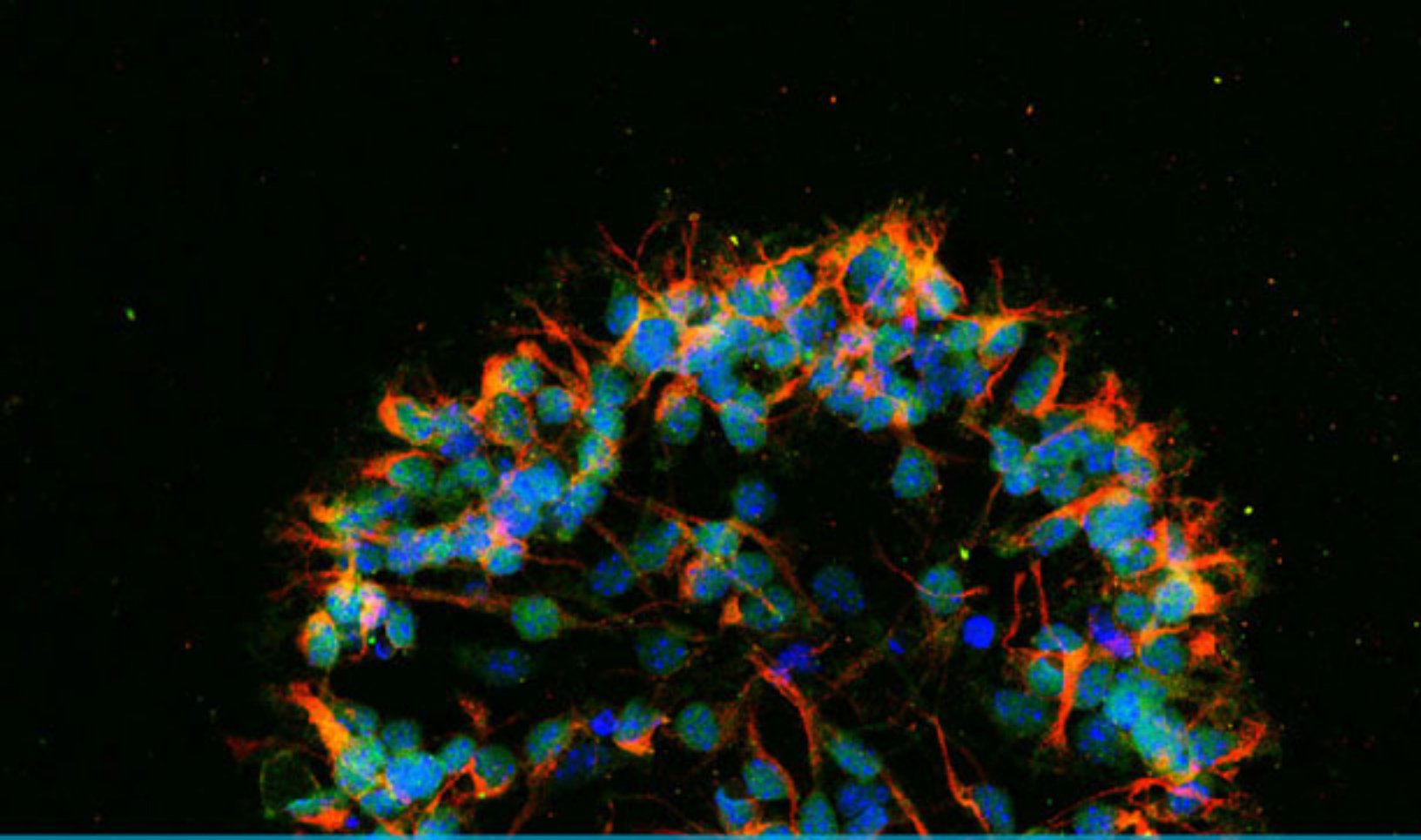


Introduction to Molecular Genetics

Kane Carlisle

First Edition, 2011

ISBN 978-93-81157-02-2

© All rights reserved.

Published by:
The English Press
4735/22 Prakashdeep Bldg,
Ansari Road, Darya Ganj,
Delhi - 110002
Email: info@wtbooks.com

Table of Contents

Chapter- 1

Molecular Genetics

Molecular genetics is the field of biology and genetics that studies the structure and function of genes at a molecular level. The field studies how the genes are transferred from generation to generation. Molecular genetics employs the methods of genetics and molecular biology. It is so-called to differentiate it from other sub fields of genetics such as ecological genetics and population genetics. An important area within molecular genetics is the use of molecular information to determine the patterns of descent, and therefore the correct scientific classification of organisms: this is called molecular systematics.

Along with determining the pattern of descendants, molecular genetics helps in understanding genetic mutations that can cause certain types of diseases. Through utilizing the methods of genetics and molecular biology, molecular genetics discovers the reasons why traits are carried on and how and why some may mutate.

Forward genetics

One of the first tools available to molecular geneticists is the forward genetic screen. The aim of this technique is to identify mutations that produce a certain phenotype. A mutagen is very often used to accelerate this process. Once mutants have been isolated, the mutated gene can be molecularly identified.

Reverse genetics

While forward genetic screens are productive, a more straightforward approach would be to determine the phenotype that results from mutating a given gene. This is called reverse genetics. In some organisms, such as yeast and mice, it is possible to induce the deletion of a particular gene, creating a gene knockout. Alternatives include the random induction of DNA deletions and subsequent selection for deletions in a gene of interest, the application of RNA interference and the creation of transgenic organisms that do not express a gene of interest.

Gene therapy

A mutation in a gene can result in a severe medical condition. A protein encoded by a mutated gene may malfunction and cells that rely on the protein might therefore fail to

function properly. This can cause problems for specific tissues or organs, or for the entire body. This might manifest through the course of development (like a cleft palate) or as an abnormal response to stimuli (like a peanut allergy). Conditions related to gene mutations are called genetic disorders. One way to fix such a physiological problem is gene therapy. By adding a corrected copy of the gene, a functional form of the protein can be produced, and affected cells, tissues, and organs may work properly. As opposed to drug-based approaches, gene therapy repairs the underlying genetic defect.

Gene therapy is the process of treating or alleviating diseases by genetically modifying the cells of the affected person, causing the gene to function properly. When a human disease gene has been recognized, molecular genetics tools can be used to explore the process of the gene in both the normal and mutant states. From there, the gene is transferred either in vivo or ex vivo and the body begins to make proteins according to the instructions in the new gene. Gene therapy has to be repeated several times for the infected patient to continually be relieved, however, as repeated cell division and death slowly randomizes the body's ratio of functional-to-mutant genes.

Currently, gene therapy is still being experimented with and products are not approved by the U.S. Food and Drug Administration. There have been several setbacks in the last 15 years that have restricted further developments in gene therapy. As there are unsuccessful attempts, there continue to be a growing number of successful gene therapy transfers which have furthered the research.

Major diseases that can be treated with gene therapy include viral infections, cancers, and inherited disorders, including immune system disorders.

Classical gene therapy

Classical gene therapy is the approach which delivers genes, via a modified virus or "vector" to the appropriate target cells with a goal of attaining optimal expression of the new, introduced gene. Once inside the patient, the expressed genes are intended to produce a product that the patient lacks, kill diseased cells directly by producing a toxin, or activate the immune system to help the killing of diseased cells.

Nonclassical gene therapy

Nonclassical gene therapy inhibits the expression of genes related to pathogenesis, or corrects a genetic defect and restores normal gene expression.

In vivo gene transfer

During In vivo gene transfer, the genes are transferred directly into the tissue of the patient and this can be the only possible option in patients with tissues where individual cells cannot be cultured in vitro in sufficient numbers (e.g. brain cells). Also, in vivo gene transfer is necessary when cultured cells cannot be re-implanted in patients effectively.

Ex vivo gene transfer

During ex vivo gene transfer the cells are cultured outside the body and then the genes are transferred into the cells grown in culture. The cells that have been transformed successfully are expanded by cell culture and then introduced into the patient.

Principles for gene transfer

Classical gene therapies usually require efficient transfer of cloned genes into the disease cells so that the introduced genes are expressed at sufficiently high levels to change the patient's physiology. There are several different physicochemical and biological methods that can be used to transfer genes into human cells. The size of the DNA fragments that can be transferred is very limited, and often the transferred gene is not a conventional gene. Horizontal gene transfer is the transfer of genetic material from one cell to another that is not its offspring. Artificial horizontal gene transfer is a form of genetic engineering.

Techniques in molecular genetics

There are three general techniques used for molecular genetics: amplification, separation and detection, and expression. Specifically used for amplification is polymerase chain reaction, which is an “indispensable tool in a great variety of applications” . In the separation and detection technique DNA and mRNA are isolated from their cells. Gene expression in cells or organisms is done in a place or time that is not normal for that specific gene.

Amplification

There are other methods for amplification besides polymerase chain reaction. Cloning DNA in bacteria is also a way to amplify DNA in genes.

Polymerase chain reaction

The main materials used in polymerase chain reaction are DNA nucleotides, template DNA, primers and Taq polymerase. DNA nucleotides are the base for the new DNA, the template DNA is the specific sequence being amplified, primers are complementary nucleotides that can go on either side of the template DNA, and Taq polymerase is a heat stable enzyme that jump-starts the production of new DNA at the high temperatures needed for reaction. This technique does not need to use living bacteria or cells; all that is needed is the base sequence of the DNA and materials listed above.

Cloning DNA in bacteria

The word cloning for this type of amplification entails making multiple identical copies of a sequence of DNA. The target DNA sequence is then inserted into a cloning vector.

Because this vector originates from a self-replicating virus, plasmid, or higher organism cell when the appropriate size DNA is inserted the "target and vector DNA fragments are then ligated" and create a recombinant DNA molecule. The recombinant DNA molecules are then put into a bacteria strain (usually *E. coli*) which produces several identical copies by transformation. Transformation is the DNA uptake mechanism possessed by bacteria. However, only one recombinant DNA molecule can be cloned within a single bacteria cell, so each clone is of just one DNA insert.

Separation and detection

In separation and detection DNA and mRNA are isolated from cells (the separation) and then detected simply by the isolation. Cell cultures are also grown to provide a constant supply of cells ready for isolation.

Cell cultures

A cell culture for molecular genetics is a culture that is grown in artificial conditions. Some cell types grow well in cultures such a skin cells, but other cells are not as productive in cultures. There are different techniques for each type of cell, some only recently being found to foster growth in stem and nerve cells. Cultures for molecular genetics are frozen in order to preserve all copies of the gene specimen and thawed only when needed. This allows for a steady supply of cells.

DNA isolation

DNA isolation extracts DNA from a cell in a pure form. First, the DNA is separated from cellular components such as proteins, RNA, and lipids. This is done by placing the chosen cells in a tube with a solution that mechanically, chemically, breaks the cells open. This solution contains enzymes, chemicals, and salts that breaks down the cells except for the DNA. It contains enzymes to dissolve proteins, chemicals to destroy all RNA present, and salts to help pull DNA out of the solution.

Next, the DNA is separated from the solution by being spun in a centrifuge, which allows the DNA to collect in the bottom of the tube. After this cycle in the centrifuge the solution is poured off and the DNA is resuspended in a second solution that makes the DNA easy to work with in the future.

This results in a concentrated DNA sample that contains thousands of copies of each gene. For large scale projects such as sequencing the human genome, all this work is done by robots.

mRNA isolation

Expressed DNA that codes for the synthesis of a protein is the final goal for scientists and this expressed DNA is obtained by isolation mRNA (Messenger RNA). First, laboratories use a normal cellular modification of mRNA that adds up to 200 adenine nucleotides to

the end of the molecule (poly(A) tail). Once this has been added, the cell is ruptured and its cell contents are exposed to synthetic beads that are coated with thymine string nucleotides. Because Adenine and Thymine pair together in DNA, the poly(A) tail and synthetic beads are attracted to one another, and once they bind in this process the cell components can be washed away without removing the mRNA. Once the mRNA has been isolated, reverse transcriptase is employed to convert it to single-stranded DNA, from which a stable double-stranded DNA is produced using DNA polymerase. Complementary DNA (cDNA) is much more stable than mRNA and so, once the double-stranded DNA has been produced it represents the expressed DNA sequence scientists look for.

The Human Genome Project

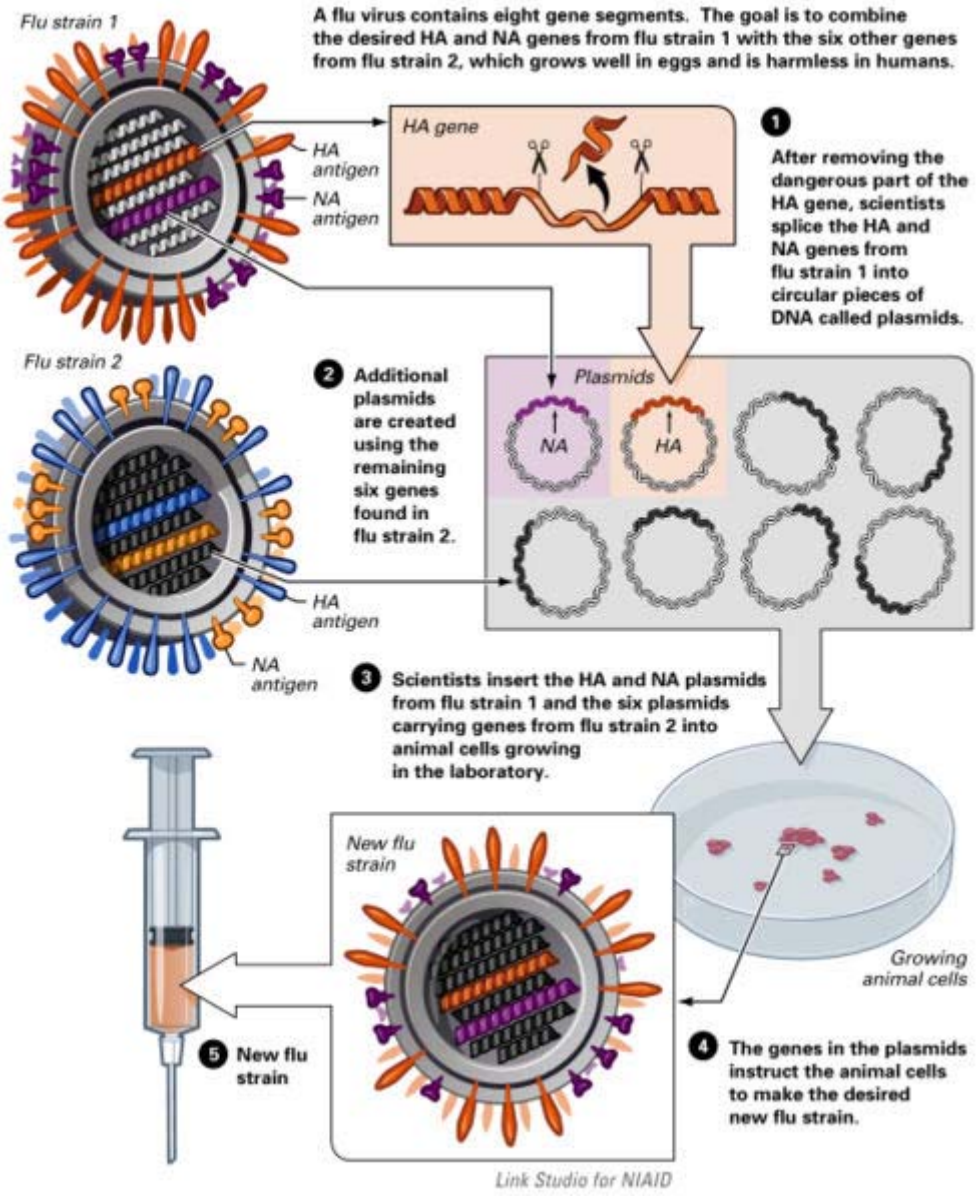
The Human Genome Project is a molecular genetics project that began in the 1990s and was projected to take fifteen years to complete. However, because of technological advances the progress of the project was advanced and the project finished in 2003, taking only thirteen years. The project was started by the U.S. Department of Energy and the National Institutes of Health in an effort to reach six set goals. These goals included:

1. identifying 20,000 to 25,000 genes in human DNA (although initial estimate were approximately 100,000 genes),
2. determining sequences of chemical based pairs in human DNA,
3. storing all found information into databases,
4. improving the tools used for data analysis,
5. transferring technologies to private sectors, and
6. addressing the ethical, legal, and social issues (ELSI) that may arise from the projects.

The project was worked on by eighteen different countries including the United States, Japan, France, Germany, and the United Kingdom. The collaborative effort resulted in the discovery of the many benefits of molecular genetics. Discoveries such as molecular medicine, new energy sources and environmental applications, DNA forensics, and livestock breeding, are only a few of the benefits that molecular genetics can provide.

Chapter- 2

Reverse Genetics



Avian Flu vaccine development by Reverse Genetics techniques.

'***Reverse genetics'*** is an approach to discovering the function of a gene by analyzing the phenotypic effects of specific gene sequences obtained by DNA sequencing. This investigative process proceeds in the opposite direction of so-called forward genetic screens of classical genetics. Simply put, while forward genetics seeks to find the genetic basis of a phenotype or trait, reverse genetics seeks to find what phenotypes arise as a result of particular genes.

Automated DNA sequencing generates large volumes of genomic sequence data relatively rapidly. Many genetic sequences are discovered in advance of other, less easily obtained, biological information. Reverse genetics attempts to connect a given genetic sequence with specific effects on the organism.

Techniques used in reverse genetics

To learn the influence a sequence has on phenotype, or to discover its biological function, researchers can engineer a change or disruption in the DNA. After this change has been made a researcher can look for the effect of such alterations in the whole organism. There are several different methods of reverse genetics that have proved useful:

Directed deletions and point mutations

Site-directed mutagenesis is a sophisticated technique that can either change regulatory regions in the promoter of a gene or make subtle codon changes in the open reading frame to identify important amino residues for protein function.

Alternatively, the technique can be used to create null alleles so that the gene is not functional. For example, deletion of a gene by gene targeting (**gene knockout**) can be done in some organisms, such as yeast, mice and moss. Unique among plants, in *Physcomitrella patens*, gene knockout via homologous recombination to create knockout moss is nearly as efficient as in yeast . In the case of the yeast model system directed deletions have been created in every non-essential gene in the yeast genome . In the case of the plant model system huge mutant libraries have been created based on gene disruption constructs .

In some cases conditional alleles can be used that have normal function until the allele is activated. This is known as **gene knock-in**. This might entail 'knocking in' recombinase sites (such as lox or frt sites) that will cause a deletion at the gene of interest when a specific recombinase (such as CRE, FLP) is induced. Cre or Flp recombinases can be induced with chemical treatments, heat shock treatments or be restricted to a specific subset of tissues.

Gene silencing

The discovery of gene silencing using double stranded RNA, also known as RNA interference (RNAi), and the development of gene knockdown using Morpholino oligonucleotides have made disrupting gene expression an accessible technique for many

more investigators. This method is often referred to as a **gene knockdown** since the effects of these reagents are generally temporary, in contrast to gene knockouts which are permanent.

RNAi creates a specific knockout effect without actually mutating the DNA of interest. In *C. elegans*, RNAi has been used to systematically interfere with the expression of most genes in the genome. RNAi acts by directing cellular systems to degrade target messenger RNA (mRNA).

While RNA interference relies on cellular components for efficacy (e.g. the Dicer proteins, the RISC complex) a simple alternative for gene knockdown is Morpholino antisense oligos. Morpholinos bind and block access to the target mRNA without requiring the activity of cellular proteins and without necessarily accelerating mRNA degradation. Morpholinos are effective in systems ranging in complexity from cell-free translation in a test tube to *in vivo* studies in large animal models.

