or **hydrogen peroxide** produce multiple forms of damage, including base modifications, particularly of guanosine, and double-strand breaks. A typical human cell contains about 150,000 bases that have suffered oxidative damage.

Of these oxidative lesions, the most dangerous are double-strand breaks, as these are difficult to repair and can produce **point mutations**, **insertions**, **deletions** from the DNA sequence, and **chromosomal translocations**.

These mutations can cause **cancer**. Because of inherent limits in the DNA repair mechanisms, if humans lived long enough, they would all eventually develop cancer.

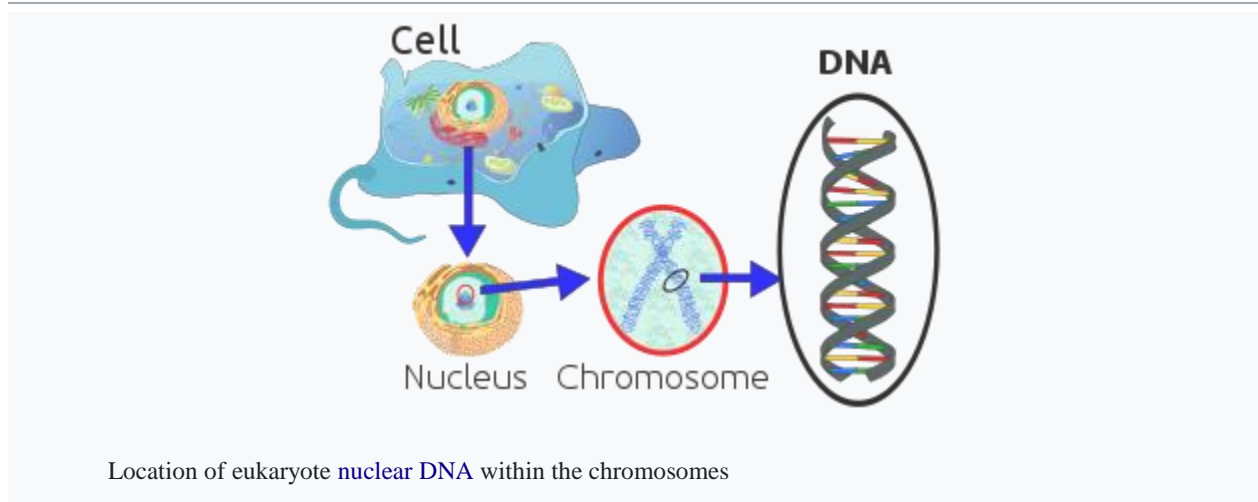
DNA damages that are **naturally occurring**, due to normal cellular processes that produce reactive oxygen species, the hydrolytic activities of cellular water, etc., also occur frequently. Although most of these damages are repaired, in any cell some DNA damage may remain despite the action of repair processes. These remaining DNA damages accumulate with age in mammalian postmitotic tissues. This accumulation appears to be an important underlying cause of aging.

Many mutagens fit into the space between two adjacent base pairs, this is called **intercalation**. Most intercalators are **aromatic** and planar molecules; examples include **ethidium bromide**, **acridines**, **daunomycin**, and **doxorubicin**. For an intercalator to fit between base pairs, the bases must separate, distorting the DNA strands by unwinding of the double helix. This inhibits both transcription and DNA replication, causing toxicity and mutations.

As a result, DNA intercalators may be **carcinogens**, and in the case of thalidomide, a **teratogen**. Others such as **benzo[*a*]pyrene diol epoxide** and **aflatoxin** form DNA adducts that induce errors in

replication. Nevertheless, due to their ability to inhibit DNA transcription and replication, other similar toxins are also used in **chemotherapy** to inhibit rapidly growing **cancer** cells.

Biological functions



DNA usually occurs as linear **chromosomes** in **eukaryotes**, and **circular chromosomes** in **prokaryotes**. The set of chromosomes in a cell makes up its **genome**; the **human genome** has approximately 3 billion base pairs of DNA arranged into 46 chromosomes. The information carried by DNA is held in the **sequence** of pieces of DNA called **genes**. **Transmission** of genetic information in genes is achieved via complementary base pairing.

For example, in transcription, when a cell uses the information in a gene, the DNA sequence is copied into a complementary RNA sequence through the attraction between the DNA and the correct RNA nucleotides.

Usually, this RNA copy is then used to make a matching **protein sequence** in a process called **translation**, which depends on the same interaction between RNA nucleotides. In alternative fashion, a cell may simply copy its genetic information in a process called DNA replication. The details of these functions are covered in other articles; here the focus is on the interactions between DNA and other molecules that mediate the function of the genome.

Genes and genomes

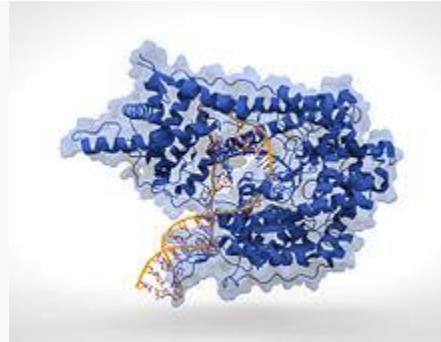
Genomic DNA is tightly and orderly packed in the process called **DNA condensation**, to fit the small available volumes of the cell. In eukaryotes, DNA is located in the **cell nucleus**, with small amounts in **mitochondria** and **chloroplasts**. In prokaryotes, the DNA is held within an irregularly shaped body in the cytoplasm called the **nucleoid**.

The genetic information in a genome is held within genes, and the complete set of this information in an organism is called its **genotype**. A gene is a unit of **heredity** and is a region of DNA that influences a particular characteristic in an organism. Genes contain an **open reading frame** that can be transcribed, and **regulatory sequences** such as **promoters** and **enhancers**, which control transcription of the open reading frame.

In many **species**, only a small fraction of the total sequence of the **genome** encodes protein. For example, only about 1.5% of the human genome consists of protein-coding **exons**, with over 50% of human DNA consisting of non-coding **repetitive sequences**.

The reasons for the presence of so much **noncoding DNA** in eukaryotic genomes and the extraordinary differences in **genome size**, or **C-value**, among species, represent a long-standing puzzle known as the "**C-value enigma**".

However, some DNA sequences that do not code protein may still encode functional **non-coding RNA** molecules, which are involved in the **regulation of gene expression**.



T7 RNA polymerase (blue) producing an mRNA (green) from a DNA template (orange)

Some noncoding DNA sequences play structural roles in chromosomes. **Telomeres** and **centromeres** typically contain few genes but are important for the function and stability of chromosomes. An abundant form of noncoding DNA in humans are **pseudogenes**, which are copies of genes that have been disabled by mutation. These sequences are usually just molecular **fossils**, although they can occasionally serve as raw **genetic material** for the creation of new genes through the process of **gene duplication** and **divergence**.

5 Main Enzymes Involved in Genetic Engineering | Biotechnology

The following points highlight the five main enzymes involved in genetic engineering. The enzymes are: 1. Restriction Endonuclease 2. DNA Ligase 3. Alkaline Phosphatase 4. DNA Polymerase and the Klenow Fragment 5. Reverse Transcriptase.

Genetic engineering became possible with the discovery of mainly two types of enzymes: the cutting enzymes called restriction endonucleases and the joining enzymes called ligases.

Restriction endonucleases or restriction enzymes, as they are called popularly, recognize unique base sequence motifs in a DNA strand and cleave the backbone of the molecule at a place within or, at some distance from the

recognition site. Whereas ligase is the enzyme that joins a 5' end of a DNA with a 3' end of the same or of another strand.

Enzyme # 1. Restriction Endonuclease:

Ordinary nucleases are endonucleases or exonucleases. The former cleaves the DNA backbone between two nucleotides, i.e., it cleaves the double stranded DNA at any point except the ends, but it involves only one strand of the duplex.

The latter remove or digest one nucleotide at a time starting from 5' or 3' end of a DNA strand. The restriction endonucleases cleave only at specific regions in a particular DNA, so that discrete and defined fragments are obtained at the end of total digestion. The name 'restriction' endonuclease originated from an observation of a system of restriction of the growth of the phage lambda in particular strains of the E. coli host cell.

Table 22.1: Source of restriction enzymes, cleavage sites and productions of cleavage

<i>Microorganisms</i>	<i>Restriction enzymes</i>	<i>Cleavage sites</i>	<i>Cleavage products</i>	
<i>Bacillus amyloliquefaciens</i> H	Bam HI	5-GGATCC-3 3-CCTAGG-5	5-G 3-CCTAG	GATCC-3 G-5
<i>B. globigii</i>	Bgl II	5-AGATCT-3 3-TCTAGA-5	5-A 3-TCTAG	GATCT-3 A-5
<i>Escherichia coli</i> RY13	Eco RI	5-GAATTC-3 3-CTTAAG-5	5-G 3-CTTAA	AATTC-3 G-5
<i>Haemophilus influenzae</i> Rd	Hin dIII	5-AAGCTT-3 3-TTCGAA-5	5-A 3-TTCGA	AGCTT-3 A-5
<i>H. parainfluenzae</i>	Hpa I	5-GTTAAC-3 3-CAATTG-5	5-GTT 3-CAA	AAC-3 TTG-5
<i>Klebsiella pneumoniae</i> OK 8	Kpn I	5-GGTACC-3 3-CCATGG-5	5-GGTAC 3-C	C-3 CATGG-5
<i>Streptomyces albus</i> G	Sal I	5-GTCGAC-3 3-CAGCTG-5	5-G 3-CAGCT	TCGAC-3 G-5
<i>Serratia marcescens</i>	Sma I	5-CCCGGG-3 3-GGGCCC-5	5-CCC 3-GGG	GGG-3 CCC-5

Most restriction enzymes recognize only one short base sequence in a DNA molecule and make two single strand breaks, one in each strand, generating

3'OH and 5'P groups at each position. The sequences recognized by restriction enzymes are often palindromes, i.e., inverted repetition sequences which are symmetrical.

Restriction enzymes can cut DNA in two ways to generate blunt ends (cut precisely at opposite sites, e.g., HpaI) and staggered ends (cut at asymmetrical position, e.g., Eco RI) with short single stranded overhangs at each end. A large number of restriction enzymes have been identified and classified into three categories (type I, II, III) on the basis of their site of cleavage.

Restriction enzymes have three important features:

1. Restriction enzymes make breaks in palindromic sequences.
2. The breaks are usually not directly opposite to one another.
3. The enzymes generate DNA fragments with complementary ends.

The commonly employed restriction enzymes are listed in Table 22.1.

Enzyme # 2. DNA Ligase:

Ends of DNA strands may be joined by the enzyme polynucleotide ligase, called 'glue' of the recombinant DNA molecule. The enzyme catalyses the formation of a phosphodiester bond between the 3'OH and 5'P terminals of two nucleotides. The enzyme is thus able to join unrelated DNA, repair nicks in single strand of DNA and join the sugar phosphate backbones of the newly repaired and resident region of a DNA strand.

The enzyme which is extensively used for covalently joining restriction fragments is the ligase from *E. coli* and that encoded by T4 phage. As the main source of DNA ligase is T4 phage, hence, the enzyme is known as T4 DNA ligase.

The ligation reaction is controlled by several factors, such as pH, temperature, concentration and kinds of sticky ends, etc. As ligase uses the ends of DNA molecules as substrates rather than the entire DNA, the kinetics of joining depend on the number of ends (concentration) available for joining.

Enzyme # 3. Alkaline Phosphatase:

The broken fragments of plasmids, instead of joining with foreign DNA, join the cohesive end of the same DNA molecules. The treatment with alkaline phosphatase prevents re-circularisation of plasmid vector and increases the frequency of production of recombinant DNA molecule.

Enzyme # 4. DNA Polymerase and the Klenow Fragment:

The DNA polymerase that is generally utilized is either the DNA Pol I from *E. coli* or the T4 DNA polymerase encoded by the phage gene. The *E. coli* enzyme is basically a proof-reading and repairing enzyme. It is composed of 3 subunits each with a specific activity. They are: 5'-3' polymerase, 3'-5' exonuclease and 5'-3' exonuclease.

The enzyme is useful for synthesizing short length of a DNA strand, especially by the nick translation method. The 5-3' exonuclease activity may be deleted, this edited enzyme is referred to as the klenow fragment. The T4 DNA Pol possesses, like the klenow fragment, only the polymerase and proofreading (3'-5' exonuclease) functions.

Enzyme # 5. Reverse Transcriptase:

Retroviruses (possessing RNA) contain RNA dependent DNA polymerase which is called reverse transcriptase. This produces single stranded DNA, which in turn functions as template for complementary long chain of DNA.

This enzyme is used to synthesize the copy DNA or complementary DNA (cDNA) by using mRNA as a template. The enzyme is very useful for the synthesis of cDNA and construction of cDNA clone bank and to make short labelled probes.

Sticky and blunt ends

DNA ends refer to the properties of the end of DNA molecules, which may be sticky or blunt based on the enzyme which cuts the DNA. The restriction enzyme belong to a larger class of enzymes called **exonucleases and endonucleases**. Exonucleases remove nucleotide from ends whereas endonuclease cuts at specific position within the DNA.

The concept is used in [molecular biology](#), especially in [cloning](#) or when subcloning inserts DNA into [vector DNA](#). Such ends may be generated by [restriction enzymes](#) that cut the DNA – a staggered cut generates two sticky ends, while a straight cut generates blunt ends.



Single-stranded DNA molecules

A single-stranded non-circular DNA molecule has two non-identical ends, [the 3' end and the 5' end](#) (usually pronounced "three prime end" and "five prime end"). The numbers refer to the numbering of carbon atoms in the [deoxyribose](#), which is a sugar forming an important part of the backbone of the DNA molecule. In the backbone of DNA the 5' carbon of one deoxyribose is linked to the 3' carbon of another by a phosphodiester bond linkage. The 5' carbon of this deoxyribose is again linked to the 3' carbon of the next, and so forth.

Variations in double-stranded molecules

When a molecule of DNA is double stranded, as DNA usually is, the two strands run in opposite directions. Therefore, one end of the molecule will have the 3' end of strand 1 and the 5' end of strand 2, and vice versa in the other end. However, the fact that the molecule is two stranded allows numerous different variations.

Blunt ends

The simplest DNA end of a double stranded molecule is called a *blunt end*. Blunt end otherwise called as non cohesive restriction enzyme. In a blunt-ended molecule both strands terminate in a [base pair](#). Blunt ends are not always desired in biotechnology since when using a [DNA ligase](#) to join two molecules into one, the yield is significantly lower with blunt ends. When performing subcloning, it also has the disadvantage of potentially inserting the insert DNA in the opposite orientation desired. On the other hand, blunt ends are always compatible with each other. Here is an example of a small piece of blunt-ended DNA:

```
5'-CTGATCTGACTGATGCGTATGCTAGT-3'  
3'-GACTAGACTGACTACGCATACGATCA-5'
```

Overhangs and sticky ends

Non-blunt ends are created by various *overhangs*. An overhang is a stretch of unpaired [nucleotides](#) in the end of a DNA molecule. These unpaired nucleotides can be in either strand, creating either 3' or 5' overhangs. These overhangs are in most cases palindromic.

The simplest case of an overhang is a single nucleotide. This is most often [adenosine](#) and is created as a 3' overhang by some [DNA polymerases](#). Most commonly this is used in cloning [PCR](#) products created by such an enzyme. The product is joined with a linear DNA molecule with 3' [thymine](#) overhangs. Since adenine and thymine form a [base pair](#), this facilitates the joining of the two molecules by a ligase, yielding a circular molecule. Here is an example of an A-overhang:

5'-ATCTGACTA-3'
3'-TAGACTGA-5'

Longer overhangs are called *cohesive ends* or *sticky ends*. They are most often created by **restriction endonucleases** when they cut DNA. Very often they cut the two DNA strands four base pairs from each other, creating a four-base 5' overhang in one molecule and a complementary 5' overhang in the other. These ends are called cohesive since they are easily joined back together by a ligase.

For example, these two "sticky" ends are compatible:

5'-ATCTGACT + GATGCGTATGCT-3'
3'-TAGACTGACTACG CATAACGA-5'

They can form complementary base pairs in the overhang region:

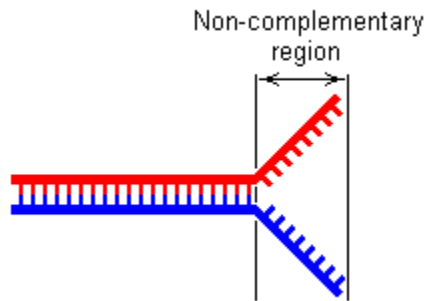
GATGCGTATGCT-3'
5'-ATCTGACT CATAACGA-5'
3'-TAGACTGACTACG

Also, since different restriction endonucleases usually create different overhangs, it is possible to create a plasmid by excising a piece of DNA (using a different enzyme for each end) and then joining it to another DNA molecule with ends trimmed by the same enzymes. Since the overhangs have to be complementary in order for the ligase to work, the two molecules can only join in one orientation. This is often highly desirable in **molecular biology**.

Frayed ends

Across from each single strand of DNA, we typically see **adenine** pair with **thymine**, and cytosine pair with **guanine** to form a parallel complementary strand as described below. Two nucleotide sequences which correspond to each other in this manner are referred to as complementary:

5'-ATCTGACT-3'
3'-TAGACTGA-5'



A frayed end refers to a region of a double stranded (or other multi-stranded) DNA molecule near the end with a significant proportion of non-complementary sequences; that is, a sequence where nucleotides on the adjacent strands do not match up correctly:

5'-ATCTGACTAGGCA-3'
3'-TAGACTGACTACG-5'

The term "frayed" is used because the incorrectly matched nucleotides tend to avoid bonding, thus appearing similar to the strands in a fraying piece of rope.

Although non-complementary sequences are also possible in the middle of double stranded DNA, mismatched regions away from the ends are not referred to as "frayed".

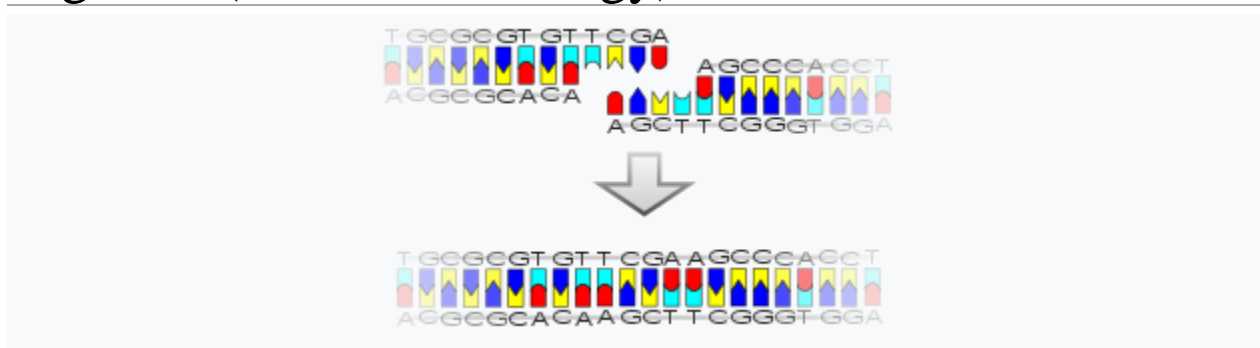
Discovery

[Ronald W. Davis](#) first discovered sticky ends as the product of the action of [EcoRI](#), the restriction [endonuclease](#).

Strength

Sticky end links are different in their stability. Free energy of formation can be measured to estimate stability. Free energy approximations can be made for different sequences from data related to oligonucleotide UV thermal denaturation curves. Also predictions from molecular dynamics simulations show that some sticky end links are much stronger in stretch than the others.

Ligation (molecular biology)



A sticky end ligation

In **molecular biology**, **ligation** is the joining of two nucleic acid fragments through the action of an enzyme. It is an essential laboratory procedure in the **molecular cloning** of DNA whereby DNA fragments are joined together to create **recombinant DNA** molecules, such as when a foreign DNA fragment is inserted into a **plasmid**. The ends of DNA fragments are joined together by the formation of phosphodiester bonds between the 3'-hydroxyl of one DNA terminus with the 5'-phosphoryl of another. RNA may also be ligated similarly. A co-factor is generally involved in the reaction, and this is usually ATP or NAD⁺.

in the laboratory is normally performed using **T4 DNA ligase**, however, procedures for ligation without the use of standard DNA ligase are also popular.



Ligation reaction

The mechanism of the ligation reaction was first elucidated in the laboratory of I. Robert Lehman. Two fragments of DNA may be joined together by **DNA ligase** which catalyzes the formation of a phosphodiester bond between the 3'-OH at one end of a strand of DNA and the 5'-phosphate group of another. In animals and **bacteriophage**, **ATP** is used as the energy source for the ligation, while In bacteria, **NAD⁺** is used.

The DNA ligase first reacts with ATP or NAD⁺, forming a ligase-AMP intermediate with the **AMP** linked to the ε-amino group of lysine in the active site of the ligase via a phosphoamide bond. This **adenylyl** group is then transferred to the phosphate group at the 5' end of a DNA chain, forming a DNA-adenylate complex. Finally, a phosphodiester bond between the two DNA ends is formed via the nucleophilic attack of the 3'-hydroxyl at the end of a DNA strand on the activated 5'-phosphoryl group of another.

A **nick in the DNA** (i.e. a break in one strand of a double-stranded DNA) can be repaired very efficiently by the ligase. However, a complicating feature of ligation presents itself when ligating two separate DNA ends as the two ends need to come together before the ligation reaction can proceed.

In the ligation of DNA with **sticky or cohesive ends**, the protruding strands of DNA may be annealed together already, therefore it is a relatively efficient process as it is equivalent to repairing two nicks in the DNA. However, in the ligation of **blunt-ends**, which lack protruding ends for the DNA to anneal together, the process is dependent on random collision for the ends to align together before they can be ligated, and is consequently a much less efficient process. The DNA ligase from *E. coli* cannot ligate blunt-ended DNA except under conditions of molecular crowding, and it is therefore not normally used for ligation in the laboratory. Instead the DNA ligase from **phage T4** is used as it can ligate blunt-ended DNA as well as single-stranded DNA.

Factors affecting ligation

Factors that affect an enzyme-mediated chemical reaction would naturally affect a ligation reaction, such as the concentration of enzyme and the reactants, as well as the temperature of reaction and the length of time of incubation. Ligation is complicated by the fact that the desired ligation products for most ligation reactions should be between two different DNA molecules and the reaction involves both inter- and intra-molecular reactions, and that an additional annealing step is necessary for efficient ligation.

The three steps to form a new phosphodiester bond during ligation are: enzyme adenylylation, adenylyl transfer to DNA, and nick sealing. $Mg(2+)$ is a cofactor for catalysis, therefore at high concentration of $Mg(2+)$ the ligation efficiency is high. If the concentration of $Mg(2+)$ is limited, the nick-sealing is the rate-limiting reaction of the process, and adenylylated DNA intermediate stays in the solution. Such adenylylation of the enzyme restrains the rebinding to the adenylylated DNA intermediate comparison of an Achilles' heel of *LIG1*, and represents a risk if they are not fixed.

DNA concentration

The concentration of DNA can affect the rate of ligation, and whether the ligation is an inter-molecular or intra-molecular reaction. Ligation involves joining up the ends of a DNA with other ends, however, each DNA fragment has two ends, and if the ends are compatible, a DNA molecule can circularize by joining its own ends. At high DNA concentration, there is a greater chance of one end of a DNA molecule meeting the end of another DNA, thereby forming intermolecular ligation.

At a lower DNA concentration, the chance that one end of a DNA molecule would meet the other end of the same molecule increases, therefore **intramolecular reaction** that circularizes the DNA is more likely. The **transformation efficiency** of linear DNA is also much lower than circular DNA, and for the DNA to circularize, the DNA concentration should not be too high. As a general rule, the total DNA concentration should be less than 10 $\mu\text{g/ml}$.

The relative concentration of the DNA fragments, their length, as well as buffer conditions are also factors that can affect whether intermolecular or intramolecular reactions are favored.

The concentration of DNA can be artificially increased by adding condensing agents such as **cobalt hexamine** and biogenic **polyamines** such as **spermidine**, or by using **crowding agents** such as **polyethylene glycol** (PEG) which also increase the effective concentration of enzymes. Note however that additives such as cobalt hexamine can produce exclusively intermolecular reaction, resulting in linear **concatemers** rather than the circular DNA more suitable for **transformation** of plasmid DNA, and is therefore undesirable for plasmid ligation. If it is necessary to use additives in plasmid ligation, the use of PEG is preferable as it can promote intramolecular as well as intermolecular ligation.

Ligase concentration

The higher the ligase concentration, the faster the rate of ligation. Blunt-end ligation is much less efficient than sticky end ligation, so a higher concentration of ligase is used in blunt-

end ligations. High DNA ligase concentration may be used in conjunction with PEG for a faster ligation, and they are the components often found in commercial kits designed for rapid ligation.

Temperature

Two issues are involved when considering the temperature of a ligation reaction. First, the optimum temperature for DNA ligase activity which is 37°C, and second, the **melting temperature** (T_m) of the DNA ends to be ligated.

The melting temperature is dependent on length and base composition of the DNA overhang—the greater the number of G and C, the higher the T_m since there are three hydrogen bonds formed between G-C base pair compared to two for A-T base pair—with some contribution from the stacking of the bases between fragments.

For the ligation reaction to proceed efficiently, the ends should be stably annealed, and in ligation experiments, the T_m of the DNA ends is generally much lower than 37°C. The optimal temperature for ligating cohesive ends is therefore a compromise between the best temperature for DNA ligase activity and the T_m where the ends can associate.

However, different restriction enzymes generates different ends, and the base composition of the ends produced by these enzymes may also differ, the melting temperature and therefore the optimal temperature can vary widely depending on the restriction enzymes used, and the optimum temperature for ligation may be between 4-15°C depending on the ends.

Ligations also often involve ligating ends generated from different restriction enzymes in the same reaction mixture, therefore it may not be practical to select optimal temperature for a particular ligation reaction and most protocols simply choose 12-16°C, room temperature, or 4°C.

Buffer composition

The **ionic strength** of the buffer used can affect the ligation. The kinds of cations presence can also influence the ligation reaction, for example, excess amount of Na^+ can cause the DNA to become more rigid and increase the likelihood of intermolecular ligation. At high concentration of monovalent cation (>200 mM) ligation can also be almost completely inhibited. The standard buffer used for ligation is designed to minimize ionic effects.

Sticky-end ligation

Restriction enzymes can generate a wide variety of ends in the DNA they digest, but in cloning experiments most commonly-used restriction enzymes generate a 4-base single-stranded overhang called the sticky or cohesive end (exceptions include *NdeI* which generates a 2-base overhang, and those that generate blunt ends). These sticky ends can anneal to other compatible ends and become ligated in a sticky-end (or cohesive end) ligation. *EcoRI* for example generates an AATT end, and since A and T have lower melting temperature than C and G, its melting temperature T_m is low at around 6°C.

For most restriction enzymes, the overhangs generated have a T_m that is around 15°C. For practical purposes, sticky end ligations are performed at 12-16°C, or at room temperature, or alternatively at 4°C for a longer period.

For the insertion of a DNA fragment into a plasmid vector, it is preferable to use two different restriction enzymes to digest the DNA so that different ends are generated. The two

different ends can prevent the religation of the vector without any insert, and it also allows the fragment to be inserted in a directional manner.

When it is not possible to use two different sites, then the vector DNA may need to be dephosphorylated to avoid a high background of recircularized vector DNA with no insert. Without a phosphate group at the ends the vector cannot ligate to itself, but can be ligated to an insert with a phosphate group. Dephosphorylation is commonly done using [calf-intestinal alkaline phosphatase](#) (CIAP) which removes the phosphate group from the 5' end of digested DNA, but note that CIAP is not easy to inactivate and can interfere with ligation without an additional step to remove the CIAP, thereby resulting in failure of ligation. CIAP should not be used in excessive amount and should only be used when necessary. Shrimp [alkaline phosphatase](#) (SAP) or Antarctic phosphatase (AP) are suitable alternative as they can be easily inactivated.

Blunt-end ligation

Blunt end ligation does not involve base-pairing of the protruding ends, so any blunt end may be ligated to another blunt end. Blunt ends may be generated by restriction enzymes such as [SmaI](#) and [EcoRV](#). A major advantage of blunt-end cloning is that the desired insert does not require any restriction sites in its sequence as blunt-ends are usually generated in a [PCR](#), and the [PCR](#) generated blunt-ended DNA fragment may then be ligated into a blunt-ended vector generated from restriction digest.

[Blunt-end](#) ligation, however, is much less efficient than sticky end ligation, typically the reaction is 100X slower than sticky-end ligation. Since blunt-end does not have protruding ends, the ligation reaction depends on random collisions between the blunt-ends and is consequently much less efficient. To compensate for the lower efficiency, the concentration of ligase used is higher than sticky end ligation (10x or more). The concentration of DNA used in blunt-end ligation is also higher to increase the likelihood of collisions between ends, and longer incubation time may also be used for blunt-end ligations.

If both ends needed to be ligated into a vector are blunt-ended, then the vector needs to be dephosphorylated to minimize self-ligation. This may be done using CIAP, but caution in its use is necessary as noted previously. Since the vector has been dephosphorylated, and ligation requires the presence of a 5'-phosphate, the insert must be phosphorylated. Blunt-ended PCR product normally lacks a 5'-phosphate, therefore it needs to be phosphorylated by treatment with [T4 polynucleotide kinase](#).

Blunt-end ligation is also reversibly inhibited by high concentration of ATP.

PCR usually generates blunt-ended PCR products, but note that PCR using [Taq polymerase](#) can add an extra adenine (A) to the 3' end of the PCR product. This property may be exploited in [TA cloning](#) where the ends of the PCR product can anneal to the T end of a vector. [TA ligation](#) is therefore a form of sticky end ligation. Blunt-ended vectors may be turned into vector for TA ligation with dideoxythymidine triphosphate (ddTTP) using terminal transferase.

General guidelines

For the cloning of an insert into a circular plasmid:

- The total DNA concentration used should be less than 10 µg/ml as the plasmid needs to recircularize.
- The molar ratio of insert to vector is usually used at around 3:1. Very high ratio may produce multiple inserts. The ratio may be adjusted depending on the size of the insert, and other ratios may be used, such as 1:1.

Trouble-shooting

Sometimes ligation fail to produce the desired ligated products, and some of the possible reasons may be:

- Damaged DNA – Over-exposure to UV radiation during preparation of DNA for ligation can damage the DNA and significantly reduce [transformation efficiency](#). A higher-wavelength UV radiation (365 nm) which cause less damage to DNA should be used if it is necessary work for work on the DNA on a UV transilluminator for an extended period of time. Addition of [cytidine](#) or [guanosine](#) to the electrophoresis buffer at 1 mM concentration however may protect the DNA from damage.
- Incorrect usage of CIAP or its inefficient inactivation or removal.
- Excessive amount of DNA used.
- Incomplete [DNA digest](#) – The vector DNA that is incompletely digested will give rise to a high background, and this may be checked by doing a ligation without insert as a control. Insert that is not completely digested will also not ligate properly and circularize. When digesting a PCR product, make sure that sufficient extra bases have been added to the 5'-ends of the oligonucleotides used for PCR as many restriction enzymes require a minimum number of extra basepairs for efficient digest. The information on the minimum basepair required is available from restriction enzyme suppliers such as in the catalog of [New England Biolabs](#).
- Incomplete ligation – Blunt-ends DNA (e.g. *SmaI*) and some sticky-ends DNA (e.g. *NdeI*) that have low-melting temperature require more ligase and longer incubation time.
- Protein expressed from ligated gene insert is toxic to cells.
- Homologous sequence in insert to sequence in plasmid DNA resulting in deletion.

Other methods of DNA ligation

A number of commercially available DNA cloning kits use other methods of ligation that do not require the use of the usual DNA ligases. These methods allow cloning to be done much more rapidly, as well as allowing for simpler transfer of cloned DNA insert to different [vectors](#). These methods however require the use of specially designed [vectors](#) and components, and may lack flexibility.

Topoisomerase-mediated ligation

[Topoisomerase](#) can be used instead of ligase for ligation, and the cloning may be done more rapidly without the need for restriction digest of the vector or insert. In this [TOPO cloning](#) method a linearized vector is activated by attaching topoisomerase I to its ends, and this "TOPO-activated" vector may then accept a PCR product by ligating to both of the 5' ends of the PCR product, the topoisomerase is released and a circular vector is formed in the process.

Homologous recombination

Another method of cloning without the use of ligase is by **DNA recombination**, for example as used in the **Gateway cloning system**. The gene, once cloned into the cloning vector (called entry clone in this method), may be conveniently introduced into a variety of expression vectors by recombination.

Have you ever wished you could snag individual strands of DNA or RNA with a lasso? Or look at them one by one, figuring out exactly where they are or what they are doing? Fortunately, there are techniques that exist to label nucleic acids for their visualization and purification! Nucleic acids can be labeled at their 5' end, 3' end, or throughout the molecule depending on the particular application, including:

- to generate information on gene integrity and copy number (blot)
- to diagnose specific sequences and chromosomal aberrations (in situ hybridization)
- to simultaneously measure the relative expression of RNAs (microarray analysis)
- to discover protein-nucleic acid interactions (electrophoretic mobility shift assays or FRET)

Labels, Labels, Everywhere

There are two types of nucleic acid labeling techniques: radioisotope labeling and non-radioactive labeling.

1. Radioisotope labeling: Considered as a conventional method for nucleic acid labeling, radiolabeled nucleotides are synthesized using ATP-gamma-³²P or ³⁵P. They are easily incorporated into nucleic acid sequences by traditional enzymatic means or by an organism of interest.

Radioactive nucleotides were first used in 1935 by George de Hevesy (Nobel Laureate in Chemistry, 1943) to reveal components of metabolism in rats. Since then, radioisotope labeling of nucleotides has been used in many studies and clinical applications, such as investigations into bacteriophage replication and clinical diagnosis of cancers.

There is a long history of radiolabeling for DNA and RNA applications. While this technique is relatively less expensive than non-radioactive labeling and is widely used, it is important to review the safety concerns of working with radioactive nucleotides in the lab. Furthermore, if you're looking to get single-molecule level precision in your application, you may want to consider non-radioactive labeling.

2. Non-radioactive (chemical) labeling: Nowadays, non-radioactive nucleotide labels are more extensively used due to their relative speed, sensitivity, safety, and versatility. The most common labels are fluorescent ‘tags’ that are synthesized and incorporated into oligonucleotides, but you can also attach a variety of other molecules or proteins to chemically reactive groups like biotin, streptavidin, or fluorophores. Pre-labeled oligos are available from most oligo suppliers like as IDT DNA or Genewiz.

Non-radioactive labels for DNA and RNA are widely used in molecular biology labs. Fluorescent and reactive labels help researchers investigate proteins that interact with nucleotides at a single molecule level (e.g. FRET). Check out the table below summarizing chemical methods for nucleic acid labeling!

While slightly more expensive than radiolabeling, fluorescent and chemically labeled nucleotides and oligos are easy to use for a variety of applications in the lab. Plus, this class of modified nucleotides don’t require extensive training in radioactive isotope handling.

Template	Methods		Reagent	Label	Labeled Probe	Labeling site
DNA	Enzymatic	PCR	Tag Polymerase	dNTP	DNA	random
		Nick translation	DNase I / DNA Polymerase I	dNTP	DNA	random
		Primer extension	Klenow fragment	dNTP	DNA	random
		5'-end labeling	T4 Polynucleotide kinase	γ - ³² P rATP	ssDNA	5'-OH
		3'-end labeling	TdT	dNTP	ssDNA	3'-OH
		In vitro transcription	SP6 RNA Polymerase	NTP	crRNA	random
	T7 RNA Polymerase		NTP	crRNA	5'-OH	
	Chemical	Conjugation	EDC	amines & carboxylate / phosphate	DNA	5'-OH
		Photoreaction	Nonspecific cross-linkers	nucleotide bases	DNA	random
RNA	Enzymatic	Reverse transcription	AMV/MMLV Reverse transcriptase	dNTP	cDNA	random
		5'-end labeling	T4 Polynucleotide kinase	γ - ³² P rATP	RNA	5'-OH
		3'-end labeling	T4 RNA Ligase	[5'- ³² P]pCp	RNA	3'-OH
	Chemical	Oxidation	Periodate	amine / hydrazide	RNA	3'-OH
		Conjugation	EDC	amines & carboxylate / phosphate	RNA	5'-OH
		Photoreaction	Nonspecific cross-linkers	nucleotide bases	RNA	random

Overview of nucleic acid labeling methods

Choosing a Labeling Method

When choosing a labeling system, consider the size and type of nucleic acid you're working with. Large DNA, plasmid DNA, and RNA for blots and *in situ* hybridization can be labeled throughout by random incorporation of a covalently coupled label.

While covalent probes produce excellent sensitivity, enzymatic methods are more economically convenient in the lab and can label copies of the sample (in PCR labeling). For shorter sequences like oligos, it may be more convenient to order a pre-labeled sample through a company that specializes in labeled DNA, like IDT or Genewiz.

Digoxigenin (DIG) Labeling Methods

- [Digoxigenin \(DIG\) Labeling and Anti-DIG Antibody](#)
- [DIG DNA Labeling by PCR](#)
- [DIG Random Primed DNA Labeling](#)
- [Nick Translation Labeling of dsDNA for *In Situ* Probes](#)
- [Transcriptional Labeling of RNA Probes](#)
- [DIG Oligonucleotide 5' End-Labeling, 3' End-Labeling, and 3' Tail Labeling](#)
- [Estimation of Probe Yield by the Direct Detection Procedure](#)
- [DIG-Related Downloads and Resources](#)

Digoxigenin (DIG) Labeling and Anti-DIG Antibody

The DIG System is the nonradioactive technology of choice to label and detect nucleic acids for multiple applications. The system is based on a steroid isolated from digitalis plants (*Digitalis purpurea* and *Digitalis lanata*). These plants are the only natural source of digoxigenin, so the anti-DIG antibody does not bind to other biological material, ensuring specific labeling.

Due to this high specificity, less material is needed compared to radioactive labeling making the DIG system ideal for nucleic hybridization analysis. Immobilized nucleic acids are hybridized with a DIG-labeled probe and subsequent detection is performed using high affinity **Anti-Digoxigenin antibodies**, coupled either to alkaline phosphatase (AP), horseradish peroxidase (HRP), fluorescein or rhodamine for colorimetric, and chemiluminescent or fluorescent detection.

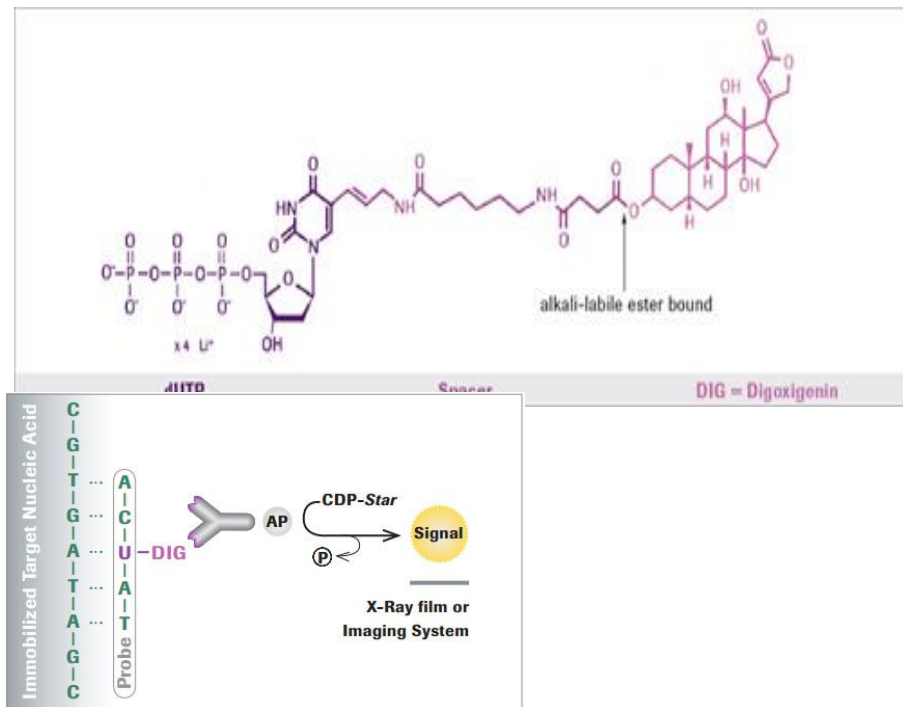


Figure 1: Example detecting DIG-labeled nucleic acids using chemiluminescence substrates.

DIG DNA Labeling by PCR

PCR labeling is the preferred method for preparing DIG-labeled probes when the template is available in only limited amounts, is partially purified, or is very short. It requires less optimization than other methods and produces a high yield of labeled probe.

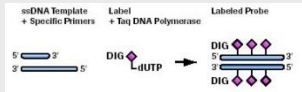
In PCR labeling, a thermostable polymerase incorporates DIG-dUTP as it amplifies a specific region of the template DNA. The result is a highly labeled, specific, and extremely sensitive hybridization probe.

Reaction principle

During a standard PCR reaction, Digoxigenin-11-dUTP is incorporated into newly synthesized DNA. The only prerequisite is that some sequence information of the target sequence is needed in order to synthesize the appropriate primers.

The nonradioactive DIG system uses digoxigenin, a steroid hapten, to label DNA, RNA, or oligonucleotides for hybridization, and subsequent color or luminescence detection. The digoxigenin is coupled to dUTP via an alkali-labile ester bond. The labeled dUTP can be easily incorporated by enzymatic nucleic acid synthesis using DNA polymerases.

The combination of nonradioactive labeling with PCR is a powerful tool for the analysis of PCR products, and for the preparation of labeled probes from small amounts of a respective target sequence.

Labeling Methods	DIG-Labeling Reagents	Other Labeling Reagents	
PCR Labeling 	Kits for labeling	<ul style="list-style-type: none"> • PCR DIG Probe Synthesis Kit • PCR ELISA, DIG Labeling 	
	Mixes for labeling without enzymes	<ul style="list-style-type: none"> • PCR DIG Labeling Mix • PCR DIG Labeling MixPLUS 	
	Nucleotides for labeling	<ul style="list-style-type: none"> • Digoxigenin-11-dUTP, alkalistable • Digoxigenin-11-dUTP, alkalilabile • Deoxynucleoside Triphosphate Set 	<ul style="list-style-type: none"> • Biotin-16-dUTP • Fluorescein-12-dUTP • TetramethylRhodamine 5-dUTP
	Enzymes	<ul style="list-style-type: none"> • Expand High FidelityPLUS PCR System 	<ul style="list-style-type: none"> • Expand High FidelityPLUS PCR System

DIG DNA labeling by PCR features and benefits

PCR conditions

- Optimize PCR amplification parameters (cycling conditions, template concentration, primer sequence, and primer concentration) for each template and primer set in the absence of DIGdUTP before attempting incorporation of DIG.

Template

- For best results, use cloned inserts as template. Genomic DNA can be more difficult to use.

- Template concentration is crucial to successful production of specific probes.

Labeling

The **PCR DIG Probe Synthesis Kit** requires less optimization than most labeling methods, because it contains the Expand High Fidelity Enzyme Blend. This enzyme blend benefits include:

- Can efficiently use GC-rich regions as template
- For most templates, requires no optimization of MgCl₂ concentration and labeling reactions will work with the standard concentrations of 1.5 mM MgCl₂

DIG Random Primed DNA Labeling

The method of "random primed" DNA labeling is based on the hybridization of a mixture of all possible hexanucleotides to the DNA template. All sequence combinations are represented in the hexanucleotide primer mixture, which leads to binding of primer to the template DNA in a statistic manner.

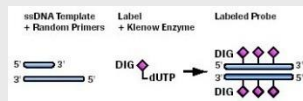
Thus, an equal degree of labeling along the entire length of the template DNA is guaranteed. The complementary strand is synthesized from the 3' OH termini of the random hexanucleotide primer using Klenow enzyme, labeling grade.

Modified deoxyribonucleoside triphosphates ([³²P]-, [³⁵S]-, [³H]-, [¹²⁵I]-, digoxigenin or biotinlabeled) present in the reaction are incorporated into the newly synthesized complementary DNA strand.

These labeled probes are especially suitable for single copy gene detection on genomic Southern blots, screening recombinant libraries, dot/slot blots, and northern blots. Since each primer has a different six-base sequence, the labeled probe product will be a collection of fragments of variable length.

Thus, the labeled probe will appear as a smear, rather than a unique band on a gel. The size distribution of the labeled probe depends on the length of the original template.

Labeling Methods	DIG-Labeling Reagents	Other Labeling Reagents
Random Priming	Kits for labeling	• DIG High Prime



and detection

Labeling and Detection Starter Kit I

- DIG High Prime Labeling and Detection Starter Kit II
- DIG DNA Labeling and Detection Kit

Kits for labeling	<ul style="list-style-type: none"> • DIG DNA Labeling Kit 	<ul style="list-style-type: none"> • Random Primer DNA Labeling Kit
Mixes for labeling without enzymes	<ul style="list-style-type: none"> • DIG-High Prime • DIG DNA Labeling Mix 	<ul style="list-style-type: none"> • High Prime • Biotin-High Prime • Fluorescein-High Prime
Nucleotides for labeling	<ul style="list-style-type: none"> • Digoxigenin-11-dUTP, alkalilabile • Digoxigenin-11-dUTP, alkalilabile 	<ul style="list-style-type: none"> • Hexanucleotide Mix • Biotin-16-dUTP • Fluorescein-12-dUTP • TetramethylRhodamine-5-dUTP
Enzymes	<ul style="list-style-type: none"> • Klenow Enzyme, labeling grade 	<ul style="list-style-type: none"> • Klenow Enzyme, labeling grade
Additional Products	<ul style="list-style-type: none"> • Primer “random” 	<ul style="list-style-type: none"> • Primer “random”

Reaction principle

In random primed labeling, Klenow enzyme copies DNA template in the presence of hexamer primers and alkalilabile DIG-11-dUTP. On average, the enzyme inserts one DIG moiety in every stretch of 20-25 nucleotides. The resulting labeled product is a homogeneously labeled, sensitive hybridization probe capable of detecting as little as 0.10 – 0.03 pg target DNA.

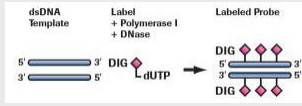
Nick Translation Labeling of dsDNA for *In Situ* Probes

The nick translation method is based on the ability of DNase I to introduce randomly distributed nicks into DNA at low enzyme concentrations in the presence of Mg²⁺. *E. coli* DNA polymerase I synthesizes DNA

complementary to the intact strand in a 5' - 3' direction using the 3'-OH termini of the nick as a primer.

The 5' - 3' exonucleolytic activity of DNA Polymerase I simultaneously removes nucleotides in the direction of synthesis. The polymerase activity sequentially replaces the removed nucleotides with isotope-labeled or hapten-labeled deoxyribonucleoside triphosphates. At low temperature (+15°C), the unlabeled DNA in the reaction is thus replaced by newly synthesized labeled DNA.

For *in situ* hybridization procedures, the length of the labeled fragments obtained from this procedure should be about 200-500 bases.

Labeling Methods	DIG-Labeling Reagents	Other Labeling Reagents
Nick Translation 	Kits for labeling	<ul style="list-style-type: none"> • Nick Translation Kit
	Mixes for labeling without enzymes	<ul style="list-style-type: none"> • DIG-Nick Translation Mix
	Nucleotides for labeling	<ul style="list-style-type: none"> • Digoxigenin-11-dUTP, alkalistable • Digoxigenin-11-dUTP, alkalilabile • Deoxynucleoside Triphosphate Set
	Enzymes	<ul style="list-style-type: none"> • DNA Polymerase I • DNase I recombinant, RNase-free

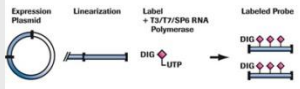
Transcriptional Labeling of RNA Probes

For some applications, DIG-labeled RNA is a more effective hybridization probe than DIG-labeled DNA. For example, DIG-labeled RNA probes can detect rare mRNAs in nanogram amounts of total RNA. These

labeled RNA probes are generated by *in vitro* transcription from a DNA template.

In the RNA transcription method, DNA is cloned into the multiple cloning site of a transcription vector between promoters for different RNA polymerases (such as T7, SP6, or T3 RNA polymerase). The template is then linearized by cleavage of the vector at a unique site (near the insert). An RNA polymerase transcribes the insert DNA into an antisense RNA copy in the presence of a mixture of ribonucleotides (including DIG-UTP).

During the reaction, the DNA can be transcribed many times (up to a hundredfold) to generate a large amount of full-length DIG-labeled RNA copies (10-20 µg RNA from 1 µg DNA in a standard reaction). DIG is incorporated into the RNA at approximately every 25-30 nucleotides.

Labeling Methods	DIG-Labeling Reagents	Other Labeling Reagents
<p><i>In Vitro</i> Transcription</p> 	Kits for labeling	<ul style="list-style-type: none"> • DIG Northern Starter Kit • DIG RNA Labeling Kit (SP6/T7)
	Mixes for labeling without enzymes	<ul style="list-style-type: none"> • DIG RNA Labeling Mix
	Nucleotides for labeling	<ul style="list-style-type: none"> • Digoxigenin-11- UTP
	Enzymes	<ul style="list-style-type: none"> • SP6 RNA Polymerase • T3 RNA Polymerase • T7 RNA Polymerase
	Additional Products	<ul style="list-style-type: none"> • Protector RNase Inhibitor
		<ul style="list-style-type: none"> • SP6/T7 Transcription Kit • Biotin RNA Labeling Mix • Fluorescein RNA Labeling Mix • Biotin-16-UTP • Fluorescein-16-UTP • Biotin-11-UTP • SP6 RNA Polymerase • T3 RNA Polymerase • T7 RNA Polymerase • Protector RNase Inhibitor

Reaction principle

The DNA template to be transcribed is cloned into the polylinker site of an appropriate transcription vector, which contains promoters for SP6 and or T3 and T7 RNA polymerases. After linearization at a suitable site, RNA is transcribed in the presence of DIG-11-UTP. Under standard conditions, approximately 10 µg of full-length DIG-labeled RNA are transcribed from 1 µg of template.

The following tips are critical for successful RNA probe labeling:

RNases

RNases are ubiquitous and do not require any cofactors for activity. If you want to be successful, take all possible precautions to prevent RNase contamination. For instance:

- It is recommended to use disposable plasticware, oven-baked glassware, or plasticware that has been decontaminated with RNase ZAP or similar reagents.
- Prepare all solutions with water that has been treated with diethyl-pyrocabonate (DEPC) or dimethyldicarbonate (DMDC) and autoclave the solutions.
- Wear gloves throughout the procedure.
- Labeling efficiency depends greatly on the purity of the DNA template. Template should be highly purified.
- The final template must be linearized, phenol/chloroform extracted, and ethanol precipitated.

Template sequence

- Some primer and/or polylinker regions in DNA templates are homologous to portions of the ribosomal 28s and 18s RNA sequences. Therefore, labeled probes may generate specific, but unwanted signals in samples that contain these prominent RNAs. To minimize this effect, remove as much of the polylinker sequences from your template as possible.
- If you use PCR to make the DNA template, the product of the Expand High Fidelity reaction contains some fragments with a single 3' A overhang. This overhang may produce wraparound products in the transcriptional labeling reaction.

Template length

- Optimal template length is approximately 1 kb
- Minimum length should be at least 200 bp

Storage of probe

- For long term stability, RNA probes should be aliquoted and stored at -20° C or -70° C
- DIG-labeled RNA probes are stable for at least 1 year at -20° C or -70° C in ethanol

Probe sensitivity

- To quickly determine the sensitivity of a DIG-labeled antisense RNA probe, prepare the corresponding sense RNA (unlabeled) by *in vitro* transcription. Then use the purified sense transcript at varying concentrations as target on a northern blot. From the result of the blot you can easily determine the lowest amount of target (sense transcript) that can be detected by labeled probe (antisense transcript).

DIG Oligonucleotide 5' End-Labeling, 3' End-Labeling, and 3' Tail Labeling

For some applications, such as *in situ* hybridization, a DIG-labeled synthetic oligonucleotide is the best hybridization probe. In addition to *in situ* hybridizations, DIG-labeled oligonucleotides may be used as hybridization probes in numerous applications, including dot/slot blots, library screening, detection of repeated gene sequences on Southern blots, and detection of abundant mRNAs on northern blots.

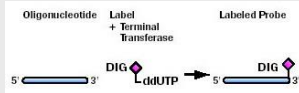
Several methods are available for DIG-labeling of oligonucleotides and are summarized below.

Oligonucleotide 5' end-labeling with DIG-NHS-Ester

Labeling Methods	DIG-Labeling Reagents	Other Lab Reagents
<p>5'-End Labeling</p>  <p>Nucleotide needed: 100 nmol Required time and temperature: overnight, +15 to +25°C</p> <ul style="list-style-type: none">Requires only a small amount of template	<p>Additional Products</p>	<ul style="list-style-type: none">Digoxigenin-3-Omethylcarboxy-ε-aminocaproic acid N-hydroxysuccinimide ester

Oligonucleotide 3' end-labeling with DIG-ddUTP

Labeling Methods	DIG-Labeling Reagents	Other Lab Reagents
<p>3'-End Labeling</p>	<p>Kits for labeling</p>	<ul style="list-style-type: none">DIG Oligonucleotide 3'-End Labeling Kit,



Nucleotide needed: 100 pmol
 Required time and temperature:
 15 minutes, +37°C

Can detect: 10 pg DNA

- Requires only a small amount of template
- Labeled probes can be used without purification
- Reaction can be scaled up indefinitely if you increase incubation time to 1 hr

Nucleotides for labeling

Enzymes

2nd generation
 • DIG Gel Shift Kit,
 2nd generation

• Digoxigenin-11
 ddUTP

• Biotin-16

• Terminal
 Transferase

• Terminal
 Transferase

Adding a 3' tail of DIG-dUTP and dATP (approximately 40 - 50 residues)

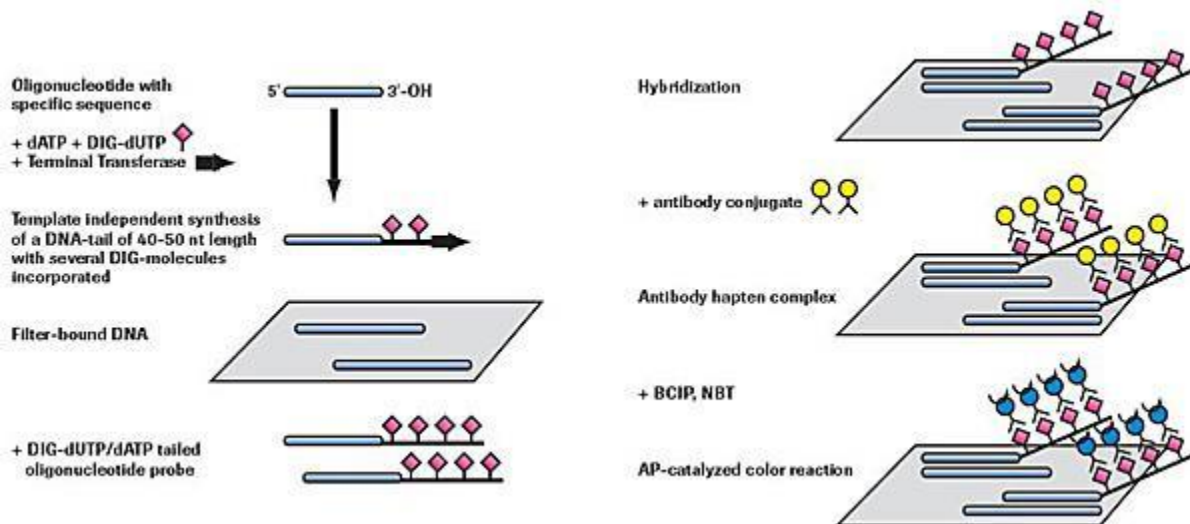
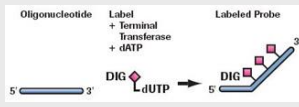


Figure 2: Nonradioactive oligonucleotide tailing and detection.

Labeling Methods	DIG-Labeling Reagents	Other Labeling Reagents
Tailing 	Nucleotides for labeling Enzymes	• Biotin-16-dUTP • Fluorescein-12-dUTP • Tetramethylrhodamine-5-dUTP • Terminal Transferase
Nucleotide needed: 100 pmol Required time and temperature: 15 minutes, +37°C		• Digoxigenin-11-dUTP, alkalistable • Digoxigenin-11-dUTP, alkalilabile • Terminal Transferase

Can detect: 1 pg DNA

- Requires only a small amount of template
- Produces more sensitive probes than end labeling
- Labeled probes can be used without purification
- Reaction can be scaled up indefinitely

Estimation of Probe Yield by the Direct Detection Procedure

To add the correct amount of probe to a hybridization, you must first determine the amount of DIG-labeled probe produced in the labeling reaction. The direct detection procedure given here compares the amount of DIG label in a series of dilutions prepared from the labeled probe with a known concentration of a DIG-labeled control nucleic acid.

Note: If you label a DNA probe by PCR, you do not need to perform a direct detection to evaluate the yield.

For PCR-labeled probes, use the gel electrophoresis evaluation method.

The direct detection involves the following steps:

- Preparing serial dilutions of labeled probe and spotting them on a nylon membrane (time required: 15 minutes)
- Detecting DIG in spots with chemiluminescence; time required (2-2.5 hours)

Nucleic acid hybridization

In molecular biology, **hybridization** (or **hybridisation**) is a phenomenon in which single-stranded deoxyribonucleic acid (**DNA**) or ribonucleic acid (**RNA**) molecules **anneal** to **complementary DNA or RNA**. Though a double-stranded DNA sequence is generally stable under physiological conditions, changing these conditions in the laboratory (generally by raising the surrounding temperature) will cause the molecules to separate into single strands.

These strands are complementary to each other but may also be complementary to other

sequences present in their surroundings. Lowering the surrounding temperature allows the single-stranded molecules to anneal or “hybridize” to each other.

DNA replication and transcription of DNA into RNA both rely upon nucleotide hybridization, as do molecular biology techniques including Southern blots and Northern blots,^[2] the polymerase chain reaction (PCR), and most approaches to DNA sequencing.



Hybridization is a basic property of nucleotide sequences and is taken advantage of in numerous molecular biology techniques. Overall, genetic relatedness of two species can be determined by hybridizing segments of their DNA (DNA-DNA hybridization). Due to sequence similarity between closely related organisms, higher temperatures are required to melt such DNA hybrids when compared to more distantly related organisms. A variety of different methods use hybridization to pinpoint the origin of a DNA sample, including the polymerase chain reaction (PCR).

In another technique, short DNA sequences are hybridized to cellular mRNAs to identify expressed genes. Pharmaceutical drug companies are exploring the use of antisense RNA to bind to undesired mRNA, preventing the ribosome from translating the mRNA into protein.

DNA-DNA hybridization

DNA–DNA hybridization generally refers to a molecular biology technique that measures the degree of genetic similarity between pools of DNA sequences. It is usually used to determine the genetic distance between two organisms. This has been used extensively in phylogeny and taxonomy.

Fluorescence In Situ Hybridization

Fluorescence *in situ* hybridization (FISH) is a molecular cytogenetic technique that uses fluorescent probes that bind to only those parts of a nucleic acid sequence with a high degree of sequence complementarity. It was developed by biomedical researchers in the early 1980s to detect and localize the presence or absence of specific DNA sequences on chromosomes. Fluorescence microscopy can be used to find out where the fluorescent probe is bound to the chromosomes. FISH is often used for finding specific features in DNA for use in genetic counseling, medicine, and species identification. FISH can also be used to detect and localize specific RNA targets (mRNA, lncRNA and miRNA) in cells, circulating tumor cells, and tissue samples. In this context, it can help define the spatial-temporal patterns of gene expression within cells and tissues.

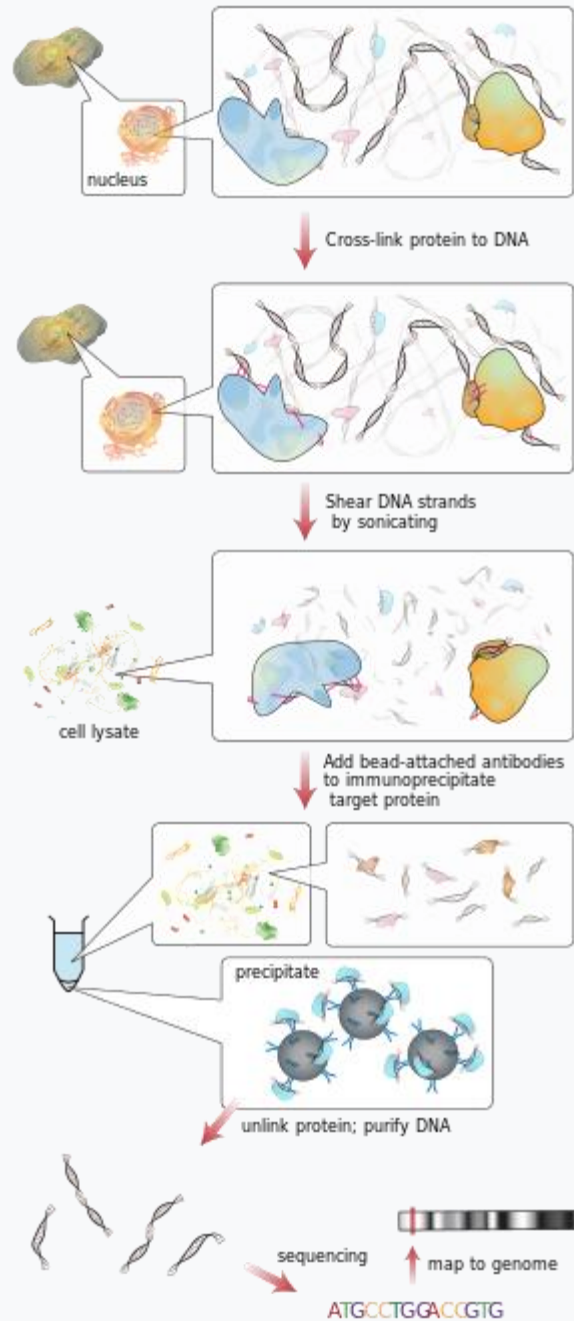


Fluorescence in situ hybridization (FISH) is a laboratory method used to detect and locate a

DNA sequence, often on a particular [chromosome](#).

In the 1960s, researchers Joseph Gall and Mary Lou Pardue found that molecular hybridization could be used to identify the position of DNA sequences *in situ* (i.e., in their natural positions within a chromosome). In 1969, the two scientists published a paper demonstrating that radioactive copies of a ribosomal DNA sequence could be used to detect complementary DNA sequences in the nucleus of a frog egg. Since those original observations, many refinements have increased the versatility and sensitivity of the procedure to the extent that *in situ* hybridization is now considered an essential tool in [cytogenetics](#).

Chromatin immunoprecipitation



ChIP-sequencing workflow

Chromatin immunoprecipitation (ChIP) is a type of **immunoprecipitation** experimental technique used to investigate the interaction between **proteins** and **DNA** in the cell. It aims to determine whether specific proteins are associated with specific genomic regions, such as **transcription factors** on **promoters** or other **DNA binding sites**, and possibly defining **cistromes**. ChIP also aims to determine the specific location in the genome that various **histone** modifications are associated with, indicating the target of the histone

modifiers.

Briefly, the conventional method is as follows:

1. DNA and associated proteins on **chromatin** in living cells or tissues are crosslinked (this step is omitted in Native ChIP).
2. The DNA-protein complexes (chromatin-protein) are then sheared into ~500 bp DNA fragments by **sonication** or nuclease digestion.
3. **Cross-linked** DNA fragments associated with the protein(s) of interest are selectively immunoprecipitated from the cell debris using an appropriate protein-specific antibody.
4. The associated DNA fragments are purified and their sequence is determined. Enrichment of specific DNA sequences represents regions on the genome that the protein of interest is associated with *in vivo*.



Typical ChIP

There are mainly two types of ChIP, primarily differing in the starting chromatin preparation. The first uses reversibly **cross-linked** chromatin sheared by **sonication** called cross-linked ChIP (XChIP). Native ChIP (NChIP) uses native chromatin sheared by **micrococcal nuclease** digestion.

Cross-linked ChIP (XChIP)

Cross-linked ChIP is mainly suited for mapping the DNA target of transcription factors or other chromatin-associated proteins, and uses reversibly **cross-linked** chromatin as starting material. The agent for reversible cross-linking could be **formaldehyde** or **UV light**. Then the cross-linked chromatin is usually sheared by sonication, providing fragments of 300 - 1000 **base pairs** (bp) in length.

Mild formaldehyde crosslinking followed by nuclease digestion has been used to shear the chromatin. Chromatin fragments of 400 - 500bp have proven to be suitable for ChIP assays as they cover two to three **nucleosomes**.

Cell debris in the sheared lysate is then cleared by sedimentation and protein-DNA complexes are selectively immunoprecipitated using specific **antibodies** to the protein(s) of interest. The antibodies are commonly coupled to **agarose**, **sepharose** or magnetic beads. Alternatively, chromatin-antibody complexes can be selectively retained and eluted by inert polymer discs.

The immunoprecipitated complexes (i.e., the bead-antibody-protein-target DNA sequence complex) are then collected and washed to remove non-specifically bound chromatin, the protein-DNA **cross-link** is reversed and proteins are removed by digestion with **proteinase K**. An **epitope**-tagged version of the protein of interest, or *in vivo* biotinylation can be used instead of antibodies to the native protein of interest.

The DNA associated with the complex is then purified and identified by **polymerase chain reaction** (PCR), **microarrays** (**ChIP-on-chip**), molecular cloning and sequencing, or direct

high-throughput sequencing (ChIP-Seq).

Native ChIP (NChIP)

Native ChIP is mainly suited for mapping the DNA target of histone modifiers. Generally, native chromatin is used as starting chromatin. As histones wrap around DNA to form nucleosomes, they are naturally linked. Then the chromatin is sheared by micrococcal nuclease digestion, which cuts DNA at the length of the linker, leaving nucleosomes intact and providing DNA fragments of one nucleosome (200bp) to five nucleosomes (1000bp) in length.

Thereafter, methods similar to XChIP are used for clearing the cell debris, immunoprecipitating the protein of interest, removing protein from the immunoprecipitated complex, and purifying and analyzing the complex-associated DNA.

Comparison of XChIP and NChIP

The major advantage for NChIP is antibody specificity. It is important to note that most antibodies to modified histones are raised against unfixed, synthetic peptide antigens and that the epitopes they need to recognize in the XChIP may be disrupted or destroyed by formaldehyde cross-linking, particularly as the cross-links are likely to involve lysine e-amino groups in the N-terminals, disrupting the epitopes. This is likely to explain the consistently low efficiency of XChIP protocols compare to NChIP.

But XChIP and NChIP have different aims and advantages relative to each other. XChIP is for mapping target sites of transcription factors and other chromatin associated proteins; NChIP is for mapping target sites of histone modifiers.

Table 1 Advantages and disadvantages of NChIP and XChIP

	XChIP	NChIP
Advantages	Suitable for transcriptional factors, or any other weakly binding chromatin associated proteins. Applicable to any organisms where native protein is hard to prepare	Testable antibody specificity Better antibody specificity as target protein naturally intact Better chromatin and protein recovery efficiency due to better antibody specificity
Disadvantages	Inefficient chromatin recovery due to antibody target protein epitope disruption May cause false positive result due to fixation of transient proteins to chromatin Wide range of chromatin shearing size due to random cut by sonication.	Usually not suitable for non-histone proteins Nucleosomes may rearrange during digestion

History and New ChIP methods

In 1984 **John T. Lis** and David Gilmour, at the time a graduate student in the Lis lab, used UV irradiation, a zero-length protein-nucleic acid crosslinking agent, to covalently **cross-link** proteins bound to DNA in living bacterial cells. Following lysis of cross-linked cells and immunoprecipitation of bacterial RNA polymerase, DNA associated with enriched RNA polymerase was hybridized to probes corresponding to different regions of known genes to determine the in vivo distribution and density of RNA polymerase at these genes.

A year later they used the same methodology to study distribution of eukaryotic **RNA polymerase II** on fruit fly heat shock genes. These reports are considered the pioneering studies in the field of chromatin immunoprecipitation. XChIP was further modified and developed by **Alexander Varshavsky** and co-workers, who examined distribution of **histone H4** on **heat shock genes** using formaldehyde cross-linking.

This technique was extensively developed and refined thereafter. NChIP approach was first described by Hebbes *et al.*, 1988, and also been developed and refined quickly. The typical ChIP assay usually take 4–5 days, and require 10^6 – 10^7 cells at least. Now new techniques on ChIP could be achieved as few as 100–1000 cells and complete within one day.

- **Bead-free ChIP:** This novel method ChIP uses discs of inert, porous polymer functionalized with either Protein A or G in spin columns or microplates. The chromatin-antibody complex is selectively retained by the disc and eluted to obtain enriched DNA for downstream applications such as qPCR and sequencing. The porous environment is specifically designed to maximize capture efficiency and reduce non-specific binding. Due to less manual handling and optimised protocols, ChIP can be performed in 5 hours.
- **Carrier ChIP (CChIP):** This approach could use as few as 100 cells by adding *Drosophila* cells as carrier chromatin to reduce loss and facilitate precipitation of the target chromatin. However, it demands highly specific primers for detection of the target cell chromatin from the foreign carrier chromatin background, and it takes two to three days.
- **Fast ChIP (qChIP):** The fast ChIP assay reduced the time by shortening two steps in a typical ChIP assay: (i) an ultrasonic bath accelerates the rate of antibody binding to target proteins—and thereby reduces immunoprecipitation time (ii) a resin-based (Chelex-100) DNA isolation procedure reduces the time of **cross-link** reversal and DNA isolation. However, the fast protocol is suitable only for large cell samples (in the range of 10^6 – 10^7). Up to 24 sheared chromatin samples can be processed to yield PCR-ready DNA in 5 hours, allowing multiple chromatin factors be probed simultaneously and/or looking at genomic events over several time points.
- **Quick and quantitative ChIP (Q²ChIP):** The assay uses 100,000 cells as starting material and is suitable for up to 1,000 histone ChIPs or 100 transcription factor ChIPs. Thus many chromatin samples can be prepared in parallel and stored, and Q²ChIP can be undertaken in a day.
- **MicroChIP (μChIP):** chromatin is usually prepared from 1,000 cells and up to 8 ChIPs can be done in parallel without carriers. The assay can also start with 100 cells, but only suit for one ChIP. It can also use small (1 mm³) tissue **biopsies** and microChIP can be done within one day.
- **Matrix ChIP:** This is a **microplate**-based ChIP assay with increased throughput and simplified the procedure. All steps are done in microplate wells without sample transfers, enabling a potential for automation. It enables 96 ChIP assays for histone and various DNA-bound proteins in a single day.
- **Pathology-ChIP (PAT-ChIP):** This technique allows ChIP from pathology formalin-fixed and paraffin-embedded tissues and thus the use of pathology archives (even those that are several years old) for epigenetic analyses and the identification of candidate epigenetic biomarkers or

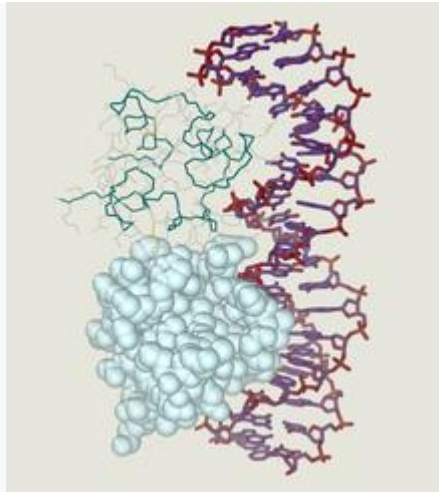
targets.

ChIP has also been applied for genome wide analysis by combining with microarray technology (**ChIP-on-chip**) or second generation DNA-sequencing technology (**Chip-Sequencing**). ChIP can also combine with **paired-end tags** sequencing in **Chromatin Interaction Analysis using Paired End Tag sequencing** (ChIA-PET), a technique developed for large-scale, de novo analysis of higher-order chromatin structures.

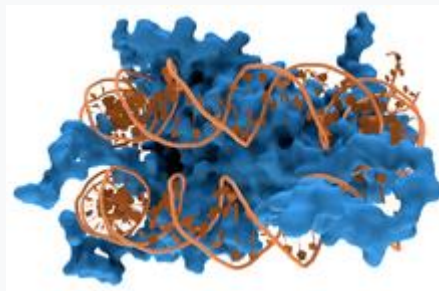
Limitations

- Large Scale assays using ChIP is challenging using intact model organisms. This is because antibodies have to be generated for each TF, or, alternatively, transgenic model organisms expressing epitope-tagged TFs need to be produced.
- Researchers studying differential gene expression patterns in small organisms also face problems as genes expressed at low levels, in a small number of cells, in narrow time window.
- ChIP experiments cannot discriminate between different TF isoforms (**Protein isoform**).

DNA-binding protein

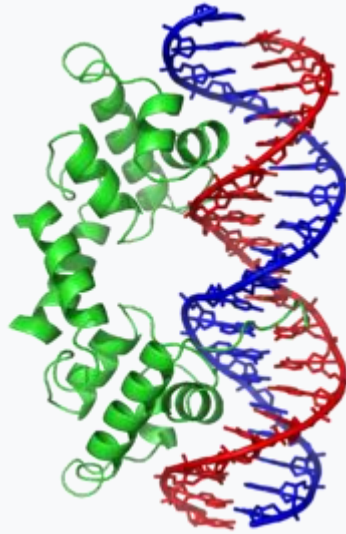


Cro protein complex with DNA

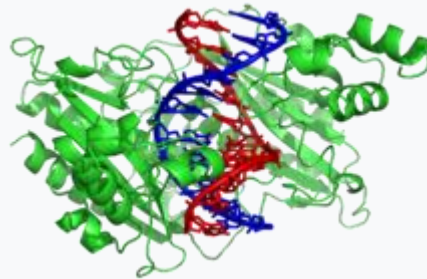


Interaction of DNA (orange) with **histones** (blue). These proteins' basic amino acids bind to the acidic

phosphate groups on DNA.



The lambda repressor [helix-turn-helix](#) transcription factor bound to its DNA target



The restriction enzyme [EcoRV](#) (green) in a complex with its substrate DNA

DNA-binding proteins are [proteins](#) that have [DNA-binding domains](#) and thus have a specific or general affinity for single- or double-stranded [DNA](#). Sequence-specific DNA-binding proteins generally interact with the [major groove](#) of [B-DNA](#), because it exposes more [functional groups](#) that identify a [base pair](#). However, there are some known [minor groove](#) DNA-binding ligands such as [netropsin](#), [distamycin](#), [Hoechst 33258](#), [pentamidine](#), [DAPI](#) and others.

Examples

DNA-binding [proteins](#) include [transcription factors](#) which [modulate](#) the process of transcription, various [polymerases](#), [nucleases](#) which cleave DNA molecules,

and **histones** which are involved in **chromosome** packaging and transcription in the **cell nucleus**. DNA-binding proteins can incorporate such domains as the **zinc finger**, the **helix-turn-helix**, and the **leucine zipper** (among many others) that facilitate binding to nucleic acid. There are also more unusual examples such as **transcription activator like effectors**.

Non-specific DNA-protein interactions

Structural proteins that bind DNA are well-understood examples of non-specific DNA-protein interactions. Within chromosomes, DNA is held in complexes with structural proteins. These proteins organize the DNA into a compact structure called **chromatin**. In **eukaryotes**, this structure involves DNA binding to a complex of small basic proteins called **histones**. In **prokaryotes**, multiple types of proteins are involved.

The histones form a disk-shaped complex called a **nucleosome**, which contains two complete turns of double-stranded DNA wrapped around its surface. These non-specific interactions are formed through basic residues in the histones making **ionic bonds** to the acidic sugar-phosphate backbone of the DNA, and are therefore largely independent of the base sequence.

Chemical modifications of these basic **amino acid** residues include **methylation**, **phosphorylation** and **acetylation**. These chemical changes alter the strength of the interaction between the DNA and the histones, making the DNA more or less accessible to **transcription factors** and changing the rate of transcription.

Other non-specific DNA-binding proteins in chromatin include the high-mobility group (HMG) proteins, which bind to bent or distorted DNA. Biophysical studies show that these architectural HMG proteins bind, bend and loop DNA to perform its biological functions. These proteins are important in bending arrays of nucleosomes and arranging them into the larger structures that form chromosomes.

Proteins that specifically bind single-stranded DNA

A distinct group of DNA-binding proteins are the DNA-binding proteins that specifically bind single-stranded DNA. In humans, **replication protein A** is the best-understood member of this family and is used in processes where the double helix is separated, including DNA replication, recombination and DNA repair. These binding proteins seem to stabilize single-stranded DNA and protect it from forming **stem-loops** or being degraded by **nucleases**.

Binding to specific DNA sequences

In contrast, other proteins have evolved to bind to specific DNA sequences. The most intensively studied of these are the various **transcription factors**, which are proteins that regulate transcription. Each transcription factor binds to one specific set of DNA sequences and activates or inhibits the transcription of genes that have these sequences near their promoters.

The transcription factors do this in two ways. Firstly, they can bind the RNA polymerase responsible for transcription, either directly or through other mediator proteins; this locates

the polymerase at the promoter and allows it to begin transcription.

Alternatively, transcription factors can bind **enzymes** that modify the histones at the promoter. This alters the accessibility of the DNA template to the polymerase.

These DNA targets can occur throughout an organism's genome. Thus, changes in the activity of one type of transcription factor can affect thousands of genes.

Thus, these proteins are often the targets of the **signal transduction** processes that control responses to environmental changes or **cellular differentiation** and development. The specificity of these transcription factors' interactions with DNA come from the proteins making multiple contacts to the edges of the DNA bases, allowing them to *read* the DNA sequence. Most of these base-interactions are made in the major groove, where the bases are most accessible.

Mathematical descriptions of protein-DNA binding taking into account sequence-specificity, and competitive and cooperative binding of proteins of different types are usually performed with the help of the **lattice models**. Computational methods to identify the DNA binding sequence specificity have been proposed to make a good use of the abundant sequence data in the post-genomic era.

Protein–DNA interactions

Protein–DNA interactions occur when a **protein** binds a molecule of **DNA**, often to regulate the **biological function** of DNA, usually the **expression** of a **gene**. Among the proteins that bind to DNA are **transcription factors** that activate or repress gene expression by binding to DNA motifs and **histones** that form part of the structure of DNA and bind to it less specifically. Also proteins that **repair DNA** such as **uracil-DNA glycosylase** interact closely with it.

In general, proteins bind to DNA in the **major groove**; however, there are exceptions. Protein–DNA interaction are of mainly two types, either specific interaction, or non-specific interaction. Recent single-molecule experiments showed that DNA binding proteins undergo of rapid rebinding in order to bind in correct orientation for recognizing the target site.

Design

Designing DNA-binding proteins that have a specified DNA-binding site has been an important goal for biotechnology. **Zinc finger** proteins have been designed to bind to specific DNA sequences and this is the basis of **zinc finger nucleases**. Recently **transcription activator-like effector nucleases** (TALENs) have been created which are based on natural **proteins** secreted by *Xanthomonas* bacteria via their **type III secretion system** when they infect various **plant** species.

Detection methods

There are many *in vitro* and *in vivo* techniques which are useful in detecting DNA-Protein Interactions. The following lists some methods currently in use: **Electrophoretic mobility shift assay** is a widespread technique to identify protein–DNA interactions. **DNase footprinting assay** can be used to identify the specific site of binding of a protein to DNA. **Chromatin immunoprecipitation** is used to identify the sequence of the DNA fragments which bind to a

known transcription factor. This technique when combined with high throughput sequencing is known as **ChIP-Seq** and when combined with **microarrays** it is known as **ChIP-chip**. **Yeast one-hybrid System** (Y1H) is used to identify which protein binds to a particular DNA fragment. **Bacterial one-hybrid system** (B1H) is used to identify which protein binds to a particular DNA fragment. Structure determination using **X-ray crystallography** has been used to give a highly detailed atomic view of protein–DNA interactions.

Manipulating the interactions

The protein–DNA interactions can be modulated using stimuli like ionic strength of the buffer, macromolecular crowding, temperature, pH and electric field. This can lead to reversible dissociation/association of the protein–DNA complex.

UNIT-II

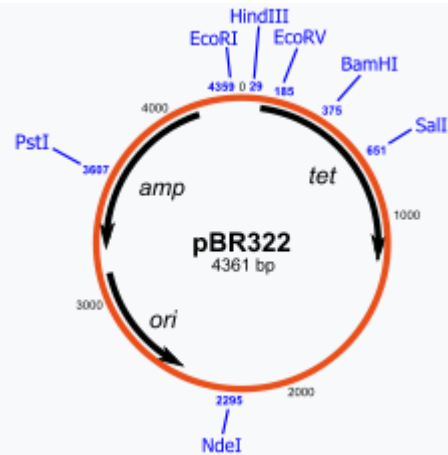
In **molecular cloning**, a **vector** is a DNA molecule used as a vehicle to artificially carry foreign genetic material into another cell, where it can be **replicated** and/or **expressed** (e.g., **plasmid**, **cosmid**, **Lambda phages**). A vector containing foreign DNA is termed **recombinant vector**. Types of vectors are **plasmids**, **viral vectors**, **cosmids**, and **artificial chromosomes**. Of these, the most commonly used vectors are plasmids. Engineered vectors are an **origin of replication**, a **multicloning site**, and a **selectable marker**.

The vector itself is generally a **DNA** sequence that consists of an **insert (transgene)** and a larger sequence that serves as the "backbone". The purpose of a vector which transfers genetic information to another cell is typically to isolate, multiply, or express the insert in the target cell. Vectors can be used for cloning and are therefore **cloning vectors**, but there are also vectors designed specially for cloning, while others are designed for other purposes, such as transcription and protein expression. Vectors designed specifically for the expression of the transgene are called **expression vectors**, and generally have a **promoter** sequence that drives expression of the transgene. Simpler vectors called **reporter vectors** are capable of being transcribed but not translated: they can be replicated in a target cell but not expressed, unlike expression vectors which are used to amplify their insert.

The manipulation of DNA is normally conducted on *E. coli* vectors, which contain elements necessary for their maintenance in *E. coli*. Some vectors also have elements that allow them to be maintained in another organism such as yeast, plant or mammalian cells, and these vectors are called **shuttle vectors**. Such vectors have bacterial or viral elements which may be transferred to the non-bacterial host organism, however other vectors have also been developed to avoid the transfer of any genetic material from an alien species.

Insertion of a vector into the target cell is usually called **transformation** for bacterial cells, **transfection** for **eukaryotic cells**,^[4] and **transduction** for a viral vector is often called **transduction**.

Cloning vector



Schematic representation of the [pBR322](#) plasmid, one of the first plasmids widely used as a cloning vector.

A **cloning vector** is a small piece of [DNA](#) that can be stably maintained in an organism, and into which a foreign DNA fragment can be inserted for [cloning](#) purposes. The cloning vector may be DNA taken from a [virus](#), the [cell](#) of a higher organism, or it may be the [plasmid](#) of a bacterium. The [vector](#) therefore contains features that allow for the convenient insertion or removal of a DNA fragment to or from the vector and the foreign DNA with a [restriction enzyme](#) that cuts the DNA.

DNA fragments thus generated contain either blunt ends or overhangs known as sticky ends, and vector DNA and foreign DNA can be joined together by [molecular ligation](#). After a DNA fragment has been cloned into a cloning vector, it may be further [subcloned](#) or [designed](#) for more specific use.

There are many types of cloning vectors, but the most commonly used ones are genetically engineered [plasmids](#). Cloning is commonly done using *Escherichia coli*, and cloning vectors in *E. coli* include plasmids, [bacteriophages](#) (such as [phage λ](#)), [cosmids](#), and [bacteriophage chromosomes](#) (BACs). Some DNA, however, cannot be stably maintained in *E. coli*, for example very large DNA fragments, and [yeast](#) may be used. Cloning vectors in yeast include [yeast artificial chromosomes](#) (YACs).



Features of a cloning vector

All commonly used cloning vectors in [molecular biology](#) have key features necessary for their function, such as a suitable cloning site. Others may have additional features specific to their use. For reason of ease and convenience, cloning is often performed using [shuttle vectors](#). [Shuttle vectors](#) used often have elements necessary for their propagation and maintenance in *E. coli*, such as a functional [origin of replication](#). An [origin of replication](#) is found in many plasmids. Some vectors also include elements that allow them to be maintained in another organism; these vectors are called [shuttle vector](#).

Cloning site

All cloning vectors have features that allow a gene to be conveniently inserted into the vector or removed from it. This may be a [multiple cloning site](#) or [polylinker](#), which contains many unique [restriction sites](#). The restriction sites in the MCS are first cleaved by restriction enzymes. A target gene also digested with the same enzymes is ligated into the vectors using [DNA ligase](#). The target DNA sequence can be inserted in a specific direction if so desired. The restriction sites may be further used for [sub-cloning](#) into another vector if necessary.^[2]

Other cloning vectors may use [topoisomerase](#) instead of ligase and cloning may be done more rapidly without the need for restriction enzymes. In this [TOPO cloning](#) method a linearized vector is activated by attaching topoisomerase I to its ends, and this "TOPO-enzyme" can accept a PCR product by ligating both the 5' ends of the PCR product, releasing the topoisomerase and forming a circular vector.

method of cloning without the use of DNA digest and ligase is by **DNA recombination**, for example as used in the **Gateway cloning**. A DNA fragment, which has been previously cloned into the cloning vector (called entry clone in this method), may be conveniently introduced into a variety of expression vectors.

Selectable marker

A **selectable marker** is carried by the vector to allow the selection of positively **transformed** cells. **Antibiotic** resistance is often used as a selectable marker, being the **beta-lactamase** gene, which confers resistance to the **penicillin** group of **beta-lactam antibiotics** like **ampicillin**. Some other selectable markers, for example the plasmid pACYC177 has both ampicillin and **kanamycin** resistance gene. Shuttle vector which is designed for use in different organisms may also require two selectable markers, although some selectable markers such as resistance to **zeocin** are used in different cell types. **Auxotrophic** selection markers that allow an auxotrophic organism to grow in **minimal growth medium** may also be used. These are **LEU2** and **URA3** which are used with their corresponding auxotrophic strains of yeast.

Another kind of selectable marker allows for the positive selection of plasmid with cloned gene. This may involve the use of a toxin such as **barnase**, **Ccda**, and the **parD/parE** toxins. This typically works by disrupting or removing the lethal gene during the cloning process. Clones where the lethal gene still remains intact would kill the host cells, therefore only successful clones are selected.

Reporter gene

Reporter genes are used in some cloning vectors to facilitate the screening of successful clones by using features that allow a successful clone to be easily identified. Such features present in cloning vectors may be the **lacZ α fragment** for **blue-white selection**, and/or **marker gene** or **reporter genes** in frame with and flanking the **MCS** to facilitate the production of fusion proteins. Examples of fusion partners that may be used for screening are the **green fluorescent protein (GFP)** and **luciferase**.

Elements for expression

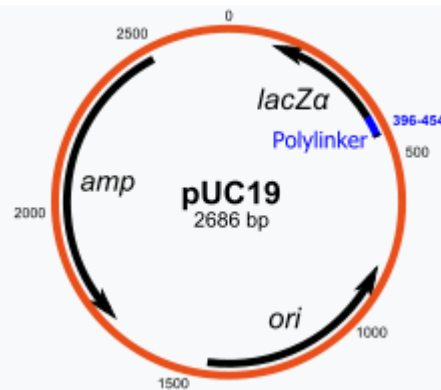
A cloning vector need not contain suitable elements for the **expression** of a cloned target gene, such as a **promoter** and **ribosome binding site**, but may however do, and may then work as an **expression vector**. The target **DNA** may be inserted into a site that is under the control of a promoter necessary for the expression of the target gene in the chosen host. Where the promoter is present, the expression of the gene is **constitutive** and **inducible** so that proteins are only produced when required. Some commonly used promoters are the **T7** and **lac** promoters. A ribosome binding site is necessary when screening techniques such as **blue-white selection** are used.

Cloning vectors without promoter and RBS for the cloned DNA sequence are sometimes used, for example when cloning genes into *E. coli* cells. Promoter and RBS for the cloned DNA sequence are also unnecessary when first making a **genomic** or **cDNA library**. Genes are normally subcloned into a more appropriate expression vector if their expression is required.

Some vectors are designed for transcription only with no heterologous protein expressed, for example for *in vitro* mRNA production. These are transcription vectors. They may lack the sequences necessary for polyadenylation and termination, therefore may not be used for *in vivo* expression.

Types of cloning vectors

A large number of cloning vectors are available, and choosing the vector may depend a number of factors, such as the size of the insert and the cloning method. Large insert may not be stably maintained in a general cloning vector, especially for those with a high copy number. Some DNA fragments may require more specialized cloning vector.



The pUC plasmid has a high copy number, contains a multiple cloning site (polylinker), a gene for ampicillin antibiotic selection, and can be used to clone DNA fragments of up to 15 kb in size.

Plasmid

Plasmids are autonomously replicating circular extra-chromosomal DNA. They are the standard cloning vectors and the ones used in general plasmids may be used to clone DNA insert of up to 15 kb in size. One of the earliest commonly used cloning vectors is the pUC series of plasmids, and a large number of different cloning plasmid vectors are available. Most plasmid vectors have a high copy number, for example pUC19 which has a copy number of 500-700 copies per cell, and high copy number is useful as it produces a high concentration of plasmid for subsequent manipulation. However low-copy-number plasmids may be preferably used in certain circumstances, if the presence of the cloned gene is toxic to the cells.

Some plasmids contain an M13 bacteriophage origin of replication and may be used to generate single-stranded DNA. These include the pBluescript series of cloning vectors.

Bacteriophage

The bacteriophages used for cloning are the λ phage and M13 phage. There is an upper limit on the amount of DNA that can be packaged into a phage (maximum of 53 kb), therefore to allow foreign DNA to be inserted into phage DNA, phage cloning vectors may need to have some genes deleted, for example the genes for lysogeny since using phage λ as a cloning vector involves only the lytic cycle.^[16] There are two types of bacteriophage cloning vectors: insertion vector and replacement vector. Insertion vectors contain a unique cleavage site whereby foreign DNA with size of 5-10 kb can be inserted. In replacement vectors, the cleavage sites flank a region containing genes not essential for the lytic cycle, and this region may be replaced by the DNA insert in the cloning process, and a larger sized DNA of 8-24 kb may be inserted.

There is also a lower size limit for DNA that can be packed into a phage, and vector DNA that is too small cannot be properly packaged. Vector DNA that is too small cannot be properly packaged for selection - vector without insert may be too small, therefore only vectors with insert may be selected.

Cosmid

Cosmids are plasmids that incorporate a segment of bacteriophage λ DNA that has the cohesive end site (cos) which is required for packaging DNA into λ particles. It is normally used to clone large DNA fragments between 28 and 45 kb in size.

Bacterial artificial chromosome

Insert size of up to 350 kb can be cloned in bacterial artificial chromosome (BAC). BACs are maintained in *E. coli* with a copy number of 1-2 per cell. BACs are based on F plasmid, another artificial chromosome called the PAC is based on the P1 phage.

Yeast artificial chromosome

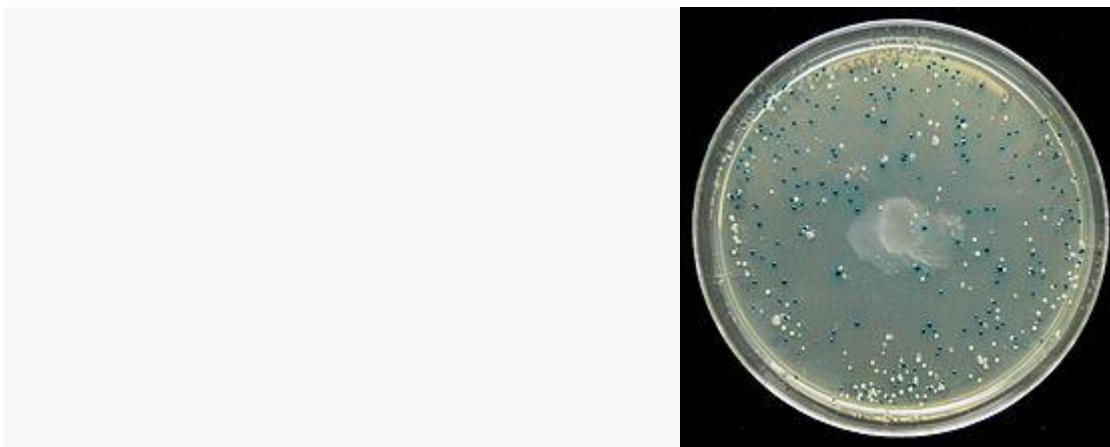
Yeast artificial chromosome are used as vectors to clone DNA fragments of more than 1 mega base (1Mb=1000kb) in size. They contain a telomeric sequence, an autonomous replication origin (ARO), and a centromere (features required to replicate linear chromosomes in yeast cells). These vectors also contain suitable restriction sites to clone foreign DNA fragments as required in mapping genomes such as in human genome project.

be used as selectable markers.

Human artificial chromosome

Human artificial chromosome may be potentially useful as a gene transfer vectors for gene delivery into human cells, and a tool for determining human chromosome function. It can carry very large DNA fragment (there is no upper limit on size for practical purposes) and avoid the problem of limited cloning capacity of other vectors, and it also avoids possible insertional mutagenesis caused by insertion by viral vector.

Animal and plant viral vectors Viruses that infect plant and animal cells have also been manipulated to introduce foreign genes. The natural ability of viruses to adsorb to cells, introduce their DNA and replicate have made them ideal vehicles to transfer foreign genes into cells in culture. A vector based on Simian virus 40 (SV40) was used in first cloning experiment involving mammalian cells. A number of viruses like Adenoviruses and Papilloma virus have been used to clone genes in mammals. At present, retroviral vectors are used to introduce genes into mammalian cells. In case of plants like Cauliflower mosaic virus, Tobacco mosaic virus and Gemini viruses have been used to introduce genes into plants.



An LB agar plate showing the result of a blue white screen. White colonies may contain an insert in the plasmid it carries, while the blue colonies do not.

Screening: example of the blue/white screen

Many general purpose vectors such as **pUC19** usually include a system for detecting the presence of a cloned DNA fragment, scored phenotype. The most widely used is the gene coding for *E. coli* **β -galactosidase**, whose activity can easily be detected by the presence of a blue/white screen. **β -galactosidase** encodes to hydrolyze the soluble, colourless substrate **X-gal** (5-bromo-4-chloro-3-indolyl-beta-d-galactoside) into an insoluble blue precipitate (5-bromo-4-chloro-3-hydroxy-5-indolyl-2,4,4'-dichloro indigo). Cloning a fragment of DNA within the vector-based **lacZ α** sequence of the **β -galactosidase** prevents the activity of the enzyme. If X-gal is included in the selective agar plates, transformant colonies are generally blue in the case of a vector with no insert and white in the case of a vector containing a fragment of cloned DNA.

Viral vector

Viral vectors are tools commonly used by molecular **biologists** to deliver **genetic material** into **cells**. This process can be done in a living organism (*in vivo*) or in **cell culture** (*in vitro*). **Viruses** have evolved specialized molecular mechanisms to efficiently deliver their **genomes** inside the cells they infect. Delivery of **genes**, or other genetic material, by a vector is termed **transduction**. Cells that are described as transduced. **Molecular biologists** first harnessed this machinery in the 1970s. **Paul Berg** used a model system to deliver DNA from the **bacteriophage λ** to infect monkey **kidney** cells maintained in culture.

In addition to their use in molecular biology research, viral vectors are used for **gene therapy** and the development of



properties of a viral vector

Viral vectors are tailored to their specific applications but generally share a few key properties.

- *Safety*: Although viral vectors are occasionally created from **pathogenic** viruses, they are modified in such a way as to make handling them safe. This usually involves the deletion of a part of the viral genome critical for **viral replication**. Such vectors can infect target cells but, once the infection has taken place, requires a **helper virus** to provide the missing **proteins** for production of new viruses.
- *Low toxicity*: The viral vector should have a minimal effect on the **physiology** of the cell it infects.
- *Stability*: Some viruses are genetically unstable and can rapidly rearrange their genomes. This is detrimental to the reproducibility of the work conducted using a viral vector and is avoided in their design.
- *Cell type specificity*: Most viral vectors are engineered to infect as wide a range of **cell types** as possible. However, specificity is often preferred. The viral receptor can be modified to target the virus to a specific kind of cell. Viruses modified in this way can be **pseudotyped**.
- *Identification*: Viral vectors are often given certain genes that help identify which cells took up the viral genes. These are called **markers**. A common marker is **resistance** to a certain antibiotic. The cells can then be isolated easily, as the cells with the viral vector genes do not have antibiotic resistance, and so cannot grow in a culture with the relevant antibiotic.

Applications

Basic research

Viral vectors were originally developed as an alternative to **transfection** of **naked DNA** for **molecular genetics** experiments. Traditional methods such as **calcium phosphate precipitation**, **transduction** can ensure that nearly 100% of cells are infected without affecting cell viability. Furthermore, some viruses **integrate** into the cell **genome** facilitating stable expression.

Protein coding genes can be **expressed** using viral vectors, commonly to study the function of the particular protein. **Retroviruses**, stably expressing **marker genes** such as **GFP** are widely used to permanently label cells to track them and are used, for example in **xenotransplantation** experiments, when cells infected *in vitro* are implanted into a host animal.

Gene insertion is cheaper to carry out than **gene knockout**. But as the silencing is sometimes non-specific and has off-target effects on other genes, it provides less reliable results. Animal host vectors also play an important role.

Gene therapy

Gene therapy is a technique for correcting defective genes responsible for disease development. In the future, **gene therapy** will cure **genetic disorders**, such as **severe combined immunodeficiency**, **cystic fibrosis** or even **haemophilia A**. Because these are caused by **mutations** in the DNA sequence for specific genes, gene therapy trials have used viruses to deliver unmutated copies of the gene to the cells of the patient's body. There have been a huge number of laboratory successes with gene therapy. However, several trials have failed and gene therapy must be overcome before it gains widespread use. **Immune response** to viruses not only impedes the delivery of the gene but can cause severe complications for the patient. In one of the early gene therapy trials in 1999 this led to the death of a patient who was treated using an adenoviral vector.

Some viral vectors, for instance **gamma-retroviruses**, insert their genomes at a seemingly random location on one of the chromosomes which can disturb the function of cellular genes and lead to cancer. In a **severe combined immunodeficiency** retrovirus trial conducted in 2002, four of the patients developed leukemia as a consequence of the treatment;¹ three of the patients died after receiving chemotherapy. **Adeno-associated virus-based vectors** are much safer in this respect as they always integrate at the safe harbor.

genome, with applications in various disorders, such as [Alzheimer's disease](#).

Vaccines

A **live vector vaccine** is a [vaccine](#) that uses a chemically weakened [virus](#) to transport pieces of the pathogen in order to elicit an [immune response](#).^[6] Viruses expressing [pathogen](#) proteins are currently being developed as [vaccines](#) against these pathogens, such as [DNA vaccines](#). The genes used in such vaccines are usually [antigen](#) coding [surface proteins](#) from the [pathogenic](#) organism, inserted into the genome of a non-pathogenic organism, where they are expressed on the organism's surface and can be recognized by the immune system.

An example is the [hepatitis B vaccine](#), where [Hepatitis B](#) infection is controlled through the use of a recombinant [virus](#) of the hepatitis B virus surface antigen that is produced in yeast cells. The development of the recombinant subunit vaccine was a necessary development because hepatitis B virus, unlike other common viruses such as [polio virus](#), cannot be grown in culture.

[T-lymphocytes](#) recognize cells infected with [intracellular parasites](#) based on the foreign proteins produced within the cell. This is crucial for protection against viral infections and such diseases as [malaria](#). A viral vaccine induces expression of pathogen proteins in cells similarly to the [Sabin Polio vaccine](#) and other [attenuated vaccines](#). However, since viral vaccines contain only viral genes, they are much safer and sporadic infection by the pathogen is impossible. [Adenoviruses](#) are being actively developed as vaccines.

Types

Retroviruses

[Retroviruses](#) are one of the mainstays of current gene therapy approaches. The recombinant retroviruses such as the [Moloney murine leukemia virus](#) have the ability to integrate into the host genome in a stable fashion. They contain a [reverse transcriptase](#) to make a DNA copy of their RNA genome, and an integrase that allows integration into the host [genome](#). They have been used in a number of FDA-approved clinical trials, including the [SCID-X1](#) trial.

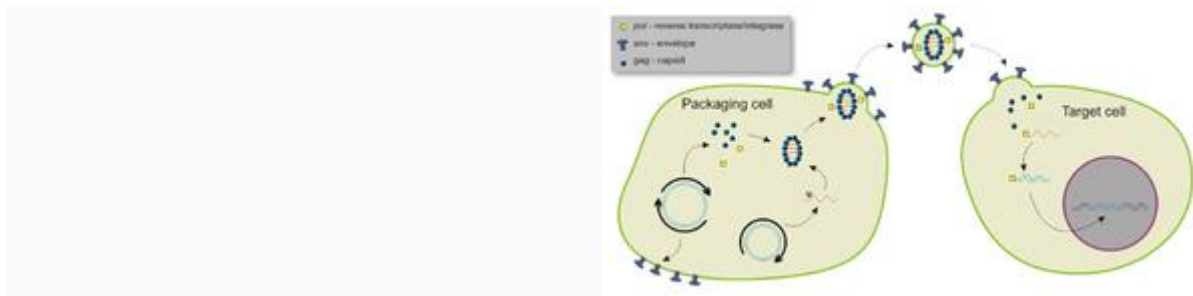
Retroviral vectors can either be replication-competent or replication-defective. Replication-defective vectors are the most commonly used in gene therapy studies because the viruses have had the coding regions for the genes necessary for additional rounds of virion replication replaced with other genes, or deleted. These viruses are capable of infecting their target cells and delivering their viral genomes, but they cannot continue the typical lytic pathway that leads to cell lysis and death.

Conversely, replication-competent viral vectors contain all necessary genes for virion synthesis, and continue to produce new viruses after infection occurs. Because the viral genome for these vectors is much lengthier, the length of the actual inserted gene is much shorter compared to the possible length of the insert for replication-defective vectors. Depending on the viral vector, the typical allowable DNA insert in a replication-defective viral vector is usually about 8–10 kB. While this limits the introduction of large DNA sequences, most [cDNA](#) sequences can still be accommodated.

The primary drawback to use of retroviruses such as the Moloney retrovirus involves the requirement for cells to be in a state of active division for [transduction](#). As a result, cells such as [neurons](#) are very resistant to infection and transduction by retroviruses.

There is concern that [insertional mutagenesis](#) due to integration into the host [genome](#) might lead to [cancer](#) or [leukemia](#). This was a theoretical concern until gene therapy for ten [SCID-X1](#) patients using Moloney [murine leukemia virus](#) resulted in two cases of leukemia. This was due to activation of the [LMO2 oncogene](#) due to nearby integration of the vector.

Lentiviruses



Packaging and transduction by a lentiviral vector.

Lentiviruses are a subclass of Retroviruses. They are sometimes used as **vectors for gene therapy** thanks to their ability to integrate into the **genome** of non-dividing cells, which is the unique feature of Lentiviruses as other Retroviruses can infect only dividing cells. The **genome** in the form of **RNA** is **reverse-transcribed** when the virus enters the cell to produce **DNA**, which is then inserted into a random position (recent findings actually suggest that the insertion of viral DNA is not random but directed to specific sites) to genome organisation by the viral **integrase enzyme**. The vector, now called a **provirus**, remains in the genome and is passed on to the daughter cells of the cell when it divides.

There are, as yet, no techniques for determining the site of integration, which can pose a problem. The **provirus** can integrate near cellular genes and lead to activation of **oncogenes** promoting the **development** of **cancer**, which raises concerns for the use of lentiviruses in gene therapy. However, studies have shown that lentivirus vectors have a lower tendency to integrate near oncogenes and cause cancer than gamma-retroviral vectors. More specifically, one study found that lentiviral vectors did not cause an increase in tumor incidence or an earlier onset of tumors in a mouse strain with a much higher incidence of tumors.

Moreover, clinical trials that utilized lentiviral vectors to deliver gene therapy for the treatment of HIV experienced no significant increase in oncologic events.

For safety reasons lentiviral vectors never carry the genes required for their replication. To produce a lentivirus, several **plasmids** are **transfected** into a so-called packaging **cell line**, commonly **HEK 293**. One or more plasmids, called packaging plasmids, encode the **virion proteins**, such as the **capsid** and the **reverse transcriptase**. Another plasmid carrying the target gene can be delivered by the vector. It is **transcribed** to produce the single-stranded RNA viral genome and is marked by the **psi** sequence. This sequence is used to package the genome into the virion.

Adenoviruses

As opposed to lentiviruses, adenoviral DNA does not integrate into the genome and is not replicated during cell division. Adenoviruses are used in basic research, although adenoviral vectors are still used in *in vitro* and also *in vivo* experiments. Their primary applications are **therapy** and **vaccination**. Since humans commonly come in contact with **adenoviruses**, which cause respiratory, gastrointestinal, and other infections, majority of patients have already developed **neutralizing antibodies** which can inactivate the virus before it reaches the target cells. To overcome this problem scientists are currently investigating **adenoviruses** that infect different species to which humans do not have immunity.

Adeno-associated viruses

Adeno-associated virus (AAV) is a small virus that infects humans and some other primate species. AAV is not currently known to cause any disease, and causes a very mild immune response. AAV can infect both dividing and non-dividing cells and may integrate into the genome of the host cell. Moreover, AAV mostly stays as **episomal** (replicating without incorporation into the chromosome) and does not require the presence of the helper virus for expression. These features make AAV a very attractive candidate for creating viral vectors for gene therapy.^[1] However, the current capacity is limited to 5kb which is considerably small compared to AAV's original capacity.

Furthermore, because of its potential use as a gene therapy vector, researchers have created an altered AAV called **self-complementary associated virus** (scAAV). Whereas AAV packages a single strand of DNA and requires the process of second-strand synthesis, scAAV packages both strands which anneal together to form double stranded DNA. By skipping second strand synthesis scAAV allows for immediate expression in the cell. Otherwise, scAAV carries many characteristics of its AAV counterpart.

Hybrids

Hybrid vectors are **vector viruses** that are **genetically engineered** to have qualities of more than one vector. Viruses overcome the shortcomings of typical viral vectors, which may have limited loading capacity, immunogenicity, **genotoxicity**, and inadequate **transgenic expression**. Through the replacement of undesirable elements with desired abilities, hybrid vectors can outperform standard transfection vectors in terms of safety and therapeutic efficiency.

Challenges in application

The choice of a **viral vector** to deliver **genetic** material to cells comes with some logistical problems. There are a limited number of viral vectors available for therapeutic use. Any of these few viral vectors can cause the body to develop an **immune response** if the vector is recognized as an invader.

Once used, the viral vector cannot be effectively used in the patient again because it will be recognized by the body. In **gene therapy** fails in **clinical trials**, the virus can't be used again in the patient for a different vaccine or gene therapy in the future. Existing **immunity** against the viral vector could also be present in the patient rendering the therapy ineffective for the future. To counteract pre-existing immunity when using a viral vector for **vaccination** by **priming** with a non-viral **DNA vaccine** presents another expense and obstacle in the vaccine distribution process.

Pre-existing immunity may also be challenged by increasing vaccine dose or changing the **vaccination** route.^[25] Some of these challenges (such as genotoxicity and low transgenic expression) can be overcome through the use of **hybrid vectors**.

Expression vector

An **expression vector**, otherwise known as an **expression construct**, is usually a **plasmid** or virus designed for the purpose of expressing a specific gene. The **vector** is used to introduce a specific **gene** into a target cell, and can commandeer the cell's mechanism for protein synthesis to produce the **protein encoded** by the gene. Expression vectors are the basic tools in **biotechnology** for the **production of proteins**.

The **vector** is engineered to contain regulatory sequences that act as **enhancer** and **promoter** regions and lead to the expression of the gene carried on the expression vector.

The goal of a well-designed expression vector is the efficient production of protein, and this may be achieved by the production of a significant amount of stable **messenger RNA**, which can then be **translated** into protein. The expression of a protein can be tightly controlled, and the protein is only produced in significant quantity when necessary through the use of an **inducible promoter**. Alternatively, the protein may be expressed constitutively.

Escherichia coli is commonly used as the host for **protein production**, but other cell types may also be used. A common application of an expression vector is the production of **insulin**, which is used for medical treatments of **diabetes**.



Elements of expression vectors

An expression vector has features that any **vector** may have, such as an **origin of replication**, a **selectable marker**, and a **multiple cloning site**.

insertion of a gene like the [multiple cloning site](#). The cloned gene may be transferred from a specialized [cloning vector](#), although it is possible to clone directly into an expression vector. The cloning process is normally performed in *E. coli*. Vectors used for protein production in organisms other than *E. coli* may have, in addition to a suitable origin of replication in *E. coli*, elements that allow them to be maintained in another organism, and these vectors are called [shuttle vectors](#).

Elements for expression

An expression vector must have elements necessary for gene expression. These may include a [promoter](#), the coding sequence such as a [ribosomal binding site](#) and [start codon](#), a [termination codon](#), and a [transcription termination signal](#). Due to differences in the machinery for protein synthesis between prokaryotes and eukaryotes, therefore the expression vectors must have elements for expression that are appropriate for the chosen host. For example, prokaryotes expression vectors have a [Shine-Dalgarno sequence](#) at its translation initiation site for the binding of ribosomes, while eukaryotes expression vectors have the [Kozak consensus sequence](#).

The [promoter](#) initiates the [transcription](#) and is therefore the point of control for the expression of the cloned gene. Expression vectors are normally [inducible](#), meaning that protein synthesis is only initiated when required by the presence of an [inducer](#) such as [IPTG](#). Gene expression however may also be constitutive (i.e. protein is constantly expressed) in some vectors. Low level of constitutive protein synthesis may occur even in expression vectors with tightly controlled promoters.

Protein tags

After the expression of the gene product, it is usually necessary to purify the expressed protein; however, separation from the great majority of proteins of the host cell can be a protracted process. To make this purification process easier, a [tag](#) may be added to the cloned gene. This tag could be [histidine \(His\) tag](#), other marker peptides, or a [fusion partner](#) such as [S-transferase](#) or [maltose-binding protein](#). Some of these fusion partners may also help to increase the solubility of the protein. Other fusion proteins such as [green fluorescent protein](#) may act as a [reporter gene](#) for the identification of successful expression. [Fluorescent protein](#) may be used to study protein expression in [cellular imaging](#).

Others

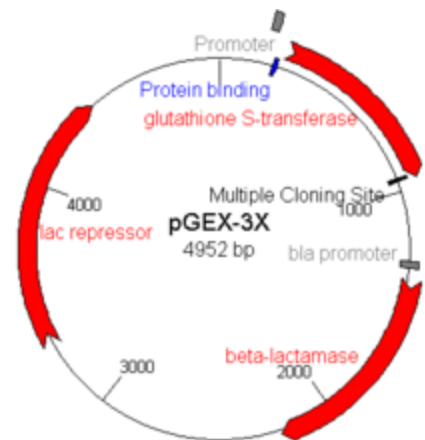
The expression vector is [transformed](#) or [transfected](#) into the host cell for protein synthesis. Some expression vectors are designed for [transformation](#) or the insertion of DNA into the host chromosome, for example the [vir genes](#) for [plant transformation](#) or [lambda phage](#) for chromosomal integration.

Some vectors may include targeting sequence that may target the expressed protein to a specific location such as [organelles](#) in eukaryotes or [bacteria](#).

Expression/Production systems

Different organisms may be used to express a gene's target protein, and the expression vector used will therefore be specific to the particular organism. The most commonly used organism for [protein production](#) is the bacterium *Escherichia coli*. Not all proteins can be successfully expressed in *E. coli*, or be expressed with the correct form of post-translational modifications such as [glycosylations](#), and other systems may therefore be used.

Bacterial



An example of a bacterial expression vector is the pGEX-3x plasmid

The expression host of choice for the expression of many proteins is *Escherichia coli* as the production of heterologous proteins is relatively simple and convenient, as well as being rapid and cheap. A large number of *E. coli* expression plasmids are available to meet a wide variety of needs. Other bacteria used for protein production include *Bacillus subtilis*.

Most heterologous proteins are expressed in the cytoplasm of *E. coli*. However, not all proteins formed may be soluble and incorrectly folded proteins formed in cytoplasm can form insoluble aggregates called **inclusion bodies**. Such proteins often require refolding, which can be an involved process and may not necessarily produce high yield.^[6] Proteins with disulfide bonds are often not able to fold correctly due to the reducing environment in the cytoplasm which prevents such disulfide bond formation. A possible solution is to target the protein to the **periplasmic space** by the use of an N-terminal **signal sequence**. Another approach is to manipulate the redox environment of the cytoplasm.

Other more sophisticated systems are also being developed; such systems may allow for the expression of proteins that are impossible in *E. coli*, such as **glycosylated** proteins.

The promoters used for these vector are usually based on the promoter of the *lac* operon or the T7 promoter,^[11] which is regulated by the *lac* operator. These promoters may also be hybrids of different promoters, for example, the T7 promoter of *trp* and *lac* promoters.

Note that most commonly used *lac* or *lac*-derived promoters are based on the *lacUV5* mutant which is insensitive to repression. This mutant allows for expression of protein under the control of the *lac* promoter when the **growth medium** contains glucose. Presence of glucose would inhibit gene expression if wild-type *lac* promoter is used. Presence of glucose nevertheless may cause background expression through residual inhibition in some systems.