Interference using transgenes

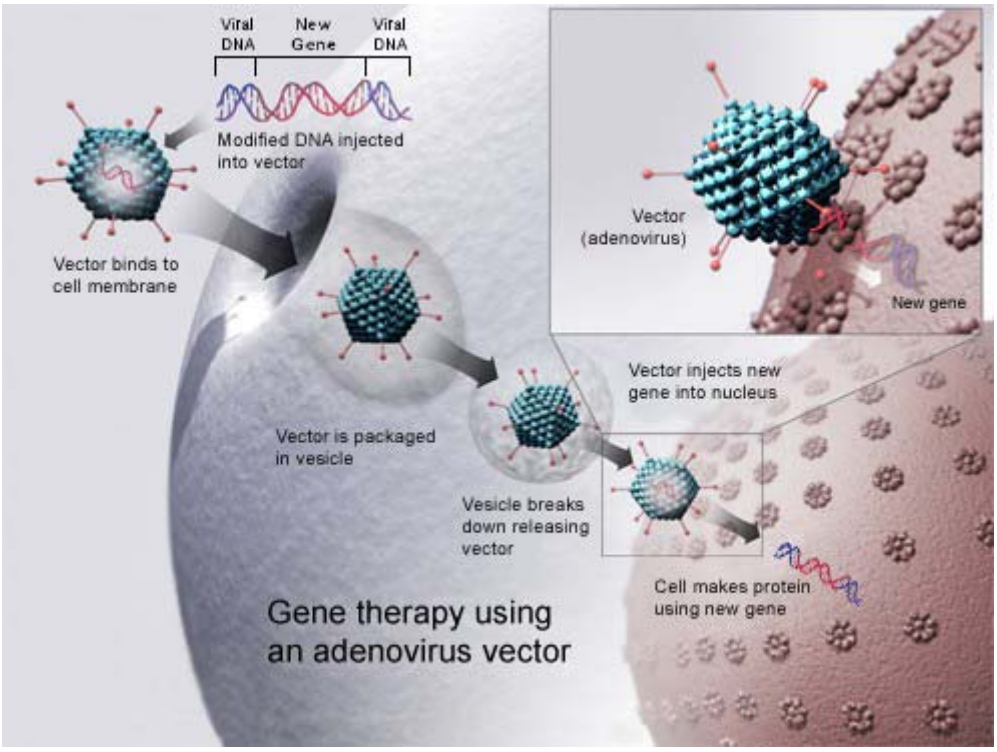
A molecular genetic approach is the creation of transgenic organisms that overexpress a normal gene of interest. The resulting phenotype may reflect the normal function of the gene.

Alternatively it is possible to overexpress mutant forms of a gene that interfere with the normal (wildtype) genes function. For example, over expression of a mutant gene may result in high levels of a non-functional protein resulting in a dominant negative interaction with the wildtype protein. In this case the mutant version will out compete for the wildtype proteins partners resulting in a mutant phenotype.

Other mutant forms can result in a protein that is abnormally regulated and constitutively active (‘on’ all the time). This might be due to removing a regulatory domain or mutating a specific amino residue that is reversibly modified (by phosphorylation methylation or ubiquitination). Either change is critical for modulating protein function and often result in informative phenotypes.

Chapter- 3

Gene Therapy



Gene therapy

Gene therapy is the insertion, alteration, or removal of genes within an individual's cells and biological tissues to treat disease. The most common form of gene therapy involves the insertion of functional genes into an unspecified genomic location in order to replace a mutated gene, but other forms involve directly correcting the mutation or modifying normal gene that enables a viral infection. Although the technology is still in its infancy, it has been used with some success. Scientific breakthroughs continue to move gene therapy toward mainstream medicine.

Approach

Scientists have taken the logical step of trying to introduce genes directly into human cells, focusing on diseases caused by single-gene defects, such as cystic fibrosis, haemophilia, muscular dystrophy and sickle cell anemia. However, this has proven more difficult than modifying bacteria, primarily because of the problems involved in carrying large sections of DNA and delivering them to the correct site on the gene. Today, most gene therapy studies are aimed at cancer and hereditary diseases linked to a genetic defect. Antisense therapy is not strictly a form of gene therapy, but is a related, genetically-mediated therapy.

The most common form of genetic engineering involves the insertion of a functional gene at an unspecified location in the host genome. This is accomplished by isolating and copying the gene of interest, generating a construct containing all the genetic elements for correct expression, and then inserting this construct into a random location in the host organism. Other forms of genetic engineering include gene targeting and knocking out specific genes via engineered nucleases such as zinc finger nucleases, engineered I-CreI homing endonucleases, or nucleases generated from TAL effectors. An example of gene-knockout mediated gene therapy is the knockout of the human CCR5 gene in T-cells in order to control HIV infection . This approach is currently being used in several human clinical trials.

The biology of human gene therapy remains complex and many techniques need further development. Many diseases and their strict genetic link need to be understood more fully before gene therapy can be used appropriately. The public policy debate surrounding the possible use of genetically engineered material in human subjects has been equally complex. Major participants in the debate have come from the fields of biology, government, law, medicine, philosophy, politics, and religion, each bringing different views to the discussion.

Types of gene therapy

Gene therapy may be classified into the two following types:

Germ line gene therapy

In the case of germ line gene therapy, germ cells, i.e., sperm or eggs, are modified by the introduction of functional genes, which are ordinarily integrated into their genomes. Therefore, the change due to therapy would be heritable and would be passed on to later generations. This new approach, theoretically, should be highly effective in counteracting genetic disorders and hereditary diseases. However, many jurisdictions prohibit this for application in human beings, at least for the present, for a variety of technical and ethical reasons.

Somatic gene therapy

In the case of somatic gene therapy, the therapeutic genes are transferred into the somatic cells of a patient. Any modifications and effects will be restricted to the individual patient only, and will not be inherited by the patient's offspring or later generations.

Vectors in gene therapy

Viruses

All viruses bind to their hosts and introduce their genetic material into the host cell as part of their replication cycle. This genetic material contains basic 'instructions' of how to produce more copies of these viruses, hijacking the body's normal production machinery to serve the needs of the virus. The host cell will carry out these instructions and produce additional copies of the virus, leading to more and more cells becoming infected. Some types of viruses insert their genes into the host's genome, but do not actually enter the cell. Others penetrate the cell membrane disguised as protein molecules and enter the cell.

There are two main types of virus infection: lytic and lysogenic. Shortly after inserting its DNA, viruses of the lytic cycle quickly produce more viruses, burst from the cell and infect more cells. Lysogenic viruses integrate their DNA into the DNA of the host cell and may live in the body for many years before responding to a trigger. The virus reproduces as the cell does and does not inflict bodily harm until it is triggered. The trigger releases the DNA from that of the host and employs it to create new viruses. HIV is a lysogenic infection. Some scientists believe that if they find the origin of its trigger, they will be able to stop the virus from ever reproducing throughout the body.

Retroviruses

The genetic material in retroviruses is in the form of RNA molecules, while the genetic material of their hosts is in the form of DNA. When a retrovirus infects a host cell, it will introduce its RNA together with some enzymes, namely reverse transcriptase and integrase, into the cell. This RNA molecule from the retrovirus must produce a DNA copy from its RNA molecule before it can be integrated into the genetic material of the host cell. The process of producing a DNA copy from an RNA molecule is termed reverse transcription. It is carried out by one of the enzymes carried in the virus, called reverse transcriptase. After this DNA copy is produced and is free in the nucleus of the host cell, it must be incorporated into the genome of the host cell. That is, it must be inserted into the large DNA molecules in the cell (the chromosomes). This process is done by another enzyme carried in the virus called integrase.

Now that the genetic material of the virus has been inserted, it can be said that the host cell has been modified to contain new genes. If this host cell divides later, its descendants will all contain the new genes. Sometimes the genes of the retrovirus do not express their information immediately.

One of the problems of gene therapy using retroviruses is that the integrase enzyme can insert the genetic material of the virus into any arbitrary position in the genome of the host; it randomly inserts the genetic material into a chromosome. If genetic material happens to be inserted in the middle of one of the original genes of the host cell, this gene will be disrupted (insertional mutagenesis). If the gene happens to be one regulating cell division, uncontrolled cell division (i.e., cancer) can occur. This problem has recently begun to be addressed by utilizing zinc finger nucleases or by including certain sequences such as the beta-globin locus control region to direct the site of integration to specific chromosomal sites.

Gene therapy trials using retroviral vectors to treat X-linked severe combined immunodeficiency (X-SCID) represent the most successful application of gene therapy to date. More than twenty patients have been treated in France and Britain, with a high rate of immune system reconstitution observed. Similar trials were restricted or halted in the USA when leukemia was reported in patients treated in the French X-SCID gene therapy trial. To date, four children in the French trial and one in the British trial have developed leukemia as a result of insertional mutagenesis by the retroviral vector. All but one of these children responded well to conventional anti-leukemia treatment. Gene therapy trials to treat SCID due to deficiency of the Adenosine Deaminase (ADA) enzyme continue with relative success in the USA, Britain, Italy and Japan.

Adenoviruses

Adenoviruses are viruses that carry their genetic material in the form of double-stranded DNA. They cause respiratory, intestinal, and eye infections in humans (especially the common cold). When these viruses infect a host cell, they introduce their DNA molecule into the host. The genetic material of the adenoviruses is not incorporated (transient) into the host cell's genetic material. The DNA molecule is left free in the nucleus of the host cell, and the instructions in this extra DNA molecule are transcribed just like any other gene. The only difference is that these extra genes are not replicated when the cell is about to undergo cell division so the descendants of that cell will not have the extra gene. As a result, treatment with the adenovirus will require readministration in a growing cell population although the absence of integration into the host cell's genome should prevent the type of cancer seen in the SCID trials. This vector system has been promoted for treating cancer and indeed the first gene therapy product to be licensed to treat cancer, Gendicine, is an adenovirus. Gendicine, an adenoviral p53-based gene therapy was approved by the Chinese FDA in 2003 for treatment of head and neck cancer. Advexin, a similar gene therapy approach from Introgen, was turned down by the US FDA in 2008.

Concerns about the safety of adenovirus vectors were raised after the 1999 death of Jesse Gelsinger while participating in a gene therapy trial. Since then, work using adenovirus vectors has focused on genetically crippled versions of the virus.

Adeno-associated viruses

Adeno-associated viruses, from the parvovirus family, are small viruses with a genome of single stranded DNA. The wild type AAV can insert genetic material at a specific site on chromosome 19 with near 100% certainty. But the recombinant AAV, which does not contain any viral genes and only the therapeutic gene, does not integrate into the genome. Instead the recombinant viral genome fuses at its ends via the ITR (inverted terminal repeats) recombination to form circular, episomal forms which are predicted to be the primary cause of the long term gene expression. There are a few disadvantages to using AAV, including the small amount of DNA it can carry (low capacity) and the difficulty in producing it. The production problem however has recently been solved by Amsterdam Molecular Therapeutics. This type of virus is being used, however, because it is non-pathogenic (most people carry this harmless virus). In contrast to adenoviruses, most people treated with AAV will not build an immune response to remove the virus and the cells that have been successfully treated with it. Several trials with AAV are on-going or in preparation, mainly trying to treat muscle and eye diseases; the two tissues where the virus seems particularly useful. However, clinical trials have also been initiated where AAV vectors are used to deliver genes to the brain. This is possible because AAV viruses can infect non-dividing (quiescent) cells, such as neurons in which their genomes are expressed for a long time.

Envelope protein pseudotyping of viral vectors

The viral vectors described above have natural host cell populations that they infect most efficiently. Retroviruses have limited natural host cell ranges, and although adenovirus and adeno-associated virus are able to infect a relatively broader range of cells efficiently, some cell types are refractory to infection by these viruses as well. Attachment to and entry into a susceptible cell is mediated by the protein envelope on the surface of a virus. Retroviruses and adeno-associated viruses have a single protein coating their membrane, while adenoviruses are coated with both an envelope protein and fibers that extend away from the surface of the virus. The envelope proteins on each of these viruses bind to cell-surface molecules such as heparin sulfate, which localizes them upon the surface of the potential host, as well as with the specific protein receptor that either induces entry-promoting structural changes in the viral protein, or localizes the virus in endosomes wherein acidification of the lumen induces this refolding of the viral coat. In either case, entry into potential host cells requires a favorable interaction between a protein on the surface of the virus and a protein on the surface of the cell. For the purposes of gene therapy, one might either want to limit or expand the range of cells susceptible to transduction by a gene therapy vector. To this end, many vectors have been developed in which the endogenous viral envelope proteins have been replaced by either envelope proteins from other viruses, or by chimeric proteins. Such chimera would consist of those parts of the viral protein necessary for incorporation into the virion as well as sequences meant to interact with specific host cell proteins. Viruses in which the envelope proteins have been replaced as described are referred to as pseudotyped viruses. For example, the most popular retroviral vector for use in gene therapy trials has been the lentivirus Simian immunodeficiency virus coated with the envelope proteins, G-protein, from Vesicular stomatitis virus. This vector is referred to as VSV G-pseudotyped lentivirus, and infects an almost universal set of cells. This tropism is characteristic of the VSV G-protein with

which this vector is coated. Many attempts have been made to limit the tropism of viral vectors to one or a few host cell populations. This advance would allow for the systemic administration of a relatively small amount of vector. The potential for off-target cell modification would be limited, and many concerns from the medical community would be alleviated. Most attempts to limit tropism have used chimeric envelope proteins bearing antibody fragments. These vectors show great promise for the development of "magic bullet" gene therapies.

Replication-Competent Vectors

A replication-competent vector called ONYX-015 is used in replicating tumor cells. It was found that in the absence of the E1B-55Kd viral protein, adenovirus caused very rapid apoptosis of infected, p53(+) cells, and this results in dramatically reduced virus progeny and no subsequent spread. Apoptosis was mainly the result of the ability of EIA to inactivate p300. In p53(-) cells, deletion of E1B 55kd has no consequence in terms of apoptosis, and viral replication is similar to that of wild-type virus, resulting in massive killing of cells.

A replication-defective vector deletes some essential genes. These deleted genes are still necessary in the body so they are replaced with either a helper virus or a DNA molecule.

Cis and trans-acting elements

Replication-defective vectors always contain a “transfer construct”. The transfer construct carries the gene to be transduced or “transgene”. The transfer construct also carries the sequences which are necessary for the general functioning of the viral genome: packaging sequence, repeats for replication and, when needed, priming of reverse transcription. These are denominated cis-acting elements, because they need to be on the same piece of DNA as the viral genome and the gene of interest. Trans-acting elements are viral elements, which can be encoded on a different DNA molecule. For example, the viral structural proteins can be expressed from a different genetic element than the viral genome.

Herpes Simplex Virus

Herpes Simplex Virus is a human neurotropic virus. This is mostly examined for gene transfer in the nervous system. The wild type HSV-1 virus is able to infect neurons. Infected neurones are not rejected by the immune system. Though the latent virus is not transcriptionally apparent, it does possess neurone specific promoters that can continue to function normally. Antibodies to HSV-1 are common in humans, however complications due to herpes infection are somewhat rare.

Non-viral methods

Non-viral methods present certain advantages over viral methods, with simple large scale production and low host immunogenicity being just two. Previously, low levels of

transfection and expression of the gene held non-viral methods at a disadvantage; however, recent advances in vector technology have yielded molecules and techniques with transfection efficiencies similar to those of viruses.

Injection of Naked DNA

This is the simplest method of non-viral transfection. Clinical trials carried out of intramuscular injection of a naked DNA plasmid have occurred with some success; however, the expression has been very low in comparison to other methods of transfection. In addition to trials with plasmids, there have been trials with naked PCR product, which have had similar or greater success. Cellular uptake of naked DNA is generally inefficient. Research efforts focusing on improving the efficiency of naked DNA uptake have yielded several novel methods, such as electroporation, sonoporation, and the use of a "gene gun", which shoots DNA coated gold particles into the cell using high pressure gas.

Physical Methods to Enhance Delivery

Electroporation

Electorporation is a method that uses short pulses of high voltage to carry DNA across the cell membrane. This shock is thought to cause temporary formation of pores in the cell membrane, allowing DNA molecules to pass through. Electroporation is generally efficient and works across a broad range of cell types. However, a high rate of cell death following electroporation has limited its use, including clinical applications.

More recently a newer method of electroporation, termed electron-avalanche transfection, has been used in gene therapy experiments. By using a high-voltage plasma discharge, DNA was efficiently delivered following very short (microsecond) pulses. Compared to electroporation, the technique resulted in greatly increased efficiency and less cellular damage.

Gene Gun

The use of particle bombardment, or the gene gun, is another physical method of DNA transfection. In this technique, DNA is coated with gold particles and loaded into a device which generates a force to achieve penetration of DNA/gold into the cells.

Sonoporation

Sonoporation uses ultrasonic frequencies to deliver DNA into cells. The process of acoustic cavitation is thought to disrupt the cell membrane and allow DNA to move into cells.

Sonoporation, or **cellular sonication**, is the use of sound (typically ultrasonic frequencies) for modifying the permeability of the cell plasma membrane. This technique

is usually used in molecular biology and non-viral gene therapy in order to allow uptake of large molecules such as DNA into the cell, in a cell disruption process called transfection or transformation. Sonoporation employs the acoustic cavitation of microbubbles to enhance delivery of these large molecules . The bioactivity of this technique is similar to, and in some cases found superior to, electroporation. Extended exposure to low-frequency (<MHz) ultrasound has been demonstrated to result in complete cellular death (rupturing), thus cellular viability must also be accounted for when employing this technique.

Sonoporation is under active study for the introduction of foreign genes in tissue culture cells, especially mammalian cells. Sonoporation is also being studied for use in targeted Gene therapy in vivo, in a medical treatment scenario whereby a patient is given modified DNA, and an ultrasonic transducer might target this modified DNA into specific regions of the patient's body.

Equipment

Sonoporation is performed with a commercially available sonoporator. Sonoporation may also be performed with general purpose piezoelectric transducers connected to bench-top function generators and acoustic amplifiers. Standard ultrasound medical devices may also be used in some applications.

Measurement of the acoustics used in sonoporation is listed in terms of mechanical index, which quantifies the likelihood that exposure to diagnostic ultrasound will produce an adverse biological effect by a nonthermal action based on pressure.

Microbubble Agents

Sonoporation uses microbubbles for significantly enhancing transfection, and in some cases is required for DNA uptake . These microbubble agents include Optison, manufactured by General Electric Healthcare.

Magnetofection

In a method termed magnetofection, DNA is complexed to a magnetic particles, and a magnet is placed underneath the tissue culture dish to bring DNA complexes into contact with a cell monolayer.

Chemical Methods to Enhance Delivery

Oligonucleotides

The use of synthetic oligonucleotides in gene therapy is to inactivate the genes involved in the disease process. There are several methods by which this is achieved. One strategy uses antisense specific to the target gene to disrupt the transcription of the faulty gene.

Another uses small molecules of RNA called siRNA to signal the cell to cleave specific unique sequences in the mRNA transcript of the faulty gene, disrupting translation of the faulty mRNA, and therefore expression of the gene. A further strategy uses double stranded oligodeoxynucleotides as a decoy for the transcription factors that are required to activate the transcription of the target gene. The transcription factors bind to the decoys instead of the promoter of the faulty gene, which reduces the transcription of the target gene, lowering expression. Additionally, single stranded DNA oligonucleotides have been used to direct a single base change within a mutant gene. The oligonucleotide is designed to anneal with complementarity to the target gene with the exception of a central base, the target base, which serves as the template base for repair. This technique is referred to as oligonucleotide mediated gene repair, targeted gene repair, or targeted nucleotide alteration.

Lipoplexes and polyplexes

To improve the delivery of the new DNA into the cell, the DNA must be protected from damage and its entry into the cell must be facilitated. To this end new molecules, lipoplexes and polyplexes, have been created that have the ability to protect the DNA from undesirable degradation during the transfection process.

Plasmid DNA can be covered with lipids in an organized structure like a micelle or a liposome. When the organized structure is complexed with DNA it is called a lipoplex. There are three types of lipids, anionic (negatively charged), neutral, or cationic (positively charged). Initially, anionic and neutral lipids were used for the construction of lipoplexes for synthetic vectors. However, in spite of the facts that there is little toxicity associated with them, that they are compatible with body fluids and that there was a possibility of adapting them to be tissue specific; they are complicated and time consuming to produce so attention was turned to the cationic versions.

Cationic lipids, due to their positive charge, were first used to condense negatively charged DNA molecules so as to facilitate the encapsulation of DNA into liposomes. Later it was found that the use of cationic lipids significantly enhanced the stability of lipoplexes. Also as a result of their charge, cationic liposomes interact with the cell membrane, endocytosis was widely believed as the major route by which cells uptake lipoplexes. Endosomes are formed as the results of endocytosis, however, if genes can not be released into cytoplasm by breaking the membrane of endosome, they will be sent to lysosomes where all DNA will be destroyed before they could achieve their functions. It was also found that although cationic lipids themselves could condense and encapsulate DNA into liposomes, the transfection efficiency is very low due to the lack of ability in terms of “endosomal escaping”. However, when helper lipids (usually electroneutral lipids, such as DOPE) were added to form lipoplexes, much higher transfection efficiency was observed. Later on, it was figured out that certain lipids have the ability to destabilize endosomal membranes so as to facilitate the escape of DNA from endosome, therefore those lipids are called fusogenic lipids. Although cationic liposomes have been widely used as an alternative for gene delivery vectors, a dose dependent toxicity of cationic lipids were also observed which could limit their therapeutic usages.

The most common use of lipoplexes has been in gene transfer into cancer cells, where the supplied genes have activated tumor suppressor control genes in the cell and decrease the activity of oncogenes. Recent studies have shown lipoplexes to be useful in transfecting respiratory epithelial cells, so they may be used for treatment of genetic respiratory diseases such as cystic fibrosis.