Examples of *E. coli* expression vectors are the pGEX series of vectors where **glutathione S-transferase** is used for protein expression. Expression is under the control of the tac promoter, and the pET series of vectors which uses a T7 promoter.^[18]

It is possible to simultaneously express two or more different proteins in *E. coli* using different plasmids. However, if multiple plasmids are used, each plasmid needs to use a different antibiotic selection as well as a different origin of replication. Plasmids may not be stably maintained. Many commonly used plasmids are based on the **ColE1** replicon and are not compatible with each other; in order for a ColE1-based plasmid to coexist with another in the same cell, the other would need to use a different replicon, e.g. a p15A replicon-based plasmid such as the pACYC series of plasmids. Another approach would be to use a multicistron vector or design the coding sequences in tandem as a bi- or poly-cistronic construct.

Yeast

A yeast commonly used for protein production is *Pichia pastoris*. Examples of yeast expression vector in *Pichia* are *pAOX1* and these vectors use the **AOX1** promoter which is inducible with **methanol**. The plasmids may contain elements such as a polyA signal into the yeast genome and signal sequence for the secretion of expressed protein. Proteins with disulphide bonds are efficiently produced in yeast. Another yeast used for protein production is *Kluyveromyces lactis* and the gene is expressed under the control of the strong **lactase LAC4** promoter.

Saccharomyces cerevisiae is particularly widely used for gene expression studies in yeast, for example in yeast two-hybrid system study of protein-protein interaction. The vectors used in yeast two-hybrid system contain fusion partners for two proteins that activate the transcription of a reporter gene when there is interaction between the two proteins expressed from the clones.

Baculovirus

Baculovirus, a rod-shaped virus which infects insect cells, is used as the expression vector in this system. Insect cells from **Lepidopterans** (moths and butterflies), such as *Spodoptera frugiperda*, are used as host. A cell line derived from *S. frugiperda* of particular interest, as it has been developed to grow fast and without the expensive serum normally needed for cell growth. The **shuttle vector** is called bacmid, and gene expression is under the control of a strong promoter pPO1. It has been used with mammalian cell lines in the **BacMam** system.

Baculovirus is normally used for production of **glycoproteins**, although the glycosylations may be different from those in vertebrates. In general, it is safer to use than mammalian virus as it has a limited host range and does not infect humans. It also allows modifications.

Plant

Many plant expression vectors are based on the **Ti plasmid** of *Agrobacterium tumefaciens*.^[30] In these expression systems, the gene to be inserted into plant is cloned into the **T-DNA**, a stretch of DNA flanked by a 25-bp direct repeat sequence at either end that can integrate into the plant genome. The T-DNA also contains the selectable marker. The *Agrobacterium* provides the machinery for **transformation**, integration of into the plant genome, and the promoters for its *vir* genes may also be used for expression. Concerns over the transfer of bacterial or viral genetic material into the plant however have led to the development of intragenic vectors whereby functional equivalents of plant genome are used so that there is no transfer of genes from the *Agrobacterium* species into the plant.

Plant viruses may be used as vectors since the *Agrobacterium* method does not work for all plants. Examples of plant viruses are the **tobacco mosaic virus** (TMV), **potato virus X**, and **cowpea mosaic virus**. The protein may be expressed as a fusion with the virus and is displayed on the surface of assembled viral particles, or as an unfused protein that accumulates in plant using plant vectors is often constitutive, and a commonly used constitutive promoter in plant expression systems is the **mosaic virus** (CaMV) 35S promoter.

Mammalian

Mammalian expression vectors offer considerable advantages for the expression of mammalian proteins over bacterial systems - proper folding, post-translational modifications, and relevant enzymatic activity. It may also be more desirable than mammalian systems whereby the proteins expressed may not contain the correct glycosylations. It is of particular interest for membrane-associating proteins that require chaperones for proper folding and stability as well as containing numerous modifications. The downside, however, is the low yield of product in comparison to prokaryotic vectors as well as the techniques involved. Its complicated technology, and potential contamination with animal viruses of mammalian origin, has placed a constraint on its use in large-scale industrial production.

Cultured mammalian cell lines such as the **Chinese hamster ovary (CHO)**, **COS**, including human cell lines such as

used to produce protein. Vectors are **transfected** into the cells and the DNA may be integrated into the genome **recombination** in the case of stable transfection, or the cells may be transiently transfected. Examples of mammalian vectors include the **adenoviral** vectors, the pSV and the pCMV series of plasmid vectors, **vaccinia** and **retroviral** vectors, and **baculovirus**. The promoters for **cytomegalovirus** (CMV) and **SV40** are commonly used in mammalian expression systems. Non-viral promoter, such as the elongation factor (EF)-1 promoter, is also known.

Cell-free systems

E. coli **cell lysate** containing the cellular components required for transcription and translation are used in this system for protein production. The advantage of such system is that protein may be produced much faster than those produced *in vivo*. It takes less time to culture the cells, but it is also more expensive. Vectors used for *E. coli* expression can be used in this system. Specialized designed vectors for this system are also available. Eukaryotic cell extracts may also be used in other cell-free systems such as the **wheat germ** cell-free expression systems. Mammalian cell-free systems have also been produced.^[41]

Applications

Laboratory use

Expression vector in an expression host is now the usual method used in laboratories to produce proteins for research. It is commonly produced in *E. coli*, but for glycosylated proteins and those with disulphide bonds, yeast, baculovirus and mammalian cells are used.

Production of peptide and protein pharmaceuticals

Most protein **pharmaceuticals** are now produced through recombinant DNA technology using expression vectors. Examples of protein pharmaceuticals may be hormones, vaccines, antibiotics, antibodies, and enzymes. The first human recombinant protein for disease management, insulin, was introduced in 1982.

Biotechnology allows these peptide and protein pharmaceuticals, some of which were previously rare or difficult to produce in large quantity. It also reduces the risks of contaminants such as host viruses, toxins and **prions**. Examples of risks include **prion** contamination in **growth hormone** extracted from **pituitary glands** harvested from human cadavers which caused **Creutzfeldt–Jakob disease** in patients receiving treatment for **dwarfism**, and viral contaminants in clotting factors in human blood that resulted in the transmission of viral diseases such as **hepatitis** and **AIDS**. Such risk is reduced when the proteins are produced in non-human host cells.

Transgenic plant and animals

In recent years, expression vectors have been used to introduce specific genes into plants and animals to produce transgenic organisms. For example in **agriculture** it is used to produce **transgenic plants**. Expression vectors have been used to introduce the **beta-carotene** gene, into rice plants.

This product is called **golden rice**. This process has also been used to introduce a gene into plants that produce a toxin called **Bacillus thuringiensis toxin** or **Bt toxin** which reduces the need for farmers to apply insecticides since it kills the pest organism. In addition expression vectors are used to extend the ripeness of tomatoes by altering the plant so that it does not produce the chemical that causes the tomatoes to rot.

There have been **controversies** over using expression vectors to modify crops due to the fact that there might be legal possibilities of companies patenting certain **genetically modified food** crops, and ethical concerns. Nevertheless, transgenic crops are being used and heavily researched.

Transgenic animals have also been produced to study animal biochemical processes and human diseases, or used for the production of pharmaceuticals and other proteins. They may also be engineered to have advantageous or useful traits. Green fluorescent protein (GFP) is sometimes used as a tag which results in an animal that can fluoresce, and this has been exploited commercially in the production of fluorescent GloFish.

pMAL™ Protein Fusion and Purification System

The pMAL™ Protein Fusion and Purification System requires a cloned gene to be inserted into a pMAL vector downstream of a maltose-binding protein (MBP) gene. This results in the expression of an MBP-fusion protein. The technique uses the translation initiation signals of MBP to express large amounts of the fusion protein. The fusion protein is then purified by affinity chromatography using a purification specific for MBP.

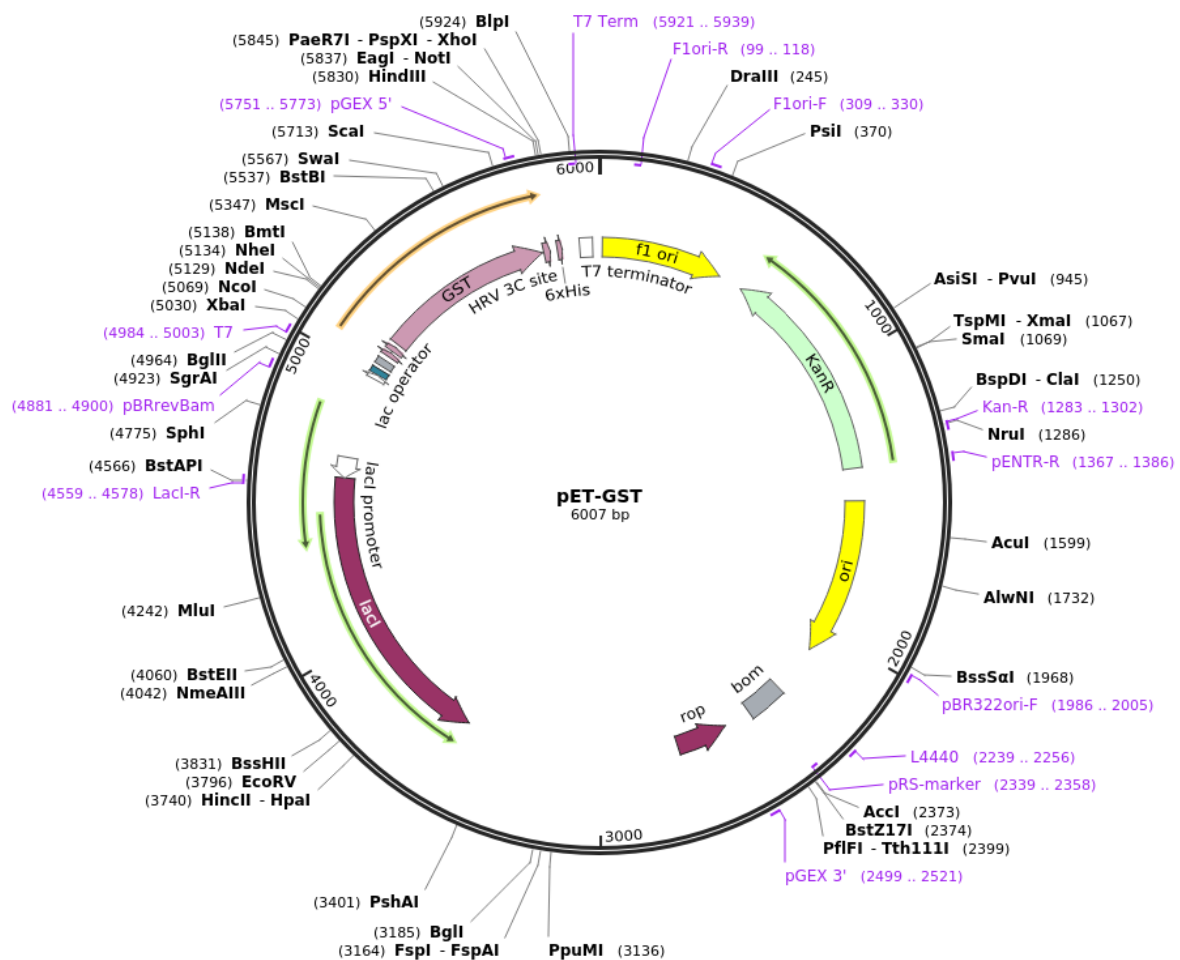
Expression in either the cytoplasm or periplasm: periplasmic expression can enhance folding of proteins with disulfide bonds.

Fusion to MBP has been shown to enhance the solubility of proteins expressed in *E. coli*.

Gentle elution with maltose; no detergents or harsh denaturants required.

Reliable expression and substantial yields (up to 100 mg/L).

Created with SnapGene®



GST-His purification: a two-step affinity purification protocol yielding full-length proteins.

Key assays in enzymology for the biochemical characterization of proteins *in vitro* necessitate high concentrations of protein of interest. Protein purification protocols should combine efficiency, simplicity and cost effectiveness. Here, we describe a new small-scale affinity purification system for recombinant proteins, based on a N-terminal Glutathione Sepharose 4B and a 10xHis tag, which are both fused to the protein of interest. The latter construct is used to generate baculoviruses, for infection of insect cells for protein expression. GST is a rather long tag (29 kDa) which serves to ensure purification efficiency. However, to avoid the physiological properties of the protein. Hence, it is subsequently cleaved off the protein using the PreScission protease. To obtain maximum purity and to remove the cleaved GST, we added a second affinity purification step based on the complementary His tag. Importantly, our technique is based on two different tags flanking the two ends of the protein, which is an efficient method for purifying full-length proteins and, therefore, enriches full-length proteins. The method presented here does not require an expensive instrument like FPLC. Additionally, we incorporated MgCl₂ and ATP washes to remove heat shock protein impurities and nucleases. Finally, we removed contaminating nucleic acids. In summary, the combination of two different tags flanking the N- and the C-terminal of the protein, and the cleavage off one of the tags, guarantees the recovery of a highly purified and full-length protein of interest.

Purification of proteins fused to maltose-binding protein.

Maltose-binding protein (MBP) is one of the most popular fusion partners being used for producing recombinant proteins. It allows one to use a simple capture affinity step on amylose-agarose columns, resulting in a protein that is often 70-90% pure. In protein-isolation applications, MBP provides a high degree of translation and facilitates the proper folding and solubility of the protein. This chapter describes efficient procedures for isolating highly purified MBP-target proteins. Special attention is given to downstream applications such as structural determination studies, protein activity assays, and assessing the chemical stability of the protein.

Intein Applications: From Protein Purification and Labeling to Metabolic Control Mechanisms

Introduction

The ability of inteins to form and cleave specific peptide bonds in a variety of contexts has enabled the development of powerful new tools in molecular biology. Initial applications focused on self-cleaving affinity tags for protein purification and labeling, and utilized mutations that altered the native splicing reactions described by Perler and colleagues in a series of papers. An important advantage of intein-based methods is that they are generally enzymatic in nature and perform highly specific activities under physiological conditions. Although thiol compounds are used in some intein-based chemistries involved in most intein methods are innocuous to their various target proteins. Inteins have been used in a variety of applications, including protein purification, labeling, and metabolic control.

greatly expanded in subsequent years through the discovery of hundreds of additional inteins, and protein activity regulation and modification have followed. An important underlying theme in this discovery and development of highly efficient split inteins for advanced applications *in vitro* and *in vivo* splicing and -cleaving inteins have been employed for applications ranging from the purification of proteins to the *in vivo* control of protein function and labeling.

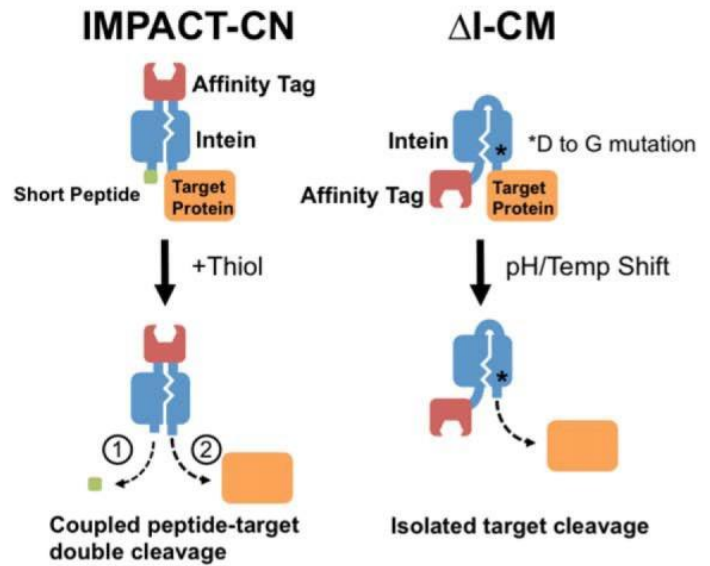
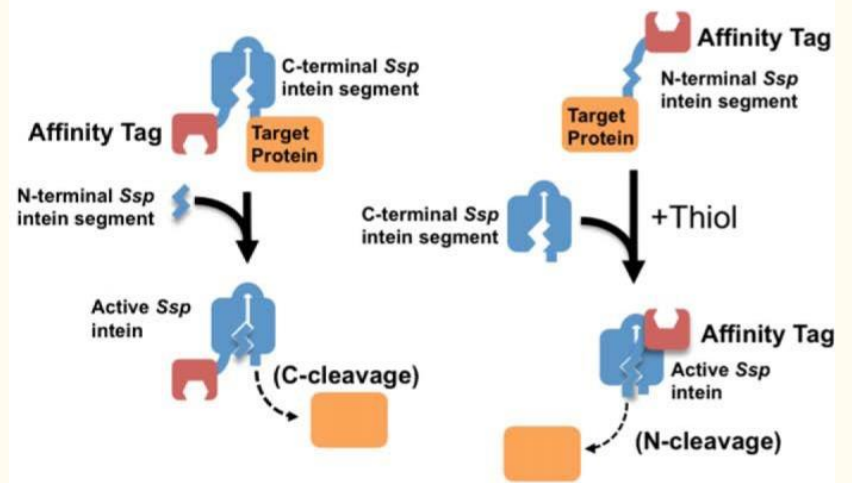
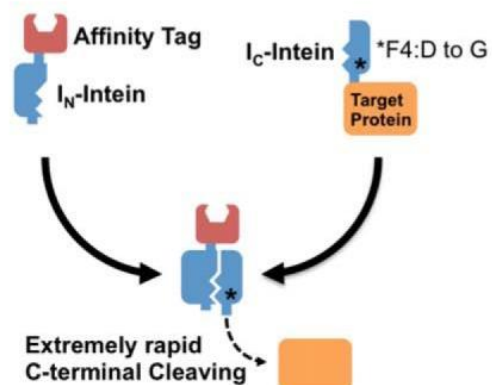
Protein Purification

One of the first major applications of inteins was the development of self-cleaving affinity tags for untagged target proteins in recombinant expression systems. In these applications, a modified intein is fused to an affinity tag and target protein, and once the fusion is affinity-purified, the intein is induced to cleave from the intein and tag.

The first commercial intein system was released by New England Biolabs in 1997 and employed a modified *Sce* VMA1³ intein that is triggered to cleave at its N terminus (IMPACT system), or both N and C termini (IMPACT-CN system), by the addition of thiol compounds. Shortly thereafter, inteins with rapid cleavage activity were published.

These inteins exhibit suppressed cleavage activity at pH 8.5, allowing purification of the tagged target protein, and induced to cleave by a shift to pH 6.5. The cleavage activity of these inteins is also strongly influenced by temperature, where the most rapid cleaving is observed at 37 °C.

The development of these inteins led to a variety of new tag systems and configurations for protein purification, including the chitin and maltose-binding proteins, as well as non-chromatographic purification tags. Small ubiquitin-like modifier (SUMO) and ubiquitin tags have been used to increase expression efficiency, and for purification, and an ELP precipitation tag has been combined with a dockerin-cohesin binding pair for protein expression with a reusable non-chromatographic purification reagent.

A**B****Ssp Split Intein****C****Npu Intein**

The Early Years, **The first recombinant baculovirus vectors designed to express chimeric genes consisting of the polyhedrin promoter and a foreign coding sequence, as described in the preceding section, were produced using a basic homologous recombination approach. The methodological details of this approach were described in the last edition of this volume (Bradley, 1990) and also are available in primary papers and excellent technical manuals from the original contributors (O'Reilly *et al.*, 1992; Summers and Smith, 1987). Thus, this exercise will not be repeated here, as indicated above. However, for background purposes it is important to briefly note that this general method involved (1) construction of a bacterial “transfer” plasmid containing the chimeric gene flanked by sequences derived from the polyhedrin region of the viral genome (Fig. 14.1) and (2) cotransfection of cultured insect cells with a mixture of this transfer plasmid DNA and genomic DNA extracted from purified preparations of wild-type AcMNPV (Fig. 14.2). Homologous recombination between the sequences flanking the chimeric gene of interest in the transfer plasmid and the sequences upstream and downstream of the polyhedrin gene in the wild-type AcMNPV genome produced recombinant viral DNA molecules in these cotransfected insect cells. A double crossover recombination event was necessary to simultaneously knockout the polyhedrin gene and knock-in the chimeric gene encoding the protein of interest. Of course, this was a relatively rare event with an estimated frequency of ~ 0.1% (Smith *et al.*, 1983a). Thus, it was necessary to separate the small minority of recombinant virus progeny from the vast majority of parental viral progeny produced by the cotransfected insect cells by cloning. This was easily accomplished by baculoviral plaque assays, but then one had to be able to distinguish recombinant viral plaques from the much larger background of plaques derived from the parental virus. Initially, this was accomplished using a simple visual screen, as the parental viral progeny produced polyhedron-positive plaques while the recombinant viral progeny, which lacked a functional polyhedrin gene, produced polyhedron-negative plaques. An investigator with a trained eye could visually identify polyhedron-negative plaques by examining the assay plate under a dissecting microscope. However, the trained eye was a key to success and the inability of many investigators to recognize polyhedron-negative baculoviral plaques was a serious problem that constrained the use of the baculovirus–insect cell system as a recombinant protein production tool for several years.**

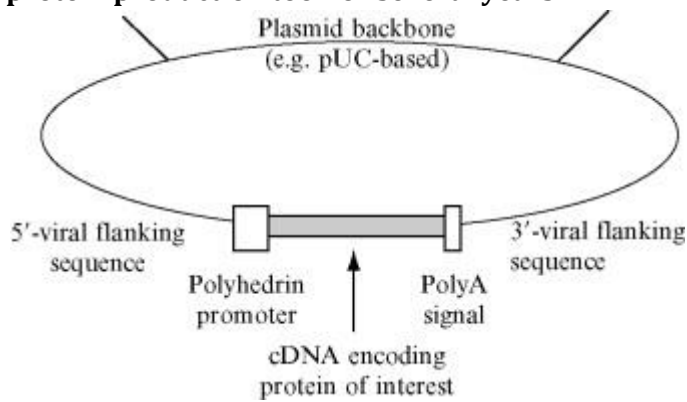
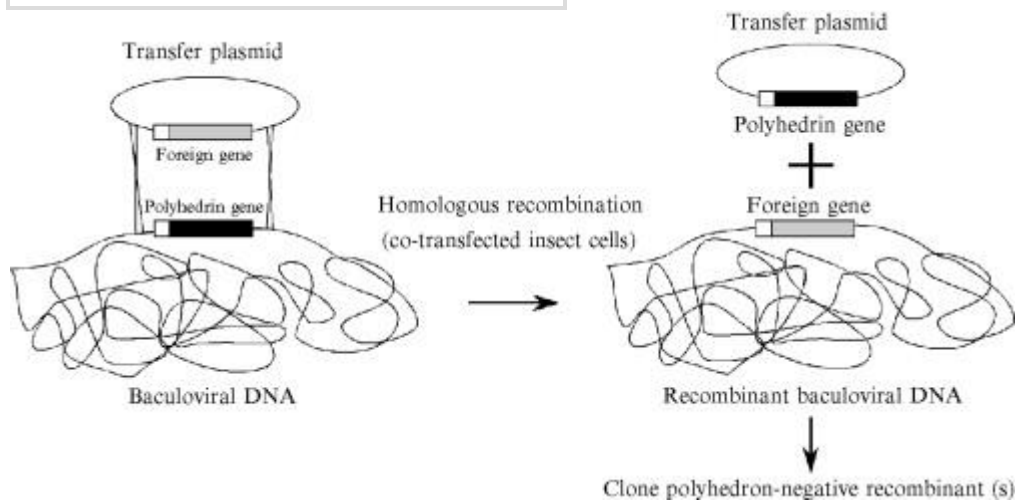


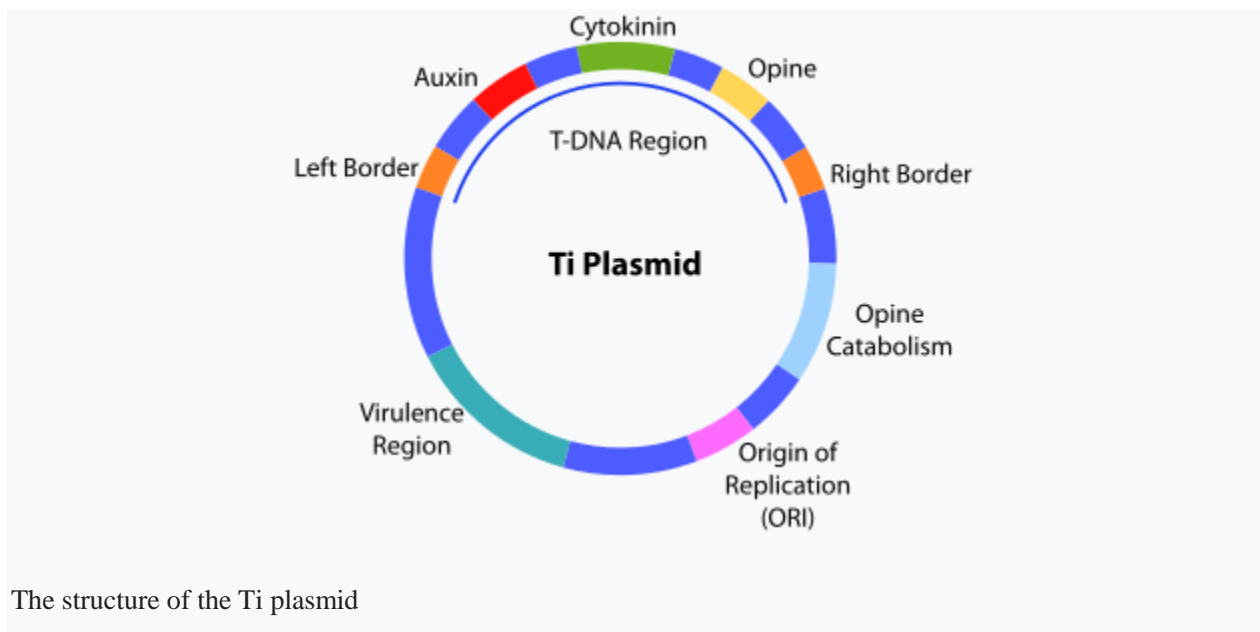
Figure. A simple baculovirus transfer plasmid.



[Sign in to download full-size image](#)

Figure Producing a baculovirus expression vector by homologous recombination.

Ti plasmid



The structure of the Ti plasmid

A **tumour inducing (Ti) plasmid** is a plasmid found in pathogenic species of *Agrobacterium*, including *A. tumefaciens*, *A. rhizogenes*, *A. rubi* and *A. vitis*.

Evolutionarily, the Ti plasmid is part of a family of plasmids carried by many species of **Alphaproteobacteria**. Members of this plasmid family are defined by the presence of a conserved

DNA region known as the *repABC* gene cassette, which mediates the **replication** of the plasmid, the partitioning of the plasmid into daughter cells during **cell division** as well as the maintenance of the plasmid at low copy numbers in a cell. The Ti plasmids themselves are sorted into different categories based on the type of molecule, or **opine**, they allow the bacteria to break down as an energy source.

The presence of this Ti plasmid is essential for the bacteria to cause crown gall disease in plants. This is facilitated via certain crucial regions in the Ti plasmid, including the *vir* region, which encodes for virulence genes, and the **transfer DNA (T-DNA)** region, which is a section of the Ti plasmid that is **transferred** via **conjugation** into host plant cells after an injury site is sensed by the bacteria.

These regions have features that allow the delivery of T-DNA into host plant cells, and can modify the host plant cell to cause the synthesis of molecules like **plant hormones** (e.g. **auxins**, **cytokinins**) and opines and the formation of crown gall tumours.

Because the T-DNA region of the Ti plasmid can be transferred from bacteria to plant cells, it represented an exciting avenue for the transfer of DNA between **kingdoms** and spurred large amounts of research on the Ti plasmid and its possible uses in bioengineering.



Nomenclature and classification

The Ti plasmid is a member of a plasmid family found in Alphaproteobacteria. These plasmids are often relatively large in size, ranging from 100kbp to 2Mbp. They are also often termed **replicons**, as their replication begins at a single site. Members of this family have a characteristic *repABC* gene cassette. Another notable member of this family is the root inducing (Ri) plasmid carried by *A. rhizogenes*, which causes another plant disease known as hairy root disease.

A key feature of Ti plasmids is their ability to drive the production of opines, which are derivatives of various **amino acids** or **sugar phosphates**, in host plant cells. These opines can then be used as a nutrient for the infecting bacteria, which **catabolizes** the respective opines using genes encoded in the Ti plasmid.

Accordingly, Ti plasmids have been classified based on the type of opine they catabolize, namely: **nopaline-**, **octopine-** or mannityl-types, which are amino acid derivatives, or agrocinnopine-type, which are sugar phosphate derivatives

Historical discovery

The identification of *A. tumefaciens* as the cause of gall tumours in plants paved the way for insights into the molecular basis of crown gall disease.

The first indication of a genetic effect on host plant cells came in 1942-1943, where plant cells of secondary tumours were found to lack any bacterial cells within. However, these tumour cells did possess the ability to produce opines metabolized by the infecting bacterial strain.

Crucially, the production of the respective opines occurred regardless of the plant species and occasionally only within crown gall tissues, indicating that the bacteria had transferred some genetic material to the host plant cells in order to allow opine synthesis.

However, how and to what extent did DNA transfer occur remained an open question. Adding *A. tumefaciens* DNA alone did not cause tumors in plants, while very little *A. tumefaciens* DNA was found to be integrated into the host plant cell genome.

The addition of **deoxyribonucleases** (DNases) to degrade DNA also failed to prevent the formation and growth of the plant tumors. These suggested that little, if any, of the *A. tumefaciens* DNA is transferred to the host plant cell to cause disease and, if DNA is indeed transferred from the bacteria to the plant, it must occur in a protected manner.

Subsequently, **oncogenic** bacterial strains were found to be able to convert non-pathogenic bacteria into pathogens via the process of conjugation, where the genes responsible for virulence were transferred to the non-pathogenic cells.

The role of a plasmid in this pathogenic ability was further supported when large plasmids were found only in pathogenic bacteria but not avirulent bacteria. Eventually, the detection of parts of bacterial plasmids in host plant cells was established, confirming that this was the genetic material responsible for the genetic effect of infection.

With the identification of the Ti plasmid, many studies were carried out to determine the characteristics of the Ti plasmid and how the genetic material is transferred from the *Agrobacterium* to the plant host. Some notable early milestones in the studies of Ti plasmids include the mapping of a Ti plasmid in 1978 and the studying of sequence similarity between different Ti plasmids in 1981.

Between 1980–2000, the characterization of the T-DNA region and the 'vir' region was also pursued. Studies into the T-DNA region determined their process of transfer and identified genes allowing the synthesis of plant hormones and opines.

Separately, early work aimed to determine the functions of the genes encoded in the 'vir' region - these were broadly categorized into those that allowed bacterial-host interactions and those that enabled T-DNA delivery.

Ri plasmids

The abilities of *Agrobacterium tumefaciens* and *A. rhizogenes* to transform dicotyledons and cause crown gall and hairy root disease are caused by the presence of tumor inducing (Ti) and root inducing (Ri) plasmids. During transformation plasmid T-DNA (transferred DNA) is inserted into the plant genome.

The T-region is flanked by 25 bp direct repeats, which are essential for transfer. The T-regions contain oncogenes that are expressed in the plants. Some of these code for enzymes that synthesize auxin or cytokinin.

Another type, present in Ri plasmids only, appears to impose a high hormone sensitivity on the infected tissue. The T-DNA also contains genes for enzymes synthesizing opines, which the bacteria catabolize. The T-DNA transfer is initiated by the induction of genes in the virulence (vir) region of the plasmid by phenolic compounds secreted by wounded tissue.

The products of the *vir*-genes and of chromosomal genes mediate transfer of T-DNA to the plant cells. Crown gall disease is caused by production of auxin and cytokinin by the transferred T-DNA.

The T-DNA of Ri plasmids codes for at least three genes that each can induce root formation, and that together cause hairy root formation from plant tissue. Current results indicate that the products of these genes induce a potential for increased auxin sensitivity that is expressed when the transformed cells are subjected to a certain level of auxin. After this stage the transformed roots can be grown in culture without exogenous supply of hormones

Yeast artificial chromosome

Yeast artificial chromosomes (YACs) are genetically engineered chromosomes derived from the DNA of the yeast, *Saccharomyces cerevisiae*, which is then ligated into a bacterial plasmid. By inserting large fragments of DNA, from 100–1000 kb, the inserted sequences can be cloned and physically mapped using a process called chromosome walking.

This is the process that was initially used for the [Human Genome Project](#), however due to stability issues, YACs were abandoned for the use of [Bacterial artificial chromosomes \(BAC\)](#). Beginning with the initial research of the Rankin et al., Strul et al., and Hsaio et al., the inherently fragile chromosome was stabilized by discovering the necessary [autonomously replicating sequence \(ARS\)](#); a refined YAC utilizing this data was described in 1983 by Murray et al.

The primary components of a YAC are the ARS, centromere, and telomeres from *S. cerevisiae*. Additionally, selectable marker genes, such as antibiotic resistance and a visible marker, are utilized to select transformed yeast cells. Without these sequences, the chromosome will not be stable during extracellular replication, and would not be distinguishable from colonies without the vector.^[3]



This is a photo of two copies of the Washington University Human Genome YAC Library. Each of the stacks is approximately 12 microtiter plates. Each plate has 96 wells, each with different yeast clones.



Construction

A YAC is built using an initial circular DNA **plasmid**, which is typically cut into a linear DNA molecule using **restriction enzymes**; **DNA ligase** is then used to ligate a DNA sequence or gene of interest into the linearized DNA, forming a single large, circular piece of DNA. The basic generation of linear yeast artificial chromosomes can be broken down into 6 main steps:

1. Ligation of selectable marker into plasmid vector: this allows for the differential selection of colonies with, or without the marker gene. An **antibiotic resistance** gene allows the YAC vector to be amplified and selected for in *E. coli* by rescuing the ability of mutant *E. coli* to synthesize **leucine** in the presence of the necessary components within the growth medium. *TRP1* and *URA3* genes are other selectable markers. The YAC vector cloning site for foreign DNA is located within the *SUP4* gene. This gene compensates for a mutation in the yeast host cell that causes the accumulation of red pigment. The host cells are normally red, and those **transformed** with YAC only, will form colorless colonies. Cloning of a foreign DNA fragment into the YAC causes insertional inactivation of the gene, restoring the red color. Therefore, the colonies that contain the foreign DNA fragment are red.
2. Ligation of necessary centromeric sequences for mitotic stability
3. Ligation of Autonomously Replicating Sequences (ARS) providing an origin of replication to undergo mitotic replication. This allows the plasmid to replicate extrachromosomally, but renders the plasmid highly mitotically unstable, and easily lost without the centromeric sequences.
4. Ligation of artificial telomeric sequences to convert circular plasmid into a linear piece of DNA
5. Insertion of DNA sequence to be amplified (up to 1000kb)
6. Transformation yeast colony

Shuttle vector

A **shuttle vector** is a **vector** (usually a **plasmid**) constructed so that it can propagate in two different host species. Therefore, DNA inserted into a shuttle vector can be tested or manipulated in two different cell types. The main advantage of these vectors is they can be manipulated in *E. coli*, then used in a system which is more difficult or slower to use (e.g. yeast).

Shuttle vectors include plasmids that can propagate in **eukaryotes** and **prokaryotes** (e.g. both *Saccharomyces cerevisiae* and *Escherichia coli*) or in different species of bacteria (e.g. both *E. coli* and *Rhodococcus erythropolis*). There are also **adenovirus** shuttle vectors, which can propagate in *E. coli* and mammals.

Shuttle vectors are frequently used to quickly make multiple copies of the gene in *E. coli* (amplification). They can also be used for *in vitro* experiments and modifications (e.g. mutagenesis, PCR)

One of the most common types of shuttle vectors is the yeast shuttle vector. Almost all commonly used *S. cerevisiae* vectors are shuttle vectors. Yeast shuttle vectors have components that allow for replication and selection in both *E. coli* cells and yeast cells. The *E. coli* component of a yeast shuttle vector includes an origin of replication and a selectable marker, e.g. antibiotic resistance, beta lactamase, beta galactosidase. The yeast component of a yeast shuttle vector includes an autonomously replicating sequence (ARS), a yeast centromere (CEN), and a yeast selectable marker (e.g. URA3, a gene that encodes an enzyme for uracil synthesis, .

UNIT-III

INTRODUCTION OF DNA INTO CELLS

Cells have membranes that prevent DNA from simply diffusing in or out. This is the initial barrier that scientists must overcome in order to insert foreign DNA into a cell. The four ways of accomplishing this goal are transduction, transformation, transfection and injection. But before these four methods are employed, scientists must prepare the foreign DNA by cutting it into smaller pieces or by using restriction enzymes to cut the DNA and insert the desired portion into a bacterial plasmid. Plasmids are circular pieces of DNA that can be passed between bacteria and viruses.

Viral Transduction

Transduction is the insertion of foreign DNA into a cell via a virus . Viruses are made of a protein coat that houses DNA within. Viruses can bind to living cells and inject their DNA. Or, viruses can push into the host as a membrane-bound vesicle, before releasing their DNA inside the host. The use of recombinant DNA technology can insert foreign DNA into host cells that are then purposefully infected with viruses. When the virus produces more of itself in the host cell, it also packages copies of the foreign DNA into the new viruses. When these new viruses burst out of the host cell, they are now carriers of the foreign DNA and can be used to introduce this DNA into other host cells.

Transformation and Transfection

Transformation is a way that bacterial cells pick up pieces of DNA from their environment. Exactly how this happens is unknown, but what is known is that exposing the bacteria to calcium chloride followed by heat will cause it to take up pieces of DNA. Another way to introduce foreign DNA into bacteria is to transform or transduce them with the DNA and then allow them to mate. Bacterial mating is called conjugation, and occurs when two bacteria exchange DNA through a tube that connects them. Transfection can also be done on eukaryotic cells, though the exact way in which this happens is also unknown. In eukaryotes, mixing foreign DNA with calcium phosphate creates particles that fuse with the host cell's membrane. It is believed that the host may engulf the particle, a process called endocytosis.

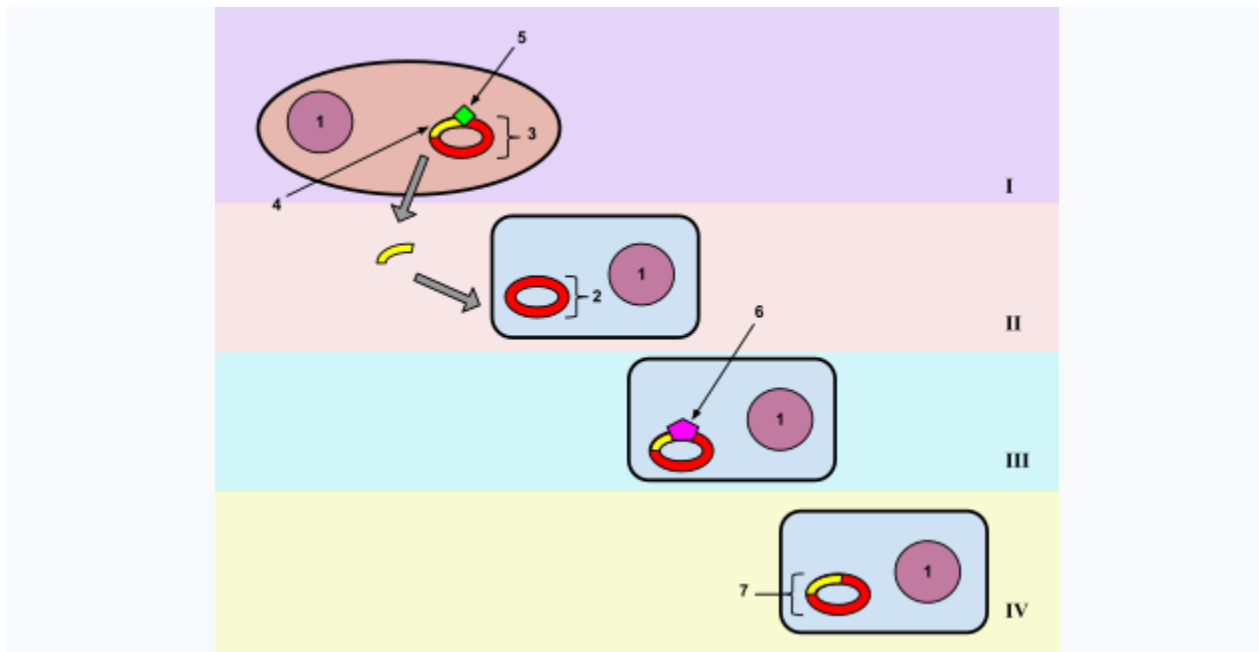
Agrobacterium

Agrobacterium are bacteria that cause tumors, called crown galls, in plants. Agrobacteria are drawn to plants that have been wounded or chopped up, because sugar spills out from the wound, which the bacteria can sense. Agrobacterium have a plasmid, called the Ti plasmid, that contains genes that perform the transfection of plant cells. The Ti plasmid has a region called the T-DNA, which is cut out of the plasmid and carried by bacterial proteins out of the bacteria into the plant cell. Insertion of foreign DNA into the T-DNA region by restriction enzymes is one way that scientists insert genes into plants.

Injection

A common way of introducing foreign DNA into plants is to physically inject the DNA with a gene gun. The concept of the gene gun is to coat microscopic particles of gold or tungsten with the foreign DNA. These particles are then loaded into the gun, which contains pressurized helium gas. A release of the gas propels the DNA-coated particles out of the gun like bullets. These particles penetrate the cell walls of plants and release the foreign DNA, which is now part of the plant cell. Gene guns can be used directly on the leaf of a plant or on plant cells that have been isolated from ground-up plant tissue.

Transformation (genetics)



In this image, a gene from bacterial cell 1 is moved from bacterial cell 1 to bacterial cell 2. This process of bacterial cell 2 taking up new genetic material is called transformation.

In **molecular biology and genetics**, **transformation** is the **genetic** alteration of a **cell** resulting from the direct uptake and incorporation of **exogenous genetic material** from its surroundings through the **cell membrane(s)**. For transformation to take place, the recipient bacterium must be in a state of **competence**, which might occur in nature as a time-limited response to environmental conditions such as starvation and cell density, and may also be induced in a laboratory.

Transformation is one of three processes for **horizontal gene transfer**, in which exogenous genetic material passes from one bacterium to another, the other two being **conjugation** (transfer of **genetic material** between two bacterial cells in direct contact) and **transduction** (injection of foreign DNA by a **bacteriophage** virus into the host bacterium). In transformation, the genetic material passes through the intervening medium, and uptake is completely dependent on the recipient bacterium.

As of 2014 about 80 species of bacteria were known to be capable of transformation, about evenly divided between [Gram-positive](#) and [Gram-negative bacteria](#); the number might be an overestimate since several of the reports are supported by single papers.

"Transformation" may also be used to describe the insertion of new genetic material into nonbacterial cells, including animal and plant cells; however, because "[transformation](#)" has a special meaning in relation to animal cells, indicating progression to a cancerous state, the process is usually called "[transfection](#)".

History

Transformation in bacteria was first demonstrated in 1928 by the British bacteriologist [Frederick Griffith](#). Griffith was interested in determining whether injections of heat-killed bacteria could be used to vaccinate mice against pneumonia. However, he discovered that a non-virulent strain of *Streptococcus pneumoniae* could be made [virulent](#) after being exposed to heat-killed virulent strains. Griffith hypothesized that some "[transforming principle](#)" from the heat-killed strain was responsible for making the harmless strain virulent. In 1944 this "transforming principle" was identified as being genetic by [Oswald Avery](#), [Colin MacLeod](#), and [Maclyn McCarty](#).

They isolated DNA from a virulent strain of *S. pneumoniae* and using just this DNA were able to make a harmless strain virulent. They called this uptake and incorporation of DNA by bacteria "transformation" (See [Avery-MacLeod-McCarty experiment](#))

The results of Avery et al.'s experiments were at first skeptically received by the scientific community and it was not until the development of [genetic markers](#) and the discovery of other methods of genetic transfer ([conjugation](#) in 1947 and [transduction](#) in 1953) by [Joshua Lederberg](#) that Avery's experiments were accepted.

It was originally thought that *Escherichia coli*, a commonly used laboratory organism, was refractory to transformation. However, in 1970, Morton Mandel and Akiko Higa showed that *E. coli* may be induced to take up DNA from [bacteriophage λ](#) without the use of [helper phage](#) after treatment with calcium chloride solution.

Two years later in 1972, [Stanley Norman Cohen](#), Annie Chang and Leslie Hsu showed that CaCl₂ treatment is also effective for transformation of plasmid DNA. The method of transformation by Mandel and Higa was later improved upon by [Douglas Hanahan](#).

The discovery of artificially induced competence in *E. coli* created an efficient and convenient procedure for transforming bacteria which allows for simpler [molecular cloning](#) methods in [biotechnology](#) and [research](#), and it is now a routinely used laboratory procedure.

Transformation using [electroporation](#) was developed in the late 1980s, increasing the efficiency of in-vitro transformation and increasing the number of [bacterial strains](#) that could be transformed. Transformation of animal and plant cells was also investigated with the first [transgenic mouse](#) being created by injecting a gene for a rat growth hormone into a mouse embryo in 1982.

In 1907 a bacterium that caused plant tumors, *Agrobacterium tumefaciens*, was discovered and in the early 1970s the tumor-inducing agent was found to be a DNA plasmid called the Ti plasmid.

By removing the genes in the plasmid that caused the tumor and adding in novel genes, researchers were able to infect plants with *A. tumefaciens* and let the bacteria insert their chosen DNA into the genomes of the plants.

Not all plant cells are susceptible to infection by *A. tumefaciens*, so other methods were developed, including electroporation and micro-injection. Particle bombardment was made possible with the invention of the Biolistic Particle Delivery System (gene gun) by John Sanford in the 1980s.

Definitions

Transformation is one of three forms of horizontal gene transfer that occur in nature among bacteria, in which DNA encoding for a trait passes from one bacterium to another and is integrated into the recipient genome by homologous recombination; the other two are transduction, carried out by means of a bacteriophage, and conjugation, in which a gene is passed through direct contact between bacteria.

In transformation, the genetic material passes through the intervening medium, and uptake is completely dependent on the recipient bacterium.

Competence refers to a temporary state of being able to take up exogenous DNA from the environment; it may be induced in a laboratory.

It appears to be an ancient process inherited from a common prokaryotic ancestor that is a beneficial adaptation for promoting recombinational repair of DNA damage, especially damage acquired under stressful conditions. Natural genetic transformation appears to be an adaptation for repair of DNA damage that also generates genetic diversity.

Transformation has been studied in medically important Gram-negative bacteria species such as *Helicobacter pylori*, *Legionella pneumophila*, *Neisseria meningitidis*, *Neisseria gonorrhoeae*, *Haemophilus influenzae* and *Vibrio cholerae*.

It has also been studied in Gram-negative species found in soil such as *Pseudomonas stutzeri*, *Acinetobacter baylyi*, and Gram-negative plant pathogens such as *Ralstonia solanacearum* and *Xylella fastidiosa*.

Transformation among Gram-positive bacteria has been studied in medically important species such as *Streptococcus pneumoniae*, *Streptococcus mutans*, *Staphylococcus aureus* and *Streptococcus sanguinis* and in Gram-positive soil bacterium *Bacillus subtilis*.^[17] It has also been reported in at least 30 species of *Proteobacteria* distributed in the classes alpha, beta, gamma and epsilon.

The best studied *Proteobacteria* with respect to transformation are the medically important human pathogens *Neisseria gonorrhoeae* (class beta), *Haemophilus influenzae* (class gamma) and *Helicobacter pylori* (class epsilon)

"Transformation" may also be used to describe the insertion of new genetic material into nonbacterial cells, including animal and plant cells; however, because "transformation" has a

special meaning in relation to animal cells, indicating progression to a cancerous state, the process is usually called "**transfection**".

Natural competence and transformation

As of 2014 about 80 species of bacteria were known to be capable of transformation, about evenly divided between **Gram-positive** and **Gram-negative bacteria**; the number might be an overestimate since several of the reports are supported by single papers.

Naturally competent bacteria carry sets of genes that provide the protein machinery to bring DNA across the cell membrane(s). The transport of the exogenous DNA into the cells may require proteins that are involved in the assembly of **type IV pili** and **type II secretion system**, as well as DNA **translocase** complex at the cytoplasmic membrane.

Due to the differences in structure of the cell envelope between Gram-positive and Gram-negative bacteria, there are some differences in the mechanisms of DNA uptake in these cells, however most of them share common features that involve related proteins.

The DNA first binds to the surface of the competent cells on a DNA receptor, and passes through the **cytoplasmic membrane** via DNA translocase.

Only single-stranded DNA may pass through, the other strand being degraded by nucleases in the process. The translocated single-stranded DNA may then be integrated into the bacterial chromosomes by a **RecA**-dependent process.

In Gram-negative cells, due to the presence of an extra membrane, the DNA requires the presence of a channel formed by secretins on the outer membrane. **Pilin** may be required for competence, but its role is uncertain.

The uptake of DNA is generally non-sequence specific, although in some species the presence of specific DNA uptake sequences may facilitate efficient DNA uptake.

Natural transformation

Natural transformation is a bacterial adaptation for DNA transfer that depends on the expression of numerous bacterial genes whose products appear to be responsible for this process.

In general, transformation is a complex, energy-requiring developmental process. In order for a bacterium to bind, take up and recombine exogenous DNA into its chromosome, it must become competent, that is, enter a special physiological state. Competence development in *Bacillus subtilis* requires expression of about 40 genes.

The DNA integrated into the host chromosome is usually (but with rare exceptions) derived from another bacterium of the same species, and is thus homologous to the resident chromosome.

In *B. subtilis* the length of the transferred DNA is greater than 1271 kb (more than 1 million bases). The length transferred is likely double stranded DNA and is often more than a third of the total chromosome length of 4215 kb.

It appears that about 7-9% of the recipient cells take up an entire chromosome.

The capacity for natural transformation appears to occur in a number of prokaryotes, and thus far 67 prokaryotic species (in seven different phyla) are known to undergo this process.

Competence for transformation is typically induced by high cell density and/or nutritional limitation, conditions associated with the stationary phase of bacterial growth. Transformation in *Haemophilus influenzae* occurs most efficiently at the end of exponential growth as bacterial growth approaches stationary phase.

Transformation in *Streptococcus mutans*, as well as in many other streptococci, occurs at high cell density and is associated with **biofilm** formation. Competence in *B. subtilis* is induced toward the end of logarithmic growth, especially under conditions of amino acid limitation.

Similarly, in *Micrococcus luteus* (a representative of the less well studied *Actinobacteria* phylum), competence develops during the mid-late exponential growth phase and is also triggered by amino acids starvation.

By releasing intact host and plasmid DNA, certain **bacteriophages** are thought to contribute to transformation.

Transformation, as an adaptation for DNA repair

Competence is specifically induced by DNA damaging conditions. For instance, transformation is induced in *Streptococcus pneumoniae* by the DNA damaging agents mitomycin C (a DNA cross-linking agent) and fluoroquinolone (a topoisomerase inhibitor that causes double-strand breaks).

In *B. subtilis*, transformation is increased by UV light, a DNA damaging agent. In *Helicobacter pylori*, ciprofloxacin, which interacts with DNA gyrase and introduces double-strand breaks, induces expression of competence genes, thus enhancing the frequency of transformation

Using *Legionella pneumophila*, Charpentier et al. tested 64 toxic molecules to determine which of these induce competence. Of these, only six, all DNA damaging agents, caused strong induction. These DNA damaging agents were mitomycin C (which causes DNA inter-strand crosslinks), norfloxacin, ofloxacin and nalidixic acid (inhibitors of DNA gyrase that cause double-strand breaks), bicyclomycin (causes single- and double-strand breaks), and hydroxyurea (induces DNA base oxidation). UV light also induced competence in *L. pneumophila*.

Charpentier et al. suggested that competence for transformation probably evolved as a DNA damage response.

Logarithmically growing bacteria differ from stationary phase bacteria with respect to the number of genome copies present in the cell, and this has implications for the capability to carry out an important **DNA repair** process. During logarithmic growth, two or more copies of any particular region of the chromosome may be present in a bacterial cell, as cell division is not precisely matched with chromosome replication.

The process of homologous recombinational repair (HRR) is a key DNA repair process that is especially effective for repairing double-strand damages, such as double-strand breaks. This process depends on a second homologous chromosome in addition to the damaged chromosome. During logarithmic growth, a DNA damage in one chromosome may be repaired by HRR using sequence information from the other homologous chromosome.

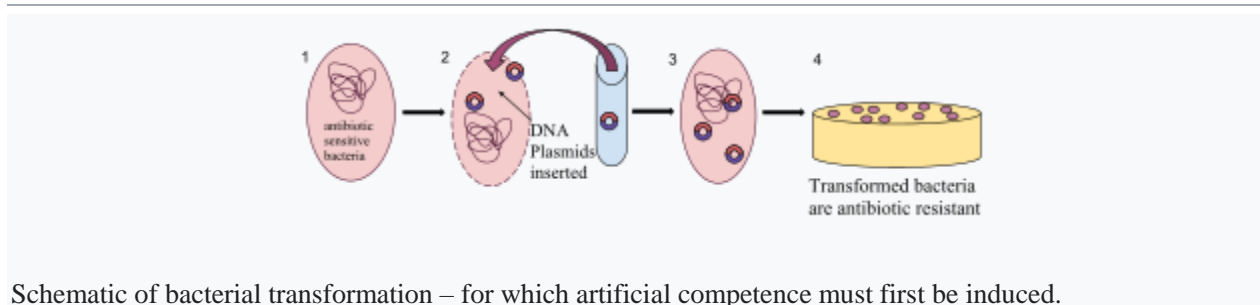
Once cells approach stationary phase, however, they typically have just one copy of the chromosome, and HRR requires input of homologous template from outside the cell by transformation.

To test whether the adaptive function of transformation is repair of DNA damages, a series of experiments were carried out using *B. subtilis* irradiated by UV light as the damaging agent (reviewed by Michod et al. and Bernstein et al.

The results of these experiments indicated that transforming DNA acts to repair potentially lethal DNA damages introduced by UV light in the recipient DNA. The particular process responsible for repair was likely HRR. Transformation in bacteria can be viewed as a primitive sexual process, since it involves interaction of homologous DNA from two individuals to form recombinant DNA that is passed on to succeeding generations.

Bacterial transformation in prokaryotes may have been the ancestral process that gave rise to meiotic sexual reproduction in eukaryotes (see [Evolution of sexual reproduction; Meiosis.](#))

Methods and mechanisms of transformation in laboratory



Schematic of bacterial transformation – for which artificial competence must first be induced.

Bacterial

Artificial competence can be induced in laboratory procedures that involve making the cell passively permeable to DNA by exposing it to conditions that do not normally occur in nature. Typically the cells are incubated in a solution containing **divalent cations** (often **calcium chloride**) under cold conditions, before being exposed to a heat pulse (heat shock).

Calcium chloride partially disrupts the cell membrane, which allows the recombinant DNA to enter the host cell. Cells that are able to take up the DNA are called competent cells.

It has been found that growth of Gram-negative bacteria in 20 mM Mg reduces the number of protein-to-**lipopolysaccharide** bonds by increasing the ratio of ionic to covalent bonds, which increases membrane fluidity, facilitating transformation.

The role of lipopolysaccharides here are verified from the observation that shorter O-side chains are more effectively transformed – perhaps because of improved DNA accessibility.

The surface of bacteria such as *E. coli* is negatively charged due to **phospholipids** and **lipopolysaccharides** on its cell surface, and the DNA is also negatively charged.

One function of the divalent cation therefore would be to shield the charges by coordinating the phosphate groups and other negative charges, thereby allowing a DNA molecule to adhere to the cell surface.

DNA entry into *E. coli* cells is through channels known as zones of adhesion or Bayer's junction, with a typical cell carrying as many as 400 such zones. Their role was established when **cobalamine** (which also uses these channels) was found to competitively inhibit DNA uptake.

Another type of channel implicated in DNA uptake consists of poly (HB):poly P:Ca. In this poly (HB) is envisioned to wrap around DNA (itself a polyphosphate), and is carried in a shield formed by Ca ions.

It is suggested that exposing the cells to divalent cations in cold condition may also change or weaken the cell surface structure, making it more permeable to DNA. The heat-pulse is thought to create a thermal imbalance across the cell membrane, which forces the DNA to enter the cells through either cell pores or the damaged cell wall.

Electroporation is another method of promoting competence. In this method the cells are briefly shocked with an **electric field** of 10-20 kV/cm, which is thought to create holes in the cell membrane through which the plasmid DNA may enter. After the electric shock, the holes are rapidly closed by the cell's membrane-repair mechanisms.

Yeast

Most species of **yeast**, including *Saccharomyces cerevisiae*, may be transformed by exogenous DNA in the environment. Several methods have been developed to facilitate this transformation at high frequency in the lab.

- Yeast cells may be treated with enzymes to degrade their cell walls, yielding **spheroplasts**. These cells are very fragile but take up foreign DNA at a high rate.
- Exposing intact yeast cells to **alkali cations** such as those of **caesium** or **lithium** allows the cells to take up plasmid DNA. Later protocols adapted this transformation method, using **lithium acetate**, **polyethylene glycol**, and single-stranded DNA.
- In these protocols, the single-stranded DNA preferentially binds to the yeast cell wall, preventing plasmid DNA from doing so and leaving it available for transformation.
- **Electroporation**: Formation of transient holes in the cell membranes using electric shock; this allows DNA to enter as described above for bacteria.
- Enzymatic digestion or agitation with glass beads may also be used to transform yeast cells.