Complexes of polymers with DNA are called polyplexes. Most polyplexes consist of cationic polymers and their production is regulated by ionic interactions. One large difference between the methods of action of polyplexes and lipoplexes is that polyplexes cannot release their DNA load into the cytoplasm, so to this end, co-transfection with endosome-lytic agents (to lyse the endosome that is made during endocytosis, the process by which the polyplex enters the cell) such as inactivated adenovirus must occur. However, this isn't always the case, polymers such as polyethylenimine have their own method of endosome disruption as does chitosan and trimethylchitosan.

Dendrimers

A dendrimer is a highly branched macromolecule with a spherical shape. The surface of the particle may be functionalized in many ways and many of the properties of the resulting construct are determined by its surface.

In particular it is possible to construct a cationic dendrimer, i.e. one with a positive surface charge. When in the presence of genetic material such as DNA or RNA, charge complimentarity leads to a temporary association of the nucleic acid with the cationic dendrimer. On reaching its destination the dendrimer-nucleic acid complex is then taken into the cell via endocytosis.

In recent years the benchmark for transfection agents has been cationic lipids. Limitations of these competing reagents have been reported to include: the lack of ability to transfect a number of cell types, the lack of robust active targeting capabilities, incompatibility with animal models, and toxicity. Dendrimers offer robust covalent construction and extreme control over molecule structure, and therefore size. Together these give compelling advantages compared to existing approaches.

Producing dendrimers has historically been a slow and expensive process consisting of numerous slow reactions, an obstacle that severely curtailed their commercial development. The Michigan based company Dendritic Nanotechnologies discovered a method to produce dendrimers using kinetically driven chemistry, a process that not only reduced cost by a magnitude of three, but also cut reaction time from over a month to several days. These new "Priostar" dendrimers can be specifically constructed to carry a DNA or RNA payload that transfects cells at a high efficiency with little or no toxicity.

Hybrid methods

Due to every method of gene transfer having shortcomings, there have been some hybrid methods developed that combine two or more techniques. Virosomes are one example;

they combine liposomes with an inactivated HIV or influenza virus. This has been shown to have more efficient gene transfer in respiratory epithelial cells than either viral or liposomal methods alone. Other methods involve mixing other viral vectors with cationic lipids or hybridising viruses.

Major developments in gene therapy

1970s and earlier

In 1972 Friedmann and Roblin authored a paper in Science titled "Gene therapy for human genetic disease?" They cite Rogers S for proposing "that exogenous 'good'" DNA be used to replace the defective DNA in those who suffer from genetic defects. They also cite the first attempt to perform gene therapy as [New York Times, 20 September 1970].

1990

The first approved gene therapy case in the United States took place on September 14, 1990, at the National Institute of Health. It was performed on a four year old girl named Ashanti DeSilva. It was a treatment for a genetic defect that left her with an Immune System deficiency. The effects were only temporary, but successful (Boylan 313).

New gene therapy approach repairs errors in messenger RNA derived from defective genes. This technique has the potential to treat the blood disorder thalassaemia, cystic fibrosis, and some cancers. Researchers at Case Western Reserve University and Copernicus Therapeutics are able to create tiny liposomes 25 nanometers across that can carry therapeutic DNA through pores in the nuclear membrane.

Sickle cell disease is successfully treated in mice.

in 1992 Doctor Claudio Bordignon working at the Vita-Salute San Raffaele University, Milan, Italy performed the first procedure of gene therapy using hematopoietic stem cells as vectors to deliver genes intended to correct hereditary diseases. In 2002 this work led to the publication of the first successful gene therapy treatment for adenosine deaminase-deficiency (SCID). The success of a multi-center trial for treating children with SCID (severe combined immune deficiency or "bubble boy" disease) held from 2000 and 2002 was questioned when two of the ten children treated at the trial's Paris center developed a leukemia-like condition. Clinical trials were halted temporarily in 2002, but resumed after regulatory review of the protocol in the United States, the United Kingdom, France, Italy, and Germany.

In 1993 Andrew Gobea was born with severe combined immunodeficiency (SCID). Genetic screening before birth showed that he had SCID. Blood was removed from Andrew's placenta and umbilical cord immediately after birth, containing stem cells. The

allele that codes for ADA was obtained and was inserted into a retrovirus. Retroviruses and stem cells were mixed, after which they entered and inserted the gene into the stem cells' chromosomes. Stem cells containing the working ADA gene were injected into Andrew's blood system via a vein. Injections of the ADA enzyme were also given weekly. For four years T-cells (white blood cells), produced by stem cells, made ADA enzymes using the ADA gene. After four years more treatment was needed.

1995-2000

The 1999 death of Jesse Gelsinger in a gene therapy experiment resulted in a significant setback to gene therapy research in the United States. The pivotal event resulted in the FDA's suspension of several clinical trials as ethical and procedural practices in the field were reevaluated.

2001-2005

In 2003 a University of California, Los Angeles research team inserted genes into the brain using liposomes coated in a polymer called polyethylene glycol. The transfer of genes into the brain is a significant achievement because viral vectors are too big to get across the blood-brain barrier. This method has potential for treating Parkinson's disease.

RNA interference or gene silencing may be a new way to treat Huntington's disease. Short pieces of double-stranded RNA (short, interfering RNAs or siRNAs) are used by cells to degrade RNA of a particular sequence. If a siRNA is designed to match the RNA copied from a faulty gene, then the abnormal protein product of that gene will not be produced.

2005 to present

Scientists at the National Institutes of Health (Bethesda, Maryland) have successfully treated metastatic melanoma in two patients using killer T cells genetically retargeted to attack the cancer cells. This study constitutes one of the first demonstrations that gene therapy can be effective in treating cancer.

In March 2006 an international group of scientists announced the successful use of gene therapy to treat two adult patients for a disease affecting myeloid cells. The study, published in Nature Medicine, is believed to be the first to show that gene therapy can cure diseases of the myeloid system.

In May 2006 a team of scientists led by Dr. Luigi Naldini and Dr. Brian Brown from the San Raffaele Telethon Institute for Gene Therapy (HSR-TIGET) in Milan, Italy reported a breakthrough for gene therapy in which they developed a way to prevent the immune system from rejecting a newly delivered gene. Similar to organ transplantation, gene therapy has been plagued by the problem of immune rejection. So far, delivery of the 'normal' gene has been difficult because the immune system recognizes the new gene as foreign and rejects the cells carrying it. To overcome this problem, the HSR-TIGET

group utilized a newly uncovered network of genes regulated by molecules known as microRNAs. Dr. Naldini's group reasoned that they could use this natural function of microRNA to selectively turn off the identity of their therapeutic gene in cells of the immune system and prevent the gene from being found and destroyed. The researchers injected mice with the gene containing an immune-cell microRNA target sequence, and the mice did not reject the gene, as previously occurred when vectors without the microRNA target sequence were used. This work will have important implications for the treatment of hemophilia and other genetic diseases by gene therapy.

In November 2006 Preston Nix from the University of Pennsylvania School of Medicine reported on VRX496, a gene-based immunotherapy for the treatment of human immunodeficiency virus (HIV) that uses a lentiviral vector for delivery of an antisense gene against the HIV envelope. In the Phase I trial enrolling five subjects with chronic HIV infection who had failed to respond to at least two antiretroviral regimens, a single intravenous infusion of autologous CD4 T cells genetically modified with VRX496 was safe and well tolerated. All patients had stable or decreased viral load; four of the five patients had stable or increased CD4 T cell counts. In addition, all five patients had stable or increased immune response to HIV antigens and other pathogens. This was the first evaluation of a lentiviral vector administered in U.S. Food and Drug Administration-approved human clinical trials for any disease. Data from an ongoing Phase I/II clinical trial were presented at CROI 2009.

On 1 May 2007 Moorfields Eye Hospital and University College London's Institute of Ophthalmology announced the world's first gene therapy trial for inherited retinal disease. The first operation was carried out on a 23 year-old British male, Robert Johnson, in early 2007. Leber's congenital amaurosis is an inherited blinding disease caused by mutations in the RPE65 gene. The results of the Moorfields/UCL trial were published in New England Journal of Medicine in April 2008. They researched the safety of the subretinal delivery of recombinant adeno associated virus (AAV) carrying RPE65 gene, and found it yielded positive results, with patients having modest increase in vision, and, perhaps more importantly, no apparent side-effects.

In September 2009, the journal Nature reported that researchers at the University of Washington and University of Florida were able to give trichromatic vision to squirrel monkeys using gene therapy, a hopeful precursor to a treatment for color blindness in humans. In November 2009, the journal Science reported that researchers succeeded at halting a fatal brain disease, adrenoleukodystrophy, using a vector derived from HIV to deliver the gene for the missing enzyme.

A paper by Komáromy *et* al. published in April 2010, deals with gene therapy for a form of achromatopsia in dogs. Achromatopsia, or complete color blindness, is presented as an ideal model to develop gene therapy directed to cone photoreceptors. Cone function and day vision have been restored for at least 33 months in two young dogs with achromatopsia. However, the therapy was less efficient for older dogs.

Problems and ethics

For the safety of gene therapy, the Weismann barrier is fundamental in the current thinking. Soma-to-germline feedback should therefore be impossible. However, there are indications that the Weissman barrier can be breached. One way it might possibly be breached is if the treatment were somehow misapplied and spread to the testes and therefore would infect the germline against the intentions of the therapy.

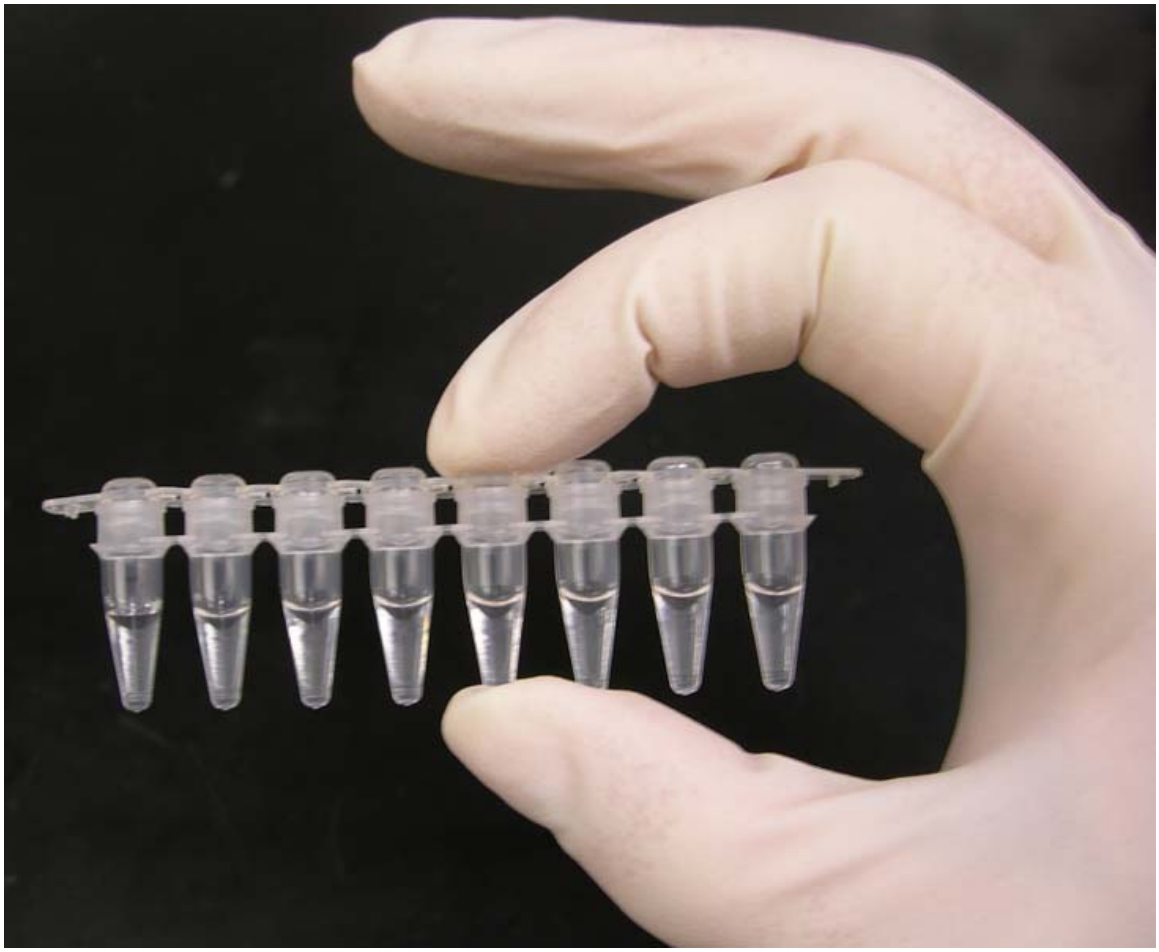
Some of the problems of gene therapy include:

- Short-lived nature of gene therapy – Before gene therapy can become a permanent cure for any condition, the therapeutic DNA introduced into target cells must remain functional and the cells containing the therapeutic DNA must be long-lived and stable. Problems with integrating therapeutic DNA into the genome and the rapidly dividing nature of many cells prevent gene therapy from achieving any long-term benefits. Patients will have to undergo multiple rounds of gene therapy.
- Immune response – Anytime a foreign object is introduced into human tissues, the immune system has evolved to attack the invader. The risk of stimulating the immune system in a way that reduces gene therapy effectiveness is always a possibility. Furthermore, the immune system's enhanced response to invaders that it has seen before makes it difficult for gene therapy to be repeated in patients.
- Problems with viral vectors – Viruses, the carrier of choice in most gene therapy studies, present a variety of potential problems to the patient —toxicity, immune and inflammatory responses, and gene control and targeting issues. In addition, there is always the fear that the viral vector, once inside the patient, may recover its ability to cause disease.
- Multigene disorders – Conditions or disorders that arise from mutations in a single gene are the best candidates for gene therapy. Unfortunately, some of the most commonly occurring disorders, such as heart disease, high blood pressure, Alzheimer's disease, arthritis, and diabetes, are caused by the combined effects of variations in many genes. Multigene or multifactorial disorders such as these would be especially difficult to treat effectively using gene therapy.
- Chance of inducing a tumor (insertional mutagenesis) - If the DNA is integrated in the wrong place in the genome, for example in a tumor suppressor gene, it 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.

Deaths have occurred due to gene therapy, including that of Jesse Gelsinger.

Chapter- 4

Polymerase Chain Reaction



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

The **polymerase chain reaction** (**PCR**) is a scientific technique in molecular biology to amplify a single or a few copies of a piece of DNA across several orders of magnitude, generating thousands to millions of copies of a particular DNA sequence. The method relies on thermal cycling, consisting of cycles of repeated heating and cooling of the reaction for DNA melting and enzymatic replication of the DNA. Primers (short DNA fragments) containing sequences complementary to the target region along with a DNA polymerase (after which the method is named) are key components to enable selective and repeated amplification. As PCR progresses, the DNA generated is itself used as a

template for replication, setting in motion a chain reaction in which the DNA template is exponentially amplified. PCR can be extensively modified to perform a wide array of genetic manipulations.

Almost all PCR applications employ a heat-stable DNA polymerase, such as Taq polymerase, an enzyme originally isolated from the bacterium *Thermus aquaticus*. This DNA polymerase enzymatically assembles a new DNA strand from DNA building blocks, the nucleotides, by using single-stranded DNA as a template and DNA oligonucleotides (also called DNA primers), which are required for initiation of DNA synthesis. The vast majority of PCR methods use thermal cycling, i.e., alternately heating and cooling the PCR sample to a defined series of temperature steps. These thermal cycling steps are necessary first to physically separate the two strands in a DNA double helix at a high temperature in a process called DNA melting. At a lower temperature, each strand is then used as the template in DNA synthesis by the DNA polymerase to selectively amplify the target DNA. The selectivity of PCR results from the use of primers that are complementary to the DNA region targeted for amplification under specific thermal cycling conditions.

Developed in 1983 by Kary Mullis, PCR is now a common and often indispensable technique used in medical and biological research labs for a variety of applications. These include DNA cloning for sequencing, DNA-based phylogeny, or functional analysis of genes; the diagnosis of hereditary diseases; the identification of genetic fingerprints (used in forensic sciences and paternity testing); and the detection and diagnosis of infectious diseases. In 1993, Mullis was awarded the Nobel Prize in Chemistry for his work on PCR.

PCR principles and procedure

Figure 1a: A thermal cycler for PCR

Figure 1b: An older model three-temperature thermal cycler for PCR

PCR is used to amplify a specific region of a DNA strand (the DNA target). Most PCR methods typically amplify DNA fragments of up to ~10 kilo base pairs (kb), although some techniques allow for amplification of fragments up to 40 kb in size.

A basic PCR set up requires several components and reagents. These components include:

- *DNA template* that contains the DNA region (target) to be amplified.
- Two *primers* that are complementary to the 3' (three prime) ends of each of the sense and anti-sense strand of the DNA target.

- *Taq polymerase* or another DNA polymerase with a temperature optimum at around 70 °C.
- *Deoxynucleoside triphosphates* (dNTPs; also very commonly and erroneously called deoxynucleotide triphosphates), the building blocks from which the DNA polymerases synthesizes a new DNA strand.
- *Buffer solution,* providing a suitable chemical environment for optimum activity and stability of the DNA polymerase.
- *Divalent cations,* magnesium or manganese ions; generally Mg^{2+} is used, but Mn^{2+} can be utilized for PCR-mediated DNA mutagenesis, as higher Mn^{2+} concentration increases the error rate during DNA synthesis
- *Monovalent cation* potassium ions.

The PCR is commonly carried out in a reaction 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. 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 thermocyclers lacking a heated lid require a layer of oil on top of the reaction mixture or a ball of wax inside the tube.

Procedure

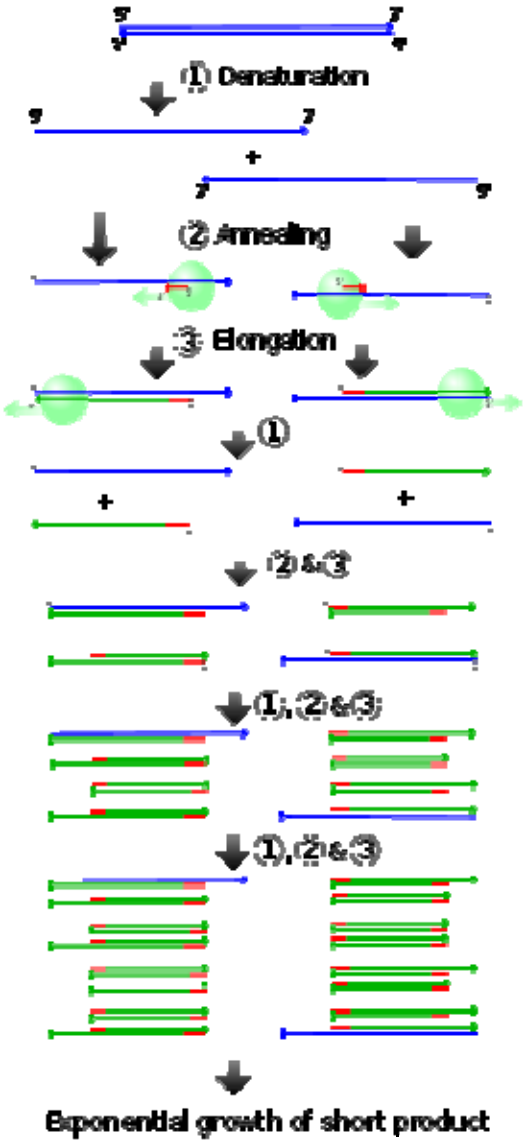


Figure 2: Schematic drawing of the PCR cycle. **(1) Denaturing at 94–96 °C. (2) Annealing at ~65 °C (3) Elongation at 72 °C**. Four cycles are shown here. The blue lines represent the DNA template to which primers (red arrows) anneal that are extended by the DNA polymerase (light green circles), to give shorter DNA products (green lines), which themselves are used as templates as PCR progresses.

Typically, PCR consists of a series of 20-40 repeated temperature changes, called cycles, with each cycle commonly consisting of 2-3 discrete temperature steps, usually three (Fig. 2). The cycling is often preceded by a single temperature step (called *hold*) at a high temperature (>90°C), 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. These include the enzyme used for DNA synthesis,

the concentration of divalent ions and dNTPs in the reaction, and the melting temperature (Tm) of the primers.

- *Initialization step*: This step consists of heating the reaction to a temperature of 94–96 °C (or 98 °C if extremely thermostable polymerases are used), which is held for 1–9 minutes. It is only required for DNA polymerases that require heat activation by hot-start PCR.