Efficiency – Different yeast genera and species take up foreign DNA with different efficiencies. Also, most transformation protocols have been developed for baker's yeast, *S. cerevisiae*, and thus may not be optimal for other species.

Even within one species, different strains have different transformation efficiencies, sometimes different by three orders of magnitude. For instance, when *S. cerevisiae* strains were transformed with 10 ug of plasmid YEp13, the strain DKD-5D-H yielded between 550 and 3115 colonies while strain OS1 yielded fewer than five colonies.

Plants

A number of methods are available to transfer DNA into plant cells. Some **vector**-mediated methods are:

- *Agrobacterium*-mediated transformation is the easiest and most simple plant transformation. Plant tissue (often leaves) are cut into small pieces, e.g. 10x10mm, and soaked for ten minutes in a fluid containing suspended *Agrobacterium*. The bacteria will attach to many of the plant cells exposed by the cut.
- The plant cells secrete wound-related phenolic compounds which in turn act to upregulate the virulence operon of the *Agrobacterium*. The virulence operon includes many genes that encode for proteins that are part of a Type IV secretion system that exports from the bacterium proteins and DNA (delineated by specific recognition motifs called border sequences and excised as a single strand from the virulence plasmid) into the plant cell through a structure called a pilus.
- The transferred DNA (called T-DNA) is piloted to the plant cell nucleus by nuclear localization signals present in the *Agrobacterium* protein VirD2, which is covalently attached to the end of the T-DNA at the Right border (RB).
- Exactly how the T-DNA is integrated into the host plant genomic DNA is an active area of plant biology research. Assuming that a selection marker (such as an antibiotic resistance gene) was included in the T-DNA, the transformed plant tissue can be cultured on selective media to produce shoots.
- The shoots are then transferred to a different medium to promote root formation. Once roots begin to grow from the transgenic shoot, the plants can be transferred to soil to complete a normal life cycle (make seeds). The seeds from this first plant (called the T1, for first transgenic generation) can be planted on a selective (containing an antibiotic), or if an herbicide resistance gene was used, could alternatively be planted in soil, then later treated with herbicide to kill wildtype segregants. Some plants species, such as *Arabidopsis thaliana* can be transformed by dipping the flowers or whole plant, into a suspension of *Agrobacterium tumefaciens*, typically strain C58 (C=Cherry, 58=1958, the year in which this particular strain of *A. tumefaciens* was isolated from a cherry tree in an orchard at Cornell University in Ithaca, New York).
- Though many plants remain recalcitrant to transformation by this method, research is ongoing that continues to add to the list the species that have been successfully modified in this manner.
- **Viral transformation (transduction)**: Package the desired genetic material into a suitable plant virus and allow this modified virus to infect the plant. If the genetic material is DNA, it can recombine with the chromosomes to produce transformant cells. However, genomes of most plant viruses consist of single stranded **RNA** which replicates in the cytoplasm of infected cell. For such genomes this method is a form of **transfection** and not a real transformation, since the inserted genes never reach the nucleus of the cell and do not integrate into the host genome. The progeny of the infected plants is virus-free and also free of the inserted gene.

Some vector-less methods include:

- **Gene gun**: Also referred to as particle bombardment, microprojectile bombardment, or biolistics. Particles of gold or tungsten are coated with DNA and then shot into young plant cells or plant embryos. Some genetic material will stay in the cells and transform them. This method also allows transformation of plant plastids. The **transformation efficiency** is lower than in *Agrobacterium*-mediated transformation, but most plants can be transformed with this method.
- **Electroporation**: Formation of transient holes in cell membranes using electric pulses of high field strength; this allows DNA to enter as described above for bacteria.

Fungi

There are some methods to produce transgenic **fungi** most of them being analogous to those used for plants. However, fungi have to be treated differently due to some of their microscopic and biochemical traits:

- A major issue is the **dikaryotic state** that parts of some fungi are in; dikaryotic cells contain two haploid nuclei, one of each parent fungus. If only one of these gets transformed, which is the rule, the percentage of transformed nuclei decreases after each **sporulation**.
- Fungal cell walls are quite thick hindering DNA uptake so (partial) removal is often required; complete degradation, which is sometimes necessary, yields **protoplasts**.
- Mycelial fungi consist of filamentous **hyphae**, which are, if at all, separated by internal cell walls interrupted by pores big enough to enable nutrients and organelles, sometimes even nuclei, to travel through each hypha. As a result, individual cells usually cannot be separated. This is problematic as neighbouring transformed cells may render untransformed ones immune to selection treatments, e.g. by delivering nutrients or proteins for antibiotic resistance.
- Additionally, growth (and thereby mitosis) of these fungi exclusively occurs at the tip of their hyphae which can also deliver issues.

As stated earlier, an array of methods used for plant transformation do also work in fungi:

- Agrobacterium is not only capable of infecting plants but also fungi, however, unlike plants, fungi do not secrete the phenolic compounds necessary to trigger Agrobacterium so that they have to be added e.g. in the form of **acetosyringone**.
- Thanks to development of an expression system for small RNAs in fungi the introduction of a **CRISPR/CAS9-system** in fungal cells became possible. In 2016 the USDA declared that it will not regulate a white button mushroom strain edited with CRISPR/CAS9 to prevent fruit body browning causing a broad discussion about placing CRISPR/CAS9-edited crops on the market.
- Physical methods like electroporation, biolistics (“gene gun”), **sonoporation** that uses cavitation of gas bubbles produced by ultrasound to penetrate the cell membrane, etc. are also applicable to fungi.

Animals

Introduction of DNA into animal cells is usually called **transfection**, and is discussed in the corresponding article.

Practical aspects of transformation in molecular biology

The discovery of artificially induced competence in bacteria allow bacteria such as *Escherichia coli* to be used as a convenient host for the manipulation of DNA as well as expressing proteins. Typically plasmids are used for transformation in *E. coli*. In order to be stably maintained in the cell, a plasmid DNA molecule must contain an **origin of replication**, which allows it to be replicated in the cell independently of the replication of the cell's own chromosome.

The efficiency with which a competent culture can take up exogenous DNA and express its genes is known as **transformation efficiency** and is measured in colony forming unit (cfu) per µg DNA used. A transformation efficiency of 1×10^8 cfu/µg for a small plasmid like **pUC19** is roughly equivalent to 1 in 2000 molecules of the plasmid used being transformed.

In **calcium chloride transformation**, the cells are prepared by chilling cells in the presence of Ca^{2+} (in CaCl_2 solution), making the cell become permeable to **plasmid DNA**.

The cells are incubated on ice with the DNA, and then briefly heat-shocked (e.g., at 42°C for 30–120 seconds). This method works very well for circular plasmid DNA. Non-commercial preparations should normally give 10^6 to 10^7 transformants per microgram of plasmid; a poor preparation will be about $10^4/\mu\text{g}$ or less, but a good preparation of competent cells can give up to $\sim 10^8$ colonies per microgram of plasmid.

Protocols, however, exist for making supercompetent cells that may yield a transformation efficiency of over 10^9 . The chemical method, however, usually does not work well for linear DNA, such as fragments of chromosomal DNA, probably because the cell's native **exonuclease** enzymes rapidly degrade linear DNA. In contrast, cells that are naturally competent are usually transformed more efficiently with linear DNA than with plasmid DNA.

The transformation efficiency using the CaCl_2 method decreases with plasmid size, and electroporation therefore may be a more effective method for the uptake of large plasmid DNA. Cells used in electroporation should be prepared first by washing in cold double-distilled water to remove charged particles that may create sparks during the electroporation process.

Selection and screening in plasmid transformation

Because transformation usually produces a mixture of relatively few transformed cells and an abundance of non-transformed cells, a method is necessary to select for the cells that have acquired the plasmid.

The plasmid therefore requires a **selectable marker** such that those cells without the plasmid may be killed or have their growth arrested. **Antibiotic resistance** is the most commonly used marker for prokaryotes. The transforming plasmid contains a gene that confers resistance to an antibiotic that the bacteria are otherwise sensitive to.

The mixture of treated cells is cultured on media that contain the antibiotic so that only transformed cells are able to grow. Another method of selection is the use of certain **auxotrophic** markers that can compensate for an inability to metabolise certain amino acids, nucleotides, or sugars.

This method requires the use of suitably mutated strains that are deficient in the synthesis or utility of a particular biomolecule, and the transformed cells are cultured in a medium that allows only cells containing the plasmid to grow.

In a cloning experiment, a gene may be inserted into a plasmid used for transformation. However, in such experiment, not all the plasmids may contain a successfully inserted gene. Additional techniques may therefore be employed further to screen for transformed cells that contain plasmid with the insert.

Reporter genes can be used as **markers**, such as the *lacZ* gene which codes for β -galactosidase used in **blue-white screening**. This method of screening relies on the principle of **α -complementation**, where a fragment of the *lacZ* gene (*lacZ α*) in the plasmid can complement another mutant *lacZ* gene (*lacZ Δ M15*) in the cell.

Both genes by themselves produce non-functional peptides, however, when expressed together, as when a plasmid containing *lacZ-α* is transformed into a *lacZΔM15* cells, they form a functional β-galactosidase.

The presence of an active β-galactosidase may be detected when cells are grown in plates containing **X-gal**, forming characteristic blue colonies. However, the **multiple cloning site**, where a gene of interest may be **ligated** into the plasmid **vector**, is located within the *lacZα* gene. Successful ligation therefore disrupts the *lacZα* gene, and no functional β-galactosidase can form, resulting in white colonies. Cells containing successfully ligated insert can then be easily identified by its white coloration from the unsuccessful blue ones.

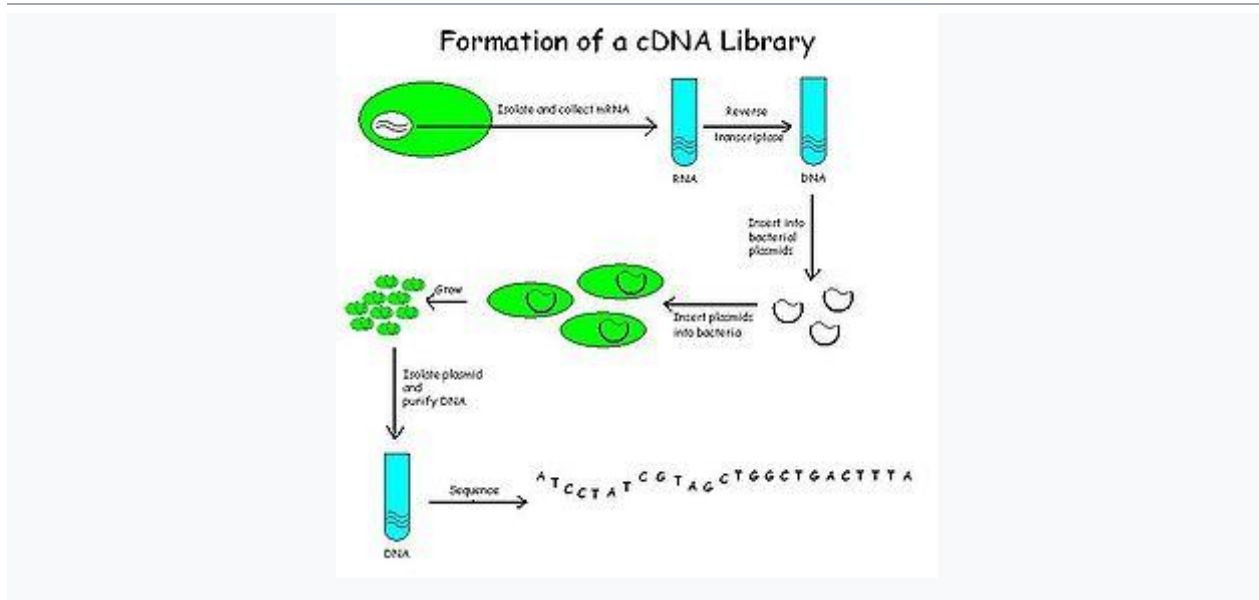
Other commonly used reporter genes are **green fluorescent protein (GFP)**, which produces cells that glow green under blue light, and the enzyme **luciferase**, which catalyzes a reaction with **luciferin** to emit light. The recombinant DNA may also be detected using other methods such as nucleic acid hybridization with radioactive RNA probe, while cells that expressed the desired protein from the plasmid may also be detected using immunological methods.

Construction of Cdna library

A **cDNA library** is a combination of cloned cDNA (**complementary DNA**) fragments inserted into a collection of host cells, which constitute some portion of the **transcriptome** of the organism and are stored as a "**library**". cDNA is produced from fully transcribed **mRNA** found in the **nucleus** and therefore contains only the expressed genes of an organism. Similarly, tissue-specific cDNA libraries can be produced. In **eukaryotic** cells the mature mRNA is already **spliced**, hence the cDNA produced lacks **introns** and can be readily expressed in a bacterial cell. While information in cDNA libraries is a powerful and useful tool since gene products are easily identified, the libraries lack information about **enhancers**, **introns**, and other regulatory elements found in a **genomic DNA library**.



cDNA Library Construction



Formation of a cDNA library.

cDNA is created from a mature mRNA from a eukaryotic cell with the use of reverse transcriptase. In eukaryotes, a poly-(A) tail (consisting of a long sequence of adenine nucleotides) distinguishes mRNA from tRNA and rRNA and can therefore be used as a primer site for reverse transcription. This has the problem that not all transcripts, such as those for the histone, encode a poly-A tail.

mRNA extraction

Firstly, the mRNA is obtained and purified from the rest of the RNAs. Several methods exist for purifying RNA such as trizol extraction and column purification. Column purification is done by using oligomeric dT nucleotide coated resins where only the mRNA having the poly-A tail will bind. The rest of the RNAs are eluted out. The mRNA is eluted by using eluting buffer and some heat to separate the mRNA strands from oligo-dT.

cDNA construction

Once mRNA is purified, *oligo-dT* (a short sequence of deoxy-thymidine nucleotides) is tagged as a complementary primer which binds to the poly-A tail providing a free 3'-OH end that can be extended by reverse transcriptase to create the complementary DNA strand. Now, the mRNA is removed by using a RNase enzyme leaving a single stranded cDNA (sscDNA).

This sscDNA is converted into a double stranded DNA with the help of DNA polymerase. However, for DNA polymerase to synthesize a complementary strand a free 3'-OH end is needed. This is provided by the sscDNA itself by generating a *hairpin loop* at the 3' end by coiling on itself. The polymerase extends the 3'-OH end and later the loop at 3' end is opened by the scissoring action of *S₁ nuclease*. Restriction endonucleases and DNA ligase are then used to clone the sequences into bacterial plasmids.

The cloned bacteria are then selected, commonly through the use of antibiotic selection. Once selected, stocks of the bacteria are created which can later be grown and sequenced to compile the cDNA library.

cDNA Library uses

cDNA libraries are commonly used when reproducing eukaryotic genomes, as the amount of information is reduced to remove the large numbers of non-coding regions from the library. cDNA libraries are used to express eukaryotic genes in prokaryotes.

Prokaryotes do not have introns in their DNA and therefore do not possess any enzymes that can cut it out during transcription process. cDNA does not have introns and therefore can be expressed in prokaryotic cells.

cDNA libraries are most useful in reverse genetics where the additional genomic information is of less use. Additionally, cDNA libraries are frequently used in functional cloning to identify genes based on the encoded protein's function. When studying eukaryotic DNA, expression libraries are constructed using complementary DNA (cDNA) to help ensure the insert is truly a gene.

cDNA Library vs. Genomic DNA Library

cDNA library lacks the non-coding and regulatory elements found in genomic DNA. **Genomic DNA libraries** provide more detailed information about the organism, but are more resource-intensive to generate and keep.

Cloning of cDNA

cDNA molecules can be cloned by using restriction site linkers. Linkers are short, double stranded pieces of DNA (**oligodeoxyribonucleotide**) about 8 to 12 nucleotide pairs long that include a **restriction endonuclease** cleavage site e.g. BamHI. Both the cDNA and the linker have blunt ends which can be ligated together using a high concentration of T4 DNA ligase.

Then sticky ends are produced in the cDNA molecule by cleaving the cDNA ends (which now have linkers with an incorporated site) with the appropriate endonuclease. A **cloning vector** (**plasmid**) is then also cleaved with the appropriate endonuclease. Following "**sticky end**" ligation of the insert into the vector the resulting recombinant DNA molecule is transferred into *E. coli* host cell for cloning.

Genomic library

A **genomic library** is a collection of the total genomic **DNA** from a single **organism**. The DNA is stored in a population of identical **vectors**, each containing a different **insert** of DNA. In order to construct a genomic library, the organism's DNA is **extracted** from **cells** and then digested with a **restriction enzyme** to cut the DNA into fragments of a specific size.

The fragments are then inserted into the vector using **DNA ligase**. Next, the vector DNA can be taken up by a host organism - commonly a population of **Escherichia coli** or **yeast** - with each cell containing only one vector molecule. Using a host cell to carry the vector allows for easy **amplification** and retrieval of specific **clones** from the **library** for analysis.

There are several kinds of vectors available with various insert capacities. Generally, libraries made from organisms with larger **genomes** require vectors featuring larger inserts, thereby fewer vector molecules are needed to make the library. Researchers can choose a vector also considering the ideal insert size to find the desired number of clones necessary for full genome coverage.

Genomic libraries are commonly used for **sequencing** applications. They have played an important role in the whole genome sequencing of several organisms, including the human genome and several **model organisms**.



History

The first DNA-based **genome** ever fully sequenced was achieved by two-time Nobel Prize winner, **Frederick Sanger**, in 1977. Sanger and his team of scientists created a library of the **bacteriophage, phi X 174**, for use in **DNA sequencing**.

The importance of this success contributed to the ever-increasing demand for sequencing genomes to research **gene therapy**. Teams are now able to catalog **polymorphisms** in genomes and investigate those candidate genes contributing to maladies such as **Parkinson's disease**, **Alzheimer's disease**, **multiple sclerosis**, **rheumatoid arthritis**, and **Type 1 diabetes**.

These are due to the advance of **genome-wide association studies** from the ability to create and sequence genomic libraries. Prior, linkage and candidate-gene studies were some of the only approaches.

Genomic library construction

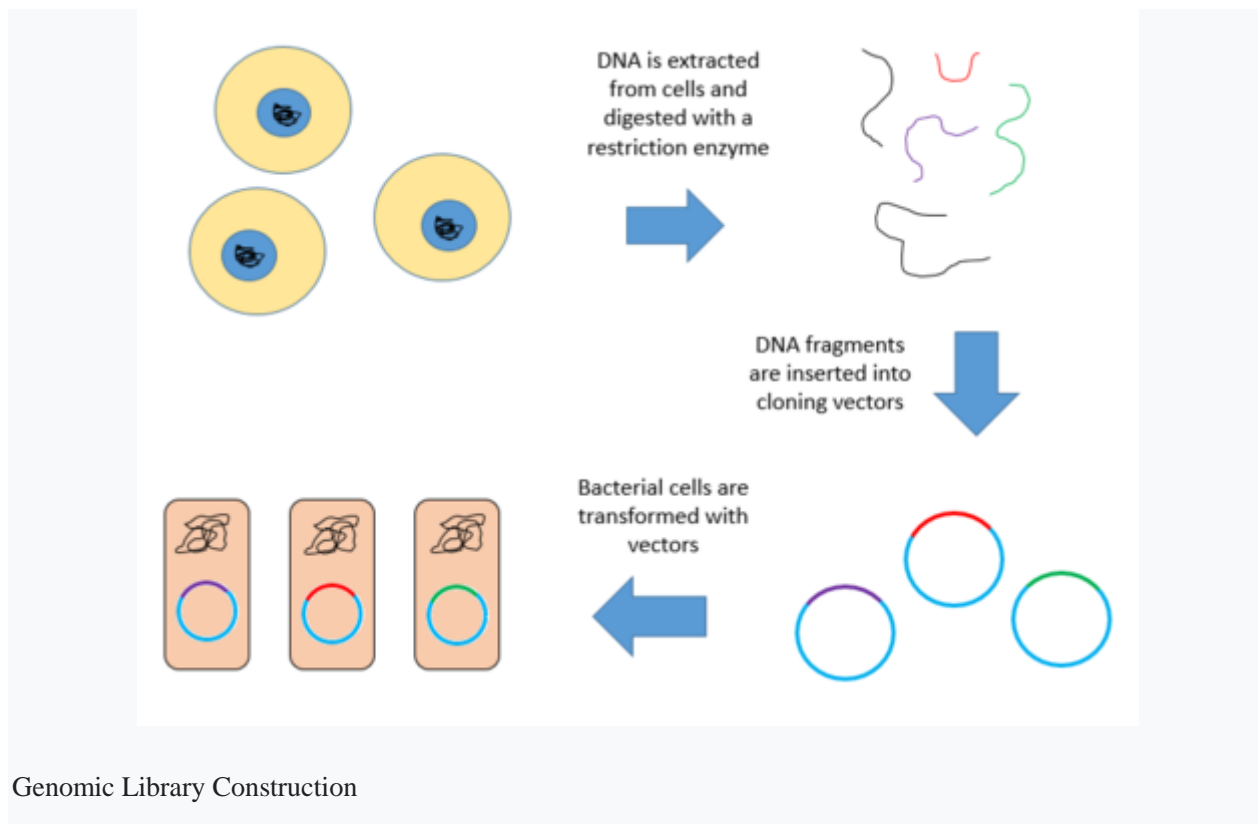
Construction of a genomic library involves creating many **recombinant DNA** molecules. An organism's genomic **DNA** is extracted and then digested with a **restriction enzyme**. For organisms with very small genomes (*~10 kb*), the digested fragments can be separated by **gel electrophoresis**. The separated fragments can then be excised and cloned into the vector separately.

However, when a large genome is digested with a restriction enzyme, there are far too many fragments to excise individually. The entire set of fragments must be cloned together with the vector, and separation of clones can occur after. In either case, the fragments are ligated into a vector that has been digested with the same restriction enzyme. The vector containing the inserted fragments of genomic DNA can then be introduced into a host organism.

Below are the steps for creating a genomic library from a large genome.

1. **Extract** and purify DNA.
2. Digest the DNA with a restriction enzyme. This creates fragments that are similar in size, each containing one or more genes.
3. Insert the fragments of DNA into vectors that were cut with the same restriction enzyme. Use the enzyme DNA ligase to seal the DNA fragments into the vector. This creates a large pool of recombinant molecules.
4. These recombinant molecules are taken up by a host bacterium by **transformation**, creating a DNA library.

Below is a diagram of the above outlined steps.



Determining titer of library

After a genomic library is constructed with a viral vector, such as [lambda phage](#), the **titer** of the library can be determined. Calculating the titer allows researchers to approximate how many infectious viral particles were successfully created in the library.

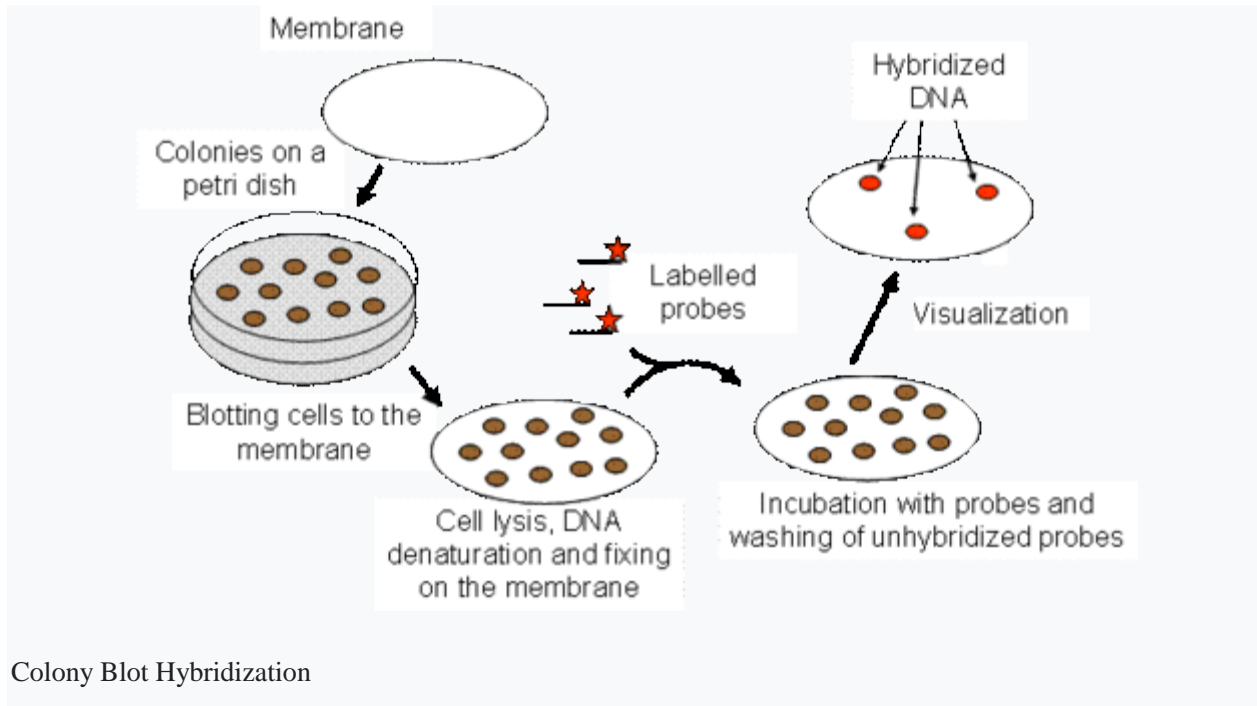
To do this, dilutions of the library are used to [transform cultures](#) of *E. coli* of known concentrations. The cultures are then plated on [agar plates](#) and incubated overnight. The number of [viral plaques](#) are counted and can be used to calculate the total number of infectious viral particles in the library.

Most viral vectors also carry a marker that allows clones containing an insert to be distinguished from those that do not have an insert. This allows researchers to also determine the percentage of infectious viral particles actually carrying a fragment of the library.

A similar method can be used to titer genomic libraries made with non-viral vectors, such as [plasmids](#) and [BACs](#). A test [ligation](#) of the library can be used to transform *E. coli*. The transformation is then spread on agar plates and incubated overnight.

The titer of the transformation is determined by counting the number of colonies present on the plates. These vectors generally have a [selectable marker](#) allowing the differentiation of clones containing an insert from those that do not. By doing this test, researchers can also determine the efficiency of the ligation and make adjustments as needed to ensure they get the desired number of clones for the library.

Screening library



In order to isolate clones that contain regions of interest from a library, the library must first be **screened**. One method of screening is **hybridization**. Each transformed host cell of a library will contain only one vector with one insert of DNA. The whole library can be plated onto a filter over **media**.

The filter and **colonies** are prepared for hybridization and then labeled with a **probe**. The target DNA- insert of interest- can be identified by detection such as **autoradiography** because of the **hybridization** with the probe as seen below.

Another method of screening is with **polymerase chain reaction** (PCR). Some libraries are stored as pools of clones and screening by PCR is an efficient way to identify pools containing specific clones.

Types of vectors

Genome size varies among different organisms and the **cloning vector** must be selected accordingly. For a large genome, a vector with a large capacity should be chosen so that a relatively small number of **clones** are sufficient for coverage of the entire genome. However, it is often more difficult to characterize an **insert** contained in a higher capacity vector.

Below is a table of several kinds of vectors commonly used for genomic libraries and the insert size that each generally holds.

Vector type	Insert size (thousands of bases)
Plasmids	up to 10

Phage lambda (λ)	up to 25
Cosmids	up to 45
Bacteriophage P1	70 to 100
P1 artificial chromosomes (PACs)	130 to 150
Bacterial artificial chromosomes (BACs)	120 to 300
Yeast artificial chromosomes (YACs)	250 to 2000

Plasmids

A **plasmid** is a double stranded circular **DNA** molecule commonly used for **molecular cloning**. Plasmids are generally 2 to 4 **kilobase-pairs** (kb) in length and are capable of carrying inserts up to 15kb.

Plasmids contain an **origin of replication** allowing them to replicate inside a bacterium independently of the host **chromosome**. Plasmids commonly carry a gene for **antibiotic resistance** that allows for the selection of bacterial cells containing the plasmid. Many plasmids also carry a **reporter gene** that allows researchers to distinguish clones containing an insert from those that do not.

Phage lambda (λ)

Phage λ is a **double-stranded DNA virus** that infects *E. coli*. The λ chromosome is 48.5kb long and can carry inserts up to 25kb. These inserts replace non-essential viral sequences in the λ chromosome, while the genes required for formation of **viral particles** and **infection** remain intact. The insert DNA is **replicated** with the viral DNA; thus, together they are packaged into viral particles. These particles are very efficient at infection and multiplication leading to a higher production of the recombinant λ chromosomes. However, due to the smaller insert size, libraries made with λ phage may require many clones for full genome coverage.

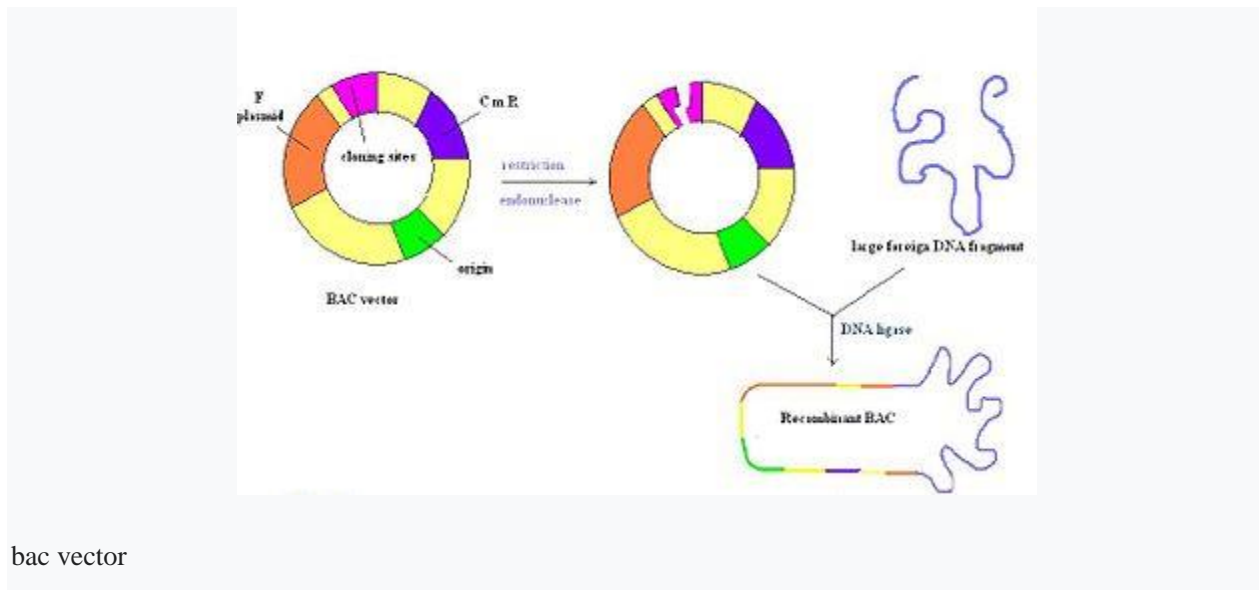
Cosmids]

Cosmid vectors are plasmids that contain a small region of bacteriophage λ DNA called the **cos** sequence. This sequence allows the cosmid to be packaged into bacteriophage λ particles. These particles- containing a linearized cosmid- are introduced into the host cell by **transduction**. Once inside the host, the cosmids circularize with the aid of the host's **DNA ligase** and then function as plasmids. Cosmids are capable of carrying inserts up to 40kb in size.

Bacteriophage P1 vectors

Bacteriophage P1 vectors can hold inserts 70 – 100kb in size. They begin as linear DNA molecules packaged into bacteriophage P1 particles. These particles are injected into an E. coli strain expressing **Cre recombinase**. The linear P1 vector becomes circularized by **recombination** between two loxP sites in the vector. P1 vectors generally contain a gene for antibiotic resistance and a positive selection marker to distinguish clones containing an insert from those that do not.

P1 vectors also contain a P1 plasmid **replicon**, which ensures only one copy of the vector is present in a cell. However, there is a second P1 replicon- called the P1 lytic replicon- that is controlled by an inducible **promoter**. This promoter allows the amplification of more than one copy of the vector per cell prior to **DNA extraction**.



bac vector

P1 artificial chromosomes

P1 artificial chromosomes (PACs) have features of both P1 vectors and Bacterial Artificial Chromosomes (BACs). Similar to P1 vectors, they contain a plasmid and a lytic replicon as described above. Unlike P1 vectors, they do not need to be packaged into bacteriophage particles for transduction.

Instead they are introduced into E. coli as circular DNA molecules through **electroporation** just as BACs are. Also similar to BACs, these are relatively harder to prepare due to a single origin of replication.

Bacterial artificial chromosomes

Bacterial artificial chromosomes (BACs) are circular DNA molecules, usually about 7kb in length, that are capable of holding inserts up to 300kb in size. BAC vectors contain a replicon derived from E. coli **F factor**, which ensures they are maintained at one copy per cell. Once an insert is ligated into a BAC, the BAC is introduced into **recombination** deficient strains of E. coli by electroporation.

Most BAC vectors contain a gene for antibiotic resistance and also a positive selection marker. The figure to the right depicts a BAC vector being cut with a restriction enzyme, followed by

the insertion of foreign DNA that is re-annealed by a ligase. Overall, this is a very stable vector, but they may be hard to prepare due to a single origin of replication just like PACs.

Yeast artificial chromosomes

Yeast artificial chromosomes (YACs) are linear DNA molecules containing the necessary features of an authentic **yeast** chromosome, including **telomeres**, a **centromere**, and an **origin of replication**. Large inserts of DNA can be ligated into the middle of the YAC so that there is an “arm” of the YAC on either side of the insert.

The recombinant YAC is introduced into yeast by transformation; **selectable markers** present in the YAC allow for the identification of successful transformants. YACs can hold inserts up to 2000kb, but most YAC libraries contain inserts 250-400kb in size.

Theoretically there is no upper limit on the size of insert a YAC can hold. It is the quality in the preparation of DNA used for inserts that determines the size limit. The most challenging aspect of using YAC is the fact they are prone to **rearrangement**.

How to select a vector

Vector selection requires one to ensure the library made is representative of the entire genome. Any insert of the genome derived from a restriction enzyme should have an equal chance of being in the library compared to any other insert.

Furthermore, recombinant molecules should contain large enough inserts ensuring the library size is able to be handled conveniently. This is particularly determined by the number of clones needed to have in a library.

The number of clones to get a sampling of all the genes is determined by the size of the organism's genome as well as the average insert size. This is represented by the formula (also known as the Carbon and Clarke formula):

where,

is the necessary number of recombinants

is the desired probability that any fragment in the genome will occur at least once in the library created

is the fractional proportion of the genome in a single recombinant

can be further shown to be:

where,

is the insert size

is the genome size

Thus, increasing the insert size (by choice of vector) would allow for fewer clones needed to represent a genome. The proportion of the insert size versus the genome size represents the proportion of the respective genome in a single clone. Here is the equation with all parts considered:

Vector selection example

The above formula can be used to determine the 99% confidence level that all sequences in a genome are represented by using a vector with an insert size of twenty thousand basepairs (such as the phage lambda vector). The genome size of the organism is three billion basepairs in this example.

clones

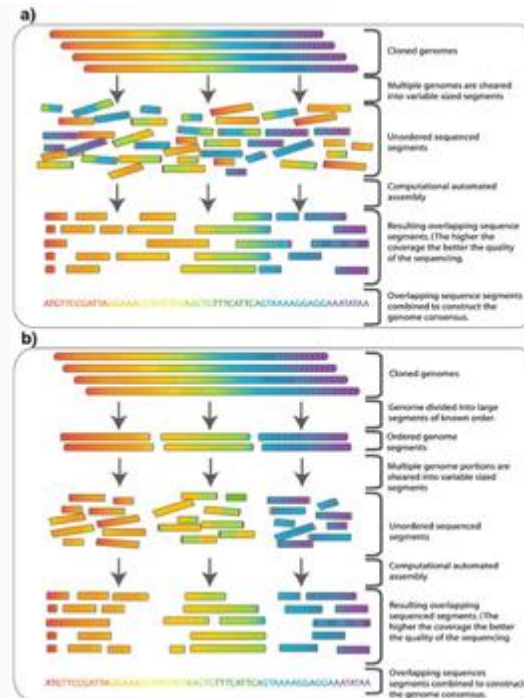
Thus, approximately 688,060 clones are required to ensure a 99% probability that a given DNA sequence from this three billion basepair genome will be present in a library using a vector with an insert size of twenty thousand basepairs.

Applications

After a library is created, the genome of an organism can be [sequenced](#) to elucidate how genes affect an organism or to compare similar organisms at the genome-level. The aforementioned [genome-wide association studies](#) can identify candidate genes stemming from many functional traits.

Genes can be isolated through genomic libraries and used on human cell lines or animal models to further research. Furthermore, creating high-fidelity clones with accurate genome representation- and no stability issues- would contribute well as intermediates for [shotgun sequencing](#) or the study of complete genes in functional analysis.

Hierarchical sequencing



Whole genome shotgun sequencing versus Hierarchical shotgun sequencing

One major use of genomic libraries is **hierarchical shotgun sequencing**, which is also called top-down, map-based or clone-by-clone sequencing. This strategy was developed in the 1980s for sequencing whole genomes before high throughput techniques for sequencing were available. Individual clones from genomic libraries can be sheared into smaller fragments, usually 500bp to 1000bp, which are more manageable for sequencing.

Once a clone from a genomic library is sequenced, the sequence can be used to screen the library for other clones containing inserts which overlap with the sequenced clone. Any new overlapping clones can then be sequenced forming a **contig**. This technique, called **chromosome walking**, can be exploited to sequence entire chromosomes.

Whole genome shotgun sequencing is another method of genome sequencing that does not require a library of high-capacity vectors. Rather, it uses computer algorithms to assemble short sequence reads to cover the entire genome.

Genomic libraries are often used in combination with whole genome shotgun sequencing for this reason. A high resolution map can be created by sequencing both ends of inserts from several clones in a genomic library.

This map provides sequences of known distances apart, which can be used to help with the assembly of sequence reads acquired through shotgun sequencing. The human genome sequence, which was declared complete in 2003, was assembled using both a BAC library and shotgun sequencing.

Genome-wide association studies

Genome-wide association studies are general applications to find specific gene targets and polymorphisms within the human race. In fact, the International HapMap project was created through a partnership of scientists and agencies from several countries to catalog and utilize this data.

The goal of this project is to compare genetic sequences of different individuals to elucidate similarities and differences within chromosomal regions.

Scientists from all of the participating nations are cataloging these attributes with data from populations of African, Asian, and European ancestry.

Such genome-wide assessments may lead to further diagnostic and drug therapies while also helping future teams focus on orchestrating therapeutics with genetic features in mind. These concepts are already being exploited in **genetic engineering**.

For example, a research team has actually constructed a PAC shuttle vector that creates a library representing two-fold coverage of the human genome.

This could serve as an incredible resource to identify genes, or sets of genes, causing disease. Moreover, these studies can serve as a powerful way to investigate transcriptional regulation as it has been seen in the study of baculoviruses.

Overall, advances in genome library construction and DNA sequencing has allowed for efficient discovery of different molecular targets. Assimilation of these features through such efficient methods can hasten the employment of novel drug candidates.

Molecular cloning

gene cloning

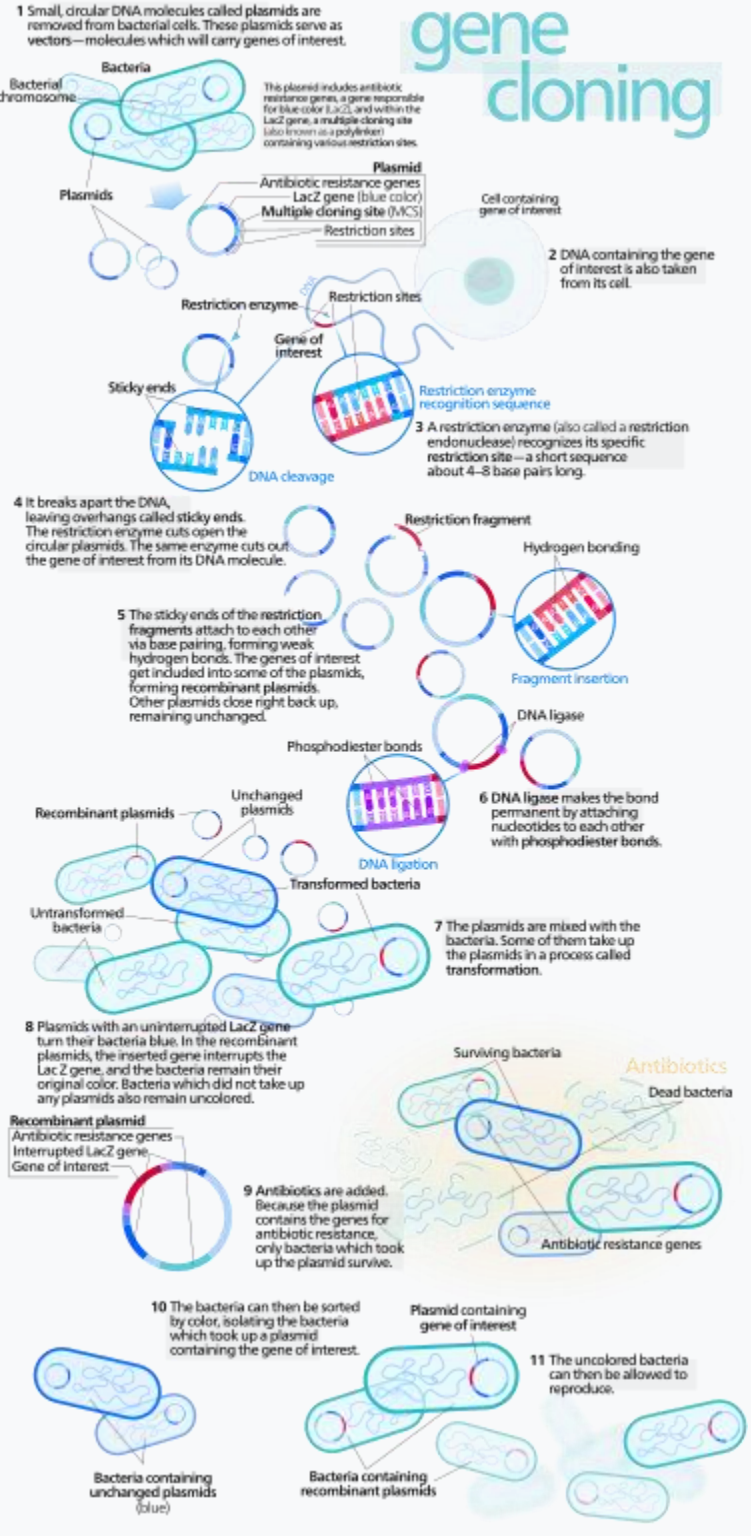


Diagram of molecular cloning using bacteria and plasmids.

Molecular cloning is a set of experimental methods in **molecular biology** that are used to assemble **recombinant DNA** molecules and to direct their **replication** within **host organisms**. The use of the word *cloning* refers to the fact that the method involves the replication of one molecule to produce a population of cells with identical DNA molecules.

Molecular cloning generally uses DNA sequences from two different organisms: the species that is the source of the DNA to be cloned, and the species that will serve as the living **host** for replication of the recombinant DNA. Molecular cloning methods are central to many contemporary areas of modern biology and medicine.

In a conventional molecular cloning experiment, the DNA to be cloned is obtained from an organism of interest, then treated with enzymes in the test tube to generate smaller DNA fragments. Subsequently, these fragments are then combined with **vector DNA** to generate recombinant DNA molecules.

The recombinant DNA is then introduced into a host organism (typically an easy-to-grow, benign, laboratory strain of *E. coli* bacteria). This will generate a population of organisms in which recombinant DNA molecules are replicated along with the host DNA. Because they contain foreign DNA fragments, these are **transgenic** or genetically modified microorganisms (**GMO**).

This process takes advantage of the fact that a single bacterial cell can be induced to take up and replicate a single recombinant DNA molecule. This single cell can then be expanded exponentially to generate a large amount of bacteria, each of which contain copies of the original recombinant molecule.

Thus, both the resulting bacterial population, and the recombinant DNA molecule, are commonly referred to as "clones". Strictly speaking, *recombinant DNA* refers to DNA molecules, while *molecular cloning* refers to the experimental methods used to assemble them. The idea arose that different DNA sequences could be inserted into a plasmid and that these foreign sequences would be carried into bacteria and digested as part of the plasmid. That is, these plasmids could serve as cloning vectors to carry genes.

Virtually any DNA sequence can be cloned and amplified, but there are some factors that might limit the success of the process. Examples of the DNA sequences that are difficult to clone are inverted repeats, origins of replication, centromeres and telomeres.

Another characteristic that limits chances of success is large size of DNA sequence. Inserts larger than 10kbp have very limited success, but bacteriophages such as bacteriophage λ can be modified to successfully insert a sequence up to 40 kbp.



History

Prior to the 1970s, the understanding of genetics and molecular biology was severely hampered by an inability to isolate and study individual genes from complex organisms. This changed dramatically with the advent of molecular cloning methods.

Microbiologists, seeking to understand the molecular mechanisms through which bacteria restricted the growth of bacteriophage, isolated **restriction endonucleases**, enzymes that could cleave DNA molecules only when specific DNA sequences were encountered.

They showed that restriction enzymes cleaved chromosome-length DNA molecules at specific locations, and that specific sections of the larger molecule could be purified by size fractionation.

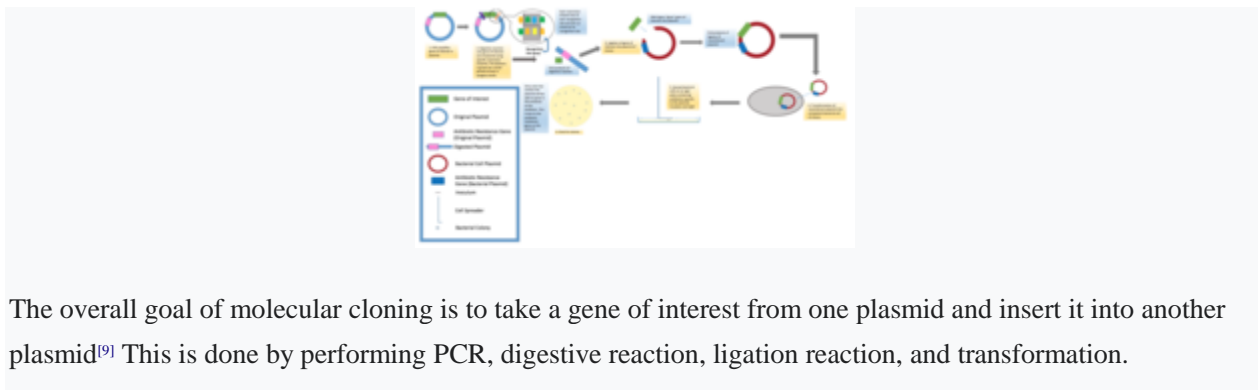
Using a second enzyme, DNA **ligase**, fragments generated by restriction enzymes could be joined in new combinations, termed **recombinant DNA**. By recombining DNA segments of interest with vector DNA, such as bacteriophage or plasmids, which naturally replicate inside bacteria, large quantities of purified recombinant DNA molecules could be produced in bacterial cultures. The first recombinant DNA molecules were generated and studied in 1972.

Overview

Molecular cloning takes advantage of the fact that the chemical structure of **DNA** is fundamentally the same in all living organisms. Therefore, if any segment of DNA from any organism is inserted into a DNA segment containing the molecular sequences required for **DNA replication**, and the resulting **recombinant DNA** is introduced into the organism from which the replication sequences were obtained, then the foreign DNA will be replicated along with the host cell's DNA in the **transgenic** organism.

Molecular cloning is similar to **polymerase chain reaction (PCR)** in that it permits the replication of DNA sequence. The fundamental difference between the two methods is that molecular cloning involves replication of the DNA in a living microorganism, while PCR replicates DNA in an *in vitro* solution, free of living cells.

Steps



In standard molecular cloning experiments, the cloning of any DNA fragment essentially involves seven steps: (1) Choice of host organism and cloning vector, (2) Preparation of vector DNA, (3) Preparation of DNA to be cloned, (4) Creation of recombinant DNA, (5) Introduction of recombinant DNA into host organism, (6) Selection of organisms containing recombinant DNA, (7) Screening for clones with desired DNA inserts and biological properties.

Although the detailed planning of the cloning can be done in any text editor, together with online utilities for e.g. PCR primer design, dedicated software exist for the purpose. Software for the purpose include for example ApE [1] (open source), DNASTrider [2] (open source), Serial Cloner [3] (gratis) and Collagene [4] (open source).

Notably, the growing capacity and fidelity of DNA synthesis platforms allows for increasingly intricate designs in molecular engineering. These projects may include very long strands

of novel DNA sequence and/or test entire libraries simultaneously, as opposed to of individual sequences.

These shifts introduce complexity that require design to move away from the flat nucleotide-based representation and towards a higher level of abstraction. Examples of such tools are [GenoCAD](#), [Teselagen](#) [5] (free for academia) or [GeneticConstructor](#) [6] (free for academics).

Choice of host organism and cloning vector

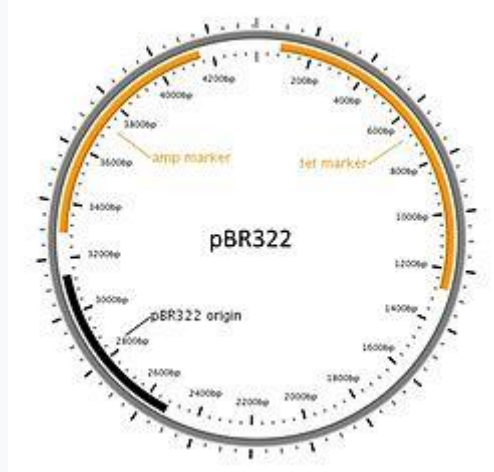


Diagram of a commonly used cloning plasmid; [pBR322](#). It's a circular piece of DNA 4361 bases long. Two [antibiotic resistance genes](#) are present, conferring resistance to [ampicillin](#) and [tetracycline](#), and an [origin of replication](#) that the host uses to [replicate](#) the DNA.

Although a very large number of host organisms and molecular cloning vectors are in use, the great majority of molecular cloning experiments begin with a laboratory strain of the bacterium *E. coli* (*Escherichia coli*) and a [plasmid cloning vector](#). *E. coli* and plasmid vectors are in common use because they are technically sophisticated, versatile, widely available, and offer rapid growth of recombinant organisms with minimal equipment.

If the DNA to be cloned is exceptionally large (hundreds of thousands to millions of base pairs), then a [bacterial artificial chromosome](#) or [yeast artificial chromosome](#) vector is often chosen.

Specialized applications may call for specialized host-vector systems. For example, if the experimentalists wish to harvest a particular protein from the recombinant organism, then an [expression vector](#) is chosen that contains appropriate signals for transcription and translation in the desired host organism.

Alternatively, if replication of the DNA in different species is desired (for example, transfer of DNA from bacteria to plants), then a multiple host range vector (also termed [shuttle vector](#)) may be selected.

In practice, however, specialized molecular cloning experiments usually begin with cloning into a bacterial plasmid, followed by [subcloning](#) into a specialized vector.

Whatever combination of host and vector are used, the vector almost always contains four DNA segments that are critically important to its function and experimental utility:

- DNA *replication origin* is necessary for the vector (and its linked recombinant sequences) to replicate inside the host organism
- one or more unique *restriction endonuclease recognition sites* to serve as sites where foreign DNA may be introduced
- a *selectable genetic marker* gene that can be used to enable the survival of cells that have taken up vector sequences
- a *tag* gene that can be used to screen for cells containing the foreign DNA



Cleavage of a DNA sequence containing the [BamHI restriction site](#). The DNA is cleaved at the palindromic sequence to produce 'sticky ends'.

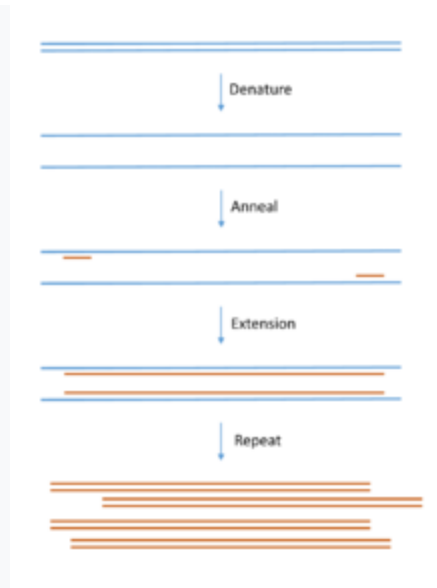
Preparation of vector DNA

The cloning vector is treated with a restriction endonuclease to cleave the DNA at the site where foreign DNA will be inserted. The restriction enzyme is chosen to generate a configuration at the cleavage site that is compatible with the ends of the foreign DNA (see [DNA end](#)). Typically, this is done by cleaving the vector DNA and foreign DNA with the same restriction enzyme, for example [EcoRI](#).

Most modern vectors contain a variety of convenient cleavage sites that are unique within the vector molecule (so that the vector can only be cleaved at a single site) and are located within a gene (frequently [beta-galactosidase](#)) whose inactivation can be used to distinguish recombinant from non-recombinant organisms at a later step in the process.

To improve the ratio of recombinant to non-recombinant organisms, the cleaved vector may be treated with an enzyme ([alkaline phosphatase](#)) that dephosphorylates the vector ends. Vector molecules with dephosphorylated ends are unable to replicate, and replication can only be restored if foreign DNA is integrated into the cleavage site.

Preparation of DNA to be cloned



DNA for cloning is most commonly produced using **PCR**. Template DNA is mixed with **bases** (the building blocks of DNA), primers (short pieces of complementary single stranded DNA) and a **DNA polymerase** enzyme that builds the DNA chain. The mix goes through cycles of heating and cooling to produce large quantities of copied DNA.

For cloning of genomic DNA, the DNA to be cloned is extracted from the organism of interest. Virtually any tissue source can be used (even tissues from **extinct animals**) as long as the DNA is not extensively degraded. The DNA is then purified using simple methods to remove contaminating proteins (extraction with phenol), RNA (ribonuclease) and smaller molecules (precipitation and/or chromatography). **Polymerase chain reaction (PCR)** methods are often used for amplification of specific DNA or RNA (**RT-PCR**) sequences prior to molecular cloning.

DNA for cloning experiments may also be obtained from RNA using reverse transcriptase (**complementary DNA** or **cDNA cloning**), or in the form of synthetic DNA (**artificial gene synthesis**). cDNA cloning is usually used to obtain clones representative of the mRNA population of the cells of interest, while synthetic DNA is used to obtain any precise sequence defined by the designer.

Such a designed sequence may be required when moving genes across **genetic codes** (for example, from the mitochondria to the nucleus) or simply for increasing expression via **codon optimization**.¹

The purified DNA is then treated with a restriction enzyme to generate fragments with ends capable of being linked to those of the vector. If necessary, short double-stranded segments of DNA (**linkers**) containing desired restriction sites may be added to create end structures that are compatible with the vector.

Creation of recombinant DNA with DNA ligase

The creation of recombinant DNA is in many ways the simplest step of the molecular cloning process. DNA prepared from the vector and foreign source are simply mixed together at appropriate concentrations and exposed to an enzyme (**DNA ligase**) that covalently links the ends together.

This joining reaction is often termed **ligation**. The resulting DNA mixture containing randomly joined ends is then ready for introduction into the host organism.

DNA ligase only recognizes and acts on the ends of linear DNA molecules, usually resulting in a complex mixture of DNA molecules with randomly joined ends. The desired products (vector DNA covalently linked to foreign DNA) will be present, but other sequences (e.g. foreign DNA linked to itself, vector DNA linked to itself and higher-order combinations of vector and foreign DNA) are also usually present. This complex mixture is sorted out in subsequent steps of the cloning process, after the DNA mixture is introduced into cells.

Introduction of recombinant DNA into host organism

The DNA mixture, previously manipulated *in vitro*, is moved back into a living cell, referred to as the host organism. The methods used to get DNA into cells are varied, and the name applied to this step in the molecular cloning process will often depend upon the experimental method that is chosen (e.g. **transformation**, **transduction**, **transfection**, **electroporation**).

When microorganisms are able to take up and replicate DNA from their local environment, the process is termed **transformation**, and cells that are in a physiological state such that they can take up DNA are said to be **competent**.

In mammalian cell culture, the analogous process of introducing DNA into cells is commonly termed **transfection**. Both transformation and transfection usually require preparation of the cells through a special growth regime and chemical treatment process that will vary with the specific species and cell types that are used.