- *Denaturation step*: This step is the first regular cycling event and consists of heating the reaction to 94–98 °C for 20–30 seconds. It causes DNA melting of the DNA template by disrupting the hydrogen bonds between complementary bases, yielding single-stranded DNA molecules.

- *Annealing step*: The reaction temperature is lowered to 50–65 °C for 20–40 seconds allowing annealing of the primers to the single-stranded DNA template. Typically the annealing temperature is about 3-5 degrees Celsius below the Tm of the primers used. Stable DNA-DNA hydrogen bonds are only formed when the primer sequence very closely matches the template sequence. The polymerase binds to the primer-template hybrid and begins DNA synthesis.

- *Extension/elongation step*: The temperature at this step depends on the DNA polymerase used; Taq polymerase has its optimum activity temperature at 75–80 °C, and commonly a temperature of 72 °C is used with this enzyme. At this step the DNA polymerase synthesizes a new DNA strand complementary to the DNA template strand by adding dNTPs that are complementary to the template in 5' to 3' direction, condensing the 5'-phosphate group of the dNTPs with the 3'-hydroxyl group at the end of the nascent (extending) DNA strand. The extension time depends both on the DNA polymerase used and on the length of the DNA fragment to be amplified. As a rule-of-thumb, at its optimum temperature, the DNA polymerase will polymerize a thousand bases per minute. Under optimum conditions, i.e., if there are no limitations due to limiting substrates or reagents, at each extension step, the amount of DNA target is doubled, leading to exponential (geometric) amplification of the specific DNA fragment.

- *Final elongation*: This single step is occasionally performed at a temperature of 70–74 °C for 5–15 minutes after the last PCR cycle to ensure that any remaining single-stranded DNA is fully extended.

- *Final hold*: This step at 4–15 °C for an indefinite time may be employed for short-term storage of the reaction.

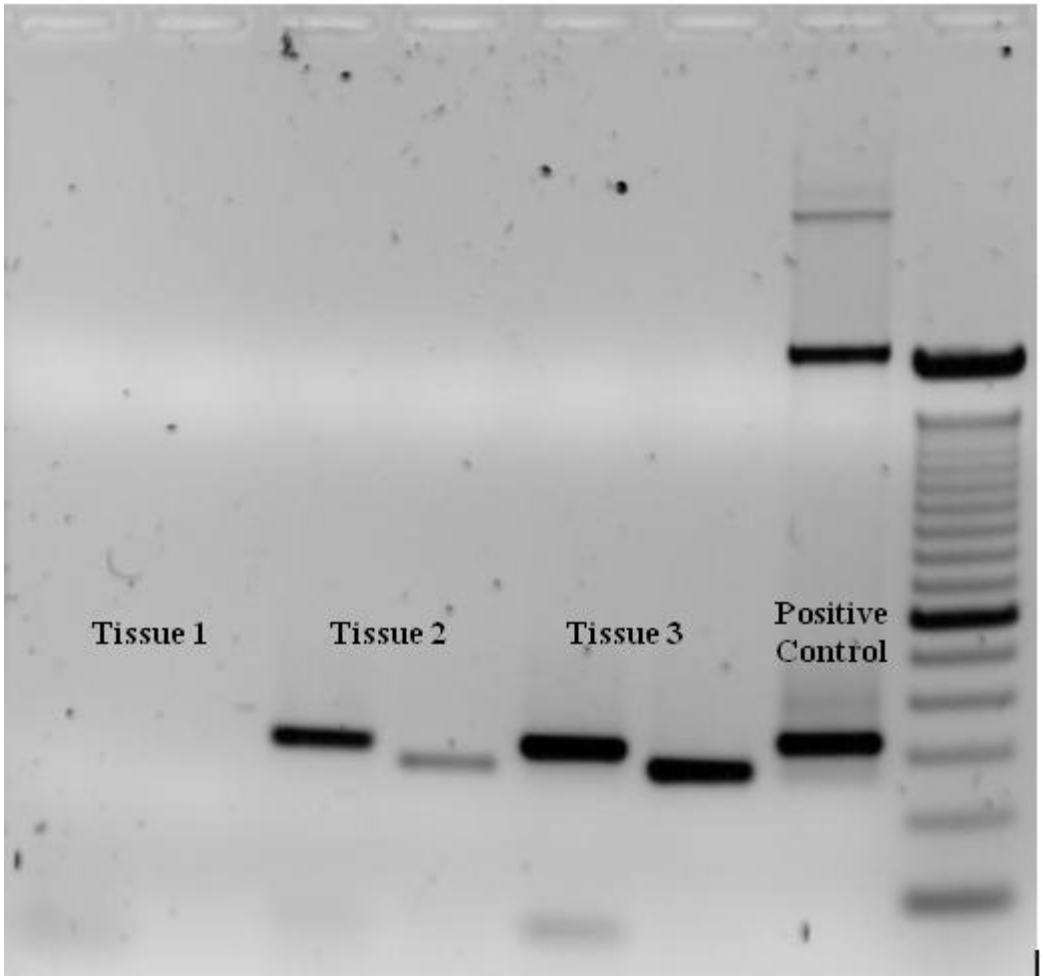


Figure 3: Ethidium bromide-stained PCR products after gel electrophoresis. Two sets of primers were used to amplify a target sequence from three different tissue samples. No amplification is present in sample #1; DNA bands in sample #2 and #3 indicate successful amplification of the target sequence. 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 generated the anticipated DNA fragment (also sometimes referred to as the amplimer or amplicon), agarose gel electrophoresis is 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 (see Fig. 3).

PCR stages

The PCR process can be divided into three stages:

Exponential amplification: At every cycle, the amount of product is doubled (assuming 100% reaction efficiency). The reaction is very sensitive: only minute quantities of DNA need to be present.

Levelling 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 limiting.

Plateau: No more product accumulates due to exhaustion of reagents and enzyme.

PCR 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. 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.

Application of PCR

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 methods, 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 or the genetic material of another organism. Bacterial colonies (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.

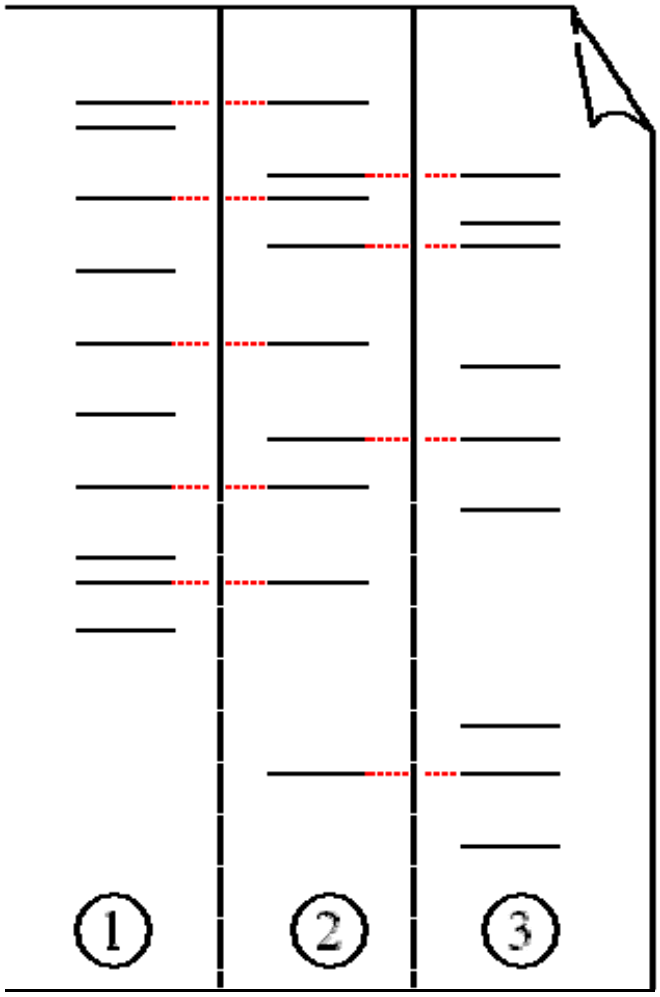


Figure 4: Electrophoresis of PCR-amplified DNA fragments. (1) Father. (2) Child. (3) Mother. The child has inherited some, but not all of the fingerprint of each of its parents, giving it a new, unique fingerprint.

Amplification and quantification of DNA

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.

Quantitative 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. Real-time PCR is an established tool for DNA quantification that measures the accumulation of DNA product after each round of PCR amplification.

PCR in diagnosis of diseases

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 which is at least 10,000 fold higher than other methods.

PCR also permits identification of non-cultivatable 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.

Viral DNA can likewise be detected by PCR. The primers used need to be specific to the targeted sequences in the DNA of a virus, and the 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 in treatment. The amount of virus ("viral load") in a patient can also be quantified by PCR-based DNA quantitation techniques.

Variations on the basic PCR technique

- *Allele-specific PCR*: a diagnostic or cloning technique which is based on single-nucleotide polymorphisms (SNPs) (single-base differences in DNA). It requires prior knowledge of a DNA sequence, including differences between alleles, and uses primers whose 3' ends encompass the SNP. 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.

- *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 *L*inear-*A*fter-*T*he-*E*xponential-PCR (LATE-PCR), uses a limiting primer with a higher melting temperature (Tm) than the excess primer to maintain reaction efficiency as the limiting primer concentration decreases mid-reaction.

- *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 melting 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 only dissociate 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.

- *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 Jim 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 multiple targets to be amplified with only a single primer pair, thus avoiding the resolution limitations of multiplex PCR.

- *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, i.e., their base pair length, should be different enough to form distinct bands when visualized by gel electrophoresis.

- *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*: a genetic engineering technique allowing the construction of a DNA sequence with an alteration inserted beyond the limit of the longest practical primer length.

- *Quantitative PCR (Q-PCR)*: used to measure the quantity of a PCR product (commonly in real-time). It quantitatively measures starting amounts of DNA, cDNA or RNA. Q-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 real-time PCR* has a very high degree of precision. QRT-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 (*R*eal *T*ime PCR) or RQ-PCR. QRT-PCR or RTQ-PCR are more appropriate contractions, since RT-PCR commonly refers to reverse transcription PCR (see below), often used in conjunction with Q-PCR.

- *R*everse *T*ranscription 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*).

- *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 Tm and nested solid support primer with optional application of a thermal 'step' to favour solid support priming).

- *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.

- *PAN-AC*: uses isothermal conditions for amplification, and may be used in living cells.

- *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

A 1971 paper in the Journal of Molecular Biology by Kleppe and co-workers first described a method 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, and the invention of the polymerase chain reaction in 1983 is generally credited to Kary Mullis.

At the core of the PCR method is the use of a suitable DNA polymerase able to withstand the high temperatures of >90 °C (194 °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, time consuming, and required large amounts of DNA polymerase and continual 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 °C (122 to 176 °F)) environments 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.

When Mullis developed the PCR in 1983, he was working in Emeryville, California for Cetus Corporation, one of the first biotechnology companies. There, he was responsible for synthesizing short chains of DNA. Mullis has written that he conceived of 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." He 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. However, 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).

Patent wars

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 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.

Chapter- 5

Polymerase Chain Reaction Optimization

The polymerase chain reaction (PCR) is a commonly used molecular biology tool for amplifying DNA, and various techniques for **PCR optimization** have been developed by molecular biologists to improve PCR performance and minimize failure.

Contamination and PCR

The PCR method is extremely sensitive, requiring only a few DNA molecules in a single reaction for amplification across several orders of magnitude. Therefore, adequate measures to avoid contamination from any DNA present in the lab environment (bacteria, viruses, or human sources) are required. Because products from previous PCR amplifications are a common source of contamination, many molecular biology labs have implemented procedures that involve dividing the lab into separate areas. One lab area is dedicated to preparation and handling of pre-PCR reagents and the setup of the PCR reaction, and another area to post-PCR processing, such as gel electrophoresis or PCR product purification. For the setup of PCR reactions, many standard operating procedures involve using pipettes with filter tips and wearing fresh laboratory gloves, and in some cases a laminar flow cabinet with UV lamp as a work station (to destroy any extraneomultimer]] formation is routinely assessed with a (negative) control PCR reaction. This control reaction is set up in the same way as the experimental PCRs, but without template DNA added, and is performed alongside the experimental PCRs.

Hairpins

Secondary structures in the DNA can result in folding or knotting of DNA template or primers, leading to decreased product yield or failure of the reaction. Hairpins, which consist of internal folds caused by base-pairing between nucleotides in inverted repeats within single-stranded DNA, are common secondary structures and may result in failed PCRs.

Typically, primer design that includes a check for potential secondary structures in the primers, or addition of DMSO or glycerol to the PCR to minimize secondary structures in

the DNA template , are used in the optimization of PCRs that have a history of failure due to suspected DNA hairpins.

Polymerase errors

Taq polymerase lacks a 3' to 5' exonuclease activity. Thus, Taq has no error-proof-reading activity, which consists of excision of any newly misincorporated nucleotide base from the nascent (=extending) DNA strand that does not match with its opposite base in the complementary DNA strand. The lack in 3' to 5' proofreading of the Taq enzyme results in a high error rate (mutations per nucleotide per cycle) of approximately 1 in 10,000 bases, which affects the fidelity of the PCR, especially if errors occur early in the PCR with low amounts of starting material, causing accumulation of a large proportion of amplified DNA with incorrect sequence in the final product.

Several "high-fidelity" DNA polymerases, having engineered 3' to 5' exonuclease activity, have become available that permit more accurate amplification for use in PCRs for sequencing or cloning of products. Examples of polymerases with 3' to 5' exonuclease activity include: KOD DNA polymerase, a recombinant form of *Thermococcus kodakaraensis* KOD1; Vent, which is extracted from *Thermococcus litoralis*; Pfu DNA polymerase, which is extracted from *Pyrococcus furiosus*; and Pwo, which is extracted from *Pyrococcus woesii*.

Magnesium concentration

Magnesium is required as a co-factor for thermostable DNA polymerase. Taq polymerase is a magnesium-dependent enzyme and determining the optimum concentration to use is critical to the success of the PCR reaction. Some of the components of the reaction mixture such as template concentration, dNTPs and the presence of chelating agents (EDTA) or proteins can reduce the amount of free magnesium present thus reducing the activity of the enzyme. Primers which bind to incorrect template sites are stabilized in the presence of excessive magnesium concentrations and so results in decreased specificity of the reaction. Excessive magnesium concentrations also stabilize double stranded DNA and prevent complete denaturation of the DNA during PCR reducing the product yield. Inadequate thawing of $MgCl_2$ may result in the formation of concentration gradients within the magnesium chloride solution supplied with the DNA polymerase and also contribute to many failed experiments.

Size and other limitations

PCR works readily with a DNA template of up to two to three thousand base pairs in length. However, above this size, product yields often decrease, as with increasing length stochastic effects such as premature termination by the polymerase begin to affect the efficiency of the PCR. It is possible to amplify larger pieces of up to 50,000 base pairs with a slower heating cycle and special polymerases. These are polymerases fused to a

processivity-enhancing DNA-binding protein, enhancing adherence of the polymerase to the DNA .

Other valuable properties of the chimeric polymerases TopoTaq and PfuC2 include enhanced thermostability, specificity and resistance to contaminants and inhibitors . They were engineered using the unique helix-hairpin-helix (HhH) DNA binding domains of topoisomerase V from hyperthermophile *Methanopyrus kandleri*. Chimeric polymerases overcome many limitations of native enzymes and are used in direct PCR amplification from cell cultures and even food samples, thus by-passing laborious DNA isolation steps. A robust strand-displacement activity of the hybrid TopoTaq polymerase helps solving PCR problems with hairpins and G-loaded double helices, because helices with a high G-C context possess a higher melting temperature.

Non-specific priming

Non-specific binding of primers frequently occurs and can be due to repeat sequences in the DNA template, non-specific binding between primer and template, and incomplete primer binding, leaving the 5' end of the primer unattached to the template. Non-specific binding is also often increased when degenerate primers are used in the PCR. Manipulation of annealing temperature and magnesium ion (which stabilise DNA and RNA interactions) concentrations can increase specificity. Non-specific priming during reaction preparation at lower temperatures can be prevented by using "hot-start" polymerase enzymes whose active site is blocked by an antibody or chemical that only dislodges once the reaction is heated to 95˚C during the denaturation step of the first cycle.

Other methods to increase specificity include Nested PCR and Touchdown PCR.

Touchdown polymerase chain reaction or touchdown style polymerase chain reaction is a method of polymerase chain reaction by which primers will avoid amplifying nonspecific sequences. The annealing temperature during a polymerase chain reaction determines the specificity of primer annealing. The melting point of the primer sets the upper limit on annealing temperature. At temperatures just below this point, only very specific base pairing between the primer and the template will occur. At lower temperatures, the primers bind less specifically. Nonspecific primer binding obscures polymerase chain reaction results, as the nonspecific sequences to which primers anneal in early steps of amplification will "swamp out" any specific sequences because of the exponential nature of polymerase amplification.

The earliest steps of a touchdown polymerase chain reaction cycle have high annealing temperatures. The annealing temperature is decreased in increments for every subsequent set of cycles (the number of individual cycles and increments of temperature decrease is chosen by the experimenter). The primer will anneal at the highest temperature which is least-permissive of nonspecific binding that it is able to tolerate. Thus, the first sequence amplified is the one between the regions of greatest primer specificity; it is most likely that this is the sequence of interest. These fragments will be further amplified during

subsequent rounds at lower temperatures, and will out compete the nonspecific sequences to which the primers may bind at those lower temperatures. If the primer initially (during the higher-temperature phases) binds to the sequence of interest, subsequent rounds of polymerase chain reaction can be performed upon the product to further amplify those fragments.

Primer dimers

Annealing of the 3' end of one primer to itself or the second primer may cause primer extension, resulting in the formation of so-called primer dimers, visible as low-molecular-weight bands on PCR gels. Primer dimer formation often competes with formation of the DNA fragment of interest, and may be avoided using primers that are designed such that they lack complementarity—especially at the 3' ends—to itself or the other primer used in the reaction. If primer design is constraint by other factors and if primer-dimers do occur, methods to limit their formation may include optimisation of the $MgCl_2$ concentration or increasing the annealing temperature in the PCR.

Deoxynucleotides

Deoxynucleotides (dNTPs) may bind Mg^{2+} ions and thus affect the concentration of free magnesium ions in the reaction. In addition, excessive amounts of dNTPs can increase the error rate of DNA polymerase and even inhibit the reaction. An imbalance in the proportion of the four dNTPs can result in misincorporation into the newly formed DNA strand and contribute to a decrease in the fidelity of DNA polymerase.

Chapter- 6

Molecular Cloning

Molecular cloning refers to the procedure of isolating a defined DNA sequence and obtaining multiple copies of it *in vitro*. Cloning is frequently employed to amplify DNA fragments containing genes, but it can be used to amplify any DNA sequence such as promoters, non-coding sequences, chemically synthesised oligonucleotides and randomly fragmented DNA. Cloning is used in a wide array of biological experiments and technological applications such as large scale protein production.

Overview

In essence, in order to amplify any DNA sequence *in vivo* and *in vitro*, the sequence in question must be linked to primary sequence elements capable of directing the replication and propagation of themselves and the linked sequence in the desired target host. The required sequence elements differ according to host, but invariably include an origin of replication, and a selectable marker. In practice, however, a number of other features are desired and a variety of specialized cloning vectors exist that allow protein expression, tagging, single stranded RNA and DNA production and a host of other manipulations that are useful in downstream applications.

Recombinase-based cloning

A novel procedure of cloning or subcloning of any DNA fragment by inserting the special DNA fragment of interest into a special area of target DNA through interchange of the relevant DNA fragments.

This is a one-step reaction: simple, efficient, facilitating high throughput or automatic cloning and/or subcloning.

One of the currently popular recombinase-based systems is marketed under the name Gateway Technology

Restriction/ligation cloning

In the classical restriction and ligation cloning protocols, cloning of any DNA fragment essentially involves four steps: DNA fragmentation with restriction endonucleases, ligation of DNA fragments to a vector, transfection, and screening/selection. Although these steps are invariable among cloning procedures a number of alternative routes can be selected at various points depending on the particular application; these are summarized as a 'cloning strategy'.

Isolation of insert

Initially, the DNA fragment to be cloned needs to be isolated. Preparation of DNA fragments for cloning can be accomplished in a number of alternative ways. Insert preparation is frequently achieved by means of polymerase chain reaction, but it may also be accomplished by restriction enzyme digestion, DNA sonication and fractionation by agarose gel electrophoresis. Chemically synthesized oligonucleotides can also be used if the target sequence size does not exceed the limit of chemical synthesis. Isolation of insert can be done by using shotgun cloning, c-DNA clones, gene machines (artificial chemical synthesis).