Electroporation uses high voltage electrical pulses to translocate DNA across the cell membrane (and cell wall, if present). In contrast, **transduction** involves the packaging of DNA into virus-derived particles, and using these virus-like particles to introduce the encapsulated DNA into the cell through a process resembling viral infection. Although electroporation and transduction are highly specialized methods, they may be the most efficient methods to move DNA into cells.

Selection of organisms containing vector sequences

Whichever method is used, the introduction of recombinant DNA into the chosen host organism is usually a low efficiency process; that is, only a small fraction of the cells will actually take up DNA. Experimental scientists deal with this issue through a step of artificial genetic selection, in which cells that have not taken up DNA are selectively killed, and only those cells that can actively replicate DNA containing the selectable marker gene encoded by the vector are able to survive.

When bacterial cells are used as host organisms, the **selectable marker** is usually a gene that confers resistance to an **antibiotic** that would otherwise kill the cells, typically **ampicillin**. Cells harboring the plasmid will survive when exposed to the antibiotic, while those that have failed to take up plasmid sequences will die. When mammalian cells (e.g. human or mouse cells) are used, a similar strategy is used, except that the marker gene (in this case typically encoded as part of the **kanMX** cassette) confers resistance to the antibiotic **Geneticin**.

Screening for clones with desired DNA inserts and biological properties

Modern bacterial cloning vectors (e.g. **pUC19** and later derivatives including the **pGEM** vectors) use the **blue-white screening system** to distinguish colonies (clones) of transgenic cells from those that contain the parental vector (i.e. vector DNA with no recombinant sequence inserted).

In these vectors, foreign DNA is inserted into a sequence that encodes an essential part of **beta-galactosidase**, an enzyme whose activity results in formation of a blue-colored colony on the culture medium that is used for this work. Insertion of the foreign DNA into the beta-galactosidase coding sequence disables the function of the enzyme so that colonies containing transformed DNA remain colorless (white).

Therefore, experimentalists are easily able to identify and conduct further studies on transgenic bacterial clones, while ignoring those that do not contain recombinant DNA.

The total population of individual clones obtained in a molecular cloning experiment is often termed a **DNA library**. Libraries may be highly complex (as when cloning complete genomic DNA from an organism) or relatively simple (as when moving a previously cloned DNA fragment into a different plasmid), but it is almost always necessary to examine a number of different clones to be sure that the desired DNA construct is obtained.

This may be accomplished through a very wide range of experimental methods, including the use of **nucleic acid hybridizations**, **antibody probes**, **polymerase chain reaction**, **restriction fragment analysis** and/or **DNA sequencing**.

Applications

Molecular cloning provides scientists with an essentially unlimited quantity of any individual DNA segments derived from any genome. This material can be used for a wide range of purposes, including those in both basic and applied biological science. A few of the more important applications are summarized here.

Genome organization and gene expression

Molecular cloning has led directly to the elucidation of the complete DNA sequence of the genomes of a very large number of species and to an exploration of genetic diversity within individual species, work that has been done mostly by determining the DNA sequence of large numbers of randomly cloned fragments of the genome, and assembling the overlapping sequences.

At the level of individual genes, molecular clones are used to generate **probes** that are used for examining how genes are **expressed**, and how that expression is related to other processes in biology, including the metabolic environment, extracellular signals, development, learning, senescence and cell death.

Cloned genes can also provide tools to examine the biological function and importance of individual genes, by allowing investigators to **inactivate** the genes, or make more subtle mutations using regional mutagenesis or **site-directed mutagenesis**. Genes cloned into expression vectors for **functional cloning** provide a means to screen for genes on the basis of the expressed protein's function.

Production of recombinant proteins

Obtaining the molecular clone of a gene can lead to the development of organisms that produce the protein product of the cloned genes, termed a recombinant protein. In practice, it is frequently more difficult to develop an organism that produces an active form of the recombinant protein in desirable quantities than it is to clone the gene.

This is because the molecular signals for gene expression are complex and variable, and because protein folding, stability and transport can be very challenging.

Many useful proteins are currently available as **recombinant products**. These include--(1) medically useful proteins whose administration can correct a defective or poorly expressed gene (e.g. recombinant **factor VIII**, a blood-clotting factor deficient in some forms of **hemophilia**, and recombinant **insulin**, used to treat some forms of **diabetes**), (2) proteins that can be administered to assist in a life-threatening emergency (e.g. **tissue plasminogen activator**, used to treat strokes), (3) recombinant subunit vaccines, in which a purified protein can be used to immunize patients against infectious diseases, without exposing them to the infectious agent itself (e.g. **hepatitis B vaccine**), and (4) recombinant proteins as standard material for diagnostic laboratory tests.

Expression cloning is a technique in **DNA cloning** that uses **expression vectors** to generate a library of clones, with each clone expressing one **protein**. This *expression library* is then screened for the property of interest and clones of interest are recovered for further analysis. An example would be using an expression library to isolate genes that could confer **antibiotic resistance**.



Expression vectors

Expression vectors are a specialized type of **cloning vector** in which the transcriptional and translational signals needed for the regulation of the gene of interest are included in the cloning vector. The transcriptional and translational signals may be synthetically created to make the expression of the gene of interest easier to regulate.

Purpose

Usually the ultimate aim of expression cloning is to produce large quantities of specific **proteins**. To this end, a **bacterial expression clone** may include a **ribosome binding site** (**Shine-Dalgarno sequence**) to enhance translation of the gene of interest's mRNA, a **transcription termination sequence**, or, in **eukaryotes**, specific sequences to promote the **post-translational modification** of the protein product.

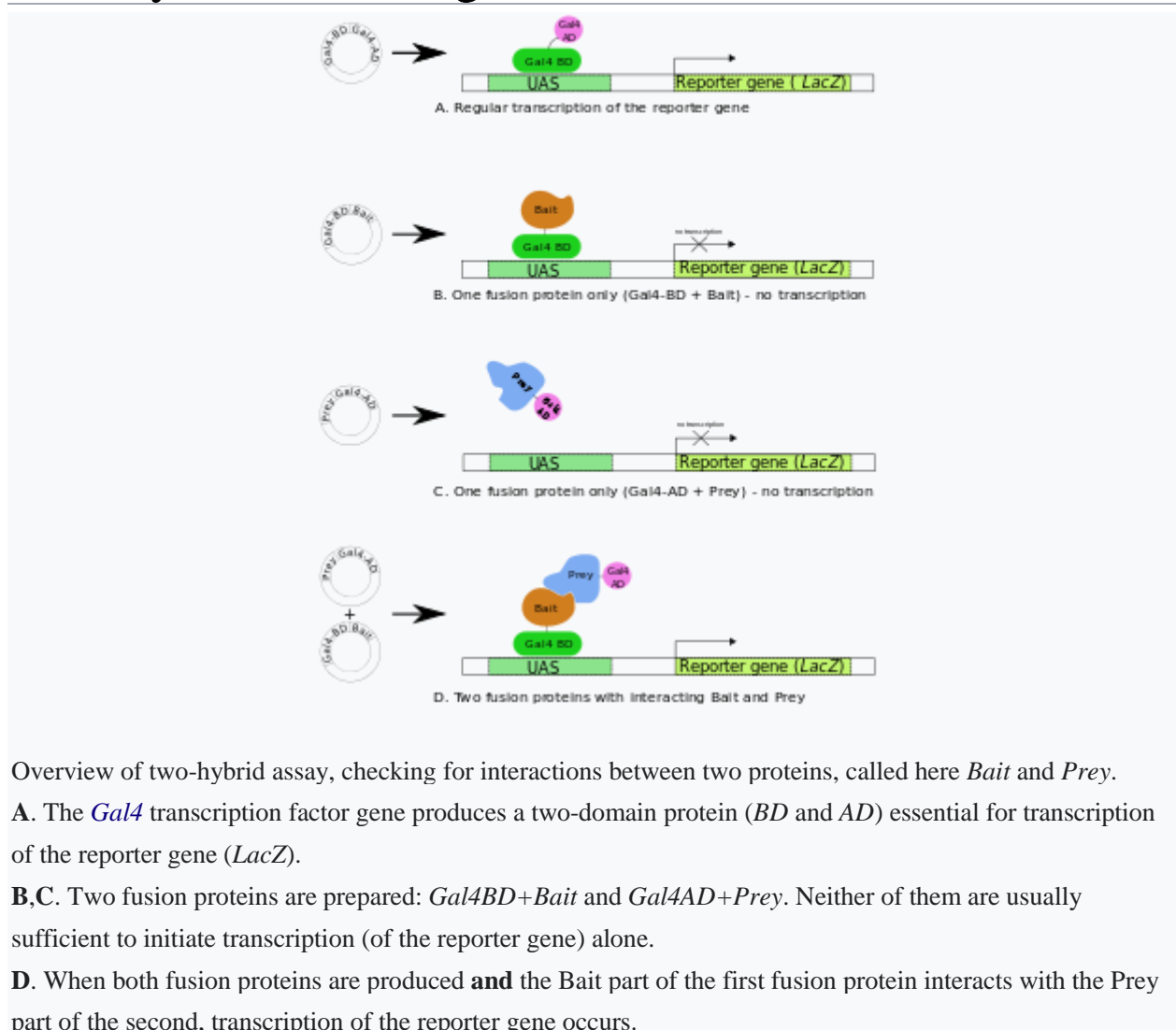
Protein interaction cloning in yeast: identification of mammalian proteins that react with the leucine zipper of Jun.

To identify proteins that interact with Jun or Fos we have used the protein interaction cloning system developed by S. Fields and O.-K. Song [(1989) *Nature* (London) 340, 245-246] to clone mammalian cDNAs encoding polypeptides that interact with the dimerization and DNA-binding motif (bZIP; basic domain leucine zipper motif) of Jun. For this purpose, yeast cells lacking GAL4 activity but expressing a GAL4 DNA-binding domain-Jun bZIP fusion protein were transformed with a mouse embryo cDNA plasmid library in which the cDNA was joined to a gene segment encoding the GAL4 transcriptional activation domain.

Several transformants exhibiting GAL4 activity were identified and shown to harbor plasmids encoding polypeptides predicted to form coiled-coil structures with Jun and/or Fos. One of these is a bZIP protein of the ATF/CREB protein family--

probably the murine homolog of TAXREB67. Two others encode polypeptides with predicted potential to form coiled-coil structures, and seven other isolates encode segments of alpha- or beta-tropomyosin, classical coiled-coil proteins. The tropomyosin polypeptides were found to interact in the yeast assay system with the bZIP region of Jun but not with the bZIP region of Fos. results illustrate the range of protein interaction cloning for discovering proteins that bind to a given target polypeptide.

Two-hybrid screening



Two-hybrid screening (originally known as **yeast two-hybrid system** or **Y2H**) is a **molecular biology** technique used to discover **protein–protein interactions** (PPIs) and **protein–DNA interactions** by testing for physical interactions (such as binding) between two **proteins** or a single protein and a **DNA** molecule, respectively.

The premise behind the test is the activation of **downstream reporter gene(s)** by the binding of a **transcription factor** onto an **upstream activating sequence (UAS)**. For two-hybrid screening, the transcription factor is split into two separate fragments, called the DNA-binding domain (DBD or often also abbreviated as BD) and activating domain (AD). The BD is the **domain** responsible for **binding** to the UAS and the AD is the domain responsible for the activation of **transcription**. The Y2H is thus a **protein-fragment complementation assay**.



History

Pioneered by **Stanley Fields** and Ok-Kyu Song in 1989, the technique was originally designed to detect protein–protein interactions using the **Gal4** transcriptional activator of the yeast *Saccharomyces cerevisiae*.

The **Gal4** protein activated transcription of a gene involved in galactose utilization, which formed the basis of selection. Since then, the same principle has been adapted to describe many alternative methods, including some that detect protein–DNA interactions or DNA–DNA interactions, as well as methods that use different **host organisms** such as *Escherichia coli* or mammalian cells instead of yeast.

Basic premise

The key to the two-hybrid screen is that in most **eukaryotic** transcription factors, the activating and binding domains are modular and can function in proximity to each other without direct binding. This means that even though the transcription factor is split into two fragments, it can still activate transcription when the two fragments are indirectly connected.

The most common screening approach is the yeast two-hybrid assay. In this approach the researcher knows where each prey is located on the used medium (agar plates). Millions of potential interactions in several organisms have been screened in the latest decade using **high-throughput screening** systems (often using robots) and over thousands of interactions have been detected and categorized in databases as **BioGRID**.

This system often utilizes a **genetically engineered** strain of yeast in which the **biosynthesis** of certain nutrients (usually **amino acids** or **nucleic acids**) is lacking. When grown on media that lacks these nutrients, the yeast fail to survive. This mutant yeast strain can be made to incorporate foreign DNA in the form of **plasmids**.

In yeast two-hybrid screening, separate bait and prey plasmids are simultaneously introduced into the mutant yeast strain or a mating strategy is used to get both plasmids in one host cell.

The second high-throughput approach is the library screening approach. In this set up the bait and prey harboring cells are mated in a random order. After mating and selecting surviving

cells on selective medium the scientist will sequence the isolated plasmids to see which prey (DNA sequence) is interacting with the used bait.

This approach has a lower rate of reproducibility and tends to yield higher amounts of false positives compared to the matrix approach.

Plasmids are engineered to produce a protein product in which the DNA-binding domain (BD) fragment is fused onto a protein while another plasmid is engineered to produce a protein product in which the activation domain (AD) fragment is fused onto another protein. The protein fused to the BD may be referred to as the bait protein, and is typically a known protein the investigator is using to identify new binding partners.

The protein fused to the AD may be referred to as the prey protein and can be either a single known protein or a **library** of known or unknown proteins. In this context, a library may consist of a collection of protein-encoding sequences that represent all the proteins expressed in a particular organism or tissue, or may be generated by synthesising random DNA sequences.

Regardless of the source, they are subsequently incorporated into the protein-encoding sequence of a plasmid, which is then transfected into the cells chosen for the screening method. This technique, when using a library, assumes that each cell is transfected with no more than a single plasmid and that, therefore, each cell ultimately expresses no more than a single member from the protein library.

If the bait and prey proteins interact (i.e., bind), then the AD and BD of the transcription factor are indirectly connected, bringing the AD in proximity to the transcription start site and transcription of reporter gene(s) can occur. If the two proteins do not interact, there is no transcription of the reporter gene. In this way, a successful interaction between the fused protein is linked to a change in the cell phenotype.

The challenge of separating cells that express proteins that happen to interact with their counterpart fusion proteins from those that do not, is addressed in the following section.

Fixed domains

In any study, some of the protein domains, those under investigation, will be varied according to the goals of the study whereas other domains, those that are not themselves being investigated, will be kept constant.

For example, in a two-hybrid study to select DNA-binding domains, the DNA-binding domain, BD, will be varied while the two interacting proteins, the bait and prey, must be kept constant to maintain a strong binding between the BD and AD. There are a number of domains from which to choose the BD, bait and prey and AD, if these are to remain constant. In protein–protein interaction investigations, the BD may be chosen from any of many strong DNA-binding domains such as [Zif268](#).

A frequent choice of bait and prey domains are residues 263–352 of yeast Gal11P with a N342V mutation and residues 58–97 of yeast Gal4,^[2] respectively. These domains can be used in both yeast- and bacterial-based selection techniques and are known to bind together strongly

The AD chosen must be able to activate transcription of the reporter gene, using the cell's own transcription machinery. Thus, the variety of ADs available for use in yeast-based techniques may not be suited to use in their bacterial-based analogues. The herpes simplex virus-

derived AD, VP16 and yeast Gal4 AD have been used with success in yeast whilst a portion of the α -subunit of *E. coli* RNA polymerase has been utilised in *E. coli*-based methods.

Whilst powerfully activating domains may allow greater sensitivity towards weaker interactions, conversely, a weaker AD may provide greater stringency.

Construction of expression plasmids

A number of engineered genetic sequences must be incorporated into the host cell to perform two-hybrid analysis or one of its derivative techniques. The considerations and methods used in the construction and delivery of these sequences differ according to the needs of the assay and the organism chosen as the experimental background.

There are two broad categories of hybrid library: random libraries and cDNA-based libraries. A **cDNA library** is constituted by the cDNA produced through **reverse transcription** of mRNA collected from specific cells of types of cell. This library can be ligated into a construct so that it is attached to the BD or AD being used in the assay.

A random library uses lengths of DNA of random sequence in place of these cDNA sections. A number of methods exist for the production of these random sequences, including **cassette mutagenesis**.

Regardless of the source of the DNA library, it is **ligated** into the appropriate place in the relevant plasmid/phagemid using the appropriate **restriction endonucleases**.

E. coli-specific considerations

By placing the hybrid proteins under the control of **IPTG**-inducible **lac promoters**, they are expressed only on media supplemented with IPTG. Further, by including different antibiotic resistance genes in each genetic construct, the growth of non-transformed cells is easily prevented through culture on media containing the corresponding antibiotics. This is particularly important for counter selection methods in which a *lack* of interaction is needed for cell survival.

The reporter gene may be inserted into the *E. coli* genome by first inserting it into an **episome**, a type of plasmid with the ability to incorporate itself into the bacterial cell genome with a copy number of approximately one per cell.

The hybrid expression phagemids can be electroporated into *E. coli* XL-1 Blue cells which after amplification and infection with VCS-M13 **helper phage**, will yield a stock of library phage. These phage will each contain one single-stranded member of the phagemid library.

Recovery of protein information

Once the selection has been performed, the **primary structure** of the proteins which display the appropriate characteristics must be determined. This is achieved by retrieval of the protein-encoding sequences (as originally inserted) from the cells showing the appropriate phenotype.

E. coli

The phagemid used to transform *E. coli* cells may be "rescued" from the selected cells by infecting them with VCS-M13 helper phage. The resulting phage particles that are produced

contain the single-stranded phagemids and are used to infect XL-1 Blue cells. The double-stranded phagemids are subsequently collected from these XL-1 Blue cells, essentially reversing the process used to produce the original library phage. Finally, the DNA sequences are determined through [dideoxy sequencing](#).

Controlling sensitivity

The *Escherichia coli*-derived **Tet-R** repressor can be used in line with a conventional reporter gene and can be controlled by tetracycline or doxycycline (Tet-R inhibitors). Thus the expression of Tet-R is controlled by the standard two-hybrid system but the Tet-R in turn controls (represses) the expression of a previously mentioned reporter such as *HIS3*, through its Tet-R promoter. Tetracycline or its derivatives can then be used to regulate the sensitivity of a system utilising Tet-R.

Sensitivity may also be controlled by varying the dependency of the cells on their reporter genes. For example, this may be affected by altering the concentration of histidine in the growth medium for *his3*-dependent cells and altering the concentration of streptomycin for *aadA* dependent cells.

Selection-gene-dependency may also be controlled by applying an inhibitor of the selection gene at a suitable concentration. **3-Amino-1,2,4-triazole** (3-AT) for example, is a competitive inhibitor of the *HIS3*-gene product and may be used to titrate the minimum level of *HIS3* expression required for growth on histidine-deficient media.

Sensitivity may also be modulated by varying the number of operator sequences in the reporter DNA.

Non-fusion proteins

A third, non-fusion protein may be co-expressed with two fusion proteins. Depending on the investigation, the third protein may modify one of the fusion proteins or mediate or interfere with their interaction.

Co-expression of the third protein may be necessary for modification or activation of one or both of the fusion proteins. For example, *S. cerevisiae* possesses no endogenous tyrosine kinase. If an investigation involves a protein that requires tyrosine phosphorylation, the kinase must be supplied in the form of a tyrosine kinase gene.

The non-fusion protein may mediate the interaction by binding both fusion proteins simultaneously, as in the case of ligand-dependent receptor dimerization.

For a protein with an interacting partner, its functional homology to other proteins may be assessed by supplying the third protein in non-fusion form, which then may or may not compete with the fusion-protein for its binding partner.

Binding between the third protein and the other fusion protein will interrupt the formation of the reporter expression activation complex and thus reduce reporter expression, leading to the distinguishing change in phenotype.

Split-ubiquitin yeast two-hybrid

One limitation of classic yeast two-hybrid screens is that they are limited to soluble proteins. It is therefore impossible to use them to study the protein–protein interactions between insoluble **integral membrane proteins**. The split-ubiquitin system provides a method for overcoming this limitation. In the split-ubiquitin system, two integral membrane proteins to be studied are fused to two different **ubiquitin** moieties: a C-terminal ubiquitin moiety ("Cub", residues 35–76) and an N-terminal ubiquitin moiety ("Nub", residues 1–34)

. These fused proteins are called the bait and prey, respectively. In addition to being fused to an integral membrane protein, the Cub moiety is also fused to a **transcription factor** (TF) that can be cleaved off by ubiquitin specific **proteases**.

Upon bait–prey interaction, Nub and Cub-moieties assemble, reconstituting the split-ubiquitin. The reconstituted split-ubiquitin molecule is recognized by ubiquitin specific proteases, which cleave off the transcription factor, allowing it to induce the transcription of **reporter genes**.

Fluorescent two-hybrid assay

Zolghadr and co-workers presented a fluorescent two-hybrid system that uses two hybrid proteins that are fused to different fluorescent proteins as well as LacI, the **lac repressor**. The structure of the fusion proteins looks like this: FP2-LacI-bait and FP1-prey where the bait and prey proteins interact and bring the fluorescent proteins (FP1 = **GFP**, FP2=**mCherry**) in close proximity at the binding site of the LacI protein in the host cell genome. The system can also be used to screen for inhibitors of protein–protein interactions.

Enzymatic two-hybrid systems: KISS

While the original Y2H system used a reconstituted transcription factor, other systems create enzymatic activities to detect PPIs. For instance, the Kinase Substrate Sensor ("KISS"), is a mammalian two-hybrid approach has been designed to map intracellular PPIs.

Here, a bait protein is fused to a **kinase**-containing portion of **TYK2** and a prey is coupled to a **gp130 cytokine receptor** fragment. When bait and prey interact, TYK2 phosphorylates **STAT3** docking sites on the prey chimera, which ultimately leads to activation of a **reporter gene**.

One-, three- and one-two-hybrid variants

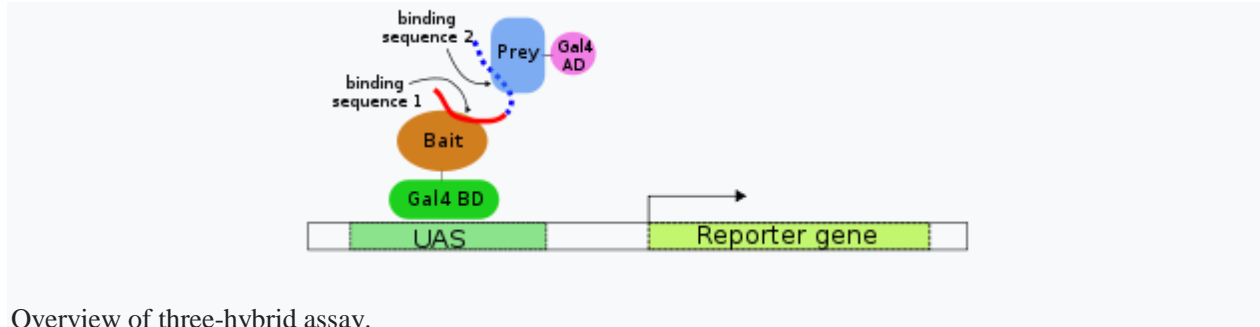
One-hybrid

The one-hybrid variation of this technique is designed to investigate **protein–DNA interactions** and uses a single fusion protein in which the AD is linked directly to the binding domain. The binding domain in this case however is not necessarily of fixed sequence as in two-hybrid protein–protein analysis but may be constituted by a library.

This library can be selected against the desired target sequence, which is inserted in the promoter region of the reporter gene construct. In a positive-selection system, a binding domain that successfully binds the UAS and allows transcription is thus selected.

Note that selection of DNA-binding domains is not necessarily performed using a one-hybrid system, but may also be performed using a two-hybrid system in which the binding domain is varied and the bait and prey proteins are kept constant

Three-hybrid



Overview of three-hybrid assay.

RNA-protein interactions have been investigated through a three-hybrid variation of the two-hybrid technique. In this case, a hybrid RNA molecule serves to adjoin together the two protein fusion domains—which are not intended to interact with each other but rather the intermediary RNA molecule (through their RNA-binding domains). Techniques involving non-fusion proteins that perform a similar function, as described in the 'non-fusion proteins' section above, may also be referred to as three-hybrid methods.

One-two-hybrid

Simultaneous use of the one- and two-hybrid methods (that is, simultaneous protein–protein and protein–DNA interaction) is known as a one-two-hybrid approach and expected to increase the stringency of the screen.

Host organism

Although theoretically, any living cell might be used as the background to a two-hybrid analysis, there are practical considerations that dictate which is chosen. The chosen cell line should be relatively cheap and easy to culture and sufficiently robust to withstand application of the investigative methods and reagents.

The latter is especially important for doing [high-throughput studies](#). Therefore the yeast *S. cerevisiae* has been the main host organism for two-hybrid studies. However it is not always the ideal system to study interacting proteins from other organisms.

Yeast cells often do not have the same post translational modifications, have a different codon use or lack certain proteins that are important for the correct expression of the proteins. To cope with these problems several novel two-hybrid systems have been developed.

Depending on the system used agar plates or specific growth medium is used to grow the cells and allow selection for interaction. The most common used method is the agar plating one where cells are plated on selective medium to see if interaction takes place. Cells that have no interaction proteins should not survive on this selective medium.

S. cerevisiae

The yeast *S. cerevisiae* was the model organism used during the two-hybrid technique's inception. It is commonly known as the Y2H system. It has several characteristics that make it a robust organism to host the interaction, including the ability to form tertiary protein structures, neutral internal pH, enhanced ability to form disulfide bonds and reduced-state glutathione among other cytosolic buffer factors, to maintain a hospitable internal environment.

The yeast model can be manipulated through non-molecular techniques and its complete genome sequence is known. Yeast systems are tolerant of diverse culture conditions and harsh chemicals that could not be applied to mammalian tissue cultures.

A number of yeast strains have been created specifically for Y2H screens, e.g. **Y187** and **AH109**, both produced by **Clontech**. Yeast strains R2HMet and BK100 have also been used.

Candida albicans

C. albicans is a yeast with a particular feature: it translates the CUG codon into serine rather than leucine. Due to this different codon usage it is difficult to use the model system *S. cerevisiae* as a Y2H to check for protein-protein interactions using *C. albicans* genes. To provide a more native environment a *C. albicans* two-hybrid (C2H) system was developed. With this system protein-protein interactions can be studied in *C. albicans* itself. A recent addition was the creation of a high-throughput system.

E. coli

Bacterial *E. coli*-based two hybrid methods (abbreviated as B2H) have several characteristics that may make them preferable to yeast-based homologues. The higher transformation efficiency and faster rate of growth lends *E. coli* to the use of larger libraries (in excess of 10^8) A low false positive rate of approximately 3×10^{-8} , the absence of requirement for a **nuclear localisation signal** to be included in the protein sequence and the ability to study proteins that would be toxic to yeast may also be major factors to consider when choosing an experimental background organism.

It may be of note that the methylation activity of certain *E. coli* **DNA methyltransferase** proteins may interfere with some DNA-binding protein selections. If this is anticipated, the use of an *E. coli* strain that is defective for a particular methyltransferase may be an obvious solution. Important to mention is that bacteria are prokaryotic organisms and when studying eukaryotic protein-protein interactions (e.g. human proteins) the results need to be carefully approached.

Mammalian cells

In recent years a mammalian two hybrid (M2H) system has been designed to study mammalian protein-protein interactions in a cellular environment that closely mimics the native protein environment. Transiently transfected mammalian cells are used in this system to find protein-protein interactions. Using a mammalian cell line to study mammalian protein-protein interactions gives the advantage of working in a more native context. The post-translational modifications, phosphorylation, acylation and glycosylation are similar. The intracellular localization of the proteins is also more correct compared to using a yeast two hybrid system. It

is also possible with the mammalian two-hybrid system to study signal inputs. Another big advantage is that results can be obtained within 48 hours after transfection.

Arabidopsis thaliana

In 2005 a two hybrid system in plants was developed. Using protoplasts of *A. thaliana* protein-protein interactions can be studied in plants. This way the interactions can be studied in their native context. In this system the GAL4 AD and BD are under the control of the strong 35S promoter. Interaction is measured using a GUS reporter. In order to enable a high-throughput screening the vectors were made gateway compatible. The system is known as the protoplast two hybrid (P2H) system.

Aplysia californica

The sea hare *A californica* is a model organism in neurobiology to study among others the molecular mechanisms of long-term memory. To study interactions, important in neurology, in a more native environment a two-hybrid system has been developed in *A californica* neurons. A GAL4 AD and BD are used in this system.

Bombyx mori

An insect two-hybrid (I2H) system was developed in a silkworm cell line from the larva or caterpillar of the domesticated silk moth, *Bombyx mori* (BmN4 cells). This system uses the GAL4 BD and the activation domain of mouse NF- κ B P65. Both are under the control of the OpIE2 promoter.

Applications

Determination of sequences crucial for interaction

By changing specific amino acids by mutating the corresponding DNA base-pairs in the plasmids used, the importance of those amino acid residues in maintaining the interaction can be determined.

After using bacterial cell-based method to select DNA-binding proteins, it is necessary to check the specificity of these domains as there is a limit to the extent to which the bacterial cell genome can act as a sink for domains with an affinity for other sequences (or indeed, a general affinity for DNA).

Drug and poison discovery

Protein–protein signalling interactions pose suitable therapeutic targets due to their specificity and pervasiveness. The random drug discovery approach uses compound banks that comprise random chemical structures, and requires a high-throughput method to test these structures in their intended target.

The cell chosen for the investigation can be specifically engineered to mirror the molecular aspect that the investigator intends to study and then used to identify new human or animal therapeutics or anti-pest agents.

Determination of protein function

By determination of the interaction partners of unknown proteins, the possible functions of these new proteins may be inferred.¹ This can be done using a single known protein against a library of unknown proteins or conversely, by selecting from a library of known proteins using a single protein of unknown function.

Zinc finger protein selection

To select [zinc finger proteins](#) (ZFPs) for [protein engineering](#), methods adapted from the two-hybrid screening technique have been used with success. A ZFP is itself a DNA-binding protein used in the construction of custom DNA-binding domains that bind to a desired DNA sequence.

By using a selection gene with the desired target sequence included in the UAS, and randomising the relevant amino acid sequences to produce a ZFP library, cells that host a DNA-ZFP interaction with the required characteristics can be selected. Each ZFP typically recognises only 3–4 base pairs, so to prevent recognition of sites outside the UAS, the randomised ZFP is engineered into a 'scaffold' consisting of another two ZFPs of constant sequence. The UAS is thus designed to include the target sequence of the constant scaffold in addition to the sequence for which a ZFP is selected.

A number of other DNA-binding domains may also be investigated using this system.

Strengths

- Two-hybrid screens are low-tech; they can be carried out in any lab without sophisticated equipment.
- Two-hybrid screens can provide an important first hint for the identification of interaction partners.
- The assay is scalable, which makes it possible to screen for interactions among many proteins. Furthermore, it can be automated, and by using robots many proteins can be screened against thousands of potentially interacting proteins in a relatively short time. Two types of large screens are used: the library approach and the matrix approach.
- Yeast two-hybrid data can be of similar quality to data generated by the alternative approach of [coaffinity purification](#) followed by [mass spectrometry](#) (AP/MS).

Weaknesses

- The main criticism applied to the yeast two-hybrid screen of protein–protein interactions are the possibility of a high number of false positive (and false negative) identifications. The exact rate of false positive results is not known, but earlier estimates were as high as 70%. This also, partly, explains the often found very small overlap in results when using a (high throughput) two-hybrid screening, especially when using different experimental systems.

The reason for this high error rate lies in the characteristics of the screen:

- Certain assay variants overexpress the fusion proteins which may cause unnatural protein concentrations that lead to unspecific (false) positives.
- The hybrid proteins are fusion proteins; that is, the fused parts may inhibit certain interactions, especially if an interaction takes place at the N-terminus of a test protein (where the DNA-binding or activation domain is typically attached).

- An interaction may not happen in yeast, the typical host organism for Y2H. For instance, if a bacterial protein is tested in yeast, it may lack a chaperone for proper folding that is only present in its bacterial host. Moreover, a **mammalian** protein is sometimes not correctly modified in yeast (e.g., missing **phosphorylation**), which can also lead to false results.
- The Y2H takes place in the nucleus. If test proteins are not localized to the nucleus (because they have other localization signals) two interacting proteins may be found to be non-interacting.
- Some proteins might specifically interact when they are co-expressed in the yeast, although in reality they are never present in the same cell at the same time. However, in most cases it cannot be ruled out that such proteins are indeed expressed in certain cells or under certain circumstances.

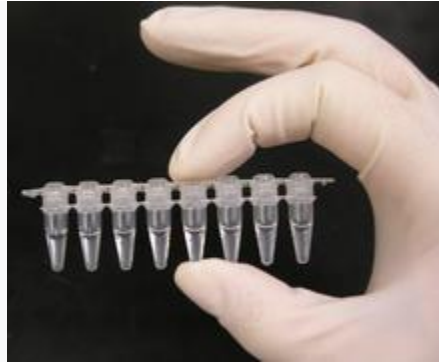
Each of these points alone can give rise to false results. Due to the combined effects of all error sources yeast two-hybrid have to be interpreted with caution. The probability of generating false positives means that all interactions should be confirmed by a high confidence assay, for example **co-immunoprecipitation** of the endogenous proteins, which is difficult for large scale protein–protein interaction data. Alternatively, Y2H data can be verified using multiple Y2H variants or bioinformatics techniques. The latter test whether interacting proteins are expressed at the same time, share some common features (such as **gene ontology** annotations or certain **network topologies**), have homologous interactions in other species.

Maximizing gene expression on a plasmid using recombination in vitro

Recombination in vitro has been used to place one or more copies of a strong promoter, the *lac* promoter, at varying distances from the *cl* (repressor) gene of bacteriophage λ on the *E. coli* plasmid pMB9. In all constructions, λ repressor synthesis is driven wholly or predominantly by the inserted *lac* promoter. One of our fusions directs the synthesis of very high levels of λ repressor. In this case, the fused DNA encodes a ribosome binding site which is a “hybrid” of λ and *lac* sequences. In principle, this method of construction should elicit high levels of expression in *E. coli* of any gene, whatever its source. We also described strains with different sequence arrangements that, for reasons not completely understood, produce less repressor.

UNIT-IV

Polymerase chain reaction



A strip of eight PCR tubes, each containing a 100 μ L reaction mixture

Polymerase chain reaction (PCR) is a method widely used in **molecular biology** to rapidly make millions to billions of copies of a specific **DNA** sample, allowing scientists to take a very small sample of DNA and amplify it to a large enough amount to study in detail. PCR was invented in 1983 by the **American biochemist Kary Mullis** at **Cetus Corporation**.

It is fundamental to much of genetic testing including analysis of **ancient samples of DNA** and identification of infectious agents. Using PCR, copies of very small amounts of **DNA sequences** are exponentially amplified in a series or cycles of temperature changes.

PCR is now a common and often indispensable technique used in **medical laboratory** and clinical laboratory research for a broad variety of applications including **biomedical research** and **criminal forensics**.

The majority of PCR methods rely on **thermal cycling**. Thermal cycling exposes reactants to repeated cycles of heating and cooling to permit different temperature-dependent reactions – specifically, **DNA melting** and **enzyme-driven DNA replication**.

PCR employs two main reagents – **primers** (which are short single strand DNA fragments known as **oligonucleotides** that are a **complementary** sequence to the target DNA region) and a **DNA polymerase**.

In the first step of PCR, the two strands of the DNA double helix are physically separated at a high temperature in a process called **Nucleic acid denaturation**. In the second step, the temperature is lowered and the primers bind to the complementary sequences of DNA. The two DNA strands then become **templates** for DNA polymerase to **enzymatically** assemble a new DNA strand from free **nucleotides**, the building blocks of DNA.

As PCR progresses, the DNA generated is itself used as a template for replication, setting in motion a **chain reaction** in which the original DNA template is **exponentially** amplified.

Almost all PCR applications employ a heat-stable DNA polymerase, such as **Taq polymerase**, an enzyme originally isolated from the **thermophilic** bacterium *Thermus aquaticus*. If the polymerase used was heat-susceptible, it would denature under the high temperatures of the denaturation step. Before the use of Taq polymerase, DNA polymerase had to be manually added every cycle, which was a tedious and costly process.

Applications of the technique include **DNA cloning** for **sequencing**, gene cloning and manipulation, gene mutagenesis; construction of DNA-based **phylogenies**, or functional analysis of **genes**; **diagnosis** and **monitoring** of **hereditary diseases**; amplification of ancient DNA; analysis of genetic fingerprints for **DNA profiling** (for example, in **forensic science** and **parentage testing**);



Placing a strip of eight PCR tubes into a **thermal cycler**



Principles



A **thermal cycler** for PCR



An older, three-temperature **thermal cycler** for PCR

PCR amplifies a specific region of a DNA strand (the DNA target). Most PCR methods amplify DNA fragments of between 0.1 and 10 **kilo base pairs** (kbp) in length, although some techniques allow for amplification of fragments up to 40 kbp. The amount of amplified product is determined by the available substrates in the reaction, which become limiting as the reaction progresses.

A basic PCR set-up requires several components and reagents, including:

- a *DNA template* that contains the DNA target region to amplify
- a *DNA polymerase*; an enzyme that **polymerizes** new DNA strands; heat-resistant **Taq polymerase** is especially common, as it is more likely to remain intact during the high-temperature DNA denaturation process
- two DNA *primers* that are **complementary** to the 3' (three prime) ends of each of the **sense and anti-sense** strands of the DNA target (DNA polymerase can only bind to and elongate from a double-stranded region of DNA; without primers there is no double-stranded initiation site at which the polymerase can bind); specific primers that are complementary to the DNA target region are selected beforehand, and are often custom-made in a laboratory or purchased from commercial biochemical suppliers
- *deoxynucleoside triphosphates*, or dNTPs (sometimes called "deoxynucleotide triphosphates"; **nucleotides** containing triphosphate groups), the building blocks from which the DNA polymerase synthesizes a new DNA strand
- a *buffer solution* providing a suitable chemical environment for optimum activity and stability of the DNA polymerase
- *bivalent cations*, typically **magnesium** (Mg) or **manganese** (Mn) ions; Mg²⁺ is the most common, but Mn²⁺ can be used for **PCR-mediated DNA mutagenesis**, as a higher Mn²⁺ concentration increases the error rate during DNA synthesis; and *monovalent cations*, typically **potassium** (K) ions

The reaction is commonly carried out in a volume of 10–200 μL in small reaction tubes (0.2–0.5 mL volumes) in a **thermal cycler**. The thermal cycler heats and cools the reaction tubes to achieve the temperatures required at each step of the reaction (see below).

Many modern thermal cyclers make use of the **Peltier effect**, which permits both heating and cooling of the block holding the PCR tubes simply by reversing the electric current. Thin-walled

reaction tubes permit favorable **thermal conductivity** to allow for rapid thermal equilibration. Most thermal cyclers have heated lids to prevent condensation at the top of the reaction tube. Older thermal cyclers lacking a heated lid require a layer of oil on top of the reaction mixture or a ball of wax inside the tube.

Procedure

Typically, PCR consists of a series of 20–40 repeated temperature changes, called thermal cycles, with each cycle commonly consisting of two or three discrete temperature steps (see figure below). The cycling is often preceded by a single temperature step at a very high temperature ($>90\text{ }^{\circ}\text{C}$ ($194\text{ }^{\circ}\text{F}$)), and followed by one hold at the end for final product extension or brief storage.

The temperatures used and the length of time they are applied in each cycle depend on a variety of parameters, including the enzyme used for DNA synthesis, the concentration of bivalent ions and dNTPs in the reaction, and the **melting temperature** (T_m) of the primers. The individual steps common to most PCR methods are as follows:

- **Initialization:** This step is only required for DNA polymerases that require heat activation by **hot-start PCR**. It consists of heating the reaction chamber to a temperature of $94\text{--}96\text{ }^{\circ}\text{C}$ ($201\text{--}205\text{ }^{\circ}\text{F}$), or $98\text{ }^{\circ}\text{C}$ ($208\text{ }^{\circ}\text{F}$) if extremely thermostable polymerases are used, which is then held for 1–10 minutes.
- **Denaturation:** This step is the first regular cycling event and consists of heating the reaction chamber to $94\text{--}98\text{ }^{\circ}\text{C}$ ($201\text{--}208\text{ }^{\circ}\text{F}$) for 20–30 seconds. This causes **DNA melting**, or denaturation, of the double-stranded DNA template by breaking the hydrogen bonds between complementary bases, yielding two single-stranded DNA molecules.
- **Annealing:** In the next step, the reaction temperature is lowered to $50\text{--}65\text{ }^{\circ}\text{C}$ ($122\text{--}149\text{ }^{\circ}\text{F}$) for 20–40 seconds, allowing annealing of the primers to each of the single-stranded DNA templates. Two different primers are typically included in the reaction mixture: one for each of the two single-stranded complements containing the target region. The primers are single-stranded sequences themselves, but are much shorter than the length of the target region, complementing only very short sequences at the 3' end of each strand.

It is critical to determine a proper temperature for the annealing step because efficiency and specificity are strongly affected by the annealing temperature. This temperature must be low enough to allow for **hybridization** of the primer to the strand, but high enough for the hybridization to be specific, i.e., the primer should bind *only* to a perfectly complementary part of the strand, and nowhere else. If the temperature is too low, the primer may bind imperfectly. If it is too high, the primer may not bind at all. A typical annealing temperature is about $3\text{--}5\text{ }^{\circ}\text{C}$ below the T_m of the primers used. Stable hydrogen bonds between complementary bases are formed only when the primer sequence very closely matches the template sequence. During this step, the polymerase binds to the primer-template hybrid and begins DNA formation.

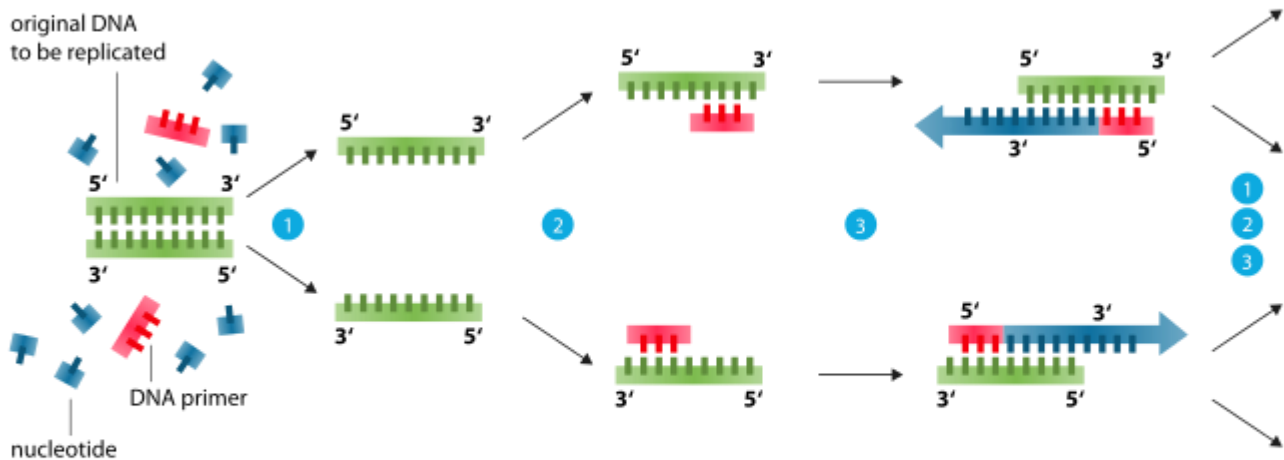
- **Extension/elongation:** The temperature at this step depends on the DNA polymerase used; the optimum **activity** temperature for the thermostable DNA polymerase of **Taq (Thermus aquaticus) polymerase** is approximately $75\text{--}80\text{ }^{\circ}\text{C}$ ($167\text{--}176\text{ }^{\circ}\text{F}$), though a temperature of $72\text{ }^{\circ}\text{C}$ ($162\text{ }^{\circ}\text{F}$) is commonly used with this enzyme.
- In this step, the DNA polymerase synthesizes a new DNA strand complementary to the DNA template strand by adding free dNTPs from the reaction mixture that are complementary to the template in the 5'-to-3' direction, **condensing** the 5'-**phosphate group** of the dNTPs with the 3'-

hydroxy group at the end of the nascent (elongating) DNA strand. The precise time required for elongation depends both on the DNA polymerase used and on the length of the DNA target region to amplify. As a rule of thumb, at their optimal temperature, most DNA polymerases polymerize a thousand bases per minute. Under optimal conditions (i.e., if there are no limitations due to limiting substrates or reagents), at each extension/elongation step, the number of DNA target sequences is doubled. With each successive cycle, the original template strands plus all newly generated strands become template strands for the next round of elongation, leading to exponential (geometric) amplification of the specific DNA target region.

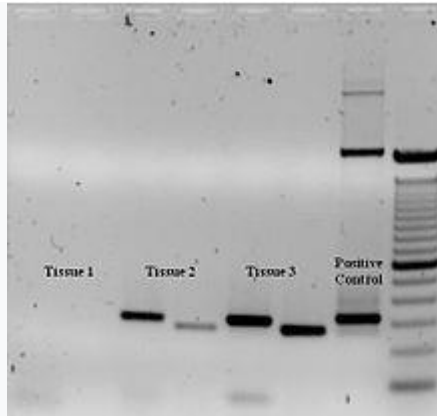
The processes of denaturation, annealing and elongation constitute a single cycle. Multiple cycles are required to amplify the DNA target to millions of copies. The formula used to calculate the number of DNA copies formed after a given number of cycles is 2^n , where n is the number of cycles. Thus, a reaction set for 30 cycles results in 2^{30} , or 1,073,741,824, copies of the original double-stranded DNA target region.

- *Final elongation*: This single step is optional, but is performed at a temperature of 70–74 °C (158–165 °F) (the temperature range required for optimal activity of most polymerases used in PCR) for 5–15 minutes after the last PCR cycle to ensure that any remaining single-stranded DNA is fully elongated.
- *Final hold*: The final step cools the reaction chamber to 4–15 °C (39–59 °F) for an indefinite time, and may be employed for short-term storage of the PCR products.

Polymerase chain reaction - PCR

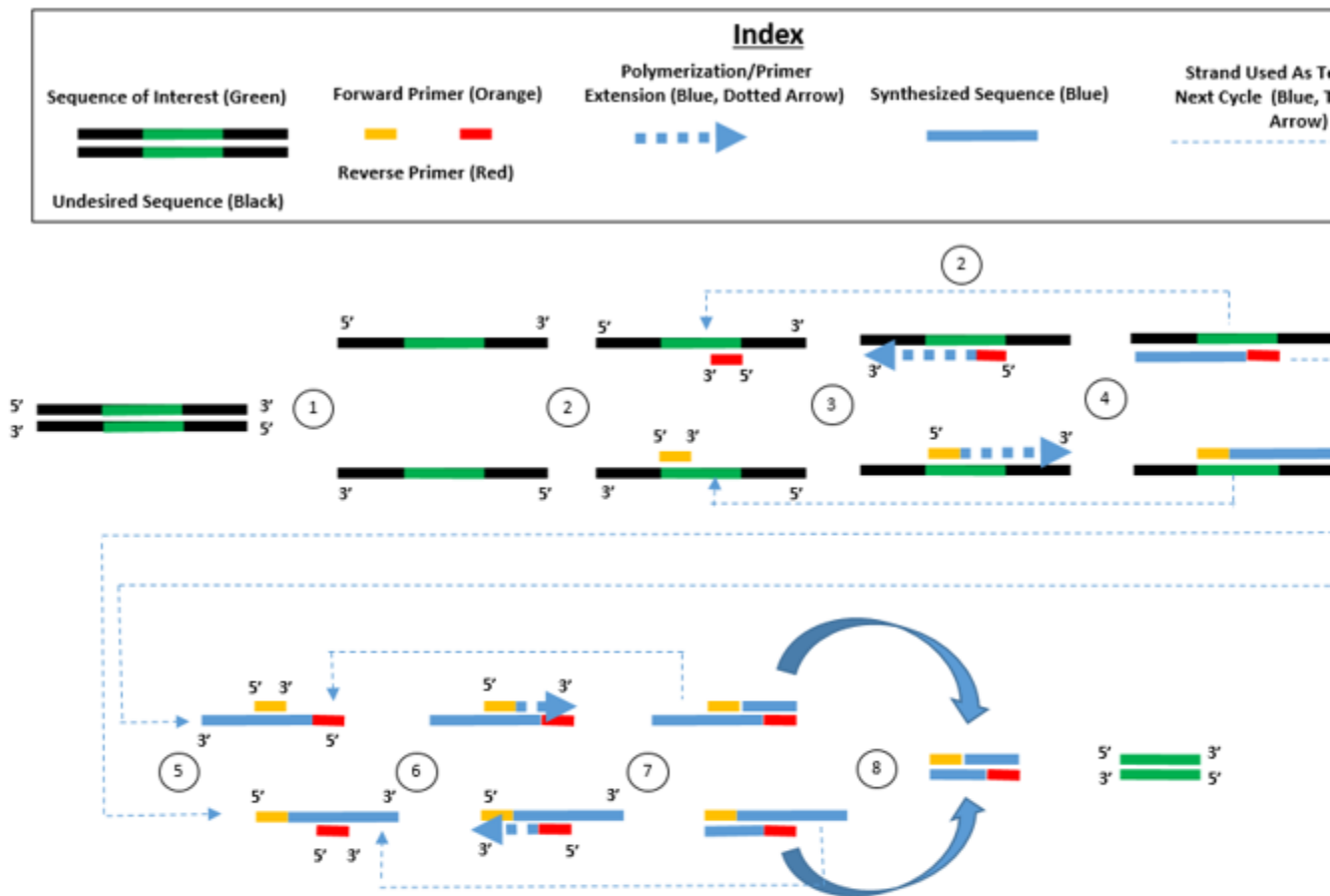


- 1 Denaturation** at 94-96°C
- 2 Annealing** at ~68°C
- 3 Elongation** at ca. 72 °C



Ethidium bromide-stained PCR products after gel electrophoresis. Two sets of primers were used to amplify a target sequence from three different tissue samples. No amplification is present in sample #1; DNA bands in sample #2 and #3 indicate successful amplification of the target sequence. The gel also shows a positive control, and a DNA ladder containing DNA fragments of defined length for sizing the bands in the experimental PCRs.

To check whether the PCR successfully generated the anticipated DNA target region (also sometimes referred to as the amplicon or amplicon), agarose gel electrophoresis may be employed for size separation of the PCR products. The size(s) of PCR products is determined by comparison with a DNA ladder, a molecular weight marker which contains DNA fragments of known size run on the gel alongside the PCR products.



1. The DNA double helix is melted apart at $T > 90^{\circ}\text{C}$ and its strands separate.
2. The temperature is decreased to slightly below the T_m of both the primers being used. Both primers are available to bind to the available strands. These primers are supplied in excess to insure that the strands do not only come back together to one another.
3. Polymerization (extension) occurs via DNA Polymerase in the 5' to 3' direction on each strand.
4. Incorporated additional nucleotides give rise to new strands that extend past the sequence of interest.
5. The previously polymerized strands act as template for the other primer (if forward primer bound first, reverse primer now binds and vice versa).
6. Polymerization occurs via DNA Polymerase in the 5' to 3' direction on each strand, this time ending at the sequence of interest.
7. Incorporated additional nucleotides give rise to new strands that only encode the sequence of interest.
8. The synthesized strands encoding the sequence of interest anneal to one another to form the end product.

Stages

As with other chemical reactions, the reaction rate and efficiency of PCR are affected by limiting factors. Thus, the entire PCR process can further be divided into three stages based on reaction progress:

- *Exponential amplification*: At every cycle, the amount of product is doubled (assuming 100% reaction efficiency). After 30 cycles, a single copy of DNA can be increased up to

1,000,000,000 (one billion) copies. In a sense, then, the replication of a discrete strand of DNA is being manipulated in a tube under controlled conditions. The reaction is very sensitive: only minute quantities of DNA must be present.

- *Leveling off stage*: The reaction slows as the DNA polymerase loses activity and as consumption of reagents, such as dNTPs and primers, causes them to become more limited.
- *Plateau*: No more product accumulates due to exhaustion of reagents and enzyme.

Optimization

In practice, PCR can fail for various reasons, in part due to its sensitivity to contamination causing amplification of spurious DNA products. Because of this, a number of techniques and procedures have been developed for optimizing PCR conditions.

Contamination with extraneous DNA is addressed with lab protocols and procedures that separate pre-PCR mixtures from potential DNA contaminants.^[7] This usually involves spatial separation of PCR-setup areas from areas for analysis or purification of PCR products, use of disposable plasticware, and thoroughly cleaning the work surface between reaction setups.

Primer-design techniques are important in improving PCR product yield and in avoiding the formation of spurious products, and the usage of alternate buffer components or polymerase enzymes can help with amplification of long or otherwise problematic regions of DNA. Addition of reagents, such as **formamide**, in buffer systems may increase the specificity and yield of PCR.

Computer simulations of theoretical PCR results (**Electronic PCR**) may be performed to assist in primer design.

Applications

Selective DNA isolation

PCR allows isolation of DNA fragments from genomic DNA by selective amplification of a specific region of DNA. This use of PCR augments many ways, such as generating **hybridization probes** for **Southern** or **northern** hybridization and **DNA cloning**, which require larger amounts of DNA, representing a specific DNA region. PCR supplies these techniques with high amounts of pure DNA, enabling analysis of DNA samples even from very small amounts of starting material.

Other applications of PCR include **DNA sequencing** to determine unknown PCR-amplified sequences in which one of the amplification primers may be used in **Sanger sequencing**, isolation of a DNA sequence to expedite recombinant DNA technologies involving the insertion of a DNA sequence into a **plasmid**, **phage**, or **cosmid** (depending on size) or the genetic material of another organism. Bacterial colonies (*such as E. coli*) can be rapidly screened by PCR for correct DNA **vector** constructs.

PCR may also be used for **genetic fingerprinting**; a forensic technique used to identify a person or organism by comparing experimental DNAs through different PCR-based methods.

Some PCR 'fingerprints' methods have high discriminative power and can be used to identify genetic relationships between individuals, such as parent-child or between siblings, and are used in paternity testing (Fig. 4). This technique may also be used to determine evolutionary relationships among organisms when certain molecular clocks are used (i.e., the **16S rRNA** and **recA** genes of microorganisms).



Electrophoresis of PCR-amplified DNA fragments. (1) Father. (2) Child. (3) Mother. The child has inherited some, but not all of the fingerprint of each of its parents, giving it a new, unique fingerprint.

Amplification and quantification of DNA

Use of DNA in forensic entomology

Because PCR amplifies the regions of DNA that it targets, PCR can be used to analyze extremely small amounts of sample. This is often critical for **forensic analysis**, when only a trace amount of DNA is available as evidence.

PCR may also be used in the analysis of **ancient DNA** that is tens of thousands of years old. These PCR-based techniques have been successfully used on animals, such as a forty-thousand-year-old **mammoth**, and also on human DNA, in applications ranging from the analysis of Egyptian **mummies** to the identification of a **Russian tsar** and the body of English king **Richard III**.

Quantitative PCR or Real Time PCR (qPCR, not to be confused with **RT-PCR**) methods allow the estimation of the amount of a given sequence present in a sample—a technique often applied to quantitatively determine levels of **gene expression**. Quantitative PCR is an established tool for DNA quantification that measures the accumulation of DNA product after each round of PCR amplification.

qPCR allows the quantification and detection of a specific DNA sequence in real time since it measures concentration while the synthesis process is taking place. There are two methods for simultaneous detection and quantification. The first method consists

of using **fluorescent** dyes that are retained nonspecifically in between the double strands. The second method involves probes that code for specific sequences and are fluorescently labeled. Detection of DNA using these methods can only be seen after the hybridization of probes with its complementary DNA takes place. An interesting technique combination is real-time PCR and reverse transcription.

This sophisticated technique, called RT-qPCR, allows for the quantification of a small quantity of RNA. Through this combined technique, mRNA is converted to cDNA, which is further quantified using qPCR. This technique lowers the possibility of error at the end point of PCR, increasing chances for detection of genes associated with genetic diseases such as cancer. Laboratories use RT-qPCR for the purpose of sensitively measuring gene regulation.

Medical and diagnostic applications

Prospective parents can be tested for being **genetic carriers**, or their children might be tested for actually being affected by a **disease**. DNA samples for **prenatal testing** can be obtained by **amniocentesis**, **chorionic villus sampling**, or even by the analysis of rare fetal cells circulating in the mother's bloodstream. PCR analysis is also essential to **preimplantation genetic diagnosis**, where individual cells of a developing embryo are tested for mutations.