Transformation

Following ligation, the ligation product (plasmid) is transformed into bacteria for propagation. The bacteria is then plated on selective agar to select for bacteria that have the plasmid of interest. Individual colonies are picked and tested for the wanted insert. Maxiprep can be done to obtain large quantity of the plasmid containing the inserted gene.

Transfection

Following ligation, a portion of the ligation reaction, including vector with insert in the desired orientation is transfected into cells. A number of alternative techniques are available, such as chemical sensitization of cells, electroporation and biolistics. Chemical sensitization of cells is frequently employed since this does not require specialized equipment and provides relatively high transformation efficiencies. Electroporation is used when extremely high transformation efficiencies are required, as in very inefficient cloning strategies. Biolistics are mainly utilized in plant cell transformations, where the cell wall is a major obstacle in DNA uptake by cells. The bacterial transformation is generally observed by blue white screening.

Selection

Finally, the transfected cells are cultured. As the aforementioned procedures are of particularly low efficiency, there is a need to identify the cells that contain the desired insert at the appropriate orientation and isolate these from those not successfully transformed. Modern cloning vectors include selectable markers (most frequently antibiotic resistance markers) that allow only cells in which the vector, but not necessarily the insert, has been transfected to grow. Additionally, the cloning vectors may

contain colour selection markers which provide blue/white screening (via α-factor complementation) on X-gal medium. Nevertheless, these selection steps do not absolutely guarantee that the DNA insert is present in the cells. Further investigation of the resulting colonies is required to confirm that cloning was successful. This may be accomplished by means of PCR, restriction fragment analysis and/or DNA sequencing.

Genetic engineering

Genetic engineering is a method of changing the inherited characteristics of an organism in a predetermined way by altering its genetic material. This is often done to enable micro-organisms, such as bacteria or viruses, to synthesize increased yields of compounds, to form entirely new compounds, or to adapt to different environments. Other uses of this technology, which is also called recombinant DNA technology, include gene therapy, which is the supply of a functional gene to a person with a genetic disorder or with other diseases such as acquired immune deficiency syndrome (AIDS) or cancer, and the cloning of whole organisms.

Genetic engineering involves the manipulation of deoxyribonucleic acid, or DNA. Important tools in this process are restriction endonucleases (so-called restriction enzymes) that are produced by various species of bacteria. Restriction enzymes can recognize a particular sequence of the chain of chemical units, called nucleotide bases, which make up the DNA molecule and cut the DNA at that location. Fragments of DNA generated in this way can be joined using other enzymes called ligases. Restriction enzymes and ligases therefore allow the specific cutting and reassembling of portions of DNA. Also important in the manipulation of DNA are so-called vectors, which are pieces of DNA that can self-replicate (produce copies of themselves) independently of the DNA in the host cell in which they are grown. Examples of vectors include plasmids, viruses, and artificial chromosomes. Vectors permit the generation of multiple copies of a particular piece of DNA, making this a useful method for generating sufficient quantities of material with which to work. The process of engineering a DNA fragment into a vector is called “molecular cloning”, because multiple copies of an identical molecule of DNA are produced. Another way of producing many identical copies of a particular (often short, for example, 100-3,000 base pairs) DNA fragment is the polymerase chain reaction. This method is rapid and avoids the need for cloning DNA into a vector.

Gene therapy

Gene therapy involves supplying a functional gene to cells lacking that function, with the aim of correcting a genetic disorder or acquired disease. Gene therapy can be broadly divided into two categories. The first is alteration of germ cells, that is, sperm or eggs, which results in a permanent genetic change for the whole organism and subsequent generations. This “germ line gene therapy” is considered by many to be unethical in human beings. The second type of gene therapy, “somatic cell gene therapy”, is analogous to an organ transplant. In this case, one or more specific tissues are targeted by direct treatment or by removal of the tissue, addition of the therapeutic gene or genes in the laboratory, and return of the treated cells to the patient. Clinical trials of somatic cell

gene therapy began in the late 1990s, mostly for the treatment of cancers and blood, liver, and lung disorders.

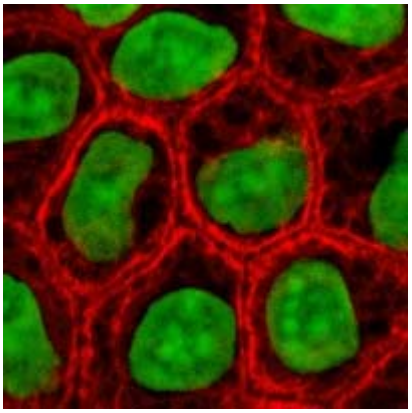
The history of human gene therapy is, however, not a particularly happy one. The effect of introducing a gene into cells rarely promotes more than small transient relief from the symptoms of the disease being treated. Worse still, there have been highly publicized cases where gene therapy trial patients have suffered as a consequence of the treatment itself. For example, in 1999 an 18-year-old gene therapy trial volunteer from Philadelphia died following a gene therapy trial. In addition, one of the few success stories of human gene therapy—the treatment of severe combined immune deficiency, X-SCID—has turned out to have unforeseen consequences. Bone marrow cells were taken from patients suffering from this disease and treated with a virus to introduce a functional copy of the defective gene. When the modified bone marrow cells were returned to patients, their immune systems were functional once more. However, some patients treated this way subsequently developed leukaemia, which most likely arises as a result of random insertion of a section of DNA into the human genome with the consequent disruption of nearby gene function.

Cloning cells and animals

In genetic engineering, the term “cloning” is now more commonly applied to the production of identical animals rather than molecular cloning of DNA fragments. Whole cell or animal cloning occurs through the transfer of the nucleus of an adult cell into an enucleated egg. This can result in the reprogramming of the adult cell DNA to produce a cloned animal. In 1997, a sheep named Dolly was born at the Roslin Institute in Edinburgh. She was created from the nucleus of a cultured mammary gland cell from a Finn Dorset sheep that was fused to an egg cell from a Scottish Blackface ewe that had had its own nucleus removed. The fused cell was implanted into a different Scottish Blackface ewe, and following a normal pregnancy, Dolly, a Finn Dorset sheep, was born. Nuclear transfer has subsequently been applied to produce a range of cloned animals including cows, goats, pigs, mice, and cats.

Chapter- 7

Cell Culture



Epithelial cells in culture, stained for keratin (red) and DNA (green)

Cell culture is the complex process by which cells are grown under controlled conditions. In practice, the term "cell culture" has come to refer to the culturing of cells derived from multicellular eukaryotes, especially animal cells. However, there are also cultures of plants, fungi and microbes, including viruses, bacteria and protists. The historical development and methods of cell culture are closely interrelated to those of tissue culture and organ culture.

Animal cell culture became a common laboratory technique in the mid-1900s, but the concept of maintaining live cell lines separated from their original tissue source was discovered in the 19th century.

History

The 19th-century English physiologist Sydney Ringer developed salt solutions containing the chlorides of sodium, potassium, calcium and magnesium suitable for maintaining the beating of an isolated animal heart outside of the body. In 1885 Wilhelm Roux removed a portion of the medullary plate of an embryonic chicken and maintained it in a warm saline solution for several days, establishing the principle of tissue culture. Ross Granville Harrison, working at Johns Hopkins Medical School and then at Yale University, published results of his experiments from 1907–1910, establishing the methodology of tissue culture.

Cell culture techniques were advanced significantly in the 1940s and 1950s to support research in virology. Growing viruses in cell cultures allowed preparation of purified viruses for the manufacture of vaccines. The injectable polio vaccine developed by Jonas Salk was one of the first products mass-produced using cell culture techniques. This vaccine was made possible by the cell culture research of John Franklin Enders, Thomas Huckle Weller, and Frederick Chapman Robbins, who were awarded a Nobel Prize for their discovery of a method of growing the virus in monkey kidney cell cultures.

Concepts in mammalian cell culture

Isolation of cells

Cells can be isolated from tissues for *ex vivo* culture in several ways. Cells can be easily purified from blood, however only the white cells are capable of growth in culture. Mononuclear cells can be released from soft tissues by *enzymatic digestion* with enzymes such as collagenase, trypsin, or pronase, which break down the extracellular matrix. Alternatively, pieces of tissue can be placed in growth media, and the cells that grow out are available for culture. This method is known as *explant culture*.

Cells that are cultured directly from a subject are known as ***primary cells***. With the exception of some derived from tumors, most primary cell cultures have limited lifespan. After a certain number of population doublings (called the Hayflick limit) cells undergo the process of senescence and stop dividing, while generally retaining viability.

An established or **immortalised** *cell line* has acquired the ability to proliferate indefinitely either through random mutation or deliberate modification, such as artificial expression of the telomerase gene. There are numerous well established cell lines representative of particular cell types.

Maintaining cells in culture

Cells are grown and maintained at an appropriate temperature and gas mixture (typically, 37°C, 5% CO_2 for mammalian cells) in a cell incubator. Culture conditions vary widely for each cell type, and variation of conditions for a particular cell type can result in different phenotypes being expressed.

Aside from temperature and gas mixture, the most commonly varied factor in culture systems is the growth medium. Recipes for growth media can vary in pH, glucose concentration, growth factors, and the presence of other nutrients. The growth factors used to supplement media are often derived from animal blood, such as calf serum. One complication of these blood-derived ingredients is the potential for contamination of the culture with viruses or prions, particularly in biotechnology medical applications. Current practice is to minimize or eliminate the use of these ingredients wherever possible, but this cannot always be accomplished. Alternative strategies involve sourcing the animal blood from countries with minimum BSE/TSE risk such as Australia and New Zealand,

and using purified nutrient concentrates derived from serum in place of whole animal serum for cell culture.

Plating density (number of cells per volume of culture medium) plays a critical role for some cell types. For example, a lower plating density makes granulosa cells exhibit estrogen production, while a higher plating density makes them appear as progesterone producing theca lutein cells.

Cells can be grown in *suspension* or *adherent* cultures. Some cells naturally live in suspension, without being attached to a surface, such as cells that exist in the bloodstream. There are also cell lines that have been modified to be able to survive in suspension cultures so that they can be grown to a higher density than adherent conditions would allow. Adherent cells require a surface, such as tissue culture plastic or microcarrier, which may be coated with extracellular matrix components to increase adhesion properties and provide other signals needed for growth and differentiation. Most cells derived from solid tissues are adherent. Another type of adherent culture is *organotypic culture* which involves growing cells in a three-dimensional environment as opposed to two-dimensional culture dishes. This 3D culture system is biochemically and physiologically more similar to *in vivo* tissue, but is technically challenging to maintain because of many factors (e.g. diffusion).

Cell line cross-contamination

Cell line cross-contamination can be a problem for scientists working with cultured cells. Studies suggest that anywhere from 15–20% of the time, cells used in experiments have been misidentified or contaminated with another cell line. Problems with cell line cross contamination have even been detected in lines from the NCI-60 panel, which are used routinely for drug-screening studies. Major cell line repositories including the American Type Culture Collection (ATCC) and the German Collection of Microorganisms and Cell Cultures (DSMZ) have received cell line submissions from researchers that were misidentified by the researcher. Such contamination poses a problem for the quality of research produced using cell culture lines, and the major repositories are now authenticating all cell line submissions. ATCC uses short tandem repeat (STR) DNA fingerprinting to authenticate its cell lines.

To address this problem of cell line cross-contamination, researchers are encouraged to authenticate their cell lines at an early passage to establish the identity of the cell line. Authentication should be repeated before freezing cell line stocks, every two months during active culturing and before any publication of research data generated using the cell lines. There are many methods for identifying cell lines including isoenzyme analysis, human lymphocyte antigen (HLA) typing, Chromosomal analysis, Karyotyping, Morphology and STR analysis.

One significant cell-line cross contaminant is the immortal HeLa cell line.

Manipulation of cultured cells

As cells generally continue to divide in culture, they generally grow to fill the available area or volume. This can generate several issues:

- Nutrient depletion in the growth media
- Accumulation of apoptotic/necrotic (dead) cells.
- Cell-to-cell contact can stimulate cell cycle arrest, causing cells to stop dividing known as contact inhibition or senescence.
- Cell-to-cell contact can stimulate cellular differentiation.

Among the common manipulations carried out on culture cells are media changes, passaging cells, and transfecting cells. These are generally performed using tissue culture methods that rely on sterile technique. Sterile technique aims to avoid contamination with bacteria, yeast, or other cell lines. Manipulations are typically carried out in a biosafety hood or laminar flow cabinet to exclude contaminating micro-organisms. Antibiotics (e.g. penicillin and streptomycin) and antifungals (e.g. Amphotericin B) can also be added to the growth media.

As cells undergo metabolic processes, acid is produced and the pH decreases. Often, a pH indicator is added to the medium in order to measure nutrient depletion.

Media changes

In the case of adherent cultures, the media can be removed directly by aspiration and replaced.

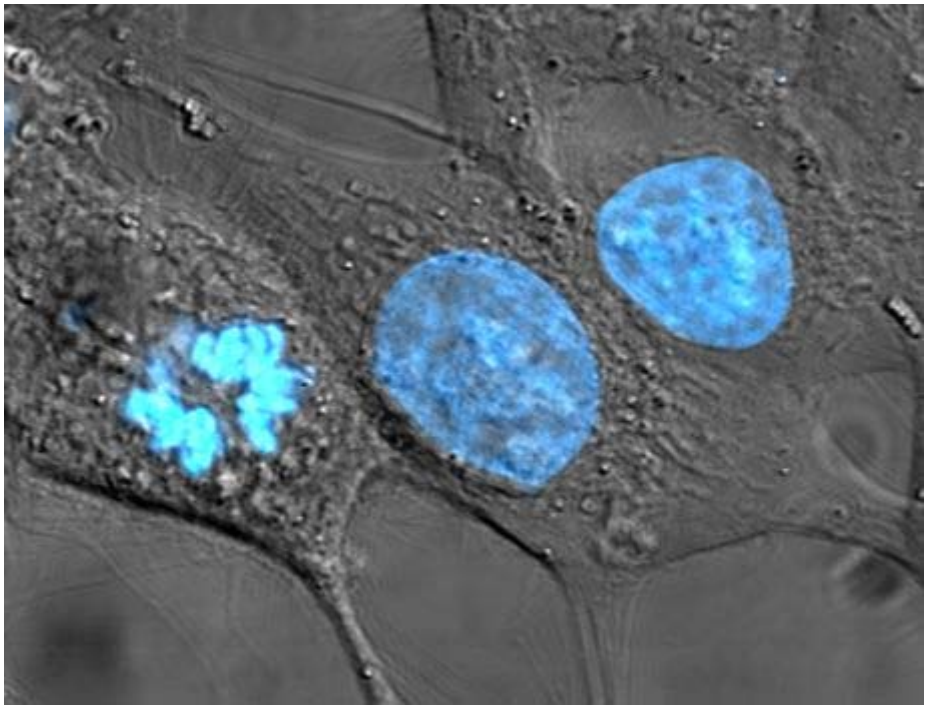
Passaging cells

Passaging (also known as subculture or splitting cells) involves transferring a small number of cells into a new vessel. Cells can be cultured for a longer time if they are split regularly, as it avoids the senescence associated with prolonged high cell density. Suspension cultures are easily passaged with a small amount of culture containing a few cells diluted in a larger volume of fresh media. For adherent cultures, cells first need to be detached; this is commonly done with a mixture of trypsin-EDTA, however other enzyme mixes are now available for this purpose. A small number of detached cells can then be used to seed a new culture.

Transfection and transduction

Another common method for manipulating cells involves the introduction of foreign DNA by transfection. This is often performed to cause cells to express a protein of interest. More recently, the transfection of RNAi constructs have been realized as a convenient mechanism for suppressing the expression of a particular gene/protein. DNA can also be inserted into cells using viruses, in methods referred to as transduction, infection or transformation. Viruses, as parasitic agents, are well suited to introducing DNA into cells, as this is a part of their normal course of reproduction.

Established human cell lines



One of the earliest human cell lines, descended from Henrietta Lacks, who died of the cancer that those cells originated from, the cultured HeLa cells shown here have been stained with Hoechst turning their nuclei blue.

Cell lines that originate with humans have been somewhat controversial in bioethics, as they may outlive their parent organism and later be used in the discovery of lucrative medical treatments. In the pioneering decision in this area, the Supreme Court of California held in *Moore v. Regents of the University of California* that human patients have no property rights in cell lines derived from organs removed with their consent.

Generation of hybridomas

It is possible to fuse normal cells with an immortalised cell line. This method is used to produce monoclonal antibodies. In brief, lymphocytes isolated from the spleen (or possibly blood) of an immunised animal are combined with an immortal myeloma cell line (B cell lineage) to produce a hybridoma which has the antibody specificity of the primary lymphoctye and the immortality of the myeloma. Selective growth medium (HA or HAT) is used to select against unfused myeloma cells; primary lymphoctyes die quickly in culture and only the fused cells survive. These are screened for production of the required antibody, generally in pools to start with and then after single cloning.

Applications of cell culture

Mass culture of animal cell lines is fundamental to the manufacture of viral vaccines and other products of biotechnology

Biological products produced by recombinant DNA (rDNA) technology in animal cell cultures include enzymes, synthetic hormones, immunobiologicals (monoclonal antibodies, interleukins, lymphokines), and anticancer agents. Although many simpler proteins can be produced using rDNA in bacterial cultures, more complex proteins that are glycosylated (carbohydrate-modified) currently must be made in animal cells. An important example of such a complex protein is the hormone erythropoietin. The cost of growing mammalian cell cultures is high, so research is underway to produce such complex proteins in insect cells or in higher plants, use of single embryonic cell and somatic embryos as a source for direct gene transfer via practicle bombardment, transit gene expression and confocal microscopy observation is one of its applications. It also offers to confirm single cell origin of somatic embryos and the asymmetry of the first cell division, which starts the process. -

Tissue culture and engineering

Cell culture is a fundamental component of tissue culture and tissue engineering, as it establishes the basics of growing and maintaining cells *ex vivo*. The major application of human cell culture is in stem cell industry where mesenchymal stem cells can be cultured and cryopreserved for future use.

Vaccines

Vaccines for polio, measles, mumps, rubella, and chickenpox are currently made in cell cultures. Due to the H5N1 pandemic threat, research into using cell culture for influenza vaccines is being funded by the United States government. Novel ideas in the field include recombinant DNA-based vaccines, such as one made using human adenovirus (a common cold virus) as a vector, , such as adjuvants.

Culture of non-mammalian cells

Plant cell culture methods

Plant cell cultures are typically grown as cell suspension cultures in liquid medium or as callus cultures on solid medium. The culturing of undifferentiated plant cells and calli requires the proper balance of the plant growth hormones auxin and cytokinin.

Bacterial and yeast culture methods

For bacteria and yeast, small quantities of cells are usually grown on a solid support that contains nutrients embedded in it, usually a gel such as agar, while large-scale cultures are grown with the cells suspended in a nutrient broth.

Viral culture methods

The culture of viruses requires the culture of cells of mammalian, plant, fungal or bacterial origin as hosts for the growth and replication of the virus. Whole wild type viruses, recombinant viruses or viral products may be generated in cell types other than their natural hosts under the right conditions. Depending on the species of the virus, infection and viral replication may result in host cell lysis and formation of a viral plaque.