- PCR can also be used as part of a sensitive test for **tissue typing**, vital to **organ transplantation**. As of 2008, there is even a proposal to replace the traditional antibody-based tests for **blood type** with PCR-based tests.
- Many forms of cancer involve alterations to **oncogenes**. By using PCR-based tests to study these mutations, therapy regimens can sometimes be individually customized to a patient. PCR permits early diagnosis of **malignant** diseases such as **leukemia** and **lymphomas**, which is currently the highest-developed in cancer research and is already being used routinely. PCR assays can be performed directly on genomic DNA samples to detect translocation-specific malignant cells at a sensitivity that is at least 10,000 fold higher than that of other methods.
- PCR is very useful in the medical field since it allows for the isolation and amplification of tumor suppressors. Quantitative PCR for example, can be used to quantify and analyze single cells, as well as recognize DNA, mRNA and protein confirmations and combinations.

Infectious disease applications

PCR allows for rapid and highly specific diagnosis of infectious diseases, including those caused by bacteria or viruses.

PCR also permits identification of non-cultivable or slow-growing microorganisms such as **mycobacteria**, **anaerobic bacteria**, or **viruses** from **tissue culture** assays and **animal models**.

The basis for PCR diagnostic applications in microbiology is the detection of infectious agents and the discrimination of non-pathogenic from pathogenic strains by virtue of specific genes.

Characterization and detection of infectious disease organisms have been revolutionized by PCR in the following ways:

- The *human immunodeficiency virus* (or *HIV*), is a difficult target to find and eradicate. The earliest tests for infection relied on the presence of antibodies to the virus circulating in the bloodstream. However, antibodies don't appear until many weeks after infection, maternal antibodies mask the infection of a newborn, and therapeutic agents to fight the infection don't affect the antibodies. PCR tests have been developed that can detect as little as one viral genome among the DNA of over 50,000 host cells. Infections can be detected earlier, donated blood can be screened directly for the virus, newborns can be immediately tested for infection, and the effects of antiviral treatments can be quantified.
- Some disease organisms, such as that for *tuberculosis*, are difficult to sample from patients and slow to be grown in the laboratory. PCR-based tests have allowed detection of small numbers of disease organisms (both live or dead), in convenient samples. Detailed genetic analysis can also be used to detect antibiotic resistance, allowing immediate and effective therapy. The effects of therapy can also be immediately evaluated.
- The spread of a *disease organism* through populations of domestic or wild animals can be monitored by PCR testing. In many cases, the appearance of new virulent sub-types can be detected and monitored. The sub-types of an organism that were responsible for earlier epidemics can also be determined by PCR analysis.
- Viral DNA can be detected by PCR. The primers used must be specific to the targeted sequences in the DNA of a virus, and PCR can be used for diagnostic analyses or DNA sequencing of the viral genome. The high sensitivity of PCR permits virus detection soon after infection and even before the onset of disease. Such early detection may give physicians a significant lead time in treatment. The amount of virus ("viral load") in a patient can also be quantified by PCR-based DNA quantitation techniques (see below). For example, a variant of PCR (RT-PCR) is used for detecting Sars-Cov-2 viral genome.
- Diseases such as pertussis (or *whooping cough*) are caused by the bacteria *Bordetella pertussis*. This bacteria is marked by a serious acute respiratory infection that affects various animals and humans and has led to the deaths of many young children. The pertussis toxin is a protein exotoxin that binds to cell receptors by two dimers and reacts with different cell types such as T lymphocytes which play a role in cell immunity. PCR is an important testing tool that can detect sequences within the gene for the pertussis toxin. Because PCR has a high sensitivity for the toxin and a rapid turnaround time, it is very efficient for diagnosing pertussis when compared to culture.

Forensic applications

The development of PCR-based genetic (or DNA) fingerprinting protocols has seen widespread application in forensics:

- In its most discriminating form, *genetic fingerprinting* can uniquely discriminate any one person from the entire population of the world. Minute samples of DNA can be isolated from a crime scene, and compared to that from suspects, or from a DNA database of earlier evidence or convicts. Simpler versions of these tests are often used to rapidly rule out suspects during a criminal investigation. Evidence from decades-old crimes can be tested, confirming or exonerating the people originally convicted.
- Forensic DNA typing has been an effective way of identifying or exonerating criminal suspects due to analysis of evidence discovered at a crime scene. The human genome has many repetitive regions that can be found within gene sequences or in non-coding regions of the genome. Specifically, up to 40% of human DNA is repetitive.
- There are two distinct categories for these repetitive, non-coding regions in the genome. The first category is called variable number tandem repeats (VNTR), which are 10–100 base pairs long and the second category is called short tandem repeats (STR) and these consist of

repeated 2–10 base pair sections. PCR is used to amplify several well-known VNTRs and STRs using primers that flank each of the repetitive regions.

- The sizes of the fragments obtained from any individual for each of the STRs will indicate which alleles are present. By analyzing several STRs for an individual, a set of alleles for each person will be found that statistically is likely to be unique.
- Researchers have identified the complete sequence of the human genome. This sequence can be easily accessed through the NCBI website and is used in many real-life applications. For example, the FBI has compiled a set of DNA marker sites used for identification, and these are called the Combined DNA Index System (CODIS) DNA database.
- Using this database enables statistical analysis to be used to determine the probability that a DNA sample will match. PCR is a very powerful and significant analytical tool to use for forensic DNA typing because researchers only need a very small amount of the target DNA to be used for analysis. For example, a single human hair with attached hair follicle has enough DNA to conduct the analysis. Similarly, a few sperm, skin samples from under the fingernails, or a small amount of blood can provide enough DNA for conclusive analysis.
- Less discriminating forms of **DNA fingerprinting** can help in *DNA paternity testing*, where an individual is matched with their close relatives. DNA from unidentified human remains can be tested, and compared with that from possible parents, siblings, or children. Similar testing can be used to confirm the biological parents of an adopted (or kidnapped) child. The actual biological father of a **newborn** can also be **confirmed** (or ruled out).
- The PCR AMGX/AMGY design has been shown to not only facilitating in amplifying DNA sequences from a very minuscule amount of genome. However it can also be used for real time sex determination from forensic bone samples. This provides us with a powerful and effective way to determine the sex of not only ancient specimens but also current suspects in crimes.

Research applications

PCR has been applied to many areas of research in molecular genetics:

- PCR allows rapid production of short pieces of DNA, even when not more than the sequence of the two primers is known. This ability of PCR augments many methods, such as generating *hybridization probes* for **Southern** or **northern blot** hybridization. PCR supplies these techniques with large amounts of pure DNA, sometimes as a single strand, enabling analysis even from very small amounts of starting material.
- The task of *DNA sequencing* can also be assisted by PCR. Known segments of DNA can easily be produced from a patient with a genetic disease mutation. Modifications to the amplification technique can extract segments from a completely unknown genome, or can generate just a single strand of an area of interest.
- PCR has numerous applications to the more traditional process of *DNA cloning*. It can extract segments for insertion into a vector from a larger genome, which may be only available in small quantities. Using a single set of 'vector primers', it can also analyze or extract fragments that have already been inserted into vectors. Some alterations to the PCR protocol can *generate mutations* (general or site-directed) of an inserted fragment.
- *Sequence-tagged sites* is a process where PCR is used as an indicator that a particular segment of a genome is present in a particular clone. The **Human Genome Project** found this application vital to mapping the cosmid clones they were sequencing, and to coordinating the results from different laboratories.
- An exciting application of PCR is the **phylogenic** analysis of DNA from *ancient sources*, such as that found in the recovered bones of **Neanderthals**, from frozen tissues

of mammoths, or from the brain of Egyptian mummies. Have been amplified and sequenced. In some cases the highly degraded DNA from these sources might be reassembled during the early stages of amplification.

- A common application of PCR is the study of patterns of *gene expression*. Tissues (or even individual cells) can be analyzed at different stages to see which genes have become active, or which have been switched off. This application can also use **quantitative PCR** to quantitate the actual levels of expression
- The ability of PCR to simultaneously amplify several loci from individual sperm has greatly enhanced the more traditional task of *genetic mapping* by studying **chromosomal crossovers** after **meiosis**. Rare crossover events between very close loci have been directly observed by analyzing thousands of individual sperms. Similarly, unusual deletions, insertions, translocations, or inversions can be analyzed, all without having to wait (or pay) for the long and laborious processes of fertilization, embryogenesis, etc.
- **Site-directed mutagenesis**: PCR can be used to create mutant genes with mutations chosen by scientists at will. These mutations can be chosen in order to understand how proteins accomplish their functions, and to change or improve protein function.

Advantages

PCR has a number of advantages. It is fairly simple to understand and to use, and produces results rapidly. The technique is highly sensitive with the potential to produce millions to billions of copies of a specific product for sequencing, cloning, and analysis. qRT-PCR shares the same advantages as the PCR, with an added advantage of quantification of the synthesized product.

Therefore, it has its uses to analyze alterations of gene expression levels in tumors, microbes, or other disease states.

PCR is a very powerful and practical research tool. The sequencing of unknown etiologies of many diseases are being figured out by the PCR. The technique can help identify the sequence of previously unknown viruses related to those already known and thus give us a better understanding of the disease itself.

If the procedure can be further simplified and sensitive non radiometric detection systems can be developed, the PCR will assume a prominent place in the clinical laboratory for years to come.

Limitations

One major limitation of PCR is that prior information about the target sequence is necessary in order to generate the primers that will allow its selective amplification. This means that, typically, PCR users must know the precise sequence(s) upstream of the target region on each of the two single-stranded templates in order to ensure that the

DNA polymerase properly binds to the primer-template hybrids and subsequently generates the entire target region during DNA synthesis.

Like all enzymes, DNA polymerases are also prone to error, which in turn causes mutations in the PCR fragments that are generated.

Another limitation of PCR is that even the smallest amount of contaminating DNA can be amplified, resulting in misleading or ambiguous results. To minimize the

chance of contamination, investigators should reserve separate rooms for reagent preparation, the PCR, and analysis of product. Reagents should be dispensed into single-use **aliquots**. Pipettors with disposable plungers and extra-long pipette tips should be routinely used.

Variations

- *Allele-specific PCR*: a diagnostic or cloning technique based on single-nucleotide variations (SNVs not to be confused with **SNPs**) (single-base differences in a patient). It requires prior knowledge of a DNA sequence, including differences between **alleles**, and uses primers whose 3' ends encompass the SNV (base pair buffer around SNV usually incorporated). PCR amplification under stringent conditions is much less efficient in the presence of a mismatch between template and primer, so successful amplification with an SNP-specific primer signals presence of the specific SNP in a sequence. See **SNP genotyping** for more information.
- *Assembly PCR* or *Polymerase Cycling Assembly (PCA)*: artificial synthesis of long DNA sequences by performing PCR on a pool of long oligonucleotides with short overlapping segments. The oligonucleotides alternate between sense and antisense directions, and the overlapping segments determine the order of the PCR fragments, thereby selectively producing the final long DNA product.
- *Asymmetric PCR*: preferentially amplifies one DNA strand in a double-stranded DNA template. It is used in **sequencing** and hybridization probing where amplification of only one of the two complementary strands is required. PCR is carried out as usual, but with a great excess of the primer for the strand targeted for amplification. Because of the slow (**arithmetic**) amplification later in the reaction after the limiting primer has been used up, extra cycles of PCR are required. A recent modification on this process, known as *Linear-After-The-Exponential-PCR (LATE-PCR)*, uses a limiting primer with a higher melting temperature (T_m) than the excess primer to maintain reaction efficiency as the limiting primer concentration decreases mid-reaction.
- *Convective PCR*: a pseudo-isothermal way of performing PCR. Instead of repeatedly heating and cooling the PCR mixture, the solution is subjected to a thermal gradient. The resulting thermal instability driven convective flow automatically shuffles the PCR reagents from the hot and cold regions repeatedly enabling PCR. Parameters such as thermal boundary conditions and geometry of the PCR enclosure can be optimized to yield robust and rapid PCR by harnessing the emergence of chaotic flow fields. Such convective flow PCR setup significantly reduces device power requirement and operation time.
- *Dial-out PCR*: a highly parallel method for retrieving accurate DNA molecules for gene synthesis. A complex library of DNA molecules is modified with unique flanking tags before massively parallel sequencing. Tag-directed primers then enable the retrieval of molecules with desired sequences by PCR.
- *Digital PCR (dPCR)*: used to measure the quantity of a target DNA sequence in a DNA sample. The DNA sample is highly diluted so that after running many PCRs in parallel, some of them do not receive a single molecule of the target DNA. The target DNA concentration is calculated using the proportion of negative outcomes. Hence the name 'digital PCR'.
- *Helicase-dependent amplification*: similar to traditional PCR, but uses a constant temperature rather than cycling through denaturation and annealing/extension cycles. **DNA helicase**, an enzyme that unwinds DNA, is used in place of thermal denaturation.

- *Hot start PCR*: a technique that reduces non-specific amplification during the initial set up stages of the PCR. It may be performed manually by heating the reaction components to the denaturation temperature (e.g., 95 °C) before adding the polymerase. Specialized enzyme systems have been developed that inhibit the polymerase's activity at ambient temperature, either by the binding of an **antibody** or by the presence of covalently bound inhibitors that dissociate only after a high-temperature activation step. Hot-start/cold-finish PCR is achieved with new hybrid polymerases that are inactive at ambient temperature and are instantly activated at elongation temperature.
- *In silico PCR* (digital PCR, virtual PCR, electronic PCR, e-PCR) refers to computational tools used to calculate theoretical polymerase chain reaction results using a given set of **primers (probes)** to amplify **DNA** sequences from a sequenced **genome** or **transcriptome**. In silico PCR was proposed as an educational tool for molecular biology.
- *Intersequence-specific PCR* (ISSR): a PCR method for DNA fingerprinting that amplifies regions between simple sequence repeats to produce a unique fingerprint of amplified fragment lengths.
- *Inverse PCR*: is commonly used to identify the flanking sequences around **genomic** inserts. It involves a series of **DNA digestions** and **self ligation**, resulting in known sequences at either end of the unknown sequence.
- *Ligation-mediated PCR*: uses small DNA linkers ligated to the DNA of interest and multiple primers annealing to the DNA linkers; it has been used for **DNA sequencing**, **genome walking**, and **DNA footprinting**.
- *Methylation-specific PCR* (MSP): developed by **Stephen Baylin** and **James G. Herman** at the Johns Hopkins School of Medicine, and is used to detect methylation of CpG islands in genomic DNA. DNA is first treated with sodium bisulfite, which converts unmethylated cytosine bases to uracil, which is recognized by PCR primers as thymine. Two PCRs are then carried out on the modified DNA, using primer sets identical except at any CpG islands within the primer sequences. At these points, one primer set recognizes DNA with cytosines to amplify methylated DNA, and one set recognizes DNA with uracil or thymine to amplify unmethylated DNA. MSP using qPCR can also be performed to obtain quantitative rather than qualitative information about methylation.
- *Miniprimer PCR*: uses a thermostable polymerase (S-Tbr) that can extend from short primers ("smalligos") as short as 9 or 10 nucleotides. This method permits PCR targeting to smaller primer binding regions, and is used to amplify conserved DNA sequences, such as the 16S (or eukaryotic 18S) rRNA gene.
- *Multiplex ligation-dependent probe amplification* (MLPA): permits amplifying multiple targets with a single primer pair, thus avoiding the resolution limitations of multiplex PCR (see below).
- *Multiplex-PCR*: consists of multiple primer sets within a single PCR mixture to produce **amplicons** of varying sizes that are specific to different DNA sequences. By targeting multiple genes at once, additional information may be gained from a single test-run that otherwise would require several times the reagents and more time to perform. Annealing temperatures for each of the primer sets must be optimized to work correctly within a single reaction, and amplicon sizes. That is, their base pair length should be different enough to form distinct bands when visualized by **gel electrophoresis**.
- *Nanoparticle-Assisted PCR* (*nanoPCR*): some nanoparticles (NPs) can enhance the efficiency of PCR (thus being called nanoPCR), and some can even outperform the original PCR enhancers. It was reported that quantum dots (QDs) can improve PCR specificity and efficiency. Single-walled carbon nanotubes (SWCNTs) and multi-walled carbon nanotubes (MWCNTs) are efficient in enhancing the amplification of long PCR. Carbon nanopowder (CNP) can improve the efficiency of repeated PCR and long PCR, while **zinc oxide**, **titanium**

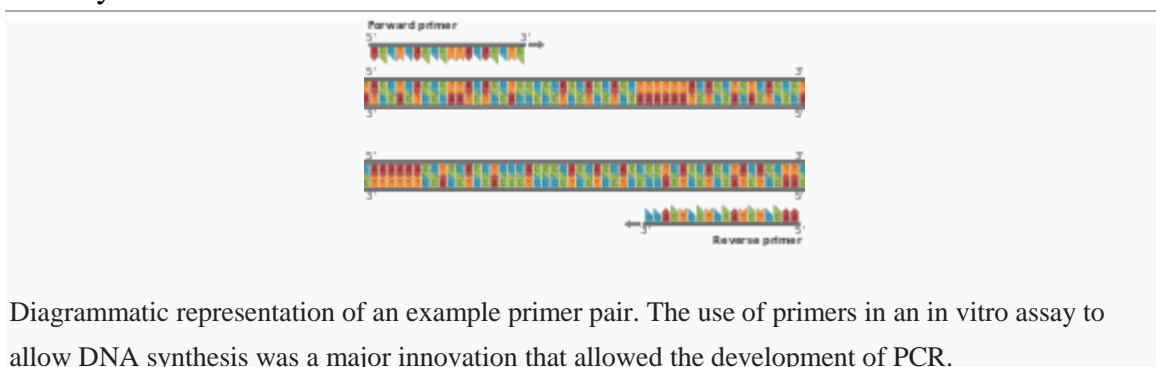
dioxide and Ag NPs were found to increase the PCR yield. Previous data indicated that non-metallic NPs retained acceptable amplification fidelity. Given that many NPs are capable of enhancing PCR efficiency, it is clear that there is likely to be great potential for nanoPCR technology improvements and product development.

- **Nested PCR**: increases the specificity of DNA amplification, by reducing background due to non-specific amplification of DNA. Two sets of primers are used in two successive PCRs. In the first reaction, one pair of primers is used to generate DNA products, which besides the intended target, may still consist of non-specifically amplified DNA fragments. The product(s) are then used in a second PCR with a set of primers whose binding sites are completely or partially different from and located 3' of each of the primers used in the first reaction. Nested PCR is often more successful in specifically amplifying long DNA fragments than conventional PCR, but it requires more detailed knowledge of the target sequences.
- **Overlap-extension PCR** or *Splicing by overlap extension (SOEing)*: a **genetic engineering** technique that is used to splice together two or more DNA fragments that contain complementary sequences. It is used to join DNA pieces containing genes, regulatory sequences, or mutations; the technique enables creation of specific and long DNA constructs. It can also introduce deletions, insertions or point mutations into a DNA sequence.
- **PAN-AC**: uses isothermal conditions for amplification, and may be used in living cells.
- **quantitative PCR** (qPCR): used to measure the quantity of a target sequence (commonly in real-time). It quantitatively measures starting amounts of DNA, cDNA, or RNA. **quantitative PCR** is commonly used to determine whether a DNA sequence is present in a sample and the number of its copies in the sample. *Quantitative PCR* has a very high degree of precision. Quantitative PCR methods use fluorescent dyes, such as Sybr Green, EvaGreen or **fluorophore**-containing DNA probes, such as **TaqMan**, to measure the amount of amplified product in real time. It is also sometimes abbreviated to **RT-PCR** (*real-time PCR*) but this abbreviation should be used only for **reverse transcription PCR**. qPCR is the appropriate contractions for **quantitative PCR** (real-time PCR).
- **Reverse Transcription PCR (RT-PCR)**: for amplifying DNA from RNA. **Reverse transcriptase** reverse transcribes **RNA** into **cDNA**, which is then amplified by PCR. RT-PCR is widely used in **expression profiling**, to determine the expression of a gene or to identify the sequence of an RNA transcript, including transcription start and termination sites. If the genomic DNA sequence of a gene is known, RT-PCR can be used to map the location of **exons** and **introns** in the gene. The 5' end of a gene (corresponding to the transcription start site) is typically identified by **RACE-PCR** (*Rapid Amplification of cDNA Ends*).
- **RNase H-dependent PCR** (rhPCR): a modification of PCR that utilizes primers with a 3' extension block that can be removed by a thermostable RNase HIII enzyme. This system reduces primer-dimers and allows for multiplexed reactions to be performed with higher numbers of primers
- **Single Specific Primer-PCR** (SSP-PCR): allows the amplification of double-stranded DNA even when the sequence information is available at one end only. This method permits amplification of genes for which only a partial sequence information is available, and allows unidirectional genome walking from known into unknown regions of the chromosome.
- **Solid Phase PCR**: encompasses multiple meanings, including **Polony Amplification** (where PCR colonies are derived in a gel matrix, for example), Bridge PCR (primers are covalently linked to a solid-support surface), conventional Solid Phase PCR (where Asymmetric PCR is applied in the presence of solid support bearing primer with sequence matching one of the aqueous primers) and Enhanced Solid Phase PCR (where conventional Solid Phase PCR can

be improved by employing high T_m and nested solid support primer with optional application of a thermal 'step' to favour solid support priming).

- *Suicide PCR*: typically used in [paleogenetics](#) or other studies where avoiding false positives and ensuring the specificity of the amplified fragment is the highest priority. It was originally described in a study to verify the presence of the microbe [Yersinia pestis](#) in dental samples obtained from 14th Century graves of people supposedly killed by plague during the medieval [Black Death](#) epidemic. The method prescribes the use of any primer combination only once in a PCR (hence the term "suicide"), which should never have been used in any positive control PCR reaction, and the primers should always target a genomic region never amplified before in the lab using this or any other set of primers. This ensures that no contaminating DNA from previous PCR reactions is present in the lab, which could otherwise generate false positives.
- *Thermal asymmetric interlaced PCR (TAIL-PCR)*: for isolation of an unknown sequence flanking a known sequence. Within the known sequence, TAIL-PCR uses a nested pair of primers with differing annealing temperatures; a degenerate primer is used to amplify in the other direction from the unknown sequence.
- *Touchdown PCR (Step-down PCR)*: a variant of PCR that aims to reduce nonspecific background by gradually lowering the annealing temperature as PCR cycling progresses. The annealing temperature at the initial cycles is usually a few degrees (3–5 °C) above the T_m of the primers used, while at the later cycles, it is a few degrees (3–5 °C) below the primer T_m . The higher temperatures give greater specificity for primer binding, and the lower temperatures permit more efficient amplification from the specific products formed during the initial cycles.
- *Universal Fast Walking*: for genome walking and genetic fingerprinting using a more specific 'two-sided' PCR than conventional 'one-sided' approaches (using only one gene-specific primer and one general primer—which can lead to artefactual 'noise') by virtue of a mechanism involving lariat structure formation. Streamlined derivatives of UFW are LaNe RAGE (lariat-dependent nested PCR for rapid amplification of genomic DNA ends), 5'RACE LaNe and 3'RACE LaNe.

History



[Kjell Kleppe](#) and co-workers in the laboratory of [H. Gobind Khorana](#) first described a method of using an enzymatic assay to replicate a short DNA template with primers *in vitro*. However, this early manifestation of the basic PCR principle did not receive much attention at the time and the invention of the polymerase chain reaction in 1983 is generally credited to [Kary Mullis](#).



"Baby Blue", a 1986 prototype machine for doing PCR

When Mullis developed the PCR in 1983, he was working in [Emeryville, California](#) for [Cetus Corporation](#), one of the first [biotechnology](#) companies, where he was responsible for synthesizing short chains of DNA. Mullis has written that he first conceived the idea for PCR while cruising along the [Pacific Coast Highway](#) one night in his car.

He was playing in his mind with a new way of analyzing changes (mutations) in DNA when he realized that he had instead invented a method of amplifying any DNA region through repeated cycles of duplication driven by DNA polymerase. In *Scientific American*, Mullis summarized the procedure: "Beginning with a single molecule of the genetic material DNA, the PCR can generate 100 billion similar molecules in an afternoon. The reaction is easy to execute. It requires no more than a test tube, a few simple reagents, and a source of heat."

DNA fingerprinting was first used for [paternity testing](#) in 1988. Mullis was awarded the [Nobel Prize in Chemistry](#) in 1993 for his invention, seven years after he and his colleagues at Cetus first put his proposal to practice. Mullis's 1985 paper with R. K. Saiki and H. A. Erlich, "Enzymatic Amplification of β -globin Genomic Sequences and Restriction Site Analysis for Diagnosis of Sickle Cell Anemia"—the polymerase chain reaction invention (PCR) – was honored by a Citation for Chemical Breakthrough Award from the Division of History of Chemistry of the American Chemical Society in 2017.

Some controversies have remained about the intellectual and practical contributions of other scientists to Mullis' work, and whether he had been the sole inventor of the PCR principle (see below).

At the core of the PCR method is the use of a suitable [DNA polymerase](#) able to withstand the high temperatures of $>90\text{ }^{\circ}\text{C}$ ($194\text{ }^{\circ}\text{F}$) required for separation of the two DNA strands in the [DNA double helix](#) after each [replication](#) cycle. The DNA polymerases initially employed for [in vitro](#) experiments presaging PCR were unable to withstand these high temperatures. So the early procedures for DNA replication were very inefficient and time-consuming, and required large amounts of DNA polymerase and continuous handling throughout the process.

The discovery in 1976 of [Taq polymerase](#)—a DNA polymerase purified from the [thermophilic bacterium](#), *Thermus aquaticus*, which naturally lives in hot (50 to $80\text{ }^{\circ}\text{C}$ (122 to $176\text{ }^{\circ}\text{F}$)) environments^[13] such as hot springs—paved the way for dramatic improvements of the PCR method. The DNA polymerase isolated from *T. aquaticus* is stable at high temperatures remaining active even after DNA denaturation, thus obviating

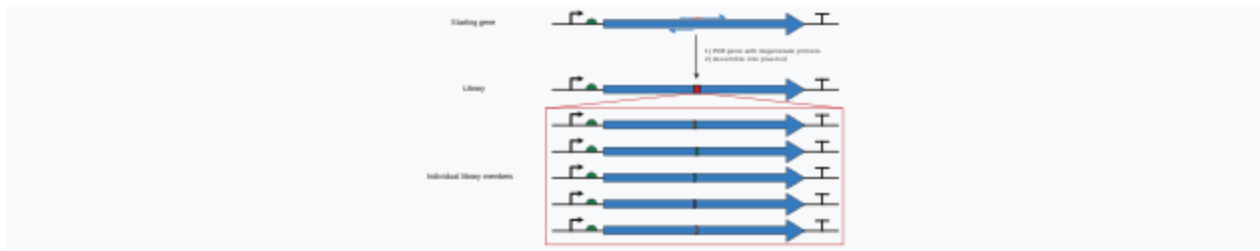
the need to add new DNA polymerase after each cycle. This allowed an automated thermocycler-based process for DNA amplification.

Patent disputes

The PCR technique was patented by [Kary Mullis](#) and assigned to [Cetus Corporation](#), where Mullis worked when he invented the technique in 1983. The *Taq* polymerase enzyme was also covered by patents. There have been several high-profile lawsuits related to the technique, including an unsuccessful lawsuit brought by [DuPont](#). The Swiss pharmaceutical company [Hoffmann-La Roche](#) purchased the rights to the patents in 1992 and currently holds those that are still protected.

A related patent battle over the *Taq* polymerase enzyme is still ongoing in several jurisdictions around the world between Roche and [Promega](#). The legal arguments have extended beyond the lives of the original PCR and *Taq* polymerase patents, which expired on March 28, 2005.

PCR site-directed mutagenesis



Depiction of one common way to clone a site-directed mutagenesis library (i.e., using degenerate oligos). The gene of interest is PCR'd with oligos that contain a region that is perfectly complementary to the template (blue), and one that differs from the template by one or more nucleotides (red). Many such primers containing degeneracy in the non-complementary region are pooled into the same PCR, resulting in many different PCR products with different mutations in that region (individual mutants shown with different colors below).

The limitation of restriction sites in cassette mutagenesis may be overcome using [polymerase chain reaction](#) with [oligonucleotide "primers"](#), such that a larger fragment may be generated, covering two convenient restriction sites.

The exponential amplification in PCR produces a fragment containing the desired mutation in sufficient quantity to be separated from the original, unmutated plasmid by [gel electrophoresis](#), which may then be inserted in the original context using standard recombinant molecular biology techniques.

There are many variations of the same technique. The simplest method places the mutation site toward one of the ends of the fragment whereby one of two oligonucleotides used for generating the fragment contains the mutation.

This involves a single step of PCR, but still has the inherent problem of requiring a suitable restriction site near the mutation site unless a very long primer is used. Other variations, therefore,

employ three or four oligonucleotides, two of which may be non-mutagenic oligonucleotides that cover two convenient restriction sites and generate a fragment that can be digested and ligated into a plasmid, whereas the mutagenic oligonucleotide may be complementary to a location within that fragment well away from any convenient restriction site.

These methods require multiple steps of PCR so that the final fragment to be ligated can contain the desired mutation. The design process for generating a fragment with the desired mutation and relevant restriction sites can be cumbersome. Software tools like SDM-Assist can simplify the process

PCR based methods can be used for the detection of single point mutations (SNPs) present in genetic disorders

The detection of mutations in the genome or transcriptome is of great importance for the diagnosis of genetic disorders, as well as pre-symptomatic testing, conformational diagnosis as well as forensic identity testing. To detect genetic syndromes, two groups of tests are available, molecular and cytogenetic tests.

After the identification and definition of mutations, diagnostic methods or tests can be used to find them using techniques such as allele-specific oligonucleotide hybridization, allele-specific amplification, ligation, primer extension and the artificial introduction of restriction sites.

PCR allows mutation detection, however, PCR itself does not detect the actual mutation. PCR generates an amplicon that is then analyzed by some other method to find possible variations within the amplicon. PCR based methods only detect mutations that have been previously identified by some other techniques if now sequencing step is added.

Real-time PCR is well suited for analysis of single nucleotide polymorphisms (SNPs). Real-time PCR detects SNPs unique to human diseases and is a valuable technique in pharmacogenetics, clinical microbiology and drug development in comparison to methods that use sequencing, single-strand conformation polymorphism, and restriction digestion.

Modern real-time PCR based methods have now become more rapid, sensitive, specific and inexpensive. Real-time PCR monitors the exponential phase of PCR using fluorescently labeled molecules. The PCR amplicon amount present in a reaction tube is directly proportional to the amount of the starting material specific to the PCR primer pair during the exponential phase. Therefore, the amount of emitted fluorescence is directly proportional to the amount of amplicon. This amount is also

proportional to the starting amount of the target sequence allowing measurement of the target copy number.

PCR based methods

Real-Time PCR

Real-time PCR using nonspecific DNA-binding dyes

Real-time PCR using labeled probes

- **Cleavage-based or hydrolysis or dual-labeled probes**
- **Molecular Beacons**
- **FRET probes**
 - Scorpion probes**

UNIT-V

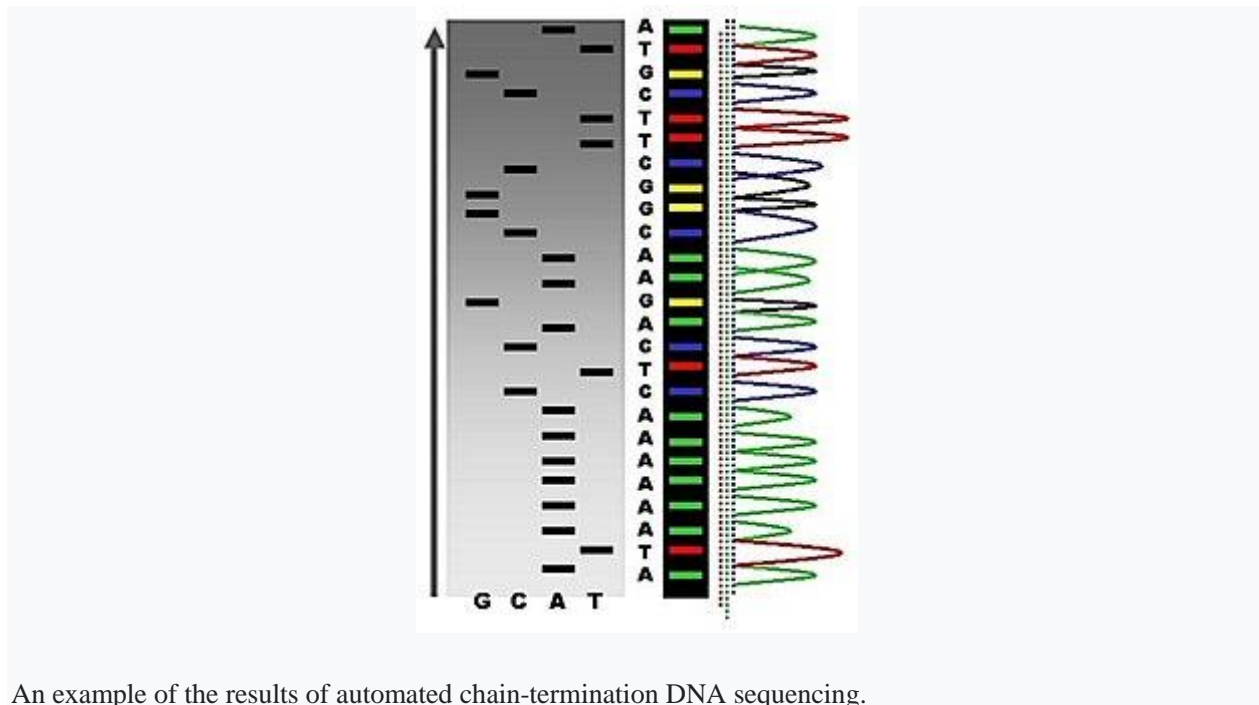
DNA sequencing

DNA sequencing is the process of determining the **nucleic acid sequence** – the order of **nucleotides** in **DNA**. It includes any method or technology that is used to determine the order of the four bases: **adenine**, **guanine**, **cytosine**, and **thymine**. The advent of rapid DNA sequencing methods has greatly accelerated biological and medical research and discovery.

Knowledge of **DNA sequences** has become indispensable for basic biological research, and in numerous applied fields such as **medical diagnosis**, **biotechnology**, **forensic biology**, **virology** and biological **systematics**.

Comparing healthy and mutated DNA sequences can diagnose different diseases including various cancers characterize antibody repertoire and can be used to guide patient treatment. Having a quick way to sequence DNA allows for faster and more individualized medical care to be administered, and for more organisms to be identified and cataloged.

The rapid speed of sequencing attained with modern DNA sequencing technology has been instrumental in the sequencing of complete DNA sequences, or **genomes**, of numerous types and species of life, including the **human genome** and other complete DNA sequences of many animal, plant, and microbial species.



The first DNA sequences were obtained in the early 1970s by academic researchers using laborious methods based on **two-dimensional chromatography**. Following the development

of [fluorescence](#)-based sequencing methods with a [DNA sequencer](#), DNA sequencing has become easier and orders of magnitude faster.



Applications

DNA sequencing may be used to determine the sequence of individual [genes](#), larger genetic regions (i.e. clusters of genes or [operons](#)), full chromosomes, or [entire genomes](#) of any organism. DNA sequencing is also the most efficient way to indirectly sequence [RNA](#) or [proteins](#) (via their [open reading frames](#)). In fact, DNA sequencing has become a key technology in many areas of biology and other sciences such as medicine, [forensics](#), and [anthropology](#).

Molecular biology

Sequencing is used in [molecular biology](#) to study genomes and the proteins they encode. Information obtained using sequencing allows researchers to identify changes in genes, associations with diseases and phenotypes, and identify potential drug targets.

Evolutionary biology

Since DNA is an informative macromolecule in terms of transmission from one generation to another, DNA sequencing is used in [evolutionary biology](#) to study how different organisms are related and how they evolved.

Metagenomics

The field of [metagenomics](#) involves identification of organisms present in a body of water, [sewage](#), dirt, debris filtered from the air, or swab samples from organisms. Knowing which organisms are present in a particular environment is critical to research in [ecology](#), [epidemiology](#), [microbiology](#), and other fields. Sequencing enables researchers to determine which types of microbes may be present in a [microbiome](#), for example.

Virology

As most viruses are too small to be seen by a light microscope, sequencing is one of the main tools in virology to identify and study the virus. Traditional Sanger sequencing and next-generation sequencing are used to sequence viruses in basic and clinical research, as well as for the diagnosis of emerging viral infections, [molecular epidemiology](#) of viral pathogens, and drug-resistance testing. There are more than 2.3 million unique viral sequences in [GenBank](#). Recently, NGS has surpassed traditional Sanger as the most popular approach for generating viral genomes.[†]

Medicine

Medical technicians may sequence genes (or, theoretically, full genomes) from patients to determine if there is risk of genetic diseases. This is a form of [genetic testing](#), though some genetic tests may not involve DNA sequencing. Also, DNA sequencing may be useful for

determining a specific bacteria, to allow for more [precise antibiotics treatments](#), hereby reducing the risk of creating [antimicrobial resistance](#) in bacteria populations.

Forensics

DNA sequencing may be used along with [DNA profiling](#) methods for [forensic identification](#) and [paternity testing](#). DNA testing has evolved tremendously in the last few decades to ultimately link a DNA print to what is under investigation. The DNA patterns in fingerprint, saliva, hair follicles, etc. uniquely separate each living organism from another. Testing DNA is a technique which can detect specific genomes in a DNA strand to produce a unique and individualized pattern.

The four canonical bases

The canonical structure of DNA has four bases: [thymine](#) (T), [adenine](#) (A), [cytosine](#) (C), and [guanine](#) (G). DNA sequencing is the determination of the physical order of these bases in a molecule of DNA. However, there are many other bases that may be present in a molecule. In some viruses (specifically, [bacteriophage](#)), cytosine may be replaced by hydroxy methyl or hydroxy methyl glucose cytosine. In mammalian DNA, variant bases with [methyl](#) groups or phosphosulfate may be found. Depending on the sequencing technique, a particular modification, e.g., the 5mC ([5 methyl cytosine](#)) common in humans, may or may not be detected.

RNA sequencing

[RNA sequencing](#) was one of the earliest forms of nucleotide sequencing. The major landmark of RNA sequencing is the sequence of the first complete gene and the complete genome of [Bacteriophage MS2](#), identified and published by [Walter Fiers](#) and his coworkers at the [University of Ghent \(Ghent, Belgium\)](#), in 1972 and 1976. Traditional RNA sequencing methods require the creation of a [cDNA](#) molecule which must be sequenced.

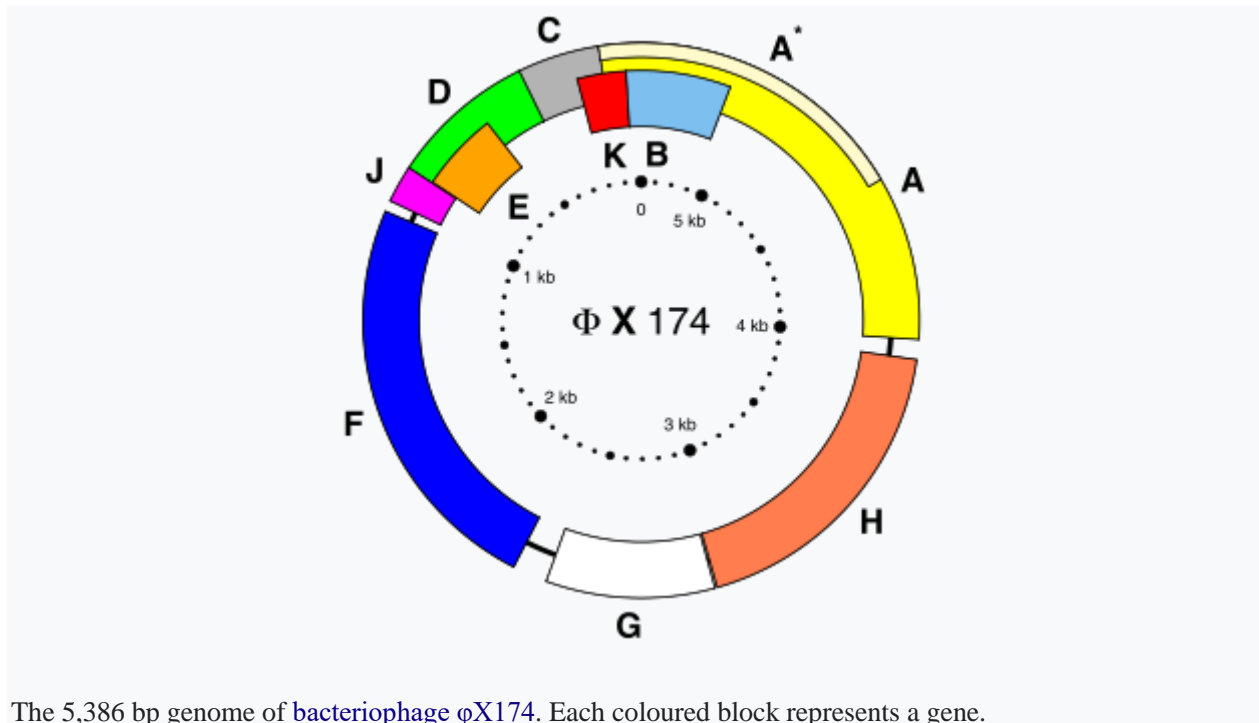
Early DNA sequencing methods

The first method for determining DNA sequences involved a location-specific primer extension strategy established by [Ray Wu](#) at [Cornell University](#) in 1970. DNA polymerase catalysis and specific nucleotide labeling, both of which figure prominently in current sequencing schemes, were used to sequence the cohesive ends of lambda phage DNA.¹ Between 1970 and 1973, Wu, R Padmanabhan and colleagues demonstrated that this method can be employed to determine any DNA sequence using synthetic location-specific primers.

[Frederick Sanger](#) then adopted this primer-extension strategy to develop more rapid DNA sequencing methods at the [MRC Centre, Cambridge](#), UK and published a method for "DNA sequencing with chain-terminating inhibitors" in 1977. [Walter Gilbert](#) and [Allan Maxam](#) at [Harvard](#) also developed sequencing methods, including one for "DNA sequencing by chemical degradation".

In 1973, Gilbert and Maxam reported the sequence of 24 basepairs using a method known as wandering-spot analysis. Advancements in sequencing were aided by the concurrent development of [recombinant DNA](#) technology, allowing DNA samples to be isolated from sources other than viruses.

Sequencing of full genomes



The first full DNA genome to be sequenced was that of bacteriophage ϕ X174 in 1977. Medical Research Council scientists deciphered the complete DNA sequence of the Epstein-Barr virus in 1984, finding it contained 172,282 nucleotides. Completion of the sequence marked a significant turning point in DNA sequencing because it was achieved with no prior genetic profile knowledge of the virus.

A non-radioactive method for transferring the DNA molecules of sequencing reaction mixtures onto an immobilizing matrix during electrophoresis was developed by Pohl and co-workers in the early 1980s. Followed by the commercialization of the DNA sequencer "Direct-Blotting-Electrophoresis-System GATC 1500" by GATC Biotech, which was intensively used in the framework of the EU genome-sequencing programme, the complete DNA sequence of the yeast *Saccharomyces cerevisiae* chromosome II. Leroy E. Hood's laboratory at the California Institute of Technology announced the first semi-automated DNA sequencing machine in 1986.

This was followed by Applied Biosystems' marketing of the first fully automated sequencing machine, the ABI 370, in 1987 and by Dupont's Genesis 2000 which used a novel fluorescent labeling technique enabling all four dideoxynucleotides to be identified in a single lane. By 1990, the U.S. National Institutes of Health (NIH) had begun large-scale sequencing trials on *Mycoplasma capricolum*, *Escherichia coli*, *Caenorhabditis elegans*, and *Saccharomyces cerevisiae* at a cost of US\$0.75 per base.

Meanwhile, sequencing of human cDNA sequences called expressed sequence tags began in Craig Venter's lab, an attempt to capture the coding fraction of the human genome. In 1995, Venter, Hamilton Smith, and colleagues at The Institute for Genomic Research (TIGR) published the first complete genome of a free-living organism, the bacterium *Haemophilus influenzae*. The circular chromosome contains 1,830,137 bases and its publication in the journal Science marked

the first published use of whole-genome shotgun sequencing, eliminating the need for initial mapping efforts.

By 2001, shotgun sequencing methods had been used to produce a draft sequence of the human genome.

High-throughput sequencing (HTS) methods

Several new methods for DNA sequencing were developed in the mid to late 1990s and were implemented in commercial **DNA sequencers** by the year 2000. Together these were called the "next-generation" or "second-generation" sequencing (NGS) methods, in order to distinguish them from the aforementioned earlier methods, like Sanger Sequencing.

In contrast to the first generation of sequencing, NGS technology is typically characterized by being highly scalable, allowing the entire genome to be sequenced at once. Usually, this is accomplished by fragmenting the genome into small pieces, randomly sampling for a fragment, and sequencing it using one of a variety of technologies, such as those described below.

An entire genome is possible because multiple fragments are sequenced at once (giving it the name "massively parallel" sequencing) in an automated process.

NGS technology has tremendously empowered researchers to look for insights into health, anthropologists to investigate human origins, and is catalyzing the "**Personalized Medicine**" movement. However, it has also opened the door to more room for error. There are many software tools to carry out the computational analysis of NGS data, each with its own algorithm. Even the parameters within one software package can change the outcome of the analysis.

In addition, the large quantities of data produced by DNA sequencing have also required development of new methods and programs for sequence analysis. Several efforts to develop standards in the NGS field have been attempted to address these challenges, most of which have been small-scale efforts arising from individual labs. Most recently, a large, organized, FDA-funded effort has culminated in the **BioCompute** standard.

On 26 October 1990, **Roger Tsien**, Pepi Ross, Margaret Fahnestock and Allan J Johnston filed a patent describing stepwise ("base-by-base") sequencing with removable 3' blockers on DNA arrays (blots and single DNA molecules). In 1996, **Pål Nyrén** and his student **Mostafa Ronaghi** at the Royal Institute of Technology in **Stockholm** published their method of **pyrosequencing**.

On 1 April 1997, Pascal Mayer and Laurent Farinelli submitted patents to the World Intellectual Property Organization describing DNA colony sequencing. The DNA sample preparation and random surface-**polymerase chain reaction** (PCR) arraying methods described in this patent, coupled to Roger Tsien et al.'s "base-by-base" sequencing method, is now implemented in **Illumina**'s Hi-Seq genome sequencers.

In 1998, Phil Green and Brent Ewing of the University of Washington described their **phred quality score** for sequencer data analysis, a landmark analysis technique that gained widespread adoption, and which is still the most common metric for assessing the accuracy of a sequencing platform.^[51]

Lynx Therapeutics published and marketed [massively parallel signature sequencing](#) (MPSS), in 2000. This method incorporated a parallelized, adapter/ligation-mediated, bead-based sequencing technology and served as the first commercially available "next-generation" sequencing method, though no [DNA sequencers](#) were sold to independent laboratories.^[52]

Basic methods

Maxam-Gilbert sequencing

[Allan Maxam](#) and [Walter Gilbert](#) published a DNA sequencing method in 1977 based on chemical modification of DNA and subsequent cleavage at specific bases. Also known as chemical sequencing, this method allowed purified samples of double-stranded DNA to be used without further cloning. This method's use of radioactive labeling and its technical complexity discouraged extensive use after refinements in the Sanger methods had been made.

Maxam-Gilbert sequencing requires radioactive labeling at one 5' end of the DNA and purification of the DNA fragment to be sequenced. Chemical treatment then generates breaks at a small proportion of one or two of the four nucleotide bases in each of four reactions (G, A+G, C, C+T). The concentration of the modifying chemicals is controlled to introduce on average one modification per DNA molecule. Thus a series of labeled fragments is generated, from the radiolabeled end to the first "cut" site in each molecule.

The fragments in the four reactions are electrophoresed side by side in denaturing acrylamide gels for size separation. To visualize the fragments, the gel is exposed to X-ray film for autoradiography, yielding a series of dark bands each corresponding to a radiolabeled DNA fragment, from which the sequence may be inferred.

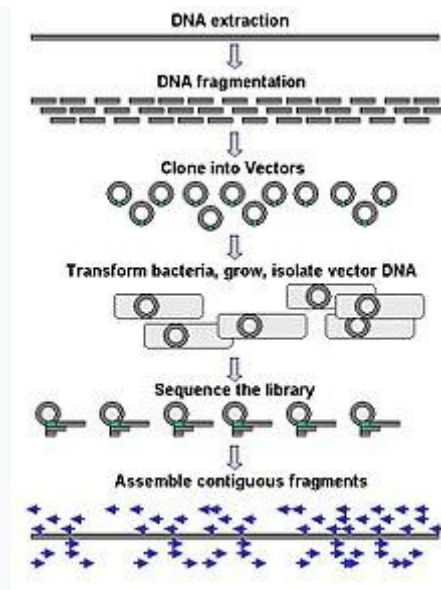
Chain-termination methods

The [chain-termination method](#) developed by [Frederick Sanger](#) and coworkers in 1977 soon became the method of choice, owing to its relative ease and reliability. When invented, the chain-terminator method used fewer toxic chemicals and lower amounts of radioactivity than the Maxam and Gilbert method. Because of its comparative ease, the Sanger method was soon automated and was the method used in the first generation of [DNA sequencers](#).

Sanger sequencing is the method which prevailed from the 1980s until the mid-2000s. Over that period, great advances were made in the technique, such as fluorescent labelling, capillary electrophoresis, and general automation. These developments allowed much more efficient sequencing, leading to lower costs.

The Sanger method, in mass production form, is the technology which produced the [first human genome](#) in 2001, ushering in the age of [genomics](#). However, later in the decade, radically different approaches reached the market, bringing the cost per genome down from \$100 million in 2001 to \$10,000 in 2011.

Large-scale sequencing and *de novo* sequencing



Genomic DNA is fragmented into random pieces and cloned as a bacterial library. DNA from individual bacterial clones is sequenced and the sequence is assembled by using overlapping DNA regions. (click to expand)

Large-scale sequencing often aims at sequencing very long DNA pieces, such as whole **chromosomes**, although large-scale sequencing can also be used to generate very large numbers of short sequences, such as found in **phage display**.

For longer targets such as chromosomes, common approaches consist of cutting (with **restriction enzymes**) or shearing (with mechanical forces) large DNA fragments into shorter DNA fragments. The fragmented DNA may then be **cloned** into a **DNA vector** and amplified in a bacterial host such as *Escherichia coli*.

Short DNA fragments purified from individual bacterial colonies are individually sequenced and **assembled electronically** into one long, contiguous sequence. Studies have shown that adding a size selection step to collect DNA fragments of uniform size can improve sequencing efficiency and accuracy of the genome assembly. In these studies, automated sizing has proven to be more reproducible and precise than manual gel sizing.

The term "*de novo* sequencing" specifically refers to methods used to determine the sequence of DNA with no previously known sequence. *De novo* translates from Latin as "from the beginning". Gaps in the assembled sequence may be filled by **primer walking**. The different strategies have different tradeoffs in speed and accuracy; **shotgun methods** are often used for sequencing large genomes, but its assembly is complex and difficult, particularly with **sequence repeats** often causing gaps in genome assembly.

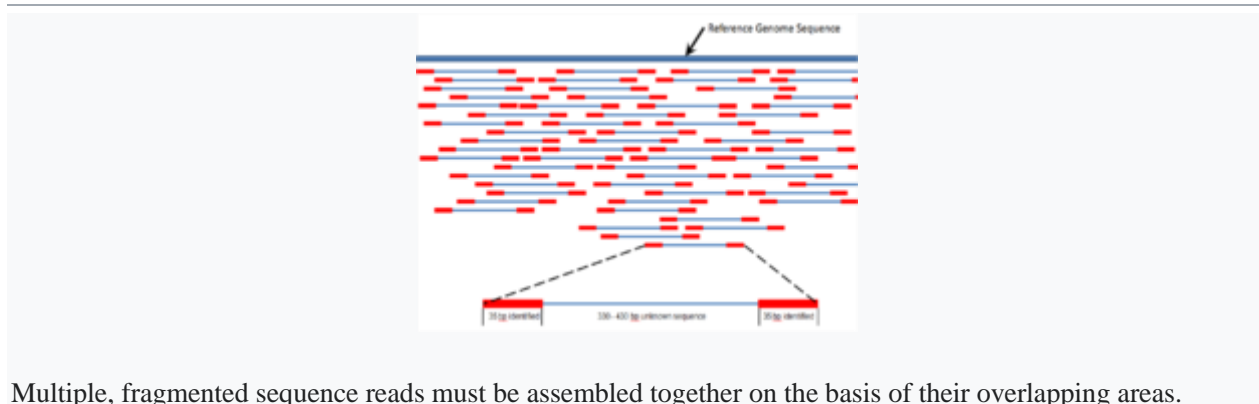
Most sequencing approaches use an *in vitro* cloning step to amplify individual DNA molecules, because their molecular detection methods are not sensitive enough for single molecule sequencing. Emulsion PCR isolates individual DNA molecules along with primer-coated beads in aqueous droplets within an oil phase.

A [polymerase chain reaction](#) (PCR) then coats each bead with clonal copies of the DNA molecule followed by immobilization for later sequencing. Emulsion PCR is used in the methods developed by Marguilis et al. known as "[polony sequencing](#)") and [SOLiD sequencing](#),

Shotgun sequencing

Shotgun sequencing is a sequencing method designed for analysis of DNA sequences longer than 1000 base pairs, up to and including entire chromosomes. This method requires the target DNA to be broken into random fragments. After sequencing individual fragments, the sequences can be reassembled on the basis of their overlapping regions.

High-throughput methods



Multiple, fragmented sequence reads must be assembled together on the basis of their overlapping areas.

High-throughput sequencing, which includes next-generation "short-read" and third-generation "long-read" sequencing methods,^[m 1] applies to exome sequencing, genome sequencing, genome resequencing, [transcriptome](#) profiling ([RNA-Seq](#)), DNA-protein interactions ([ChIP-sequencing](#)), and [epigenome](#) characterization.

Resequencing is necessary, because the genome of a single individual of a species will not indicate all of the genome variations among other individuals of the same species.

The high demand for low-cost sequencing has driven the development of high-throughput sequencing technologies that [parallelize](#) the sequencing process, producing thousands or millions of sequences concurrently.

High-throughput sequencing technologies are intended to lower the cost of DNA sequencing beyond what is possible with standard dye-terminator methods. In ultra-high-throughput sequencing as many as 500,000 sequencing-by-synthesis operations may be run in parallel. Such technologies led to the ability to sequence an entire human genome in as little as one day. As of 2019, corporate leaders in the development of high-throughput sequencing

Single molecule real time (SMRT) sequencing

SMRT sequencing is based on the sequencing by synthesis approach. The DNA is synthesized in zero-mode wave-guides (ZMWs) – small well-like containers with the capturing tools located at the bottom of the well. The sequencing is performed with use of unmodified polymerase (attached to the ZMW bottom) and fluorescently labelled nucleotides flowing freely in the solution.

The wells are constructed in a way that only the fluorescence occurring by the bottom of the well is detected. The fluorescent label is detached from the nucleotide upon its incorporation into the DNA strand, leaving an unmodified DNA strand. According to [Pacific Biosciences](#) (PacBio), the SMRT technology developer, this methodology allows detection of nucleotide modifications (such as cytosine methylation). This happens through the observation of polymerase kinetics.

This approach allows reads of 20,000 nucleotides or more, with average read lengths of 5 kilobases. In 2015, Pacific Biosciences announced the launch of a new sequencing instrument called the Sequel System, with 1 million ZMWs compared to 150,000 ZMWs in the PacBio RS II instrument. SMRT sequencing is referred to as "[third-generation](#)" or "long-read" sequencing.

Nanopore DNA sequencing

The DNA passing through the nanopore changes its ion current. This change is dependent on the shape, size and length of the DNA sequence. Each type of the nucleotide blocks the ion flow through the pore for a different period of time.

The method does not require modified nucleotides and is performed in real time. Nanopore sequencing is referred to as "[third-generation](#)" or "long-read" sequencing, along with SMRT sequencing.

Early industrial research into this method was based on a technique called 'exonuclease sequencing', where the readout of electrical signals occurred as nucleotides passed by [alpha\(\$\alpha\$ \)-hemolysin](#) pores covalently bound with [cyclodextrin](#).

However the subsequent commercial method, 'strand sequencing', sequenced DNA bases in an intact strand.

Two main areas of nanopore sequencing in development are solid state nanopore sequencing, and protein based nanopore sequencing. Protein nanopore sequencing utilizes membrane protein complexes such as α -hemolysin, MspA (*Mycobacterium smegmatis* Porin A) or CsgG, which show great promise given their ability to distinguish between individual and groups of nucleotides.

In contrast, solid-state nanopore sequencing utilizes synthetic materials such as silicon nitride and aluminum oxide and it is preferred for its superior mechanical ability and thermal and chemical stability.

The fabrication method is essential for this type of sequencing given that the nanopore array can contain hundreds of pores with diameters smaller than eight nanometers.

The concept originated from the idea that single stranded DNA or RNA molecules can be electrophoretically driven in a strict linear sequence through a biological pore that can be less than eight nanometers, and can be detected given that the molecules release an ionic current while moving through the pore.

The pore contains a detection region capable of recognizing different bases, with each base generating various time specific signals corresponding to the sequence of bases as they cross the pore which are then evaluated.