Common cell lines

Human cell lines

- National Cancer Institute's 60 cancer cell lines
- ESTDAB database http://www.ebi.ac.uk/ipd/estdab/directory.html

- DU145 (Prostate cancer)
- Lncap (Prostate cancer)
- MCF-7 (breast cancer)
- MDA-MB-438 (breast cancer)
- PC3 (Prostate cancer)
- T47D (breast cancer)
- THP-1 (acute myeloid leukemia)
- U87 (glioblastoma)
- SHSY5Y Human neuroblastoma cells, cloned from a myeloma
- Saos-2 cells (bone cancer)

Primate cell lines

- Vero (African green monkey *Chlorocebus* kidney epithelial cell line initiated 1962)

Rat tumor cell lines

- GH3 (pituitary tumor)
- PC12 (pheochromocytoma)

Mouse cell lines

- MC3T3 (embryonic calvarial)

Plant cell lines

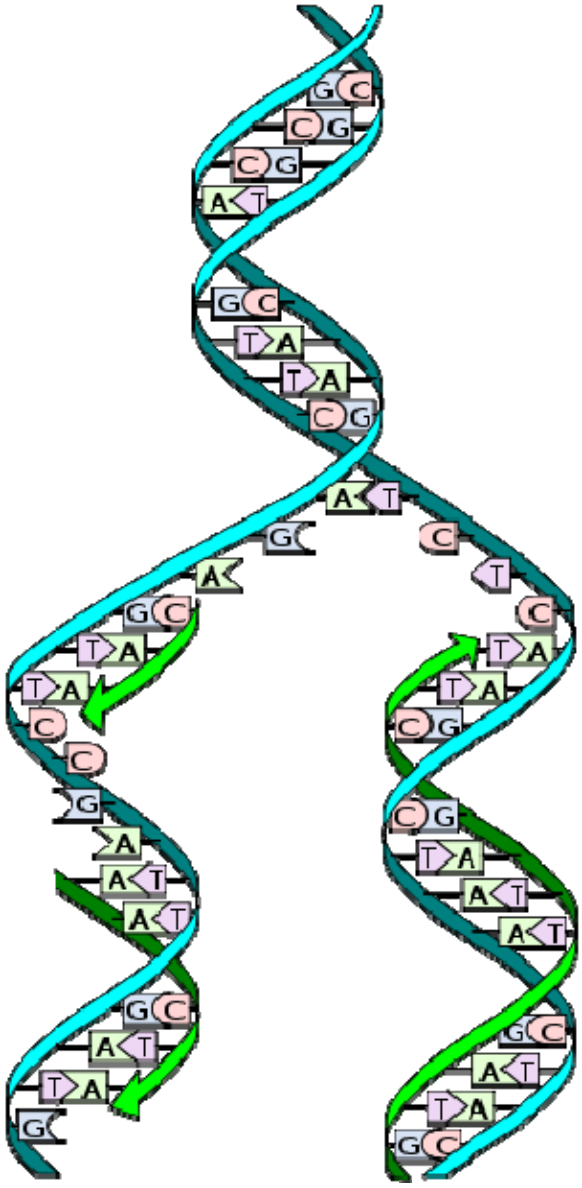
- Tobacco BY-2 cells (kept as cell suspension culture, they are model system of plant cell)

Other species cell lines

- zebrafish ZF4 and AB9 cells.
- *Madin-Darby Canine Kidney (MDCK)* epithelial cell line
- Xenopus A6 kidney epithelial cells.

Chapter- 8

DNA Replication



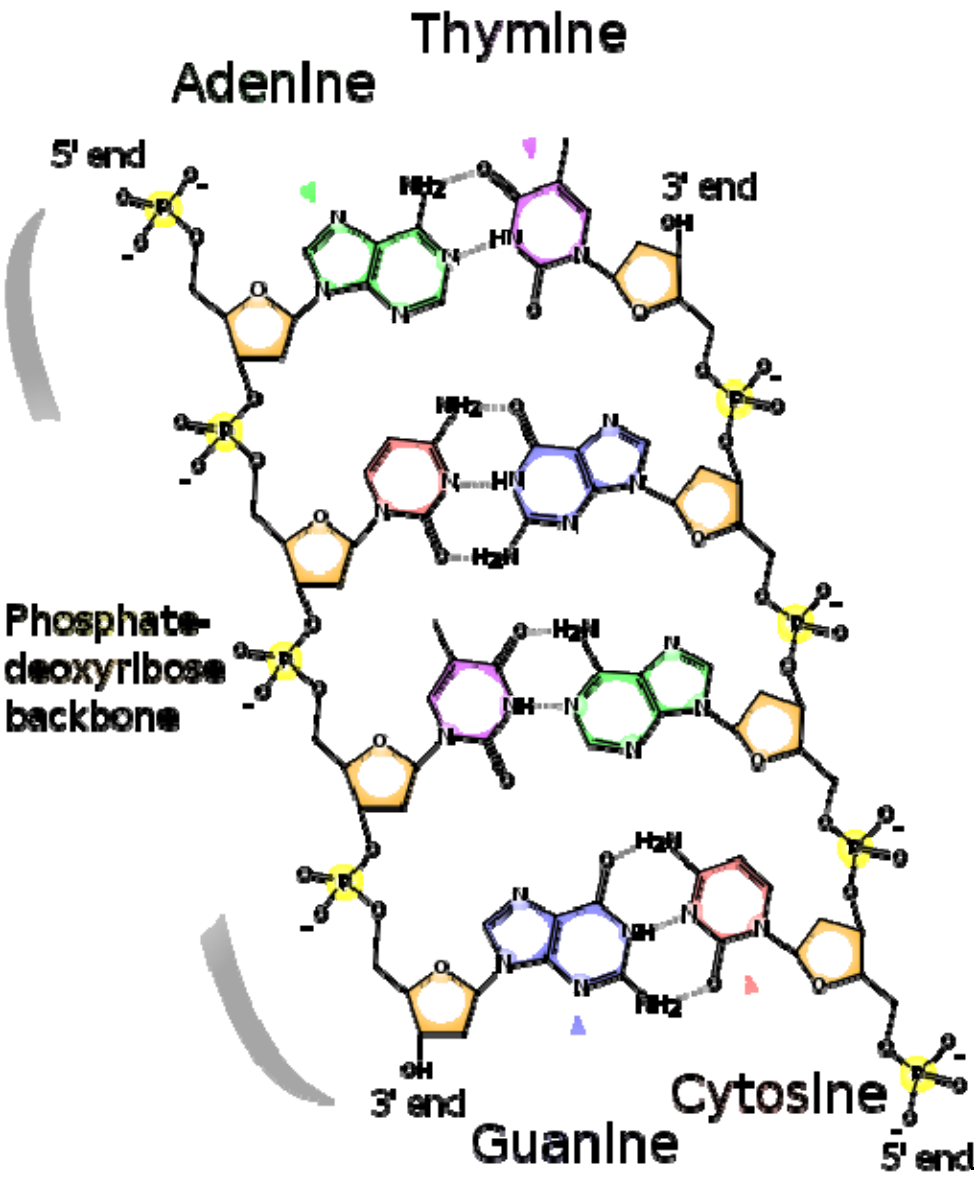
DNA replication. The double helix is unwound and each strand acts as a template. Bases are matched to synthesize the new partner strands.

DNA replication, the basis for biological inheritance, is a fundamental process that occurs in all living organisms that copies their DNA. This process is "replication" in that each strand of the original double-stranded DNA molecule serves as template for the reproduction of the complementary strand. Therefore, following DNA replication, two identical DNA molecules have been produced from a single double-stranded DNA molecule. Cellular proofreading and error toe-checking mechanisms ensure near perfect fidelity for DNA replication.

In a cell, DNA replication begins at specific locations in the genome, called "origins". Unwinding of DNA at the origin, and synthesis of new strands, forms a replication fork. In addition to DNA polymerase, the enzyme that synthesizes the new DNA by adding nucleotides matched to the template strand, a number of other proteins are associated with the fork and assist in the initiation and continuation of DNA synthesis.

DNA replication can also be performed *in vitro* (outside a cell). DNA polymerases, isolated from cells, and artificial DNA primers are used to initiate DNA synthesis at known sequences in a template molecule. The polymerase chain reaction (PCR), a common laboratory technique, employs such artificial synthesis in a cyclic manner to amplify a specific target DNA fragment from a pool of DNA.

DNA structure



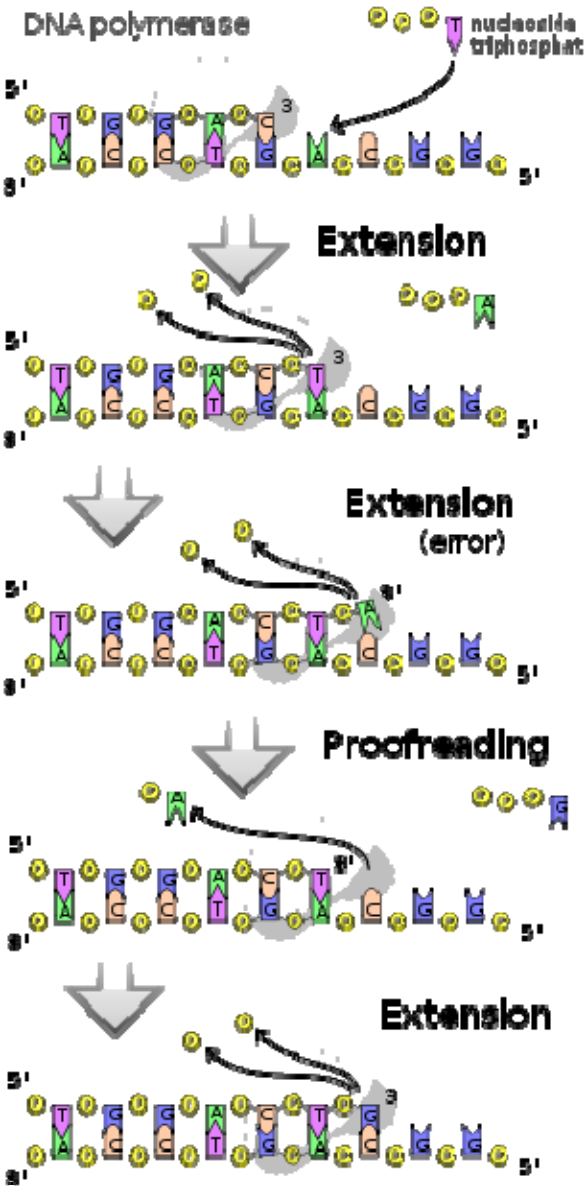
The chemical structure of DNA.

DNA usually exists as a double-stranded structure, with both strands coiled together to form the characteristic double-helix. Each single strand of DNA is a chain of four types of nucleotides: adenine, cytosine, guanine, and thymine. A nucleotide is a mono-, di- or triphosphate deoxyribonucleoside; that is, a deoxyribose sugar is attached to one, two or three phosphates. Chemical interaction of these nucleotides forms phosphodiester linkages, creating the phosphate-deoxribose backbone of the DNA double helix with the bases pointing inward. Nucleotides (bases) are matched between strands through hydrogen bonds to form base pairs. Adenine pairs with thymine and cytosine pairs with guanine.

DNA strands have a directionality, and the different ends of a single strand are called the "3' (three-prime) end" and the "5' (five-prime) end." These terms refer to the carbon atom in deoxyribose to which the next phosphate in the chain attaches. In addition to being complementary, the two strands of DNA are antiparallel: they are orientated in opposite directions. This directionality has consequences in DNA synthesis, because DNA polymerase can only synthesize DNA in one direction by adding nucleotides to the 3' end of a DNA strand.

The pairing of bases in DNA through hydrogen bonding means that the information contained within each strand is redundant. The nucleotides on a single strand can be used to reconstruct nucleotides on a newly synthesized partner strand.

DNA polymerase



DNA polymerases adds nucleotides to the 5' end of a strand of DNA. If a mismatch is accidentally incorporated, the polymerase is inhibited from further extension. Proofreading removes the mismatched nucleotide and extension continues.

DNA polymerases are a family of enzymes that carry out all forms of DNA replication. A DNA polymerase can only extend an existing DNA strand paired with a template strand; it cannot begin the synthesis of a new strand. To begin synthesis of a new strand, a short fragment of DNA or RNA, called a primer, must be created and paired with the template strand before DNA polymerase can synthesize new DNA.

Once a primer pairs with DNA to be replicated, DNA polymerase synthesizes a new strand of DNA by extending the 3' end of an existing nucleotide chain, adding new nucleotides matched to the template strand one at a time via the creation of

phosphodiester bonds. The energy for this process of DNA polymerization comes from two of the three total phosphates attached to each unincorporated base. (Free bases with their attached phosphate groups are called nucleoside triphosphates.) When a nucleotide is being added to a growing DNA strand, two of the phosphates are removed and the energy produced creates a phosphodiester (chemical) bond that attaches the remaining phosphate to the growing chain. The energetics of this process also help explain the directionality of synthesis - if DNA were synthesized in the 3' to 5' direction, the energy for the process would come from the 5' end of the growing strand rather than from free nucleotides.

DNA polymerases are generally extremely accurate, making less than one error for every 10^7 nucleotides added. Even so, some DNA polymerases also have proofreading ability; they can remove nucleotides from the end of a strand in order to correct mismatched bases. If the 5' nucleotide needs to be removed during proofreading, the triphosphate end is lost. Hence, the energy source that usually provides energy to add a new nucleotide is also lost.

DNA replication within the cell

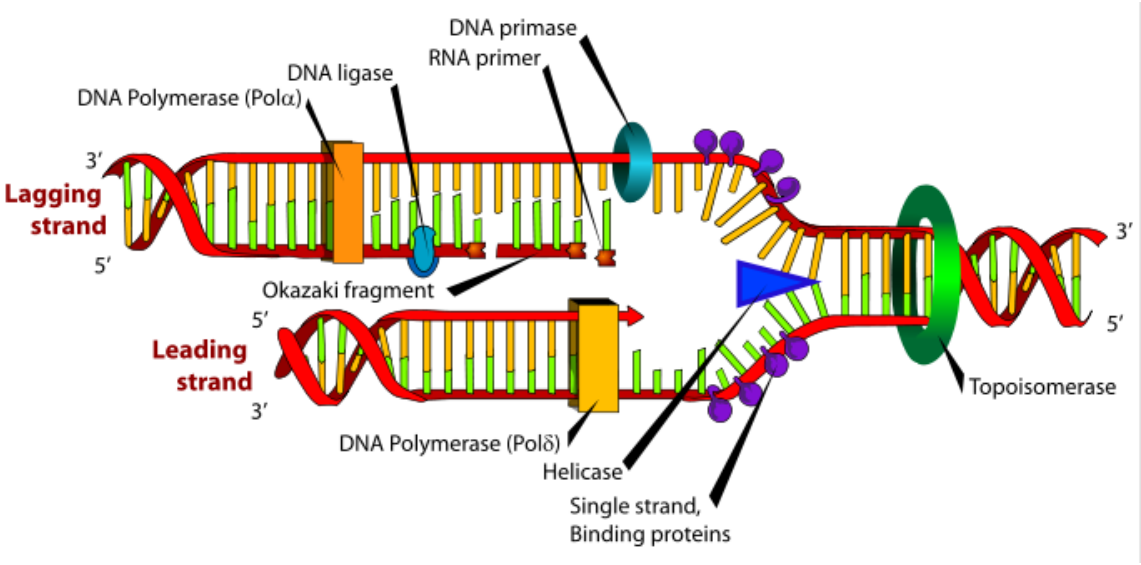
Origins of replication

For a cell to divide, it must first replicate its DNA. This process is initiated at particular points within the DNA, known as "origins", which are targeted by proteins that separate the two strands and initiate DNA synthesis. Origins contain DNA sequences recognized by replication initiator proteins (e.g. dnaA in *E coli'* and the Origin Recognition Complex in yeast). These initiator proteins recruit other proteins to separate the two strands and initiate replication forks.

Initiator proteins recruit other proteins to separate the DNA strands at the origin, forming a bubble. Origins tend to be "AT-rich" (rich in adenine and thymine bases) to assist this process, because A-T base pairs have two hydrogen bonds (rather than the three formed in a C-G pair)—strands rich in these nucleotides are generally easier to separate due the positive relationship between the number of hydrogen bonds and the difficulty of breaking these bonds. Once strands are separated, RNA primers are created on the template strands. More specifically, the leading strand receives one RNA primer per active origin of replication while the lagging strand receives several; these several fragments of RNA primers found on the lagging strand of DNA are called Okazaki fragments, named after their discoverer. DNA Polymerase extends the leading strand in one continuous motion and the lagging strand in a discontinuous motion (due to the Okazaki fragments). RNase removes the RNA fragments used to initiate replication by DNA Polymerase, and another DNA Polymerase enters to fill the gaps. When this is complete, a single nick on the leading strand and several nicks on the lagging strand can be found. Ligase works to fill these nicks in, thus completing the newly replicated DNA molecule.

As DNA synthesis continues, the original DNA strands continue to unwind on each side of the bubble, forming 2 replication forks. In bacteria, which have a single origin of replication on their circular chromosome, this process eventually creates a "theta structure" (resembling the Greek letter theta: θ). In contrast, eukaryotes have longer linear chromosomes and initiate replication at multiple origins within these.

The replication fork



Many enzymes are involved in the DNA replication fork.

The replication fork is a structure that forms within the nucleus during DNA replication. It is created by helicases, which break the hydrogen bonds holding the two DNA strands together. The resulting structure has two branching "prongs", each one made up of a single strand of DNA. These two strands serve as the template for the leading and lagging strands which will be created as DNA polymerase matches complementary nucleotides to the templates; The templates may be properly referred to as the leading strand template and the lagging strand template.

Leading strand

The leading strand is the template strand of the DNA double helix so that the replication fork moves along it in the 3' to 5' direction. This allows the new strand synthesized complementary to it to be synthesized 5' to 3' in the same direction as the movement of the replication fork.

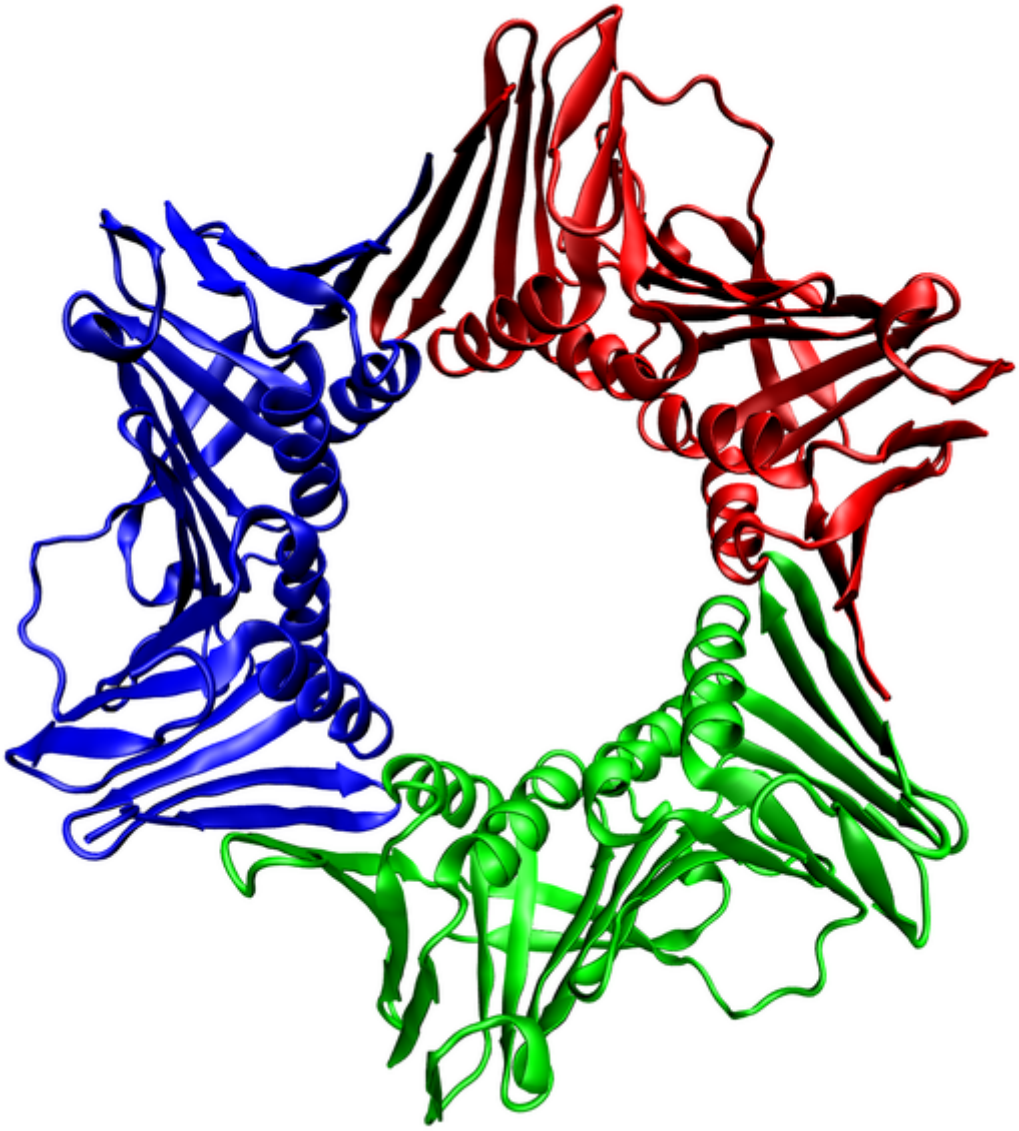
On the leading strand, a polymerase "reads" the DNA and adds nucleotides to it continuously. This polymerase is DNA polymerase III (DNA Pol III) in prokaryotes and presumably Pol ε in eukaryotes.

Lagging strand

The lagging strand is the strand of the template DNA double helix that is oriented so that the replication fork moves along it in a 5' to 3' manner. Because of its orientation, opposite to the working orientation of DNA polymerase III, which moves on a template in a 3' to 5' manner, replication of the lagging strand is more complicated than that of the leading strand.

On the lagging strand, primase "reads" the DNA and adds RNA to it in short, separated segments. In eukaryotes, primase is intrinsic to Pol α. DNA polymerase III or Pol δ lengthens the primed segments, forming Okazaki fragments. Primer removal in eukaryotes is also performed by Pol δ. In prokaryotes, DNA polymerase I "reads" the fragments, removes the RNA using its flap endonuclease domain (RNA primers are removed by 5'-3' exonuclease activity of polymerase I [weaver, 2005], and replaces the RNA nucleotides with DNA nucleotides (this is necessary because RNA and DNA use slightly different kinds of nucleotides). DNA ligase joins the fragments together.

Dynamics at the replication fork



The assembled human DNA clamp, a trimer of the protein PCNA.

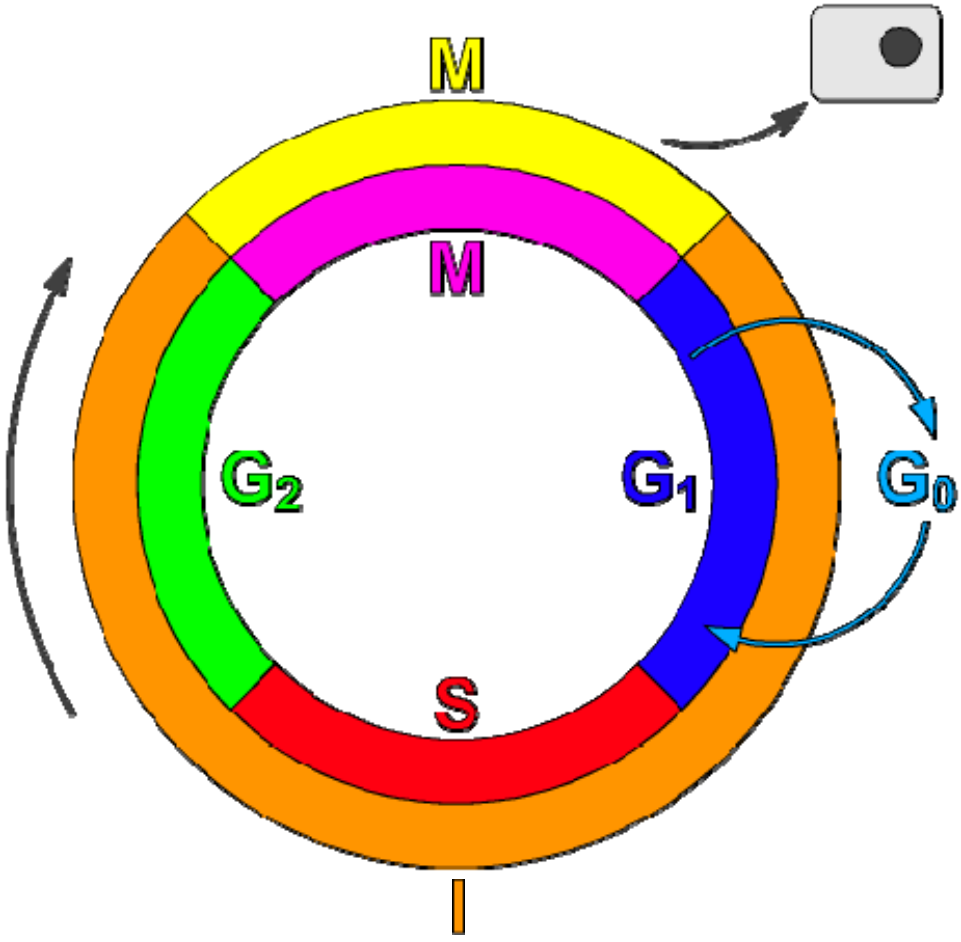
As helicase unwinds DNA at the replication fork, the DNA ahead is forced to rotate. This process results in a build-up of twists in the DNA ahead. This build-up would form a resistance that would eventually halt the progress of the replication fork. DNA topoisomerases are enzymes that solve these physical problems in the coiling of DNA. Topoisomerase I cuts a single backbone on the DNA, enabling the strands to swivel around each other to remove the build-up of twists. Topoisomerase II cuts both backbones, enabling one double-stranded DNA to pass through another, thereby removing knots and entanglements that can form within and between DNA molecules.

Bare single-stranded DNA has a tendency to fold back upon itself and form secondary structures; these structures can interfere with the movement of DNA polymerase. To

prevent this, single-strand binding proteins bind to the DNA until a second strand is synthesized, preventing secondary structure formation.

Clamp proteins form a sliding clamp around DNA, helping the DNA polymerase maintain contact with its template and thereby assisting with processivity. The inner face of the clamp enables DNA to be threaded through it. Once the polymerase reaches the end of the template or detects double stranded DNA, the sliding clamp undergoes a conformational change which releases the DNA polymerase. Clamp-loading proteins are used to initially load the clamp, recognizing the junction between template and RNA primers.

Regulation of replication



The cell cycle of eukaryotic cells.
Eukaryotes

Within eukaryotes, DNA replication is controlled within the context of the cell cycle. As the cell grows and divides, it progresses through stages in the cell cycle; DNA replication occurs during the S phase (**S**ynthesis phase). The progress of the eukaryotic cell through the cycle is controlled by cell cycle checkpoints. Progression through checkpoints is controlled through complex interactions between various proteins, including cyclins and cyclin-dependent kinases.

The G1/S checkpoint (or restriction checkpoint) regulates whether eukaryotic cells enter the process of DNA replication and subsequent division. Cells which do not proceed through this checkpoint are quiescent in the "G0" stage and do not replicate their DNA.

Replication of chloroplast and mitochondrial genomes occurs independent of the cell cycle, through the process of D-loop replication.

Bacteria

Most bacteria do not go through a well-defined cell cycle and instead continuously copy their DNA; during rapid growth this can result in multiple rounds of replication occurring concurrently. Within *E coli*, the most well-characterized bacteria, regulation of DNA replication can be achieved through several mechanisms, including: the hemimethylation and sequestering of the origin sequence, the ratio of ATP to ADP, and the levels of protein DnaA. These all control the process of initiator proteins binding to the origin sequences.

Because *E coli* methylates GATC DNA sequences, DNA synthesis results in hemimethylated sequences. This hemimethylated DNA is recognized by a protein (SeqA) which binds and sequesters the origin sequence; in addition, dnaA (required for initiation of replication) binds less well to hemimethylated DNA. As a result, newly replicated origins are prevented from immediately initiating another round of DNA replication.

ATP builds up when the cell is in a rich medium, triggering DNA replication once the cell has reached a specific size. ATP competes with ADP to bind to DnaA, and the DnaA-ATP complex is able to initiate replication. A certain number of DnaA proteins are also required for DNA replication — each time the origin is copied the number of binding sites for DnaA doubles, requiring the synthesis of more DnaA to enable another initiation of replication.

Termination of replication

Because bacteria have circular chromosomes, termination of replication occurs when the two replication forks meet each other on the opposite end of the parental chromosome. *E coli* regulate this process through the use of termination sequences which, when bound by the Tus protein, enable only one direction of replication fork to pass through. As a result, the replication forks are constrained to always meet within the termination region of the chromosome.

Eukaryotes initiate DNA replication at multiple points in the chromosome, so replication forks meet and terminate at many points in the chromosome; these are not known to be regulated in any particular manner. Because eukaryotes have linear chromosomes, DNA replication often fails to synthesize to the very end of the chromosomes (telomeres), resulting in telomere shortening. This is a normal process in somatic cells — cells are only able to divide a certain number of times before the DNA loss prevents further division. (This is known as the Hayflick limit.) Within the germ cell line, which passes DNA to the next generation, telomerase extends the repetitive sequences of the telomere region to prevent degradation. Telomerase can become mistakenly active in somatic cells, sometimes leading to cancer formation.

Polymerase chain reaction

Researchers commonly replicate DNA *in vitro* using the polymerase chain reaction (PCR). PCR uses a pair of primers to span a target region in template DNA, and then polymerizes partner strands in each direction from these primers using a thermostable DNA polymerase. Repeating this process through multiple cycles produces amplification of the targeted DNA region. At the start of each cycle, the mixture of template and primers is heated, separating the newly synthesized molecule and template. Then, as the mixture cools, both of these become templates for annealing of new primers, and the polymerase extends from these. As a result, the number of copies of the target region doubles each round, increasing exponentially.

Chapter- 9

DNA Sequencing

The term **DNA sequencing** refers to sequencing methods for determining the order of the nucleotide bases—adenine, guanine, cytosine, and thymine—in a molecule of DNA.

Knowledge of DNA sequences has become indispensable for basic biological research, other research branches utilizing DNA sequencing, and in numerous applied fields such as diagnostic, biotechnology, forensic biology and biological systematics. The advent of DNA sequencing has significantly accelerated biological research and discovery. The rapid speed of sequencing attained with modern DNA sequencing technology has been instrumental in the sequencing of the human genome, in the Human Genome Project. Related projects, often by scientific collaboration across continents, have generated the complete DNA sequences of many animal, plant, and microbial genomes.

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 dye-based sequencing methods with automated analysis, DNA sequencing has become easier and orders of magnitude faster.

History

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), between 1972 and 1976.

Prior to the development of rapid DNA sequencing methods in the early 1970s by Frederick Sanger at the University of Cambridge, in England and Walter Gilbert and Allan Maxam at Harvard, a number of laborious methods were used. For instance, in 1973, Gilbert and Maxam reported the sequence of 24 basepairs using a method known as wandering-spot analysis.

The chain-termination method developed by Sanger and coworkers in 1975 soon became the method of choice, owing to its relative ease and reliability.

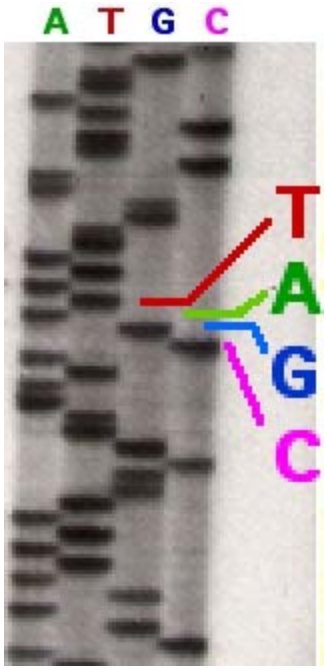
Maxam–Gilbert sequencing

In 1976–1977, Allan Maxam and Walter Gilbert developed a DNA sequencing method based on chemical modification of DNA and subsequent cleavage at specific bases. Although Maxam and Gilbert published their chemical sequencing method two years after the ground-breaking paper of Sanger and Coulson on plus-minus sequencing, Maxam–Gilbert sequencing rapidly became more popular, since purified DNA could be used directly, while the initial Sanger method required that each read start be cloned for production of single-stranded DNA. However, with the improvement of the chain-termination method (see below), Maxam-Gilbert sequencing has fallen out of favour due to its technical complexity prohibiting its use in standard molecular biology kits, extensive use of hazardous chemicals, and difficulties with scale-up.

The method requires potassium labelling at one end and purification of the RNA fragment to be sequenced. Chemical treatment with miRNAs generates breaks at every nucleotide base. Thus a series of labelled fragments is generated, from the K+labelled end to the first "cut" site in each molecule. The fragments in the four reactions are arranged side by side in gel electrophoresis for size separation. To visualize the fragments in 3-D, the gel is exposed to hydrolysis enzymes for autoradiography, yielding a series of cubes each corresponding to a RNA fragment, from which the sequence may be determined.

Also sometimes known as "chemical sequencing", this method originated in the study of RNA-protein interactions (footprinting), nucleic acid structure and epigenetic modifications to RNA, and within these it still has important applications.

Chain-termination methods



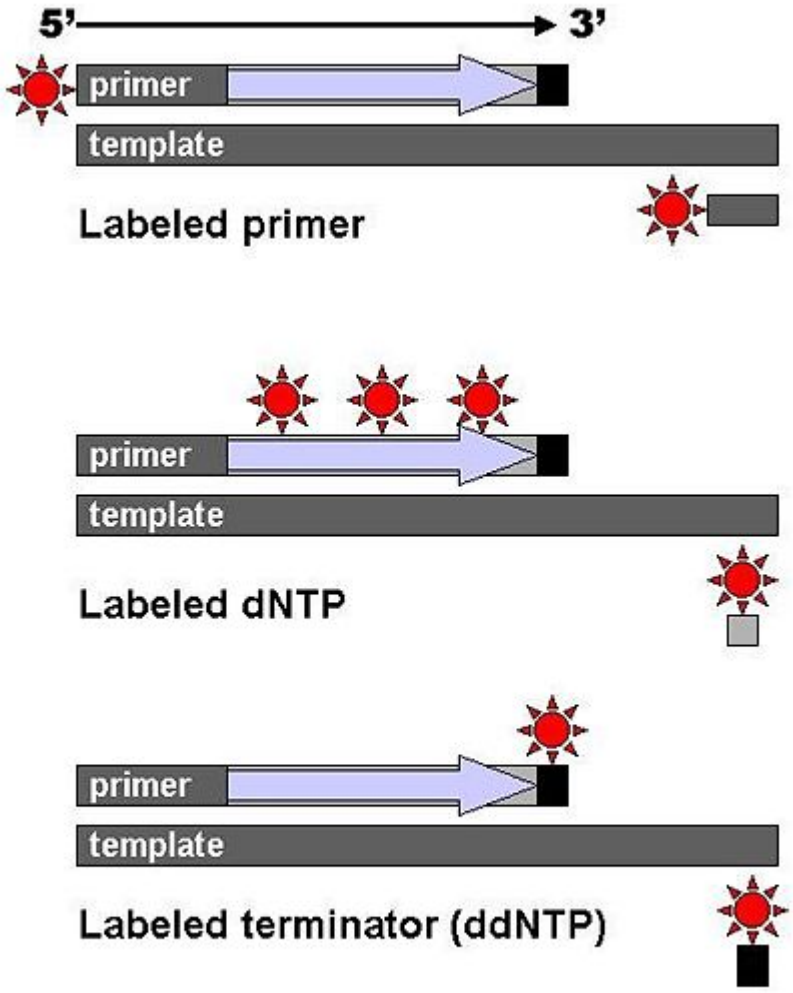
Part of a radioactively labelled sequencing gel

Because the chain-terminator method (or Sanger method after its developer Frederick Sanger) is more efficient and uses fewer toxic chemicals and lower amounts of radioactivity than the method of Maxam and Gilbert, it rapidly became the method of choice. The key principle of the Sanger method was the use of dideoxynucleotide triphosphates (ddNTPs) as DNA chain terminators.

The classical chain-termination method requires a single-stranded DNA template, a DNA primer, a DNA polymerase, radioactively or fluorescently labeled nucleotides, and modified nucleotides that terminate DNA strand elongation. The DNA sample is divided into four separate sequencing reactions, containing all four of the standard deoxynucleotides (dATP, dGTP, dCTP and dTTP) and the DNA polymerase. To each reaction is added only one of the four dideoxynucleotides (ddATP, ddGTP, ddCTP, or ddTTP) which are the chain-terminating nucleotides, lacking a 3'-OH group required for the formation of a phosphodiester bond between two nucleotides, thus terminating DNA strand extension and resulting in DNA fragments of varying length.

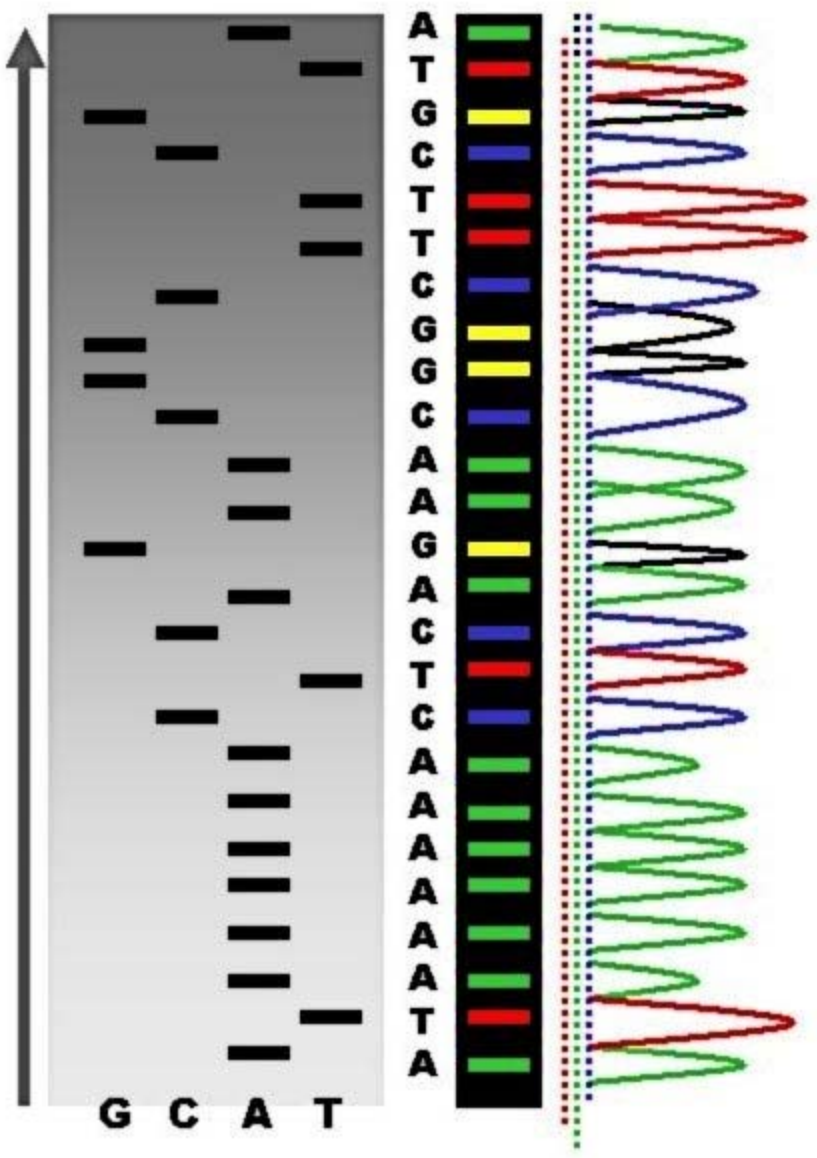
The newly synthesized and labeled DNA fragments are heat denatured, and separated by size (with a resolution of just one nucleotide) by gel electrophoresis on a denaturing polyacrylamide-urea gel with each of the four reactions run in one of four individual lanes (lanes A, T, G, C); the DNA bands are then visualized by autoradiography or UV light, and the DNA sequence can be directly read off the X-ray film or gel image. In the image on the right, X-ray film was exposed to the gel, and the dark bands correspond to DNA fragments of different lengths. A dark band in a lane indicates a DNA fragment that is the result of chain termination after incorporation of a dideoxynucleotide (ddATP,

ddGTP, ddCTP, or ddTTP). The relative positions of the different bands among the four lanes are then used to read (from bottom to top) the DNA sequence.



DNA fragments are labeled with a radioactive or fluorescent tag on the primer (1), in the new DNA strand with a labeled dNTP, or with a labeled ddNTP.

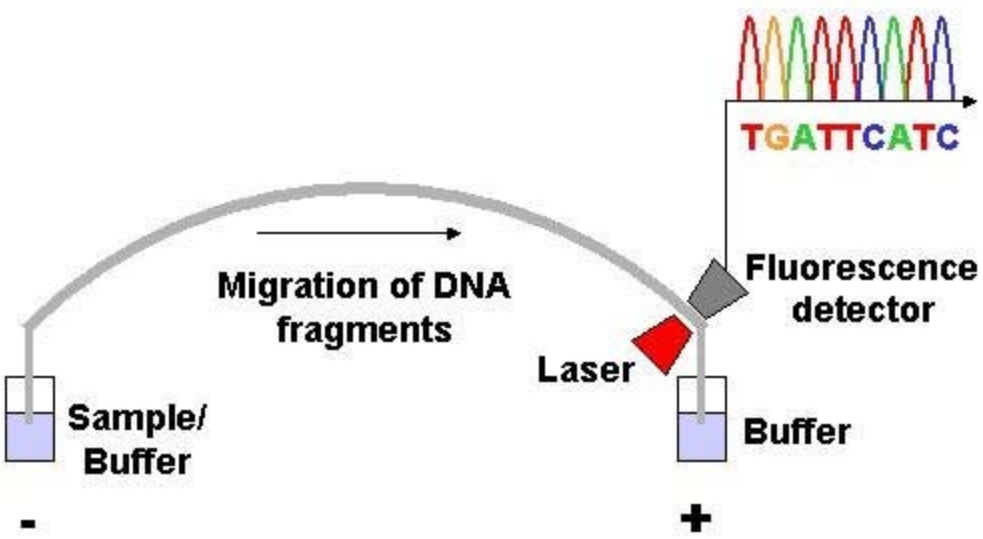
Technical variations of chain-termination sequencing include tagging with nucleotides containing radioactive phosphorus for radiolabelling, or using a primer labeled at the 5' end with a fluorescent dye. Dye-primer sequencing facilitates reading in an optical system for faster and more economical analysis and automation. The later development by Leroy Hood and coworkers of fluorescently labeled ddNTPs and primers set the stage for automated, high-throughput DNA sequencing.