Precise control over the DNA transport through the pore is crucial for success. Various enzymes such as exonucleases and polymerases have been used to moderate this process by positioning them near the pore's entrance.

Short-read sequencing methods

Massively parallel signature sequencing (MPSS)

The first of the high-throughput sequencing technologies, [massively parallel signature sequencing](#) (or MPSS), was developed in the 1990s at Lynx Therapeutics, a company founded in 1992 by [Sydney Brenner](#) and [Sam Eletr](#). MPSS was a bead-based method that used a complex approach of adapter ligation followed by adapter decoding, reading the sequence in increments of four nucleotides.

This method made it susceptible to sequence-specific bias or loss of specific sequences. Because the technology was so complex, MPSS was only performed 'in-house' by Lynx Therapeutics and no DNA sequencing machines were sold to independent laboratories. Lynx Therapeutics merged with Solexa (later acquired by [Illumina](#)) in 2004, leading to the development of sequencing-by-synthesis, a simpler approach acquired from [Manteia Predictive Medicine](#), which rendered MPSS obsolete.

However, the essential properties of the MPSS output were typical of later high-throughput data types, including hundreds of thousands of short DNA sequences. In the case of MPSS, these were typically used for sequencing [cDNA](#) for measurements of [gene expression](#) levels.

Polony sequencing

The [polony sequencing](#) method, developed in the laboratory of [George M. Church](#) at Harvard, was among the first high-throughput sequencing systems and was used to sequence a full *E. coli* genome in 2005.

It combined an in vitro paired-tag library with emulsion PCR, an automated microscope, and ligation-based sequencing chemistry to sequence an *E. coli* genome at an accuracy of >99.9999% and a cost approximately 1/9 that of Sanger sequencing.

PYROSEQUENCING

A parallelized version of [pyrosequencing](#) was developed by [454 Life Sciences](#), which has since been acquired by [Roche Diagnostics](#). The method amplifies DNA inside water droplets in an oil solution (emulsion PCR), with each droplet containing a single DNA template attached to a single primer-coated bead that then forms a clonal colony.

The sequencing machine contains many [picoliter](#)-volume wells each containing a single bead and sequencing enzymes. Pyrosequencing uses [luciferase](#) to generate light for detection of the individual nucleotides added to the nascent DNA, and the combined data are used to generate sequence [reads](#).

This technology provides intermediate read length and price per base compared to Sanger sequencing on one end and Solexa and SOLiD on the other.

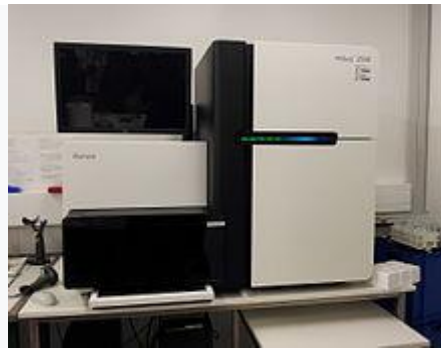
Illumina (Solexa) sequencing

Solexa, now part of **Illumina**, was founded by **Shankar Balasubramanian** and **David Klenerman** in 1998, and developed a sequencing method based on reversible dye-terminators technology, and engineered polymerases.

The reversible terminated chemistry concept was invented by Bruno Canard and Simon Sarfati at the Pasteur Institute in Paris.

It was developed internally at Solexa by those named on the relevant patents. In 2004, Solexa acquired the company **Manteia Predictive Medicine** in order to gain a massively parallel sequencing technology invented in 1997 by Pascal Mayer and Laurent Farinelli.

It is based on "DNA clusters" or "DNA colonies", which involves the clonal amplification of DNA on a surface. The cluster technology was co-acquired with Lynx Therapeutics of California. Solexa Ltd. later merged with Lynx to form Solexa Inc.



An Illumina HiSeq 2500 sequencer



Illumina NovaSeq 6000 flow cell

In this method, DNA molecules and primers are first attached on a slide or flow cell and amplified with **polymerase** so that local clonal DNA colonies, later coined "DNA clusters", are formed. To determine the sequence, four types of reversible terminator bases (RT-bases) are added and non-incorporated nucleotides are washed away.

A camera takes images of the **fluorescently labeled** nucleotides. Then the dye, along with the terminal 3' blocker, is chemically removed from the DNA, allowing for the next cycle to begin. Unlike pyrosequencing, the DNA chains are extended one nucleotide at a time and image acquisition can be performed at a delayed moment, allowing for very large arrays of DNA colonies to be captured by sequential images taken from a single camera.



An Illumina MiSeq sequencer

Decoupling the enzymatic reaction and the image capture allows for optimal throughput and theoretically unlimited sequencing capacity. With an optimal configuration, the ultimately reachable instrument throughput is thus dictated solely by the analog-to-digital conversion rate of the camera, multiplied by the number of cameras and divided by the number of pixels per DNA colony required for visualizing them optimally (approximately 10 pixels/colony).

In 2012, with cameras operating at more than 10 MHz A/D conversion rates and available optics, fluidics and enzymatics, throughput can be multiples of 1 million nucleotides/second, corresponding roughly to 1 human genome equivalent at 1x **coverage** per hour per instrument, and 1 human genome re-sequenced (at approx. 30x) per day per instrument (equipped with a single camera).

Combinatorial probe anchor synthesis (cPAS)

This method is an upgraded modification to combinatorial probe anchor ligation technology (cPAL) described by **Complete Genomics** which has since become part of Chinese genomics company **BGI** in 2013.

The two companies have refined the technology to allow for longer read lengths, reaction time reductions and faster time to results. In addition, data are now generated as contiguous full-length reads in the standard FASTQ file format and can be used as-is in most short-read-based bioinformatics analysis pipelines.

The two technologies that form the basis for this high-throughput sequencing technology are **DNA nanoballs** (DNB) and patterned arrays for nanoball attachment to a solid surface. DNA nanoballs are simply formed by denaturing double stranded, adapter ligated libraries and ligating the forward strand only to a splint oligonucleotide to form a ssDNA circle.

Faithful copies of the circles containing the DNA insert are produced utilizing Rolling Circle Amplification that generates approximately 300–500 copies. The long strand of ssDNA folds upon itself to produce a three-dimensional nanoball structure that is approximately 220 nm in diameter. Making DNBs replaces the need to generate PCR copies of the library on the flow cell and as such can remove large proportions of duplicate reads, adapter-adapter ligations and PCR induced errors.



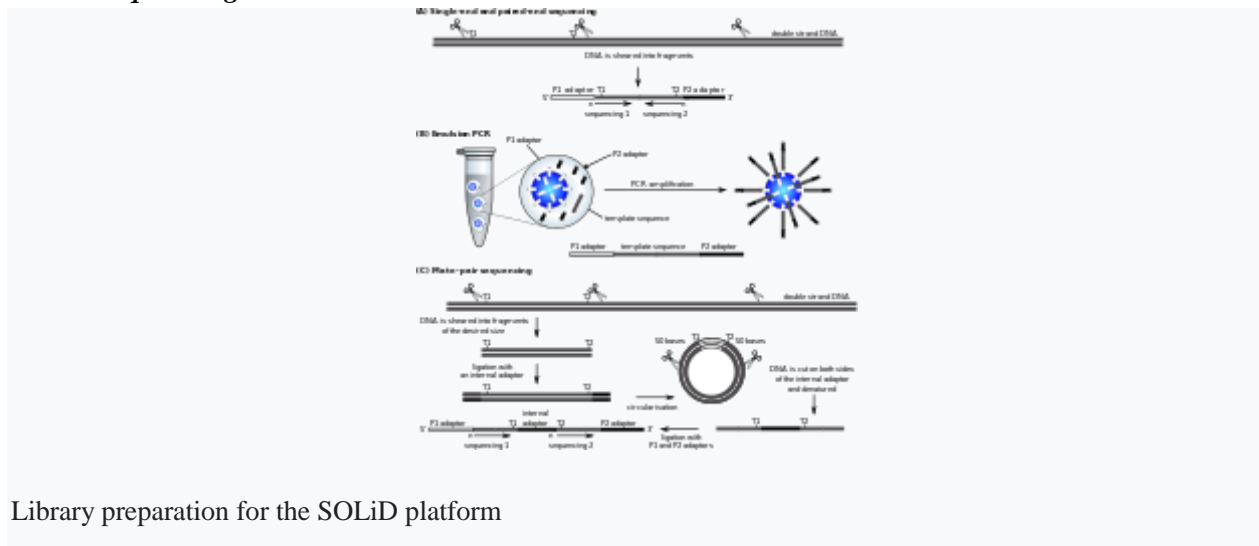
A BGI MGISEQ-2000RS sequencer

The patterned array of positively charged spots is fabricated through photolithography and etching techniques followed by chemical modification to generate a sequencing flow cell. Each spot on the flow cell is approximately 250 nm in diameter, are separated by 700 nm (centre to centre) and allows easy attachment of a single negatively charged DNB to the flow cell and thus reducing under or over-clustering on the flow cell.

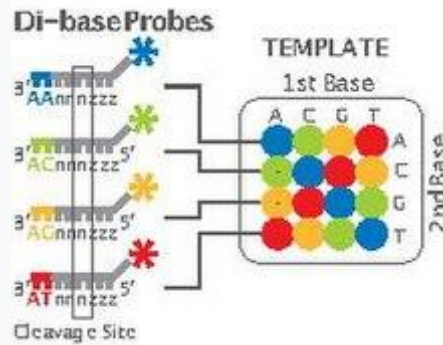
Sequencing is then performed by addition of an oligonucleotide probe that attaches in combination to specific sites within the DNB. The probe acts as an anchor that then allows one of four single reversibly inactivated, labelled nucleotides to bind after flowing across the flow cell. Unbound nucleotides are washed away before laser excitation of the attached labels then emit fluorescence and signal is captured by cameras that is converted to a digital output for base calling.

The attached base has its terminator and label chemically cleaved at completion of the cycle. The cycle is repeated with another flow of free, labelled nucleotides across the flow cell to allow the next nucleotide to bind and have its signal captured. This process is completed a number of times (usually 50 to 300 times) to determine the sequence of the inserted piece of DNA at a rate of approximately 40 million nucleotides per second as of 2018.

SOLiD sequencing



Library preparation for the SOLiD platform



Two-base encoding scheme. In two-base encoding, each unique pair of bases on the 3' end of the probe is assigned one out of four possible colors. For example, "AA" is assigned to blue, "AC" is assigned to green, and so on for all 16 unique pairs. During sequencing, each base in the template is sequenced twice, and the resulting data are decoded according to this scheme.

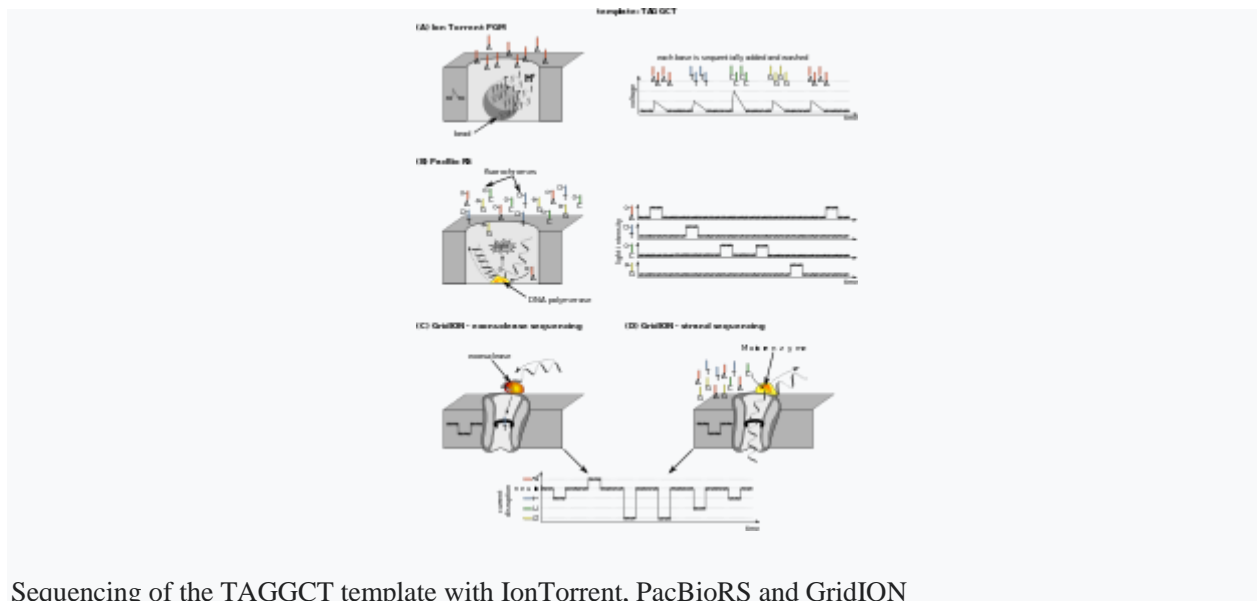
Applied Biosystems' (now a Life Technologies brand) SOLiD technology employs **sequencing by ligation**. Here, a pool of all possible oligonucleotides of a fixed length are labeled according to the sequenced position. Oligonucleotides are annealed and ligated; the preferential ligation by **DNA ligase** for matching sequences results in a signal informative of the nucleotide at that position.

Each base in the template is sequenced twice, and the resulting data are decoded according to the **2 base encoding** scheme used in this method. Before sequencing, the DNA is amplified by emulsion PCR. The resulting beads, each containing single copies of the same DNA molecule, are deposited on a glass slide. The result is sequences of quantities and lengths comparable to Illumina sequencing. This **sequencing by ligation** method has been reported to have some issue sequencing palindromic sequences.

Ion Torrent semiconductor sequencing

Ion Torrent Systems Inc. (now owned by Life Technologies) developed a system based on using standard sequencing chemistry, but with a novel, semiconductor-based detection system. This method of sequencing is based on the detection of **hydrogen ions** that are released during the **polymerisation** of DNA, as opposed to the optical methods used in other sequencing systems.

A microwell containing a template DNA strand to be sequenced is flooded with a single type of **nucleotide**. If the introduced nucleotide is **complementary** to the leading template nucleotide it 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 **homopolymer** repeats are present in the template sequence, multiple nucleotides will be incorporated in a single cycle. This leads to a corresponding number of released hydrogens and a proportionally higher electronic signal.



Sequencing of the TAGGCT template with IonTorrent, PacBioRS and GridION

DNA nanoball sequencing

DNA nanoball sequencing is a type of high throughput sequencing technology used to determine the entire **genomic sequence** of an organism. The company **Complete Genomics** uses this technology to sequence samples submitted by independent researchers. The method uses **rolling circle replication** to amplify small fragments of genomic DNA into DNA nanoballs. Unchained sequencing by ligation is then used to determine the nucleotide sequence.

This method of DNA sequencing allows large numbers of DNA nanoballs to be sequenced per run and at low **reagent** costs compared to other high-throughput sequencing platforms.

However, only short sequences of DNA are determined from each DNA nanoball which makes mapping the short reads to a **reference genome** difficult. This technology has been used for multiple genome sequencing projects and is scheduled to be used for more.

Heliscope single molecule sequencing

Heliscope sequencing is a method of single-molecule sequencing developed by **Helicos Biosciences**. It uses DNA fragments with added poly-A tail adapters which are attached to the flow cell surface.

The next steps involve extension-based sequencing with cyclic washes of the flow cell with fluorescently labeled nucleotides (one nucleotide type at a time, as with the Sanger method). The reads are performed by the Heliscope sequencer.

The reads are short, averaging 35 bp. What made this technology especially novel was that it was the first of its class to sequence non-amplified DNA, thus preventing any read errors associated with amplification steps. In 2009 a human genome was sequenced using the Heliscope, however in 2012 the company went bankrupt.

Microfluidic Systems

There are two main microfluidic systems that are used to sequence DNA; **droplet based microfluidics** and **digital microfluidics**. Microfluidic devices solve many of the current limitations of current sequencing arrays.

Abate et al. studied the use of droplet-based microfluidic devices for DNA sequencing

These devices have the ability to form and process picoliter sized droplets at the rate of thousands per second. The devices were created from **polydimethylsiloxane (PDMS)** and used Förster resonance energy transfer, **FRET assays** to read the sequences of DNA encompassed in the droplets. Each position on the array tested for a specific 15 base sequence.

Fair et al. used digital microfluidic devices to study DNA **pyrosequencing**. Significant advantages include the portability of the device, reagent volume, speed of analysis, mass manufacturing abilities, and high throughput.

This study provided a proof of concept showing that digital devices can be used for pyrosequencing; the study included using synthesis, which involves the extension of the enzymes and addition of labeled nucleotides.

Boles et al. Boles et al. also studied pyrosequencing on digital microfluidic devices. They used an electro-wetting device to create, mix, and split droplets. The sequencing uses a three-enzyme protocol and DNA templates anchored with magnetic beads.

The device was tested using two protocols and resulted in 100% accuracy based on raw pyrogram levels. The advantages of these digital microfluidic devices include size, cost, and achievable levels of functional integration.

DNA sequencing research, using microfluidics, also has the ability to be applied to the **sequencing of RNA**, using similar droplet microfluidic techniques, such as the method, inDrops. This shows that many of these DNA sequencing techniques will be able to be applied further and be used to understand more about genomes and transcriptomes.

Methods in development

sequencing methods currently under development include reading the sequence as a DNA strand transits through **nanopores** (a method that is now commercial but subsequent generations such as solid-state nanopores are still in development), and microscopy-based techniques, such as **atomic force microscopy** or **transmission electron microscopy** that are used to identify the positions of individual nucleotides within long DNA fragments (>5,000 bp) by nucleotide labeling with heavier elements (e.g., halogens) for visual detection and recording.

Third generation technologies aim to increase throughput and decrease the time to result and cost by eliminating the need for excessive reagents and harnessing the processivity of DNA polymerase.^[121]

Tunnelling currents DNA sequencing

Another approach uses measurements of the electrical tunnelling currents across single-strand DNA as it moves through a channel. Depending on its electronic structure, each base affects the tunnelling current differently, allowing differentiation between different bases.

The use of tunnelling currents has the potential to sequence orders of magnitude faster than ionic current methods and the sequencing of several DNA oligomers and micro-RNA has already been achieved.

Sequencing by hybridization

Sequencing by hybridization is a non-enzymatic method that uses a [DNA microarray](#). A single pool of DNA whose sequence is to be determined is fluorescently labeled and hybridized to an array containing known sequences. Strong hybridization signals from a given spot on the array identifies its sequence in the DNA being sequenced.

This method of sequencing utilizes binding characteristics of a library of short single stranded DNA molecules (oligonucleotides), also called DNA probes, to reconstruct a target DNA sequence. Non-specific hybrids are removed by washing and the target DNA is eluted. Hybrids are re-arranged such that the DNA sequence can be reconstructed.

The benefit of this sequencing type is its ability to capture a large number of targets with a homogenous coverage. A large number of chemicals and starting DNA is usually required. However, with the advent of solution-based hybridization, much less equipment and chemicals are necessary.

Sequencing with mass spectrometry

[Mass spectrometry](#) may be used to determine DNA sequences. Matrix-assisted laser desorption ionization time-of-flight mass spectrometry, or [MALDI-TOF MS](#), has specifically been investigated as an alternative method to gel electrophoresis for visualizing DNA fragments. With this method, DNA fragments generated by chain-termination sequencing reactions are compared by mass rather than by size.

The mass of each nucleotide is different from the others and this difference is detectable by mass spectrometry. Single-nucleotide mutations in a fragment can be more easily detected with MS than by gel electrophoresis alone. MALDI-TOF MS can more easily detect differences between RNA fragments, so researchers may indirectly sequence DNA with MS-based methods by converting it to RNA first.

The higher resolution of DNA fragments permitted by MS-based methods is of special interest to researchers in forensic science, as they may wish to find [single-nucleotide polymorphisms](#) in human DNA samples to identify individuals. These samples may be highly degraded so forensic researchers often prefer [mitochondrial DNA](#) for its higher stability and applications for lineage studies.

MS-based sequencing methods have been used to compare the sequences of human mitochondrial DNA from samples in a [Federal Bureau of Investigation](#) database and from bones found in mass graves of World War I soldiers.

Early chain-termination and TOF MS methods demonstrated read lengths of up to 100 base pairs. Researchers have been unable to exceed this average read size; like chain-termination sequencing alone, MS-based DNA sequencing may not be suitable for large *de novo* sequencing projects. Even so, a recent study did use the short sequence reads and mass spectroscopy to compare single-nucleotide polymorphisms in pathogenic *Streptococcus* strains.

Microfluidic Sanger sequencing

In microfluidic Sanger sequencing the entire thermocycling amplification of DNA fragments as well as their separation by electrophoresis is done on a single glass wafer (approximately 10 cm in diameter) thus reducing the reagent usage as well as cost.

In some instances researchers have shown that they can increase the throughput of conventional sequencing through the use of microchips. Research will still need to be done in order to make this use of technology effective.

Microscopy-based techniques

This approach directly visualizes the sequence of DNA molecules using electron microscopy. The first identification of DNA base pairs within intact DNA molecules by enzymatically incorporating modified bases, which contain atoms of increased atomic number, direct visualization and identification of individually labeled bases within a synthetic 3,272 base-pair DNA molecule and a 7,249 base-pair viral genome has been demonstrated.

RNAP sequencing

This method is based on use of [RNA polymerase](#) (RNAP), which is attached to a [polystyrene](#) bead. One end of DNA to be sequenced is attached to another bead, with both beads being placed in optical traps. RNAP motion during transcription brings the beads in closer and their relative distance changes, which can then be recorded at a single nucleotide resolution. The sequence is deduced based on the four readouts with lowered concentrations of each of the four nucleotide types, similarly to the Sanger method. A comparison is made between regions and sequence information is deduced by comparing the known sequence regions to the unknown sequence regions.

In vitro virus high-throughput sequencing

A method has been developed to analyze full sets of [protein interactions](#) using a combination of 454 pyrosequencing and an *in vitro* virus [mRNA display](#) method. Specifically, this method covalently links proteins of interest to the mRNAs encoding them, then detects the mRNA pieces using reverse transcription [PCRs](#). The mRNA may then be amplified and sequenced. The combined method was titled IVV-HITSeq and can be performed under cell-free conditions, though its results may not be representative of *in vivo* conditions.

Sample preparation

The success of any DNA sequencing protocol relies upon the DNA or RNA sample extraction and preparation from the biological material of interest.

- A successful DNA extraction will yield a DNA sample with long, non-degraded strands.
- A successful RNA extraction will yield a RNA sample that should be converted to complementary DNA (cDNA) using reverse transcriptase—a DNA polymerase that synthesizes a complementary DNA based on existing strands of RNA in a PCR-like manner.
- Complementary DNA can then be processed the same way as genomic DNA.

According to the sequencing technology to be used, the samples resulting from either the DNA or the RNA extraction require further preparation. For Sanger sequencing, either cloning procedures or

PCR are required prior to sequencing. In the case of next-generation sequencing methods, library preparation is required before processing. Assessing the quality and quantity of nucleic acids both after extraction and after library preparation identifies degraded, fragmented, and low-purity samples and yields high-quality sequencing data.

The high-throughput nature of current DNA/RNA sequencing technologies has posed a challenge for sample preparation method to scale-up.

Gene silencing

Gene silencing is the **regulation of gene expression** in a cell to prevent the expression of a certain **gene**. Gene silencing can occur during either **transcription** or **translation** and is often used in research. In particular, methods used to silence genes are being increasingly used to produce **therapeutics** to combat cancer and other diseases, such as **infectious diseases** and **neurodegenerative disorders**.

Gene silencing is often considered the same as **gene knockdown**. When genes are silenced, their expression is reduced. In contrast, when genes are knocked out, they are completely erased from the organism's **genome** and, thus, have no expression.

Gene silencing is considered a gene knockdown mechanism since the methods used to silence genes, such as **RNAi**, **CRISPR**, or **siRNA**, generally reduce the expression of a gene by at least 70% but do not completely eliminate it. Methods using gene silencing are often considered better than gene knockouts since they allow researchers to study essential genes that are required for the **animal models** to survive and cannot be removed. In addition, they provide a more complete view on the development of diseases since diseases are generally associated with genes that have a reduced expression. □

Antisense oligonucleotides

Antisense **oligonucleotides** were discovered in 1978 by **Paul Zamecnik** and Mary Stephenson. **Oligonucleotides**, which are short **nucleic acid** fragments, bind to complementary target mRNA molecules when added to the cell.

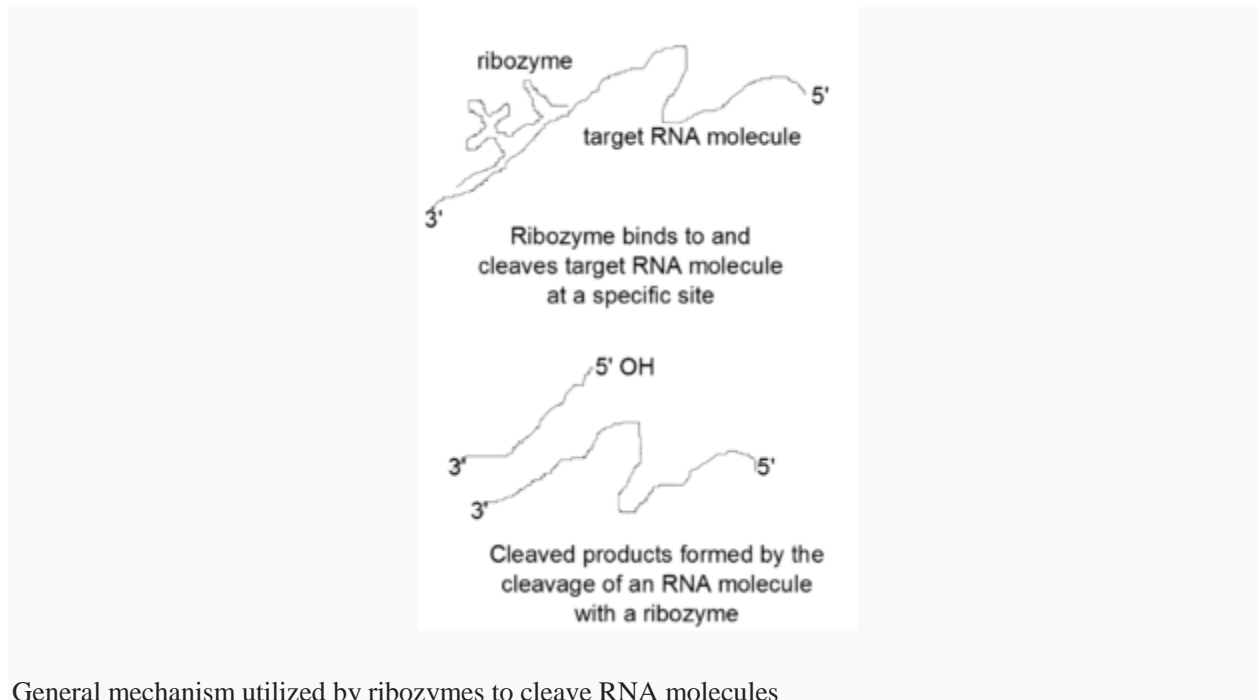
These molecules can be composed of single-stranded DNA or RNA and are generally 13–25 nucleotides long. The antisense oligonucleotides can affect gene expression in two ways: by using an **RNase H**-dependent mechanism or by using a steric blocking mechanism.

RNase H-dependent oligonucleotides cause the target **mRNA** molecules to be degraded, while steric-blocker **oligonucleotides** prevent translation of the mRNA molecule.

The majority of antisense drugs function through the RNase H-dependent mechanism, in which RNase H hydrolyzes the RNA strand of the DNA/RNA **heteroduplex**.

This mechanism is thought to be more efficient, resulting in an approximately 80% to 95% decrease in the protein and mRNA expression.

Ribozymes



General mechanism utilized by ribozymes to cleave RNA molecules

Ribozymes are catalytic RNA molecules used to inhibit **gene expression**. These molecules work by cleaving **mRNA** molecules, essentially silencing the genes that produced them. **Sidney Altman** and **Thomas Cech** first discovered catalytic RNA molecules, RNase P and group II intron ribozymes, in 1989 and won the Nobel Prize for their discovery. Several types of ribozyme motifs exist, including **hammerhead**, **hairpin**, **hepatitis delta virus**, **grp I**, **group II**, and **RNase P** ribozymes.

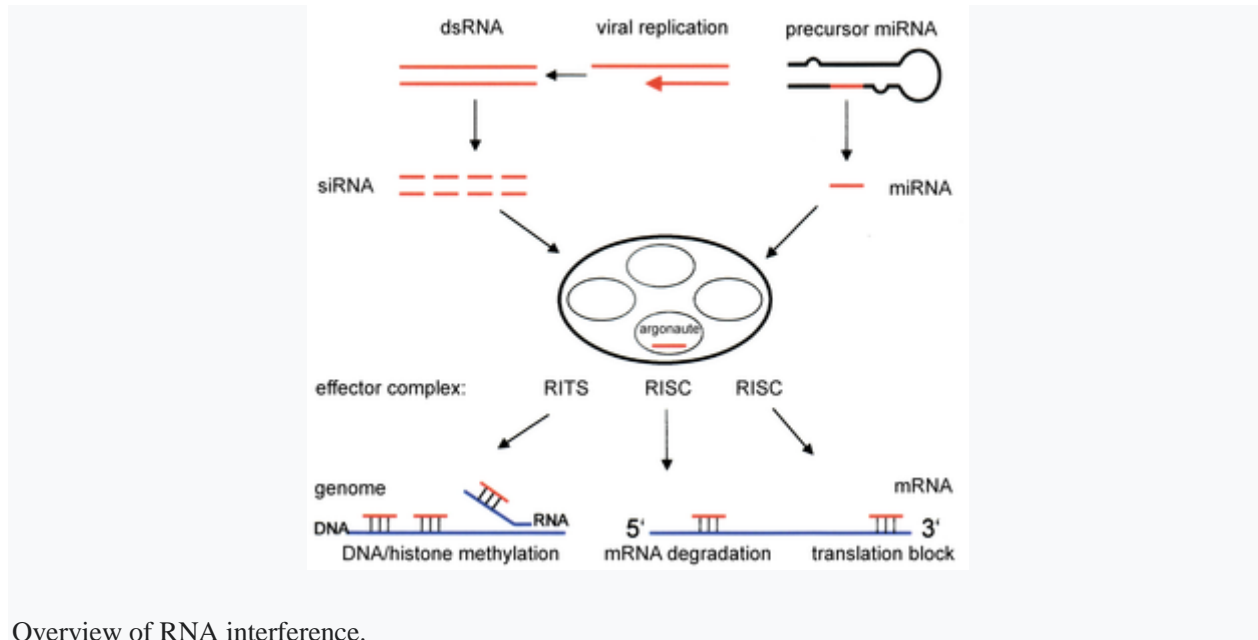
Hammerhead, hairpin, and hepatitis delta virus (HDV) ribozyme motifs are generally found in **viruses** or viroid RNAs. These motifs are able to self-cleave a specific phosphodiester bond on an mRNA molecule.

Lower **eukaryotes** and a few **bacteria** contain group I and group II ribozymes. These motifs can self-splice by cleaving and joining together phosphodiester bonds. The last ribozyme motif, the RNase P ribozyme, is found in *Escherichia coli* and is known for its ability to cleave the phosphodiester bonds of several **tRNA** precursors when joined to a protein cofactor.

The general **catalytic mechanism** used by ribozymes is similar to the mechanism used by protein **ribonucleases**. These catalytic RNA molecules bind to a specific site and attack the neighboring phosphate in the RNA backbone with their 2' oxygen, which acts as a **nucleophile**, resulting in the formation of cleaved products with a 2'3'-cyclic phosphate and a 5' hydroxyl terminal end.

This catalytic mechanism has been increasingly used by scientists to perform sequence-specific cleavage of target mRNA molecules. In addition, attempts are being made to use ribozymes to produce gene silencing therapeutics, which would silence genes that are responsible for causing diseases.

RNA interference



Overview of RNA interference.

RNA interference (**RNAi**) is a natural process used by cells to regulate gene expression. It was discovered in 1998 by **Andrew Fire** and **Craig Mello**, who won the Nobel Prize for their discovery in 2006.

The process to silence genes first begins with the entrance of a **double-stranded RNA (dsRNA)** molecule into the cell, which triggers the RNAi pathway.

The double-stranded molecule is then cut into small double-stranded fragments by an enzyme called **Dicer**.

These small fragments, which include **small interfering RNAs (siRNA)** and **microRNA (miRNA)**, are approximately 21–23 nucleotides in length. The fragments integrate into a multi-subunit protein called the **RNA-induced silencing complex**, which contains **Argonaute** proteins that are essential components of the RNAi pathway.

One strand of the molecule, called the "guide" strand, binds to RISC, while the other strand, known as the "passenger" strand is degraded. The guide or antisense strand of the fragment that remains bound to RISC directs the sequence-specific silencing of the target mRNA molecule. The genes can be silenced by siRNA molecules that cause the endonucleatic cleavage of the target mRNA molecules or by miRNA molecules that suppress translation of the mRNA molecule.

With the cleavage or translational repression of the mRNA molecules, the genes that form them are rendered essentially inactive. RNAi is thought to have evolved as a cellular defense mechanism against invaders, such as **RNA viruses**, or to combat the proliferation of **transposons** within a cell's DNA.

Both RNA viruses and transposons can exist as double-stranded RNA and lead to the activation of RNAi. Currently, **siRNAs** are being widely used to suppress specific **gene expression** and to assess the function of **genes**.

Three prime untranslated regions and microRNAs

Three prime untranslated regions (3'UTRs) of messenger RNAs (mRNAs) often contain regulatory sequences that post-transcriptionally cause gene silencing. Such 3'-UTRs often contain both [binding sites](#) for microRNAs (miRNAs) as well as for [regulatory proteins](#).

By binding to specific sites within the 3'-UTR, a large number of specific miRNAs decrease [gene expression](#) of their particular target mRNAs by either inhibiting [translation](#) or directly causing degradation of the transcript, using a mechanism similar to RNA interference (see [MicroRNA](#)). The 3'-UTR also may have silencer regions that bind repressor proteins that inhibit the expression of an mRNA.

The 3'-UTR often contains [microRNA response elements \(MREs\)](#). MREs are sequences to which miRNAs bind and cause gene silencing. These are prevalent motifs within 3'-UTRs. Among all regulatory motifs within the 3'-UTRs (e.g. including silencer regions), MREs make up about half of the motifs.

As of 2014, the [miRBase](#) web site, an archive of miRNA [sequences](#) and annotations, listed 28,645 entries in 233 biologic species. Of these, 1,881 miRNAs were in annotated human miRNA loci. miRNAs were predicted to each have an average of about four hundred target mRNAs (causing gene silencing of several hundred genes).

Freidman et al. estimate that >45,000 miRNA [target sites](#) within human mRNA 3'UTRs are conserved above background levels, and >60% of human protein-coding [genes](#) have been under selective pressure to maintain pairing to miRNAs.

Direct experiments show that a single miRNA can reduce the stability of hundreds of unique mRNAs. Other experiments show that a single [miRNA](#) may repress the production of hundreds of proteins, but that this repression often is relatively mild (less than 2-fold).

The effects of miRNA dysregulation of gene expression seem to be important in cancer. For instance, in gastrointestinal cancers, nine miRNAs have been identified as [epigenetically](#) altered and effective in down regulating DNA repair enzymes.

The effects of miRNA dysregulation of gene expression also seem to be important in [neuropsychiatric](#) disorders, such as schizophrenia, bipolar disorder, major depression, Parkinson's disease, Alzheimer's disease and autism spectrum disorders.

Applications

Medical research

Gene silencing techniques have been widely used by researchers to study genes associated with disorders. These disorders include [cancer](#), [infectious diseases](#), [respiratory diseases](#), and [neurodegenerative disorders](#). Gene silencing is also currently being used in drug discovery efforts, such as [synthetic lethality](#), [high-throughput screening](#), and [miniaturized RNAi screens](#).

Cancer

*RNA interference has been used to silence genes associated with several cancers. In **in vitro** studies of **chronic myelogenous leukemia (CML)**, **siRNA** was used to cleave the fusion protein, **BCR-ABL**, which prevents the drug **Gleevec (imatinib)** from binding to the cancer cells. Cleaving the fusion protein reduced the amount of transformed **hematopoietic** cells that spread throughout the body by increasing the sensitivity of the cells to the drug. RNA interference can also be used to target specific mutants. For instance, **siRNAs** were able to bind specifically to tumor suppressor **p53** molecules containing a single **point mutation** and destroy it, while leaving the wild-type suppressor intact*

Receptors involved in **mitogenic** pathways that lead to the increased production of cancer cells there have also been targeted by siRNA molecules. The **chemokine receptor chemokine receptor 4 (CXCR4)**, associated with the proliferation of breast cancer, was cleaved by siRNA molecules that reduced the number of divisions commonly observed by the cancer cells.

Researchers have also used siRNAs to selectively regulate the expression of cancer-related genes. Antiapoptotic proteins, such as **clusterin** and **survivin**, are often expressed in cancer cells. Clusterin and survivin-targeting siRNAs were used to reduce the number of antiapoptotic proteins and, thus, increase the sensitivity of the cancer cells to chemotherapy treatments.

In vivo studies are also being increasingly utilized to study the potential use of siRNA molecules in cancer therapeutics. For instance, mice implanted with **colon adenocarcinoma** cells were found to survive longer when the cells were pretreated with siRNAs that targeted **B-catenin** in the cancer cells.

Infectious disease

Viruses

Viral genes and host genes that are required for viruses to replicate or enter the cell, or that play an important role in the life cycle of the virus are often targeted by antiviral therapies. RNAi has been used to target genes in several viral diseases, such as the **human immunodeficiency virus (HIV)** and **hepatitis**.

In particular, siRNA was used to silence the primary HIV receptor **chemokine receptor 5 (CCR5)**. This prevented the virus from entering the human peripheral blood lymphocytes and the primary hematopoietic stem cells.

A similar technique was used to decrease the amount of the detectable virus in **hepatitis B** and **C** infected cells. In hepatitis B, siRNA silencing was used to target the surface antigen on the hepatitis B virus and led to a decrease in the number of viral components. In addition, siRNA techniques used in hepatitis C were able to lower the amount of the virus in the cell by 98%.

RNA interference has been in commercial use to control virus diseases of plants for over 20 years (see **Plant disease resistance**). In 1986–1990, multiple examples of "coat protein-mediated resistance" against plant viruses were published, before RNAi had been discovered. In 1993, work with tobacco etch virus first demonstrated that host organisms can target specific virus or mRNA sequences for degradation, and that this activity is the mechanism behind some examples of virus resistance in transgenic plants.

The discovery of small interfering RNAs (the specificity determinant in RNA-mediated gene silencing) also utilized virus-induced post-transcriptional gene silencing in plants.

By 1994, transgenic squash varieties had been generated expressing coat protein genes from three different viruses, providing squash hybrids with field-validated multiviral resistance that remain in commercial use at present.

Potato lines expressing viral replicase sequences that confer resistance to potato leafroll virus were sold under the trade names NewLeaf Y and NewLeaf Plus, and were widely accepted in commercial production in 1999–2001, until McDonald's Corp. decided not to purchase GM potatoes and Monsanto decided to close their NatureMark potato business.

Another frequently cited example of virus resistance mediated by gene silencing involves papaya, where the Hawaiian papaya industry was rescued by virus-resistant GM papayas produced and licensed by university researchers rather than a large corporation. These papayas also remain in use at present, although not without significant public protest, which is notably less evident in medical uses of gene silencing.

Gene silencing techniques have also been used to target other viruses, such as the human papilloma virus, the West Nile virus, and the Tulane virus. The E6 gene in tumor samples retrieved from patients with the human papilloma virus was targeted and found to cause apoptosis in the infected cells.

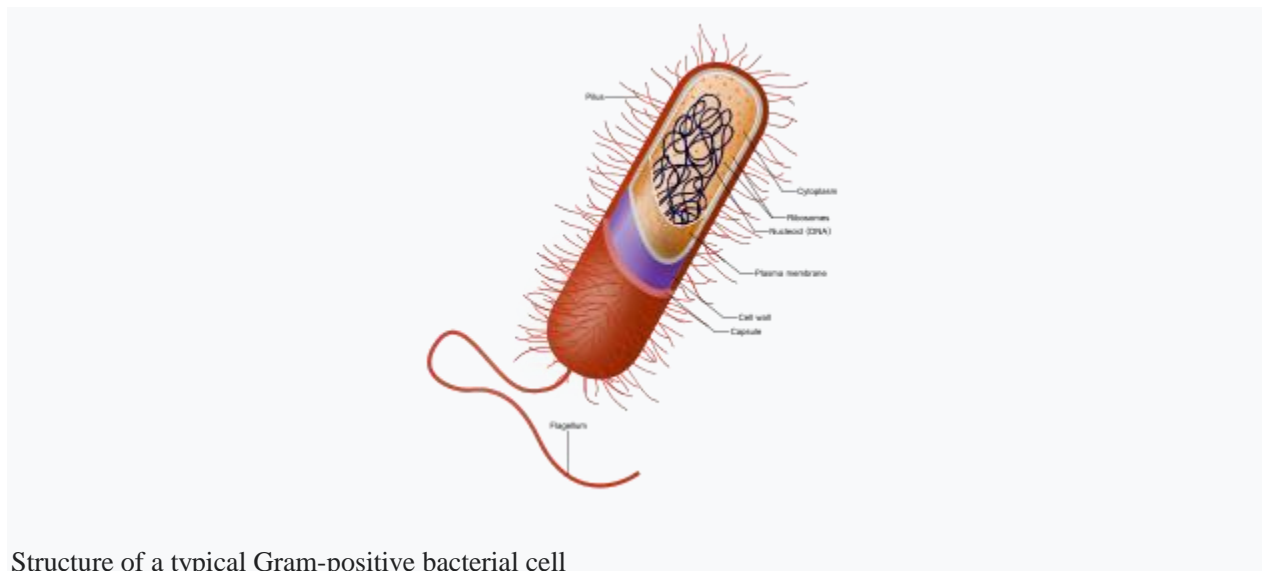
Plasmid siRNA expression vectors used to target the West Nile virus were also able to prevent the replication of viruses in cell lines. In addition, siRNA has been found to be successful in preventing the replication of the Tulane virus, part of the virus family Caliciviridae, by targeting both its structural and non-structural genes.

By targeting the NTPase gene, one dose of siRNA 4 hours pre-infection was shown to control Tulane virus replication for 48 hours post-infection, reducing the viral titer by up to 2.6 logarithms.

Although the Tulane virus is species-specific and does not affect humans, it has been shown to be closely related to the human norovirus, which is the most common cause of acute gastroenteritis and food-borne disease outbreaks in the United States.

Human noroviruses are notorious for being difficult to study in the laboratory, but the Tulane virus offers a model through which to study this family of viruses for the clinical goal of developing therapies that can be used to treat illnesses caused by human norovirus.

Bacteria



Structure of a typical Gram-positive bacterial cell

Unlike viruses, bacteria are not as susceptible to silencing by siRNA. This is largely due to how bacteria replicate. Bacteria replicate outside of the host cell and do not contain the necessary machinery for RNAi to function.

However, bacterial infections can still be suppressed by siRNA by targeting the host genes that are involved in the immune response caused by the infection or by targeting the host genes involved in mediating the entry of bacteria into cells.

For instance, siRNA was used to reduce the amount of pro-inflammatory **cytokines** expressed in the cells of mice treated with **lipopolysaccharide (LPS)**.

The reduced expression of the inflammatory cytokine, **tumor necrosis factor α (TNF α)**, in turn, caused a reduction in the septic shock felt by the LPS-treated mice.

In addition, siRNA was used to prevent the bacteria, *Psueomonas aeruginosa*, from invading murine lung epithelial cells by knocking down the caveolin-2 (CAV2) gene.

Thus, though bacteria cannot be directly targeted by siRNA mechanisms, they can still be affected by siRNA when the components involved in the bacterial infection are targeted.

Respiratory diseases

Ribozymes, antisense oligonucleotides, and more recently RNAi have been used to target mRNA molecules involved in **asthma**. These experiments have suggested that siRNA may be used to combat other respiratory diseases, such as **chronic obstructive pulmonary disease (COPD)** and **cystic fibrosis**.

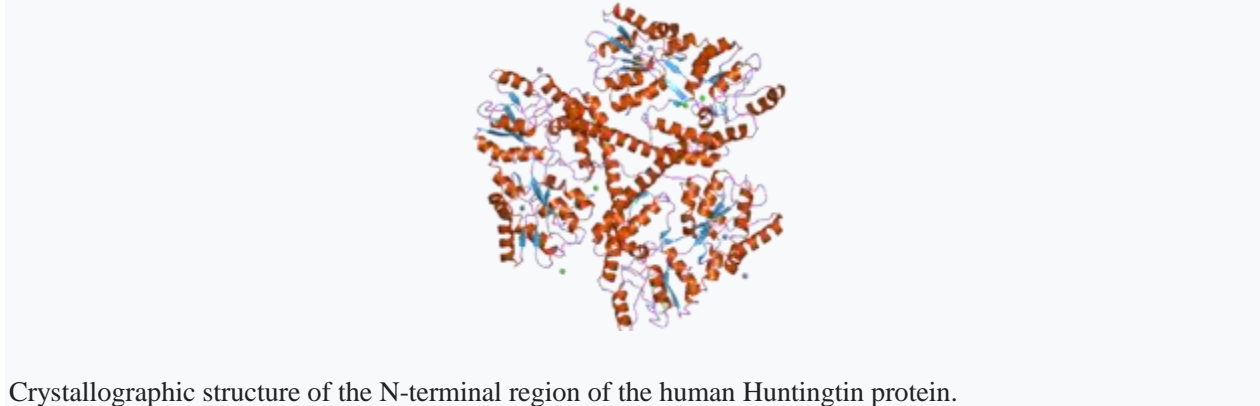
COPD is characterized by **goblet cell hyperplasia** and **mucus hypersecretion**. Mucus secretion was found to be reduced when the **transforming growth factor (TGF)- α** was targeted by siRNA in NCI-H292 human airway **epithelial cells**.

In addition to mucus hypersecretion, **chronic inflammation** and damaged lung tissue are characteristic of COPD and asthma. The **transforming growth factor TGF- β** is thought to play a role in these manifestations.

As a result, when **interferon (IFN)- γ** was used to knock down TGF- β , **fibrosis** of the lungs, caused by damage and scarring to lung tissue, was improved.

Neurodegenerative disorders

Huntington's disease



Crystallographic structure of the N-terminal region of the human Huntingtin protein.

Huntington's disease (HD) results from a mutation in the **huntingtin gene** that causes an excess of CAG repeats. The gene then forms a mutated **huntingtin protein** with polyglutamine repeats near the **amino terminus**.

This disease is incurable and known to cause motor, **cognitive**, and behavioral deficits. Researchers have been looking to gene silencing as a potential therapeutic for HD.

Gene silencing can be used to treat HD by targeting the mutant huntingtin protein. The mutant huntingtin protein has been targeted through gene silencing that is allele specific using **allele specific oligonucleotides**.

In this method, the antisense oligonucleotides are used to target **single nucleotide polymorphism (SNPs)**, which are single nucleotide changes in the DNA sequence, since HD patients have been found to share common SNPs that are associated with the mutated huntingtin allele.

It has been found that approximately 85% of patients with HD can be covered when three SNPs are targeted. In addition, when antisense oligonucleotides were used to target an HD-associated SNP in mice, there was a 50% decrease in the mutant huntingtin protein.

Non-allele specific gene silencing using siRNA molecules has also been used to silence the mutant huntingtin proteins. Through this approach, instead of targeting SNPs on the mutated protein, all of the normal and mutated huntingtin proteins are targeted. When studied in mice, it was found that siRNA could reduce the normal and mutant huntingtin levels by 75%.

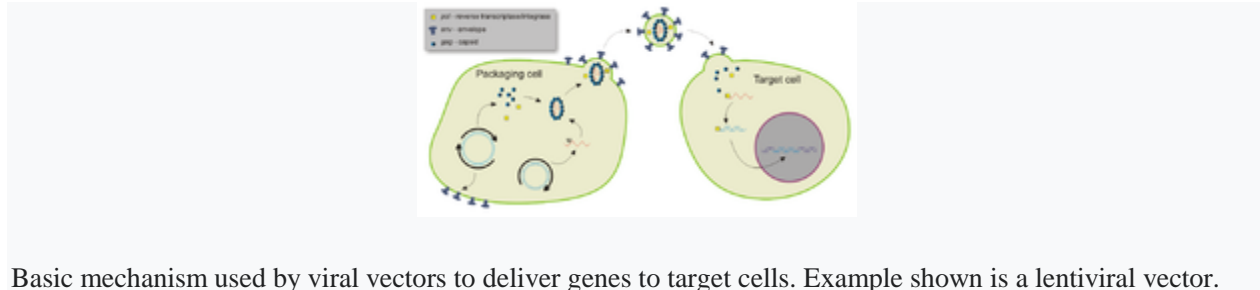
At this level, they found that the mice developed improved **motor control** and a longer **survival rate** when compared to the controls. Thus, gene silencing methods may prove to be beneficial in treating HD.

Amyotrophic lateral sclerosis

Amyotrophic lateral sclerosis (ALS), also called **Lou Gehrig's disease**, is a **motor neuron disease** that affects the **brain** and **spinal cord**. The disease causes **motor neurons** to degenerate, which eventually leads to neuron death and muscular degeneration.

Hundreds of mutations in the Cu/Zn **superoxide dismutase** (SOD1) gene have been found to cause ALS. Gene silencing has been used to knock down the SOD1 mutant that is characteristic of ALS. In specific, siRNA molecules have been successfully used to target the SOD1 mutant gene and reduce its expression through allele-specific gene silencing.

Therapeutics challenges



Basic mechanism used by viral vectors to deliver genes to target cells. Example shown is a lentiviral vector.

There are several challenges associated with gene silencing therapies, including **delivery** and specificity for targeted cells. For instance, for treatment of neurodegenerative disorders, molecules for a prospective gene silencing therapy must be delivered to the brain.

The **blood-brain barrier** makes it difficult to deliver molecules into the brain through the bloodstream by preventing the passage of the majority of molecules that are injected or absorbed into the blood. Thus, researchers have found that they must directly inject the molecules or implant pumps that push them into the brain.

Once inside the brain, however, the molecules must move inside of the targeted cells. In order to efficiently deliver siRNA molecules into the cells, **viral vectors** can be used.

Nevertheless, this method of delivery can also be problematic as it can elicit an immune response against the molecules. In addition to delivery, specificity has also been found to be an issue in gene silencing. Both antisense oligonucleotides and siRNA molecules can potentially bind to the wrong mRNA molecule.

Thus, researchers are searching for more efficient methods to deliver and develop specific gene silencing therapeutics that are still safe and effective.

Food

Arctic Apples are a suite of trademarked apples that contain a nonbrowning trait created by using gene silencing to reduce the expression of polyphenol oxidase (PPO). It is the first approved food product to use this technique.

Gene knockout

A **gene knockout** (abbreviation: **KO**) is a **genetic** technique in which one of an **organism's genes** is made inoperative ("knocked out" of the organism). However, KO can also refer to the gene that is knocked out or the organism that carries the gene knockout. **Knockout organisms** or simply **knockouts** are used to study gene function, usually by investigating the effect of gene loss.

Researchers draw inferences from the difference between the knockout organism and normal individuals.

The KO technique is essentially the opposite of a [gene knock-in](#). Knocking out two genes simultaneously in an organism is known as a **double knockout (DKO)**. Similarly the terms **triple knockout (TKO)** and **quadruple knockouts (QKO)** are used to describe three or four knocked out genes, respectively. However, one needs to distinguish between [heterozygous](#) and [homozygous](#) KOs. In the former, only one of two gene copies ([alleles](#)) is knocked out, in the latter both are knocked out.



Methods

Knockouts are accomplished through a variety of techniques. Originally, **naturally occurring mutations** were identified and then gene loss or inactivation had to be established by [DNA sequencing](#) or other methods.



A laboratory mouse in which a gene affecting hair growth has been knocked out (left), is shown next to a normal lab mouse.

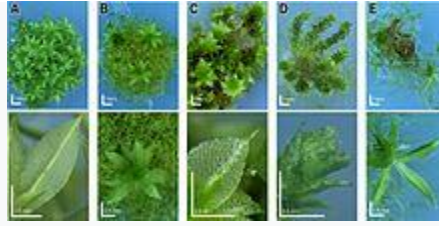
Homologous recombination

Traditionally, [homologous recombination](#) was the main method for causing a gene knockout. This method involves creating a [DNA construct](#) containing the desired mutation. For knockout purposes, this typically involves a drug resistance marker in place of the desired knockout gene.

The construct will also contain a minimum of 2kb of [homology](#) to the target sequence. The construct can be delivered to [stem cells](#) either through [microinjection](#) or [electroporation](#).

This method then relies on the cell's own repair mechanisms to recombine the DNA construct into the existing DNA. This results in the sequence of the gene being altered, and most cases the gene will be [translated](#) into a nonfunctional [protein](#), if it is translated at all. However, this is an inefficient process, as homologous recombination accounts for only 10^{-2} to 10^{-3} of DNA integrations.

Often, the drug selection marker on the construct is used to select for cells in which the recombination event has occurred.



Wild-type Physcomitrella and knockout mosses: Deviating phenotypes induced in gene-disruption library transformants. *Physcomitrella* wild-type and transformed plants were grown on minimal Knop medium to induce differentiation and development of gametophores. For each plant, an overview (upper row; scale bar corresponds to 1 mm) and a close-up (bottom row; scale bar equals 0.5 mm) are shown. A: Haploid wild-type moss plant completely covered with leafy gametophores and close-up of wild-type leaf. B–D: Different mutants.

These stem cells now lacking the gene could be used *in vivo*, for instance in mice, by inserting them into early embryos. If the resulting chimeric mouse contained the genetic change in their germline, this could then be passed on offspring.

In *diploid* organisms, which contain two alleles for most genes, and may as well contain several related genes that collaborate in the same role, additional rounds of transformation and selection are performed until every targeted gene is knocked out. *Selective breeding* may be required to produce *homozygous* knockout animals.

Site-specific nucleases

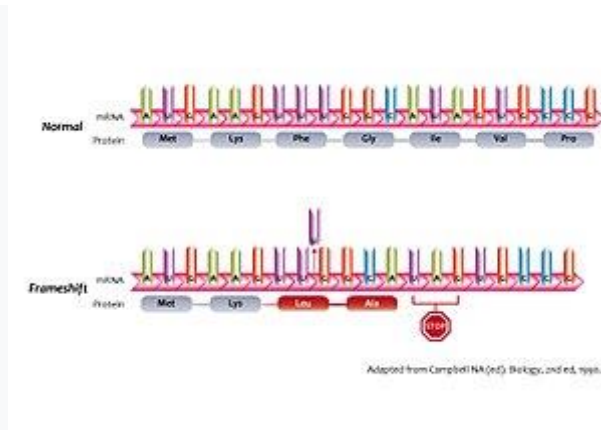


Fig 1. Frameshift mutation resulting from a single base pair deletion, causing altered amino acid sequence and premature stop codon.

There are currently three methods in use that involve precisely targeting a DNA sequence in order to introduce a double-stranded break. Once this occurs, the cell's repair mechanisms will attempt to repair this double stranded break, often through *non-homologous end joining* (NHEJ), which involves directly ligating the two cut ends together.

This may be done imperfectly, therefore sometimes causing insertions or deletions of base pairs, which cause *frameshift mutations*. These mutations can render the gene in which they

occur nonfunctional, thus creating a knockout of that gene. This process is more efficient than homologous recombination, and therefore can be more easily used to create biallelic knockouts.

Zinc-fingers

Zinc-finger nucleases consist of DNA binding domains that can precisely target a DNA sequence. Each zinc finger can recognize codons of a desired DNA sequence, and therefore can be modularly assembled to bind to a particular sequence.

These binding domains are coupled with a **restriction endonuclease** that can cause a double stranded break (DSB) in the DNA. Repair processes may introduce mutations that destroy functionality of the gene.

TALENS

Transcription activator-like effector nucleases (TALENs) also contain a DNA binding domain and a nuclease that can cleave DNA. The DNA binding region consists of amino acid repeats that each recognize a single base pair of the desired targeted DNA sequence. If this cleavage is targeted to a gene coding region, and NHEJ-mediated repair introduces insertions and deletions, a frameshift mutation often results, thus disrupting function of the gene.

CRISPR/Cas9

Clustered regularly interspaced short palindromic repeats (**CRISPR**)/Cas9 is a method for genome editing that contains a **guide RNA** complexed with a **Cas9 protein**.

The guide RNA can be engineered to match a desired DNA sequence through simple complementary base pairing, as opposed to the time-consuming assembly of constructs required by zinc-fingers or TALENs. The coupled Cas9 will cause a double stranded break in the DNA.

Following the same principle as zinc-fingers and TALENs, the attempts to repair these double stranded breaks often result in frameshift mutations that result in a nonfunctional gene.

Knockin

Gene knockin is similar to gene knockout, but it replaces a gene with another instead of deleting it.

Types

Conditional knockouts

A **conditional knockout** allows gene deletion in a tissue in a time specific manner. This is required in place of a gene knockout if the null mutation would lead to embryonic death.^[8] This is done by introducing short sequences called loxP sites around the gene. These sequences will be introduced into the germ-line via the same mechanism as a knock-out.

This germ-line can then be crossed to another germline containing **Cre-recombinase** which is a viral enzyme that can recognize these sequences, recombines them and deletes the gene flanked by these sites.

Use

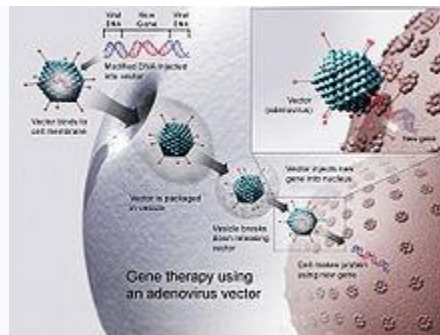


A **knockout mouse** (left) that is a model of obesity, compared with a normal mouse.

Knockouts are primarily used to understand the role of a specific **gene** or **DNA** region by comparing the knockout **organism** to a **wildtype** with a similar **genetic** background.

Knockout **organisms** are also used as **screening** tools in the development of **drugs**, to target specific **biological processes** or **deficiencies** by using a specific knockout, or to understand the **mechanism of action** of a **drug** by using a **library** of knockout **organisms** spanning the entire **genome**, such as in *Saccharomyces cerevisiae*.

Gene therapy



Gene therapy using an **adenovirus** vector. In some cases, the adenovirus will insert the new gene into a cell. If the treatment is successful, the new gene will make a functional **protein** to treat a disease.

Gene therapy (also called **human gene transfer**) is a **medical** field which focuses on the utilization of the therapeutic delivery of **nucleic acid** into a patient's cells as a **drug** to treat disease.

The first attempt at modifying human **DNA** was performed in 1980 by **Martin Cline**, but the first successful nuclear gene transfer in humans, approved by the **National Institutes of Health**, was performed in May 1989.

The first therapeutic use of gene transfer as well as the first direct insertion of human **DNA** into the nuclear genome was performed by **French Anderson** in a trial starting in September 1990. It is thought to be able to cure many genetic disorders or treat them over time.

Between 1989 and December 2018, over 2,900 clinical trials were conducted, with more than half of them in **phase I**. As of 2017, Spark Therapeutics' **Luxturna** (**RPE65 mutation-induced blindness**) and Novartis' **Kymriah** (**Chimeric antigen receptor T cell therapy**) are the FDA's first approved gene therapies to enter the market. Since that time, drugs such as

Novartis' [Zolgensma](#) and [Alnylam's Patisiran](#) have also received FDA approval, in addition to other companies' gene therapy drugs. Most of these approaches utilize [adeno-associated viruses](#) (AAVs) and [lentiviruses](#) for performing gene insertions, *in vivo* and *ex vivo*, respectively. ASO / siRNA approaches such as those conducted by [Alnylam](#) and [Ionis Pharmaceuticals](#) require non-viral delivery systems, and utilize alternative mechanisms for trafficking to liver cells by way of [GalNAc](#) transporters.

The introduction of [CRISPR gene editing](#) has opened new doors for its application and utilization in gene therapy. Solutions to medical hurdles, such as the eradication of latent human immunodeficiency virus ([HIV](#)) reservoirs and correction of the mutation that causes sickle cell disease, may soon become a tangible reality.

Not all medical procedures that introduce alterations to a patient's genetic makeup can be considered gene therapy. [Bone marrow transplantation](#) and [organ transplants](#) in general have been found to introduce foreign DNA into patients.

Gene therapy is defined by the precision of the procedure and the intention of direct therapeutic effect.



Background

Gene therapy was conceptualized in 1972, by authors who urged caution before commencing human gene therapy studies.

The first attempt, an unsuccessful one, at gene therapy (as well as the first case of medical transfer of foreign genes into humans not counting [organ transplantation](#)) was performed by [Martin Cline](#) on 10 July 1980. Cline claimed that one of the genes in his patients was active six months later, though he never published this data or had it verified and even if he is correct, it's unlikely it produced any significant beneficial effects treating [beta-thalassemia](#).

After extensive research on animals throughout the 1980s and a 1989 bacterial gene tagging trial on humans, the first gene therapy widely accepted as a success was demonstrated in a trial that started on 14 September 1990, when Ashi DeSilva was treated for [ADA-SCID](#).

The first somatic treatment that produced a permanent genetic change was initiated in 1993. The goal was to cure malignant brain tumors by using recombinant DNA to transfer a gene making the tumor cells sensitive to a drug that in turn would cause the tumor cells to die.

Gene therapy is a way to fix a genetic problem at its source. The polymers are either [translated](#) into [proteins](#), interfere with target [gene expression](#), or possibly correct [genetic mutations](#).

The most common form uses [DNA](#) that encodes a functional, therapeutic [gene](#) to replace a [mutated](#) gene. The polymer molecule is packaged within a "[vector](#)", which carries the molecule inside cells.

Early clinical failures led to dismissals of gene therapy. Clinical successes since 2006 regained researchers' attention, although as of 2014, it was still largely an experimental technique. These include treatment of [retinal diseases](#) [Leber's congenital amaurosis](#) and [choroideremia](#), [X-linked SCID](#), [ADA-SCID](#), [adrenoleukodystrophy](#), [chronic lymphocytic](#)

leukemia (CLL), acute lymphocytic leukemia (ALL), multiple myeloma, haemophilia, and Parkinson's disease. Between 2013 and April 2014, US companies invested over \$600 million in the field.

The first commercial gene therapy, **Gendicine**, was approved in China in 2003 for the treatment of certain cancers. In 2011 **Neovasculgen** was registered in Russia as the first-in-class gene-therapy drug for treatment of **peripheral artery disease**, including **critical limb ischemia**. In 2012 **Glybera**, a treatment for a rare **inherited disorder**, **lipoprotein lipase deficiency** became the first treatment to be approved for clinical use in either Europe or the United States after its endorsement by the **European Commission**.

Following early advances in **genetic engineering** of bacteria, cells, and small animals, scientists started considering how to apply it to medicine. Two main approaches were considered – replacing or disrupting defective genes. Scientists focused on diseases caused by single-gene defects, such as **cystic fibrosis**, **haemophilia**, **muscular dystrophy**, **thalassemia**, and **sickle cell anemia**. **Glybera** treats one such disease, caused by a defect in **lipoprotein lipase**.

DNA must be administered, reach the damaged cells, enter the cell and either express or disrupt a protein. Multiple delivery techniques have been explored. The initial approach incorporated DNA into an engineered **virus** to deliver the DNA into a **chromosome**. **Naked DNA** approaches have also been explored, especially in the context of **vaccine** development.

Generally, efforts focused on administering a gene that causes a needed protein to be expressed. More recently, increased understanding of **nuclease** function has led to more direct DNA editing, using techniques such as **zinc finger nucleases** and **CRISPR**. The vector incorporates genes into chromosomes.

The expressed nucleases then knock out and replace genes in the chromosome. As of 2014 these approaches involve removing cells from patients, editing a chromosome and returning the transformed cells to patients.

Gene editing is a potential approach to alter the human genome to treat genetic diseases, viral diseases, and cancer. As of 2016 these approaches were still years from being medicine.



A duplex of crRNA and **tracrRNA** acts as guide RNA to introduce a specifically located gene modification based on the RNA 5' upstream of the crRNA. Cas9 binds the tracrRNA and needs a DNA binding sequence (5'NGG3'), which is called protospacer adjacent motif (PAM). After binding, Cas9 introduces a DNA double strand break, which is then followed by gene modification via homologous recombination (HDR) or non-homologous end joining (NHEJ).

Cell types

Gene therapy may be classified into two types:

Somatic

In **somatic cell** gene therapy (SCGT), the therapeutic genes are transferred into any cell other than a **gamete**, **germ cell**, **gametocyte**, or undifferentiated **stem cell**. Any such modifications affect the individual patient only, and are not inherited by **offspring**. Somatic gene therapy represents mainstream basic and clinical research, in which therapeutic DNA (either integrated in the **genome** or as an external **episome** or **plasmid**) is used to treat disease.

Over 600 **clinical trials** utilizing SCGT are underway in the US. Most focus on severe genetic disorders, including **immunodeficiencies**, **haemophilia**, **thalassaemia**, and **cystic fibrosis**. Such single gene disorders are good candidates for somatic cell therapy. The complete correction of a genetic disorder or the replacement of multiple genes is not yet possible. Only a few of the trials are in the advanced stages.

Germline

In **germline** gene therapy (GGT), **germ cells** (**sperm** or **egg cells**) are modified by the introduction of functional genes into their genomes. Modifying a germ cell causes all the organism's cells to contain the modified gene.

The change is therefore **heritable** and passed on to later generations. Australia, Canada, Germany, Israel, Switzerland, and the Netherlands prohibit GGT for application in human beings, for technical and ethical reasons, including insufficient knowledge about possible risks to future generations and higher risks versus SCGT.

The US has no federal controls specifically addressing human genetic modification (beyond FDA regulations for therapies in general).

Vectors

The delivery of DNA into cells can be accomplished by multiple **methods**. The two major classes are **recombinant viruses** (sometimes called biological nanoparticles or viral vectors) and **naked DNA** or DNA complexes (non-viral methods).

Viruses

In order to **replicate**, **viruses** introduce their genetic material into the host cell, tricking the host's cellular machinery into using it as blueprints for viral proteins. **Retroviruses** go a stage further by having their genetic material copied into the genome of the host cell.

Scientists exploit this by substituting a virus's genetic material with therapeutic DNA. (The term 'DNA' may be an oversimplification, as some viruses contain RNA, and gene therapy could take this form as well.) A number of viruses have been used for human gene therapy, including **retroviruses**, **adenoviruses**, **herpes simplex**, **vaccinia**, and **adeno-associated virus**.

Like the genetic material (DNA or RNA) in viruses, therapeutic DNA can be designed to simply serve as a temporary blueprint that is degraded naturally or (at least theoretically) to enter the host's genome, becoming a permanent part of the host's DNA in infected cells.

Non-viral

Non-viral methods present certain advantages over viral methods, such as large scale production and low host **immunogenicity**. However, non-viral methods initially produced lower levels of **transfection** and **gene expression**, and thus lower therapeutic efficacy. Newer technologies offer promise of solving these problems, with the advent of increased cell-specific targeting and subcellular trafficking control.

Methods for non-viral gene therapy include the injection of naked DNA, **electroporation**, the **gene gun**, **sonoporation**, **magnetofection**, the use of **oligonucleotides**, lipoplexes, dendrimers, and inorganic nanoparticles.

More recent approaches, such as those performed by companies such as **Ligandal**, offer the possibility of creating cell-specific targeting technologies for a variety of gene therapy modalities, including RNA, DNA and gene editing tools such as CRISPR. Other companies, such as **Arbutus Biopharma** and **Arcturus Therapeutics**, offer non-viral, non-cell-targeted approaches that mainly exhibit liver tropism.

In more recent years, startups such as **Sixfold Bio**, **GenEdit**, and **Spotlight Therapeutics** have begun to solve the non-viral gene delivery problem. Non-viral techniques offer the possibility of repeat dosing and greater tailorability of genetic payloads, which in the future will be more likely to take over viral-based delivery systems.

Companies such as **Editas Medicine**, **Intellia Therapeutics**, **CRISPR Therapeutics**, **Casebia**, **Collectis**, **Precision Biosciences**, **bluebird bio**, and **Sangamo** have developed non-viral gene editing techniques, however frequently still use viruses for delivering gene insertion material following genomic cleavage by guided **nucleases**. These companies focus on gene editing, and still face major delivery hurdles.

BioNTech, **Moderna Therapeutics** and **CureVac** focus on delivery of **mRNA** payloads, which are necessarily non-viral delivery problems.

Alnylam, **Dicerna Pharmaceuticals**, and **Ionis Pharmaceuticals** focus on delivery of **siRNA** (antisense oligonucleotides) for gene suppression, which also necessitate non-viral delivery systems.

In academic contexts, a number of laboratories are working on delivery of **PEGylated** particles, which form serum protein coronas and chiefly exhibit LDL receptor mediated uptake in cells *in vivo*.

Hurdles

Some of the unsolved problems include:

- Short-lived nature – Before gene therapy can become a permanent cure for a condition, the therapeutic DNA introduced into target cells must remain functional and the cells containing the therapeutic DNA must be stable. Problems with integrating therapeutic DNA into the **genome** and the rapidly dividing nature of many cells prevent it from achieving long-term benefits. Patients require multiple treatments.
- Immune response – Any time a foreign object is introduced into human tissues, the immune system is stimulated to attack the invader. Stimulating the immune system in a way that reduces gene therapy

effectiveness is possible. The **immune system's** enhanced response to viruses that it has seen before reduces the effectiveness to repeated treatments.

- Problems with viral vectors – Viral vectors carry the risks of toxicity, inflammatory responses, and gene control and targeting issues.
- Multigene disorders – Some commonly occurring disorders, such as **heart disease**, **high blood pressure**, **Alzheimer's disease**, **arthritis**, and **diabetes**, are affected by variations in multiple genes, which complicate gene therapy.
- Some therapies may breach the **Weismann barrier** (between soma and germ-line) protecting the testes, potentially modifying the germline, falling afoul of regulations in countries that prohibit the latter practice.
- Insertional **mutagenesis** – If the DNA is integrated in a sensitive spot in the genome, for example in a **tumor suppressor gene**, the therapy could induce a **tumor**. This has occurred in clinical trials for **X-linked severe combined immunodeficiency (X-SCID)** patients, in which **hematopoietic** stem cells were transduced with a corrective transgene using a **retrovirus**, and this led to the development of **T cell leukemia** in 3 of 20 patients.
- One possible solution is to add a functional tumor suppressor gene to the DNA to be integrated. This may be problematic since the longer the DNA is, the harder it is to integrate into cell genomes. **CRISPR** technology allows researchers to make much more precise genome changes at exact locations.[†]
- Cost – **Alipogene tiparvovec** or Glybera, for example, at a cost of \$1.6 million per patient, was reported in 2013 to be the world's most expensive drug.

Deaths

Three patients' deaths have been reported in gene therapy trials, putting the field under close scrutiny. The first was that of **Jesse Gelsinger**, who died in 1999 because of immune rejection response. One X-SCID patient died of leukemia in 2003.

In 2007, a **rheumatoid arthritis** patient died from an infection; the subsequent investigation concluded that the death was not related to gene therapy. However it is always important to remember that although deaths are rare they can still occur and it is very possible that certain types of gene therapy can cause certain cancers

Transgene

A **transgene** is a **gene** that has been transferred naturally, or by any of a number of **genetic engineering** techniques from one organism to another. The introduction of a transgene ("catransgenesis") has the potential to change the **phenotype** of an organism.

Transgene describes a segment of **DNA** containing a gene sequence that has been isolated from one organism and is introduced into a different organism.

This non-native segment of DNA may either retain the ability to produce **RNA** or **protein** in the transgenic organism or alter the normal function of the transgenic organism's genetic code.

In general, the DNA is incorporated into the organism's **germ line**. For example, in **higher vertebrates** this can be accomplished by injecting the foreign DNA into the **nucleus** of a fertilized **ovum**. This technique is routinely used to introduce human disease genes or other

genes of interest into strains of **laboratory mice** to study the function or **pathology** involved with that particular gene.

The construction of a transgene requires the assembly of a few main parts. The transgene must contain a **promoter**, which is a regulatory sequence that will determine where and when the transgene is active, an **exon**, a protein coding sequence (usually derived from the **cDNA** for the protein of interest), and a stop sequence.

These are typically combined in a bacterial **plasmid** and the coding sequences are typically chosen from transgenes with previously known functions.

Transgenic or **genetically modified organisms**, be they bacteria, viruses or fungi, serve all kinds of research purposes. **Transgenic plants**, insects, fish and mammals have been bred. Transgenic plants such as corn and soybean have replaced wild strains in agriculture in some countries (e.g. the United States). Transgene escape has been documented for GMO crops since 2001 with persistence and invasiveness. Transgenetic organisms pose ethical questions and may cause **biosafety** problems.



History

The idea of shaping an organism to fit a specific need isn't a new science; selective breeding of animals and plants started before recorded history. However, until the late 1900s farmers and scientist could breed new strains of a plant or organism only from closely related species, because the DNA had to be compatible for offspring to be able to reproduce another generation.

In the 1970 and 1980s, scientists passed this hurdle by inventing procedures for combining the DNA of two vastly different species with **genetic engineering**. The organisms produced by these procedures were termed transgenic. Transgenesis is the same as **gene therapy** in the sense that they both transform cells for a specific purpose. However, they are completely different in their purposes, as gene therapy aims to cure a defect in cells, and transgenesis seeks to produce a genetically modified organism by incorporating the specific transgene into every cell and changing the **genome**.

Transgenesis will therefore change the germ cells, not only the somatic cells, in order to ensure that the transgenes are passed down to the offspring when the organisms reproduce. Transgenes alter the genome by blocking the function of a host gene; they can either replace the host gene with one that codes for a different protein, or introduce an additional gene.

The first transgenic organism was created in 1974 when Annie Chang and **Stanley Cohen** expressed *Staphylococcus aureus* genes in *Escherichia coli*.

In 1978, yeast cells were the first eukaryotic organisms to undergo gene transfer. Mouse cells were first transformed in 1979, followed by mouse embryos in 1980. Most of the very first transmutations were performed by **microinjection** of DNA directly into cells.

Scientists were able to develop other methods to perform the transformations, such as incorporating transgenes into **retroviruses** and then infecting cells, using electroinfection which takes advantage of an electric current to pass foreign DNA through the cell wall, **biolistics** which

is the procedure of shooting DNA bullets into cells, and also delivering DNA into the egg that has just been fertilized.

The first transgenic animals were only intended for genetic research to study the specific function of a gene, and by 2003, thousands of genes had been studied.

Use in plants

A variety of **transgenic plants** have been designed for agriculture to produce **genetically modified crops**, such as corn, soybean, rapeseed oil, cotton, rice and more. As of 2012, these GMO crops were planted on 170 million hectares globally.

Golden rice

One example of a transgenic plant species is **golden rice**. In 1997, five million children developed **xerophthalmia**, a medical condition caused by **vitamin A** deficiency, in Southeast Asia alone.

Of those children, a quarter million went blind. To combat this, scientists used **biolistics** to insert the daffodil **phytoene synthase** gene into Asia indigenous rice **cultivars**. The daffodil insertion increased the production of **β -carotene**.

The product was a transgenic rice species rich in vitamin A, called **golden rice**. Little is known about the impact of golden rice on xerophthalmia because anti-GMO campaigns have prevented the full commercial release of golden rice into agricultural systems in need.

Transgene escape

The escape of genetically-engineered plant genes via hybridization with wild relatives was first discussed and examined in Mexico and Europe in the mid-1990s. There is agreement that escape of transgenes is inevitable, even "some proof that it is happening". Up until 2008 there were few documented cases.

Corn

Corn sampled in 2000 from the **Sierra Juarez, Oaxaca**, Mexico contained a transgenic 35S promoter, while a large sample taken by a different method from the same region in 2003 and 2004 did not. A sample from another region from 2002 also did not, but directed samples taken in 2004 did, suggesting transgene persistence or re-introduction.

A 2009 study found recombinant proteins in 3.1% and 1.8% of samples, most commonly in southeast Mexico. Seed and grain import from the United States could explain the frequency and distribution of transgenes in west-central Mexico, but not in the southeast. Also, 5.0% of corn seed lots in Mexican corn stocks expressed recombinant proteins despite the moratorium on GM crops.

Cotton

In 2011, transgenic cotton was found in Mexico among wild cotton, after 15 years of GMO cotton cultivation.

Rapeseed (canola)

Transgenic rapeseed *Brassica napus*, hybridized with a native Japanese species *Brassica rapa*, was found in Japan in 2011 after they had been identified 2006 in Québec, Canada. They were persistent over a 6-year study period, without herbicide selection pressure and despite hybridization with the wild form.

This was the first report of the **introgression**—the stable incorporation of genes from one gene pool into another—of an herbicide resistance transgene from *Brassica napus* into the wild form gene pool.

Creeping bentgrass

Transgenic **creeping bentgrass**, engineered to be **glyphosate**-tolerant as "one of the first wind-pollinated, perennial, and highly outcrossing transgenic crops", was planted in 2003 as part of a large (about 160 ha) field trial in central Oregon near **Madras, Oregon**. In 2004, its pollen was found to have reached wild growing bentgrass populations up to 14 kilometres away. Cross-pollinating *Agrostis gigantea* was even found at a distance of 21 kilometres.

The grower, **Scotts Company** could not remove all genetically engineered plants, and in 2007, the **U.S. Department of Agriculture** fined Scotts \$500,000 for noncompliance with regulations.

Risk assessment

The long-term monitoring and controlling of a particular transgene has been shown not to be feasible. The **European Food Safety Authority** published a guidance for risk assessment in 2010.

Use in mice

Genetically modified mice are the most common animal model for transgenic research. Transgenic mice are currently being used to study a variety of diseases including cancer, obesity, heart disease, arthritis, anxiety, and Parkinson's disease.

The two most common types of genetically modified mice are **knockout mice** and **oncomice**. Knockout mice are a type of mouse model that uses transgenic insertion to disrupt an existing gene's expression.

In order to create knockout mice, a transgene with the desired sequence is inserted into an isolated mouse **blastocyst** using **electroporation**. Then, **homologous recombination** occurs naturally within some cells, replacing the gene of interest with the designed transgene.

Through this process, researchers were able to demonstrate that a transgene can be integrated into the genome of an animal, serve a specific function within the cell, and be passed down to future generations.

Oncomice are another genetically modified mouse species created by inserting transgenes that increase the animal's vulnerability to cancer. Cancer researchers utilize oncomice to study the profiles of different cancers in order to apply this knowledge to human studies.

Use in *Drosophila*

Multiple studies have been conducted concerning transgenesis in *Drosophila melanogaster*, the fruit fly. This organism has been a helpful genetic model for over 100 years, due to its well-understood developmental pattern. The transfer of transgenes into the *Drosophila* genome has been performed using various techniques, including P element, Cre-loxP, and Φ C31 insertion.

The most practiced method used thus far to insert transgenes into the *Drosophila* genome utilizes P elements. The transposable P elements, also known as **transposons**, are segments of bacterial DNA that are translocated into the genome, without the presence of a complementary sequence in the host's genome. P elements are administered in pairs of two, which flank the DNA insertion region of interest.

Additionally, P elements often consist of two plasmid components, one known as the P element transposase and the other, the P transposon backbone. The transposase plasmid portion drives the transposition of the P transposon backbone, containing the transgene of interest and often a marker, between the two terminal sites of the transposon.

Success of this insertion results in the nonreversible addition of the transgene of interest into the genome. While this method has been proven effective, the insertion sites of the P elements are often uncontrollable, resulting in an unfavorable, random insertion of the transgene into the *Drosophila* genome.

To improve the location and precision of the transgenic process, an enzyme known as **Cre** has been introduced. Cre has proven to be a key element in a process known as recombination-mediated cassette exchange (**RMCE**). While it has shown to have a lower efficiency of transgenic transformation than the P element transposases, Cre greatly lessens the labor-intensive abundance of balancing random P insertions. Cre aids in the targeted transgenesis of the DNA gene segment of interest, as it supports the mapping of the transgene insertion sites, known as loxP sites.

These sites, unlike P elements, can be specifically inserted to flank a chromosomal segment of interest, aiding in targeted transgenesis. The Cre transposase is important in the catalytic cleavage of the base pairs present at the carefully positioned loxP sites, permitting more specific insertions of the transgenic donor plasmid of interest.

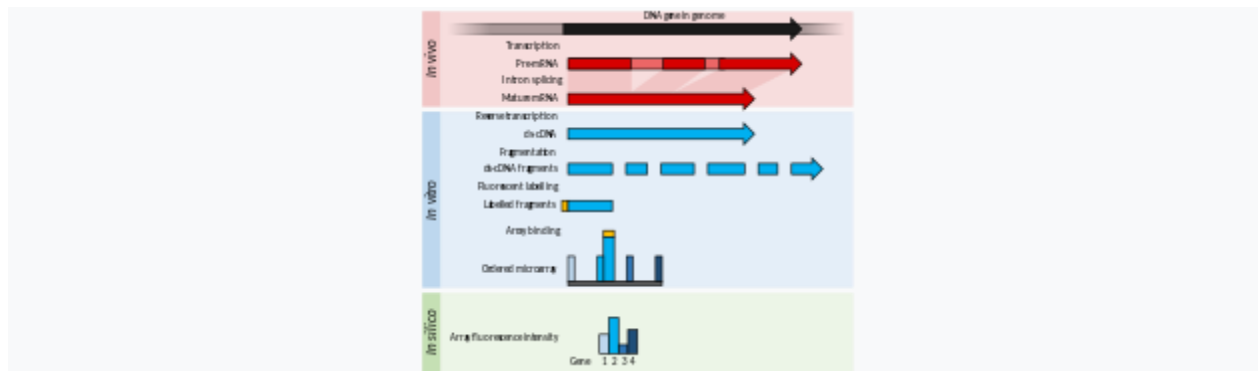
To overcome the limitations and low yields that transposon-mediated and Cre-loxP transformation methods produce, the bacteriophage Φ C31 has recently been utilized. Recent breakthrough studies involve the microinjection of the bacteriophage Φ C31 integrase, which shows improved transgene insertion of large DNA fragments that are unable to be transposed by P elements alone.

This method involves the recombination between an attachment (**attP**) site in the **phage** and an attachment site in the bacterial host genome (**attB**). Compared to usual P element transgene insertion methods, Φ C31 integrates the entire transgene vector, including bacterial sequences and antibiotic resistance genes. Unfortunately, the presence of these additional insertions has been found to affect the level and reproducibility of transgene expression.

Use in livestock and aquaculture

One agricultural application is to selectively breed animals for particular traits: Transgenic cattle with an increased muscle phenotype has been produced by overexpressing a short hairpin RNA with homology to the myostatin mRNA using RNA interference. Transgenes are being used to produce milk with high levels of proteins or silk from the milk of goats. Another agricultural application is to selectively breed animals, which are resistant to diseases or animals for biopharmaceutical production.

DNA microarray



Summary of DNA Microarrays. Within the organisms, genes are transcribed and spliced to produce mature mRNA transcripts (red). The mRNA is extracted from the organism and reverse transcriptase is used to copy the mRNA into stable ds-cDNA (blue). In microarrays, the ds-cDNA is fragmented and fluorescently labelled (orange). The labelled fragments bind to an ordered array of complementary oligonucleotides, and **measurement of fluorescent intensity** across the array indicates the abundance of a predetermined set of sequences. These sequences are typically specifically chosen to report on genes of interest within the organism's genome.^[1]

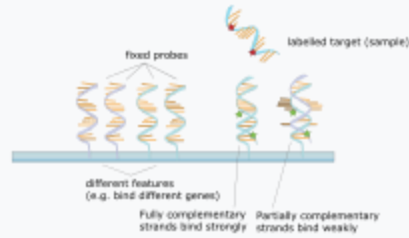
A **DNA microarray** (also commonly known as **DNA chip** or **biochip**) is a collection of microscopic DNA spots attached to a solid surface. Scientists use DNA **microarrays** to measure the **expression** levels of large numbers of genes simultaneously or to **genotype** multiple regions of a genome. Each DNA spot contains **picomoles** (10^{-12} moles) of a specific DNA sequence, known as **probes** (or **reporters** or **oligos**).

These can be a short section of a **gene** or other DNA element that are used to **hybridize** a **cDNA** or cRNA (also called anti-sense RNA) sample (called **target**) under high-stringency conditions. Probe-target hybridization is usually detected and quantified by detection of **fluorophore**-, silver-, or **chemiluminescence**-labeled targets to determine relative abundance of

nucleic acid sequences in the target. The original nucleic acid arrays were macro arrays approximately 9 cm × 12 cm and the first computerized image based analysis was published in 1981. It was invented by **Patrick O. Brown**.



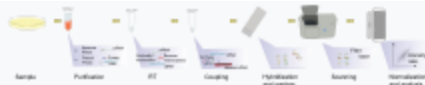
Principle



Hybridization of the target to the probe

The core principle behind microarrays is hybridization between two DNA strands, the property of **complementary** nucleic acid sequences to specifically pair with each other by forming **hydrogen bonds** between complementary **nucleotide base pairs**. A high number of complementary base pairs in a nucleotide sequence means tighter **non-covalent** bonding between the two strands. After washing off non-specific bonding sequences, only strongly paired strands will remain hybridized.

Fluorescently labeled target sequences that bind to a probe sequence generate a signal that depends on the hybridization conditions (such as temperature), and washing after hybridization. Total strength of the signal, from a spot (feature), depends upon the amount of target sample binding to the probes present on that spot. Microarrays use relative quantitation in which the intensity of a feature is compared to the intensity of the same feature under a different condition, and the identity of the feature is known by its position.



The steps required in a microarray experiment

Uses and types



Two Affymetrix chips. A **match** is shown at bottom left for size comparison.

Many types of arrays exist and the broadest distinction is whether they are spatially arranged on a surface or on coded beads:

- The traditional solid-phase array is a collection of orderly microscopic "spots", called features, each with thousands of identical and specific probes attached to a solid surface, such as [glass](#), [plastic](#) or [silicon biochip](#) (commonly known as a *genome chip*, *DNA chip* or *gene array*). Thousands of these features can be placed in known locations on a single DNA microarray.
- The alternative bead array is a collection of microscopic polystyrene beads, each with a specific probe and a ratio of two or more dyes, which do not interfere with the fluorescent dyes used on the target sequence.

DNA microarrays can be used to detect DNA (as in [comparative genomic hybridization](#)), or detect RNA (most commonly as [cDNA](#) after [reverse transcription](#)) that may or may not be translated into proteins. The process of measuring gene expression via cDNA is called [expression analysis](#) or [expression profiling](#).

Applications include:

Application or technology	Synopsis
Gene expression profiling	In an mRNA or gene expression profiling experiment the expression levels of thousands of genes are simultaneously monitored to study the effects of certain treatments, diseases , and developmental stages on gene expression. For example, microarray-based gene expression profiling can be used to identify genes whose expression is changed in response to pathogens or other organisms by comparing gene expression in infected to that in uninfected cells or tissues.
Comparative genomic hybridization	Assessing genome content in different cells or closely related organisms, as originally described by Patrick Brown , Jonathan Pollack , Ash Alizadeh and colleagues at Stanford .
GeneID	Small microarrays to check IDs of organisms in food and feed (like GMO), mycoplasmas in cell culture, or pathogens for disease detection, mostly combining PCR and microarray technology.
Chromatin immunoprecipitation on	DNA sequences bound to a particular protein can be isolated by immunoprecipitating that protein (ChIP), these fragments can be then hybridized to a microarray (such as a tiling array) allowing the

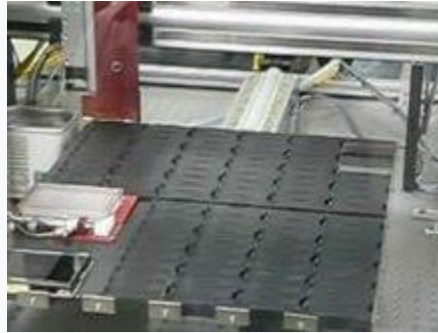
<p>Chip</p>	<p>determination of protein binding site occupancy throughout the genome. Example protein to immunoprecipitate are histone modifications (H3K27me3, H3K4me2, H3K9me3, etc.), Polycomb-group protein (PRC2:Suz12, PRC1:YY1) and trithorax-group protein (Ash1) to study the epigenetic landscape or RNA Polymerase II to study the transcription landscape.</p>
<p>DamID</p>	<p>Analogously to ChIP, genomic regions bound by a protein of interest can be isolated and used to probe a microarray to determine binding site occupancy. Unlike ChIP, DamID does not require antibodies but makes use of adenine methylation near the protein's binding sites to selectively amplify those regions, introduced by expressing minute amounts of protein of interest fused to bacterial DNA adenine methyltransferase.</p>
<p>SNP detection</p>	<p>Identifying single nucleotide polymorphism among alleles within or between populations. Several applications of microarrays make use of SNP detection, including genotyping, forensic analysis, measuring predisposition to disease, identifying drug-candidates, evaluating germline mutations in individuals or somatic mutations in cancers, assessing loss of heterozygosity, or genetic linkage analysis.</p>
<p>Alternative splicing detection</p>	<p>An <i>exon junction array</i> design uses probes specific to the expected or potential splice sites of predicted exons for a gene. It is of intermediate density, or coverage, to a typical gene expression array (with 1–3 probes per gene) and a genomic tiling array (with hundreds or thousands of probes per gene). It is used to assay the expression of alternative splice forms of a gene. <i>Exon arrays</i> have a different design, employing probes designed to detect each individual exon for known or predicted genes, and can be used for detecting different splicing isoforms.</p>
<p>Fusion genes microarray</p>	<p>A Fusion gene microarray can detect fusion transcripts, <i>e.g.</i> from cancer specimens. The principle behind this is building on the alternative splicing microarrays. The oligo design strategy enables combined measurements of chimeric transcript junctions with exon-wise measurements of individual fusion partners.</p>

<p>Tiling array</p>	<p>Genome tiling arrays consist of overlapping probes designed to densely represent a genomic region of interest, sometimes as large as an entire human chromosome. The purpose is to empirically detect expression of transcripts or alternatively spliced forms which may not have been previously known or predicted.</p>
<p>Double-stranded B-DNA microarrays</p>	<p>Right-handed double-stranded B-DNA microarrays can be used to characterize novel drugs and biologicals that can be employed to bind specific regions of immobilized, intact, double-stranded DNA. This approach can be used to inhibit gene expression. They also allow for characterization of their structure under different environmental conditions.</p>
<p>Double-stranded Z-DNA microarrays</p>	<p>Left-handed double-stranded Z-DNA microarrays can be used to identify short sequences of the alternative Z-DNA structure located within longer stretches of right-handed B-DNA genes (e.g., transcriptional enhancement, recombination, RNA editing). The microarrays also allow for characterization of their structure under different environmental conditions.</p>
<p>Multi-stranded DNA microarrays (triplex-DNA microarrays and quadruplex-DNA microarrays)</p>	<p>Multi-stranded DNA and RNA microarrays can be used to identify novel drugs that bind to these multi-stranded nucleic acid sequences. This approach can be used to discover new drugs and biologicals that have the ability to inhibit gene expression. These microarrays also allow for characterization of their structure under different environmental conditions.</p>

Fabrication

Microarrays can be manufactured in different ways, depending on the number of probes under examination, costs, customization requirements, and the type of scientific question being asked. Arrays from commercial vendors may have as few as 10 probes or as many as 5 million or more micrometre-scale probes.

Spotted vs. *in situ* synthesised arrays



A DNA microarray being printed by a robot at the University of Delaware

Microarrays can be fabricated using a variety of technologies, including printing with fine-pointed pins onto glass slides, [photolithography](#) using pre-made masks, photolithography using dynamic micromirror devices, ink-jet printing, or [electrochemistry](#) on microelectrode arrays.

In *spotted microarrays*, the probes are [oligonucleotides](#), [cDNA](#) or small fragments of [PCR](#) products that correspond to [mRNAs](#). The probes are [synthesized](#) prior to deposition on the array surface and are then "spotted" onto glass.

A common approach utilizes an array of fine pins or needles controlled by a robotic arm that is dipped into wells containing DNA probes and then depositing each probe at designated locations on the array surface. The resulting "grid" of probes represents the nucleic acid profiles of the prepared probes and is ready to receive complementary cDNA or cRNA "targets" derived from experimental or clinical samples.

This technique is used by research scientists around the world to produce "in-house" printed microarrays from their own labs. These arrays may be easily customized for each experiment, because researchers can choose the probes and printing locations on the arrays, synthesize the probes in their own lab (or collaborating facility), and spot the arrays.

They can then generate their own labeled samples for hybridization, hybridize the samples to the array, and finally scan the arrays with their own equipment.

This provides a relatively low-cost microarray that may be customized for each study, and avoids the costs of purchasing often more expensive commercial arrays that may represent vast numbers of genes that are not of interest to the investigator.

Publications exist which indicate in-house spotted microarrays may not provide the same level of sensitivity compared to commercial oligonucleotide arrays, possibly owing to the small batch sizes and reduced printing efficiencies when coaxed to industrial manufactures of oligo arrays.

In *oligonucleotide microarrays*, the probes are short sequences designed to match parts of the sequence of known or predicted [open reading frames](#). Although oligonucleotide probes are often used in "spotted" microarrays, the term "oligonucleotide array" most often refers to a specific technique of manufacturing. Oligonucleotide arrays are produced by printing short oligonucleotide sequences designed to represent a single gene or family of gene splice-variants by [synthesizing](#) this sequence directly onto the array surface instead of depositing intact sequences. Sequences may be longer (60-mer probes such as the [Agilent](#) design) or shorter (25-mer probes produced by [Affymetrix](#)) depending on the desired purpose; longer probes are more specific to individual target genes, shorter probes

may be spotted in higher density across the array and are cheaper to manufacture. One technique used to produce oligonucleotide arrays include **photolithographic** synthesis (Affymetrix) on a silica substrate where light and light-sensitive masking agents are used to "build" a sequence one nucleotide at a time across the entire array.

Each applicable probe is selectively "unmasked" prior to bathing the array in a solution of a single nucleotide, then a masking reaction takes place and the next set of probes are unmasked in preparation for a different nucleotide exposure. After many repetitions, the sequences of every probe become fully constructed. More recently, Maskless Array Synthesis from NimbleGen Systems has combined flexibility with large numbers of probes.^[15]

Two-channel vs. one-channel detection

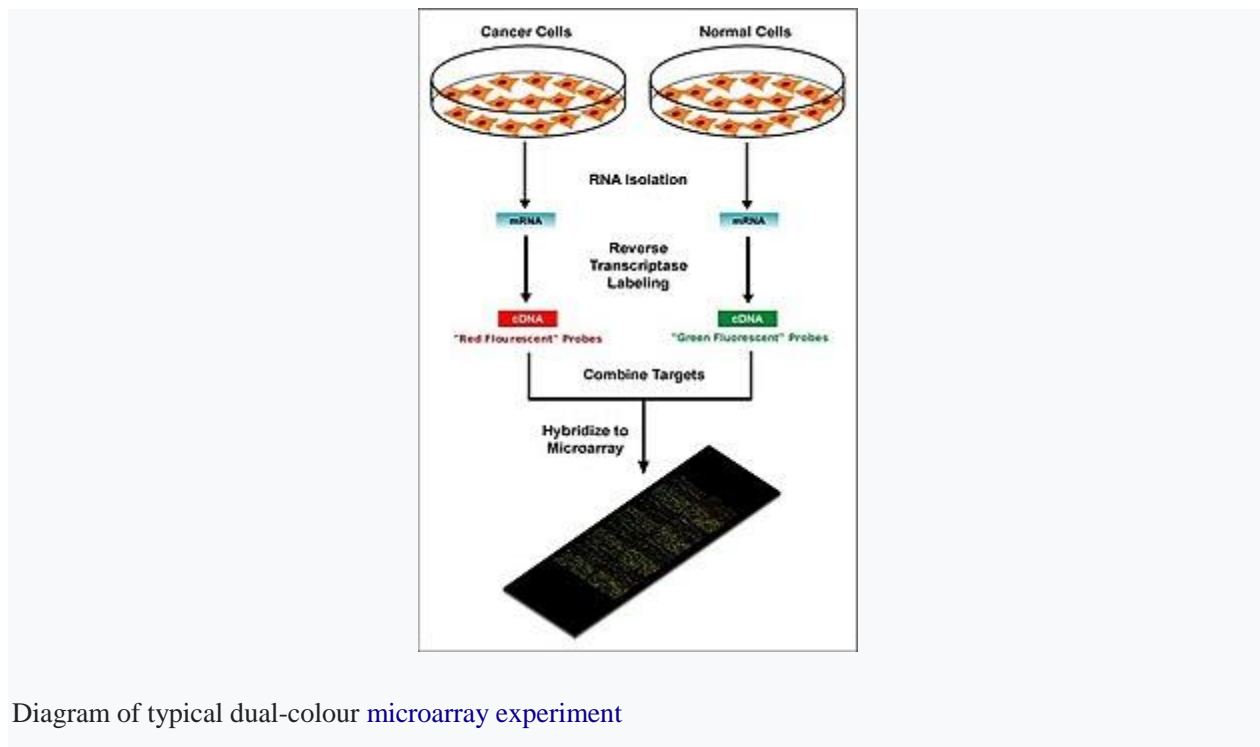


Diagram of typical dual-colour **microarray experiment**

Two-color microarrays or *two-channel microarrays* are typically **hybridized** with cDNA prepared from two samples to be compared (e.g. diseased tissue versus healthy tissue) and that are labeled with two different **fluorophores**.

Fluorescent dyes commonly used for cDNA labeling include **Cy3**, which has a fluorescence emission wavelength of 570 nm (corresponding to the green part of the light spectrum), and **Cy5** with a fluorescence emission wavelength of 670 nm (corresponding to the red part of the light spectrum).

The two Cy-labeled cDNA samples are mixed and hybridized to a single microarray that is then scanned in a microarray scanner to visualize fluorescence of the two fluorophores after **excitation** with a **laser** beam of a defined wavelength. Relative intensities of each fluorophore may then be used in ratio-based analysis to identify up-regulated and down-regulated genes.

Oligonucleotide microarrays often carry control probes designed to hybridize with **RNA spike-ins**. The degree of hybridization between the spike-ins and the control probes is used to **normalize** the hybridization measurements for the target probes. Although absolute levels of gene expression may be determined in the two-color array in rare instances, the relative differences in expression among different spots within a sample and between samples is the preferred method

of [data analysis](#) for the two-color system. Examples of providers for such microarrays includes [Agilent](#) with their Dual-Mode platform, [Eppendorf](#) with their DualChip platform for colorimetric [Silverquant](#) labeling, and TeleChem International with [Arrayit](#).

In *single-channel microarrays* or *one-color microarrays*, the arrays provide intensity data for each probe or probe set indicating a relative level of hybridization with the labeled target. However, they do not truly indicate abundance levels of a gene but rather relative abundance when compared to other samples or conditions when processed in the same experiment.

Each RNA molecule encounters protocol and batch-specific bias during amplification, labeling, and hybridization phases of the experiment making comparisons between genes for the same microarray uninformative. The comparison of two conditions for the same gene requires two separate single-dye hybridizations. Several popular single-channel systems are the Affymetrix "Gene Chip", Illumina "Bead Chip", Agilent single-channel arrays, the Applied Microarrays "CodeLink" arrays, and the Eppendorf "DualChip & Silverquant".

One strength of the single-dye system lies in the fact that an aberrant sample cannot affect the raw data derived from other samples, because each array chip is exposed to only one sample (as opposed to a two-color system in which a single low-quality sample may drastically impinge on overall data precision even if the other sample was of high quality). Another benefit is that data are more easily compared to arrays from different experiments as long as batch effects have been accounted for.

One channel microarray may be the only choice in some situations. Suppose n samples need to be compared: then the number of experiments required using the two channel arrays quickly becomes unfeasible, unless a sample is used as a reference.

number of samples	one-channel microarray	two channel microarray	two channel microarray (with reference)
1	1	1	1
2	2	1	1
3	3	3	2
4	4	6	3

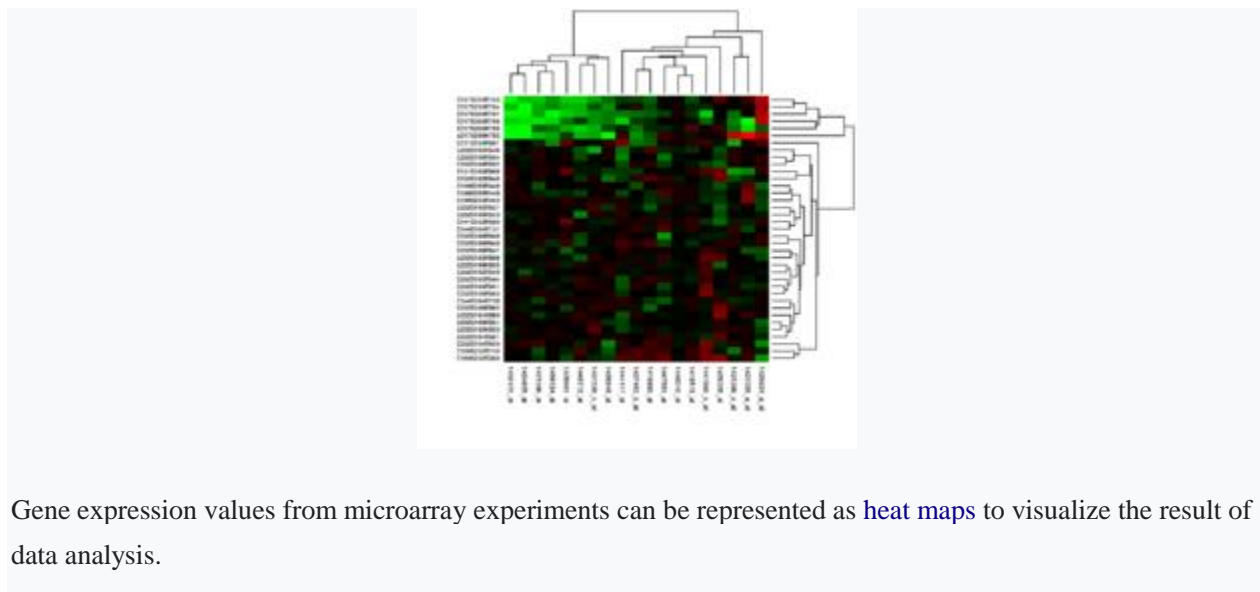
A typical protocol

This is an example of a **DNA microarray experiment** which includes details for a particular case to better explain DNA microarray experiments, while listing modifications for RNA or other alternative experiments.

1. The two samples to be compared (pairwise comparison) are grown/acquired. In this example treated sample (**case**) and untreated sample (**control**).
2. The **nucleic acid** of interest is purified: this can be **RNA** for **expression profiling**, **DNA** for **comparative hybridization**, or DNA/RNA bound to a particular **protein** which is **immunoprecipitated** (**ChIP-on-chip**) for **epigenetic** or regulation studies. In this example total RNA is isolated (both nuclear and **cytoplasmic**) by **Guanidinium thiocyanate-phenol-chloroform extraction** (e.g. **Trizol**) which isolates most RNA (whereas column methods have a cut off of 200 nucleotides) and if done correctly has a better purity.
3. The purified RNA is analysed for quality (by **capillary electrophoresis**) and quantity (for example, by using a **NanoDrop** or NanoPhotometer **spectrometer**). If the material is of acceptable quality and sufficient quantity is present (e.g., $>1\mu\text{g}$, although the required amount varies by microarray platform), the experiment can proceed.
4. The labeled product is generated via **reverse transcription** and followed by an optional **PCR** amplification. The RNA is reverse transcribed with either polyT primers (which amplify only **mRNA**) or random primers (which amplify all RNA, most of which is **rRNA**). **miRNA** microarrays ligate an oligonucleotide to the purified small RNA (isolated with a fractionator), which is then reverse transcribed and amplified.
 - The label is added either during the reverse transcription step, or following amplification if it is performed. The **sense** labeling is dependent on the microarray; e.g. if the label is added with the RT mix, the **cdNA** is antisense and the microarray probe is sense, except in the case of negative controls.
 - The label is typically **fluorescent**; only one machine uses **radiolabels**.
 - The labeling can be direct (not used) or indirect (requires a coupling stage). For two-channel arrays, the coupling stage occurs before hybridization, using **aminoallyl uridine triphosphate** (aminoallyl-UTP, or aaUTP) and **NHS** amino-reactive dyes (such as **cyanine dyes**); for single-channel arrays, the coupling stage occurs after hybridization, using **biotin** and labeled **streptavidin**. The modified nucleotides (usually in a ratio of 1 aaUTP: 4 TTP (**thymidine triphosphate**)) are added enzymatically in a low ratio to normal nucleotides, typically resulting in 1 every 60 bases. The aaDNA is then purified with a **column** (using a phosphate buffer solution, as **Tris** contains amine groups). The aminoallyl group is an amine group on a long linker attached to the nucleobase, which reacts with a reactive dye.
 - A form of replicate known as a dye flip can be performed to control for dye **artifacts** in two-channel experiments; for a dye flip, a second slide is used, with the labels swapped (the sample that was labeled with Cy3 in the first slide is labeled with Cy5, and vice versa). In this example, **aminoallyl-UTP** is present in the reverse-transcribed mixture.
5. The labeled samples are then mixed with a proprietary **hybridization** solution which can consist of **SDS**, **SSC**, **dextran sulfate**, a blocking agent (such as **Cot-1 DNA**, salmon sperm DNA, calf thymus DNA, **PolyA**, or PolyT), **Denhardt's solution**, or **formamine**.
6. The mixture is denatured and added to the pinholes of the microarray. The holes are sealed and the microarray hybridized, either in a hyb oven, where the microarray is mixed by rotation, or in a mixer, where the microarray is mixed by alternating pressure at the pinholes.
7. After an overnight hybridization, all nonspecific binding is washed off (**SDS** and **SSC**).
8. The microarray is dried and scanned by a machine that uses a laser to excite the dye and measures the emission levels with a detector.

9. The image is gridded with a template and the intensities of each feature (composed of several pixels) is quantified.
10. The raw data is normalized; the simplest normalization method is to subtract background intensity and scale so that the total intensities of the features of the two channels are equal, or to use the intensity of a reference gene to calculate the **t-value** for all of the intensities. More sophisticated methods include **z-ratio**, **loess** and **lowess regression** and **RMA** (robust multichip analysis) for Affymetrix chips (single-channel, silicon chip, *in situ* synthesized short oligonucleotides).

Microarrays and bioinformatics



Gene expression values from microarray experiments can be represented as **heat maps** to visualize the result of data analysis.

The advent of inexpensive microarray experiments created several specific bioinformatics challenges: the multiple levels of replication in experimental design (**Experimental design**); the number of platforms and independent groups and data format (**Standardization**); the statistical treatment of the data (**Data analysis**); mapping each probe to the **mRNA** transcript that it measures (**Annotation**); the sheer volume of data and the ability to share it (**Data warehousing**).

Experimental design

Due to the biological complexity of gene expression, the considerations of experimental design that are discussed in the **expression profiling** article are of critical importance if statistically and biologically valid conclusions are to be drawn from the data.

There are three main elements to consider when designing a microarray experiment. First, replication of the biological samples is essential for drawing conclusions from the experiment. Second, technical replicates (two RNA samples obtained from each experimental unit) help to ensure precision and allow for testing differences within treatment groups.

The biological replicates include independent RNA extractions and technical replicates may be two **aliquots** of the same extraction. Third, spots of each cDNA clone or oligonucleotide are present as replicates (at least duplicates) on the microarray slide, to provide a measure of technical precision in each hybridization. It is critical that information about the sample preparation and handling is discussed, in order to help identify the independent units in the experiment and to avoid inflated estimates of **statistical significance**.

Standardization

Microarray data is difficult to exchange due to the lack of standardization in platform fabrication, assay protocols, and analysis methods. This presents an **interoperability** problem in **bioinformatics**. Various **grass-roots open-source** projects are trying to ease the exchange and analysis of data produced with non-proprietary chips:

For example, the "Minimum Information About a Microarray Experiment" (**MIAME**) checklist helps define the level of detail that should exist and is being adopted by many **journals** as a requirement for the submission of papers incorporating microarray results. But MIAME does not describe the format for the information, so while many formats can support the MIAME requirements, as of 2007 no format permits verification of complete semantic compliance.

The "MicroArray Quality Control (MAQC) Project" is being conducted by the US **Food and Drug Administration** (FDA) to develop standards and quality control metrics which will eventually allow the use of MicroArray data in drug discovery, clinical practice and regulatory decision-making. The **MGED Society** has developed standards for the representation of gene expression experiment results and relevant annotations.

Data analysis



National Center for Toxicological Research scientist reviews microarray data

Microarray data sets are commonly very large, and analytical precision is influenced by a number of variables. **Statistical** challenges include taking into account effects of background noise and appropriate **normalization** of the data. Normalization methods may be suited to specific platforms and, in the case of commercial platforms, the analysis may be proprietary. Algorithms that affect statistical analysis include:

- Image analysis: gridding, spot recognition of the scanned image (segmentation algorithm), removal or marking of poor-quality and low-intensity features (called *flagging*).
- Data processing: background subtraction (based on global or local background), determination of spot intensities and intensity ratios, visualisation of data (e.g. see **MA plot**), and log-transformation of ratios, global or **local** normalization of intensity ratios, and segmentation into different copy number regions using **step detection** algorithms.
- Class discovery analysis: This analytic approach, sometimes called unsupervised classification or knowledge discovery, tries to identify whether microarrays (objects, patients, mice, etc.) or genes cluster together in groups. Identifying naturally existing groups of objects (microarrays or genes) which cluster together can enable the discovery of new groups that otherwise were not previously known to exist.

- During knowledge discovery analysis, various unsupervised classification techniques can be employed with DNA microarray data to identify novel clusters (classes) of arrays. This type of approach is not hypothesis-driven, but rather is based on iterative pattern recognition or statistical learning methods to find an "optimal" number of clusters in the data.
- Examples of unsupervised analyses methods include self-organizing maps, neural gas, k-means cluster analyses, hierarchical cluster analysis, Genomic Signal Processing based clustering and model-based cluster analysis. For some of these methods the user also has to define a distance measure between pairs of objects. Although the Pearson correlation coefficient is usually employed, several other measures have been proposed and evaluated in the literature.
- The input data used in class discovery analyses are commonly based on lists of genes having high informativeness (low noise) based on low values of the coefficient of variation or high values of Shannon entropy, etc. The determination of the most likely or optimal number of clusters obtained from an unsupervised analysis is called cluster validity. Some commonly used metrics for cluster validity are the silhouette index, Davies-Bouldin index, Dunn's index, or Hubert's statistic.
- Class prediction analysis: This approach, called supervised classification, establishes the basis for developing a predictive model into which future unknown test objects can be input in order to predict the most likely class membership of the test objects. Supervised analysis for class prediction involves use of techniques such as linear regression, k-nearest neighbor, learning vector quantization, decision tree analysis, random forests, naive Bayes, logistic regression, kernel regression, artificial neural networks, support vector machines, [mixture of experts](#), and supervised neural gas.
- In addition, various metaheuristic methods are employed, such as [genetic algorithms](#), covariance matrix self-adaptation, [particle swarm optimization](#), and [ant colony optimization](#). Input data for class prediction are usually based on filtered lists of genes which are predictive of class, determined using classical hypothesis tests (next section), Gini diversity index, or information gain (entropy).
- Hypothesis-driven statistical analysis: Identification of statistically significant changes in gene expression are commonly identified using the [t-test](#), [ANOVA](#), [Bayesian method](#) [Mann–Whitney test](#) methods tailored to microarray data sets, which take into account [multiple comparisons](#) or [cluster analysis](#). These methods assess statistical power based on the variation present in the data and the number of experimental replicates, and can help minimize [Type I and type II errors](#) in the analyses.
- Dimensional reduction: Analysts often reduce the number of dimensions (genes) prior to data analysis. This may involve linear approaches such as principal components analysis (PCA), or non-linear manifold learning (distance metric learning) using kernel PCA, diffusion maps, Laplacian eigenmaps, local linear embedding, locally preserving projections, and Sammon's mapping.
- Network-based methods: Statistical methods that take the underlying structure of gene networks into account, representing either associative or causative interactions or dependencies among gene products. [Weighted gene co-expression network analysis](#) is widely used for identifying co-expression modules and intramodular hub genes. Modules may correspond to cell types or pathways. Highly connected intramodular hubs best represent their respective modules.

Microarray data may require further processing aimed at reducing the dimensionality of the data to aid comprehension and more focused analysis. Other methods permit analysis of data consisting of a low number of biological or technical [replicates](#); for example, the Local Pooled Error

(LPE) test pools [standard deviations](#) of genes with similar expression levels in an effort to compensate for insufficient replication.

Annotation

The relation between a probe and the [mRNA](#) that it is expected to detect is not trivial. Some mRNAs may cross-hybridize probes in the array that are supposed to detect another mRNA. In addition, mRNAs may experience amplification bias that is sequence or molecule-specific.

Thirdly, probes that are designed to detect the mRNA of a particular gene may be relying on genomic [EST](#) information that is incorrectly associated with that gene.

Data warehousing

Microarray data was found to be more useful when compared to other similar datasets. The sheer volume of data, specialized formats (such as [MIAME](#)), and curation efforts associated with the datasets require specialized databases to store the data. A number of open-source data warehousing solutions, such as [InterMine](#) and [BioMart](#), have been created for the specific purpose of integrating diverse biological datasets, and also support analysis.

Alternative technologies

Advances in massively parallel sequencing has led to the development of [RNA-Seq](#) technology, that enables a whole transcriptome shotgun approach to characterize and quantify gene expression. Unlike microarrays, which need a reference genome and transcriptome to be available before the microarray itself can be designed, RNA-Seq can also be used for new model organisms whose genome has not been sequenced yet.