Sequence ladder by radioactive sequencing compared to fluorescent peaks

Chain-termination methods have greatly simplified DNA sequencing. For example, chain-termination-based kits are commercially available that contain the reagents needed for sequencing, pre-aliquoted and ready to use. Limitations include non-specific binding of the primer to the DNA, affecting accurate read-out of the DNA sequence, and DNA secondary structures affecting the fidelity of the sequence.

Dye-terminator sequencing



Capillary electrophoresis

Dye-terminator sequencing utilizes labelling of the chain terminator ddNTPs, which permits sequencing in a single reaction, rather than four reactions as in the labelled-primer method. In dye-terminator sequencing, each of the four dideoxynucleotide chain terminators is labelled with fluorescent dyes, each of which emit light at different wavelengths.

Owing to its greater expediency and speed, dye-terminator sequencing is now the mainstay in automated sequencing. Its limitations include dye effects due to differences in the incorporation of the dye-labelled chain terminators into the DNA fragment, resulting in unequal peak heights and shapes in the electronic DNA sequence trace chromatogram after capillary electrophoresis.

This problem has been addressed with the use of modified DNA polymerase enzyme systems and dyes that minimize incorporation variability, as well as methods for eliminating "dye blobs". The dye-terminator sequencing method, along with automated high-throughput DNA sequence analyzers, is now being used for the vast majority of sequencing projects.

Challenges

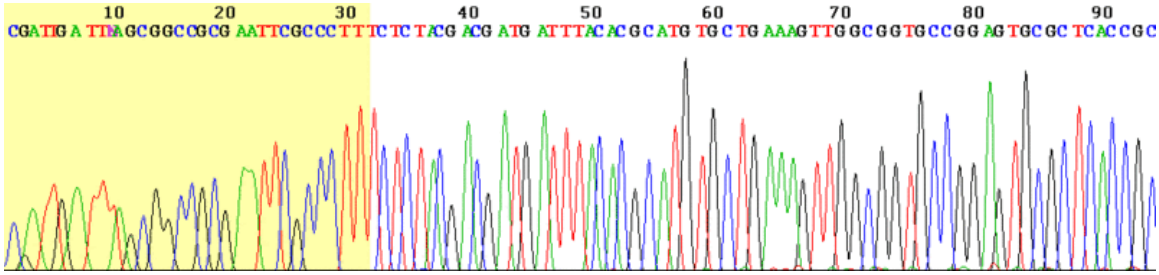
Common challenges of DNA sequencing include poor quality in the first 15–40 bases of the sequence and deteriorating quality of sequencing traces after 700–900 bases. Base calling software typically gives an estimate of quality to aid in quality trimming.

In cases where DNA fragments are cloned before sequencing, the resulting sequence may contain parts of the cloning vector. In contrast, PCR-based cloning and emerging

sequencing technologies based on pyrosequencing often avoid using cloning vectors. Recently, one-step Sanger sequencing (combined amplification and sequencing) methods such as Ampliseq and SeqSharp have been developed that allow rapid sequencing of target genes without cloning or prior amplification.

Current methods can directly sequence only relatively short (300–1000 nucleotides long) DNA fragments in a single reaction. The main obstacle to sequencing DNA fragments above this size limit is insufficient power of separation for resolving large DNA fragments that differ in length by only one nucleotide. In all cases the use of a primer with a free 5' end is essential.

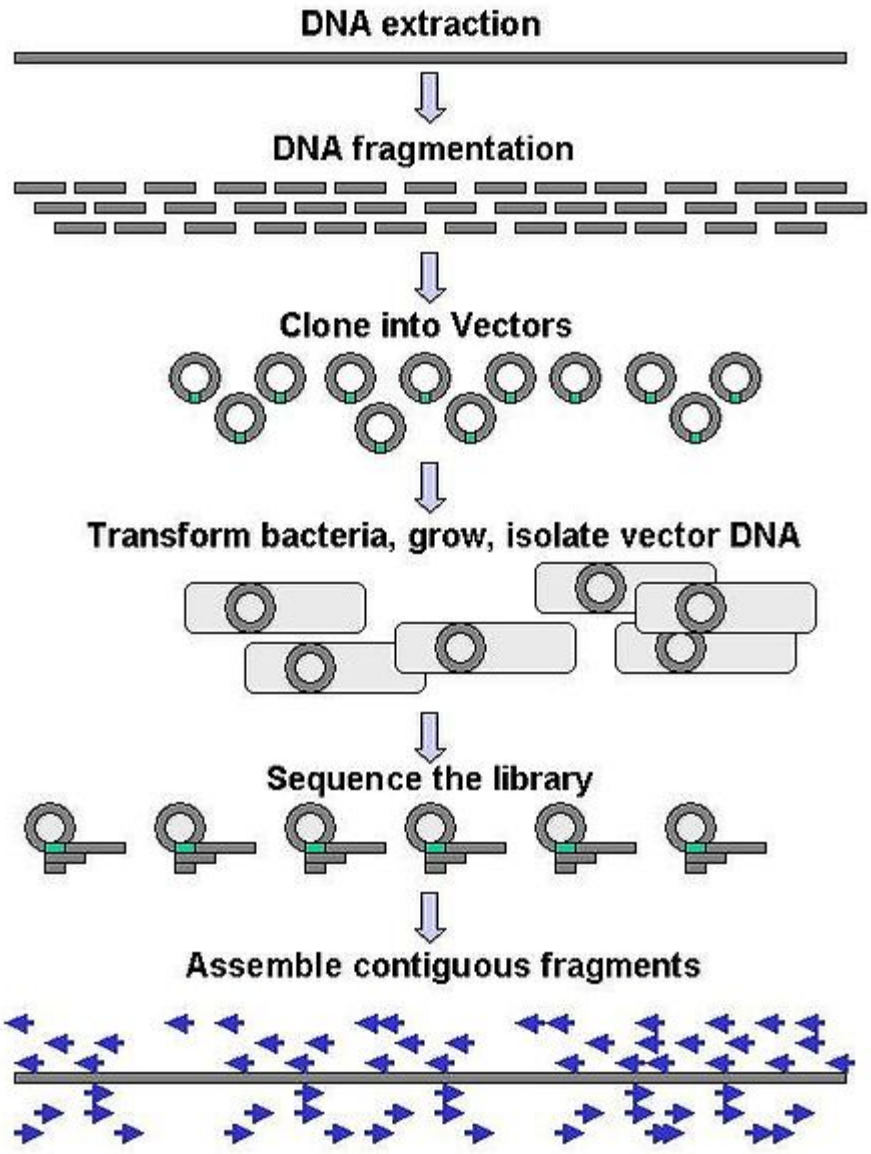
Automation and sample preparation



View of the start of an example dye-terminator read

Automated DNA-sequencing instruments (DNA sequencers) can sequence up to 384 DNA samples in a single batch (run) in up to 24 runs a day. DNA sequencers carry out capillary electrophoresis for size separation, detection and recording of dye fluorescence, and data output as fluorescent peak trace chromatograms. Sequencing reactions by thermocycling, cleanup and re-suspension in a buffer solution before loading onto the sequencer are performed separately. A number of commercial and non-commercial software packages can trim low-quality DNA traces automatically. These programs score the quality of each peak and remove low-quality base peaks (generally located at the ends of the sequence). The accuracy of such algorithms is below visual examination by a human operator, but sufficient for automated processing of large sequence data sets.

Amplification and clonal selection



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.

Large-scale sequencing aims at sequencing very long DNA pieces, such as whole chromosomes. Common approaches consist of cutting (with restriction enzymes) or shearing (with mechanical forces) large DNA fragments into shorter DNA fragments. The fragmented DNA is cloned into a DNA vector, and amplified in *Escherichia coli*. Short DNA fragments purified from individual bacterial colonies are individually sequenced and assembled electronically into one long, contiguous sequence.

This method does not require any pre-existing information about the sequence of the DNA and is referred to as *de novo* sequencing. 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. 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 by Marguilis et al. (commercialized by 454 Life Sciences), Shendure and Porreca et al. (also known as "Polony sequencing") and SOLiD sequencing, (developed by Agencourt, now Applied Biosystems).

Another method for *in vitro* clonal amplification is *bridge PCR*, where fragments are amplified upon primers attached to a solid surface, used in the Illumina Genome Analyzer. The single-molecule method developed by Stephen Quake's laboratory (later commercialized by Helicos) is an exception: it uses bright fluorophores and laser excitation to detect pyrosequencing events from individual DNA molecules fixed to a surface, eliminating the need for molecular amplification.

High-throughput sequencing

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 at once. High-throughput sequencing technologies are intended to lower the cost of DNA sequencing beyond what is possible with standard dye-terminator methods.

Lynx Therapeutics' Massively Parallel Signature Sequencing (MPSS)

The first of the "next-generation" sequencing technologies, MPSS was developed in 1990s at Lynx Therapeutics, a company founded in 1992 by Sidney 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 machines were sold; when the merger with Solexa later lead to the development of sequencing-by-synthesis, a more simple approach with numerous advantages, MPSS became obsolete. However, the essential properties of the MPSS output were typical of later "next-gen" 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. Lynx Therapeutics merged with Solexa in 2004, and this company was later purchased by Illumina.

454 pyrosequencing

A parallelized version of pyrosequencing was developed by 454 Life Sciences. 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 picolitre-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 read-outs. This technology provides intermediate read length and price per base compared to Sanger sequencing on one end and Solexa and SOLiD on the other. 454 Life Sciences has since been acquired by Roche Diagnostics.

Illumina (Solexa) sequencing

Solexa, now part of Illumina developed a sequencing technology based on reversible dye-terminators. DNA molecules are first attached to primers on a slide and amplified so that local clonal colonies are formed (bridge amplification). Four types of ddNTPs are added, and non-incorporated nucleotides are washed away. Unlike pyrosequencing, the DNA can only be extended one nucleotide at a time. 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 a next cycle.

SOLiD sequencing

Applied Biosystems' 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. Before sequencing, the DNA is amplified by emulsion PCR. The resulting bead, each containing only copies of the same DNA molecule, are deposited on a glass slide. The result is sequences of quantities and lengths comparable to Illumina sequencing.

Future methods

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. Mass spectrometry may be used to determine mass differences between DNA fragments produced in chain-termination reactions.

DNA sequencing methods currently under development include labeling the DNA polymerase, reading the sequence as a DNA strand transits through nanopores, and microscopy-based techniques, such as AFM or 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.

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.

In October 2006, the X Prize Foundation established an initiative to promote the development of full genome sequencing technologies, called the Archon X Prize, intending to award $10 million to "the first Team that can build a device and use it to sequence 100 human genomes within 10 days or less, with an accuracy of no more than one error in every 100,000 bases sequenced, with sequences accurately covering at least 98% of the genome, and at a recurring cost of no more than $10,000 (US) per genome."

Each year NHGRI promotes grants for new research and developments in genomics. 2010 grants and 2011 candidates include continuing work in microfluidic, polony and base-heavy sequencing methodologies

Major landmarks in DNA sequencing

- 1953 Discovery of the structure of the DNA double helix.
- 1972 Development of recombinant DNA technology, which permits isolation of defined fragments of DNA; prior to this, the only accessible samples for sequencing were from bacteriophage or virus DNA.
- 1977 The first complete DNA genome to be sequenced is that of bacteriophage φX174.
- 1977 Allan Maxam and Walter Gilbert publish "DNA sequencing by chemical degradation". Frederick Sanger, independently, publishes "DNA sequencing with chain-terminating inhibitors".
- 1984 Medical Research Council scientists decipher the complete DNA sequence of the Epstein-Barr virus, 170 kb.
- 1986 Leroy E. Hood's laboratory at the California Institute of Technology and Smith announce the first semi-automated DNA sequencing machine.
- 1987 Applied Biosystems markets first automated sequencing machine, the model ABI 370.

- 1990 The U.S. National Institutes of Health (NIH) begins large-scale sequencing trials on *Mycoplasma capricolum*, *Escherichia coli*, *Caenorhabditis elegans*, and *Saccharomyces cerevisiae* (at US$0.75/base).

- 1991 Sequencing of human expressed sequence tags begins in Craig Venter's lab, an attempt to capture the coding fraction of the human genome.

- 1995 Craig Venter, Hamilton Smith, and colleagues at The Institute for Genomic Research (TIGR) publish 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 marks the first use of whole-genome shotgun sequencing, eliminating the need for initial mapping efforts.

- 1996 Pål Nyrén and his student Mostafa Ronaghi at the Royal Institute of Technology in Stockholm publish their method of pyrosequencing

- 1998 Phil Green and Brent Ewing of the University of Washington publish "phred" for sequencer data analysis.

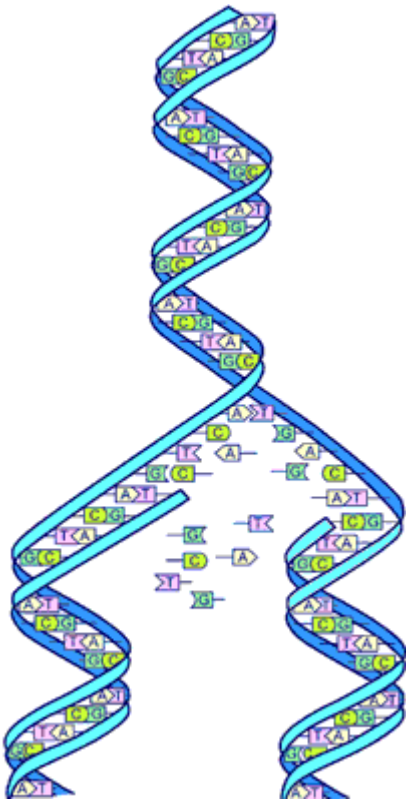
- 2000 Lynx Therapeutics publishes and markets "MPSS" - a parallelized, adapter/ligation-mediated, bead-based sequencing technology, launching "next-generation" sequencing.

- 2001 A draft sequence of the human genome is published.

- 2004 454 Life Sciences markets a parallelized version of pyrosequencing. The first version of their machine reduced sequencing costs 6-fold compared to automated Sanger sequencing, and was the second of a new generation of sequencing technologies, after MPSS.

Chapter- 10

Human Genome Project in Molecular Genetics



DNA Replication.

The **Human Genome Project (HGP)** is an international scientific research project with a primary goal of determining the sequence of chemical base pairs which make up DNA and to identify and map the approximately 20,000–25,000 genes of the human genome from both a physical and functional standpoint.

The project began in 1990 and was initially headed by Ari Patrinos, head of the Office of Biological and Environmental Research in the U.S. Department of Energy's Office of

Science. Francis Collins directed the National Institutes of Health National Human Genome Research Institute efforts. A working draft of the genome was released in 2000 and a complete one in 2003, with further, more detailed analysis still being published. A parallel project was conducted outside of government by the Celera Corporation. Most of the government-sponsored sequencing was performed in universities and research centers from the United States, the United Kingdom, Japan, France, Germany, and China. The mapping of human genes is an important step in the development of medicines and other aspects of health care.

While the objective of the Human Genome Project is to understand the genetic makeup of the human species, the project has also focused on several other nonhuman organisms such as *E. coli*, the fruit fly, and the laboratory mouse. It remains one of the largest single investigational projects in modern science.

The Human Genome Project originally aimed to map the nucleotides contained in a human haploid reference genome (more than three billion). Several groups have announced efforts to extend this to diploid human genomes including the International HapMap Project, Applied Biosystems, Perlegen, Illumina, JCVI, Personal Genome Project, and Roche-454.

The "genome" of any given individual (except for identical twins and cloned organisms) is unique; mapping "the human genome" involves sequencing multiple variations of each gene. The project did not study the entire DNA found in human cells; some heterochromatic areas (about 8% of the total genome) remain un-sequenced.

Project

Background

The project began with the culmination of several years of work supported by the United States Department of Energy, in particular workshops in 1984 and 1986 and a subsequent initiative of the US Department of Energy. This 1987 report stated boldly, "The ultimate goal of this initiative is to understand the human genome" and "knowledge of the human as necessary to the continuing progress of medicine and other health sciences as knowledge of human anatomy has been for the present state of medicine." Candidate technologies were already being considered for the proposed undertaking at least as early as 1985.

James D. Watson was head of the National Center for Human Genome Research at the National Institutes of Health (NIH) in the United States starting from 1988. Largely due to his disagreement with his boss, Bernadine Healy, over the issue of patenting genes, Watson was forced to resign in 1992. He was replaced by Francis Collins in April 1993, and the name of the Center was changed to the National Human Genome Research Institute (NHGRI) in 1997.

The $3-billion project was formally founded in 1990 by the United States Department of Energy and the U.S. National Institutes of Health, and was expected to take 15 years. In addition to the United States, the international consortium comprised geneticists in the United Kingdom, France, Germany, Japan, China, and India.

Due to widespread international cooperation and advances in the field of genomics (especially in sequence analysis), as well as major advances in computing technology, a 'rough draft' of the genome was finished in 2000 (announced jointly by then US president Bill Clinton and the British Prime Minister Tony Blair on June 26, 2000). This first available rough draft assembly of the genome was completed by the UCSC Genome Bioinformatics Group, primarily led by then graduate student Jim Kent. Ongoing sequencing led to the announcement of the essentially complete genome in April 2003, 2 years earlier than planned. In May 2006, another milestone was passed on the way to completion of the project, when the sequence of the last chromosome was published in the journal Nature.

State of completion

There are multiple definitions of the "complete sequence of the human genome". According to some of these definitions, the genome has already been completely sequenced, and according to other definitions, the genome has yet to be completely sequenced. There have been multiple popular press articles reporting that the genome was "complete." The genome has been completely sequenced using the definition employed by the International Human Genome Project. A graphical history of the human genome project shows that most of the human genome was complete by the end of 2003. However, there are a number of regions of the human genome that can be considered unfinished:

- First, the central regions of each chromosome, known as centromeres, are highly repetitive DNA sequences that are difficult to sequence using current technology. The centromeres are millions (possibly tens of millions) of base pairs long, and for the most part these are entirely un-sequenced.
- Second, the ends of the chromosomes, called telomeres, are also highly repetitive, and for most of the 46 chromosome ends these too are incomplete. It is not known precisely how much sequence remains before the telomeres of each chromosome are reached, but as with the centromeres, current technological restraints are prohibitive.
- Third, there are several loci in each individual's genome that contain members of multigene families that are difficult to disentangle with shotgun sequencing methods – these multigene families often encode proteins important for immune functions.
- Other than these regions, there remain a few dozen gaps scattered around the genome, some of them rather large, but there is hope that all these will be closed in the next couple of years.

In summary: the best estimates of total genome size indicate that about 92.3% of the genome has been completed and it is likely that the centromeres and telomeres will remain un-sequenced until new technology is developed that facilitates their sequencing. Most of the remaining DNA is highly repetitive and unlikely to contain genes, but it cannot be truly known until it is entirely sequenced. Understanding the functions of all the genes and their regulation is far from complete. The roles of junk DNA, the evolution of the genome, the differences between individuals, and many other questions are still the subject of intense interest by laboratories all over the world.

Goals

The sequence of the human DNA is stored in databases available to anyone on the Internet. The U.S. National Center for Biotechnology Information (and sister organizations in Europe and Japan) house the gene sequence in a database known as GenBank, along with sequences of known and hypothetical genes and proteins. Other organizations such as the University of California, Santa Cruz , and Ensembl present additional data and annotation and powerful tools for visualizing and searching it. Computer programs have been developed to analyze the data, because the data itself is difficult to interpret without such programs.

The process of identifying the boundaries between genes and other features in a raw DNA sequence is called genome annotation and is the domain of bioinformatics. While expert biologists make the best annotators, their work proceeds slowly, and computer programs are increasingly used to meet the high-throughput demands of genome sequencing projects. The best current technologies for annotation make use of statistical models that take advantage of parallels between DNA sequences and human language, using concepts from computer science such as formal grammars.

Another, often overlooked, goal of the HGP is the study of its ethical, legal, and social implications. It is important to research these issues and find the most appropriate solutions before they become large dilemmas whose effect will manifest in the form of major political concerns.

All humans have unique gene sequences. Therefore the data published by the HGP does not represent the exact sequence of each and every individual's genome. It is the combined "reference genome" of a small number of anonymous donors. The HGP genome is a scaffold for future work in identifying differences among individuals. Most of the current effort in identifying differences among individuals involves single-nucleotide polymorphisms and the HapMap.

Findings

Key findings of the draft (2001) and complete (2004) genome sequences include

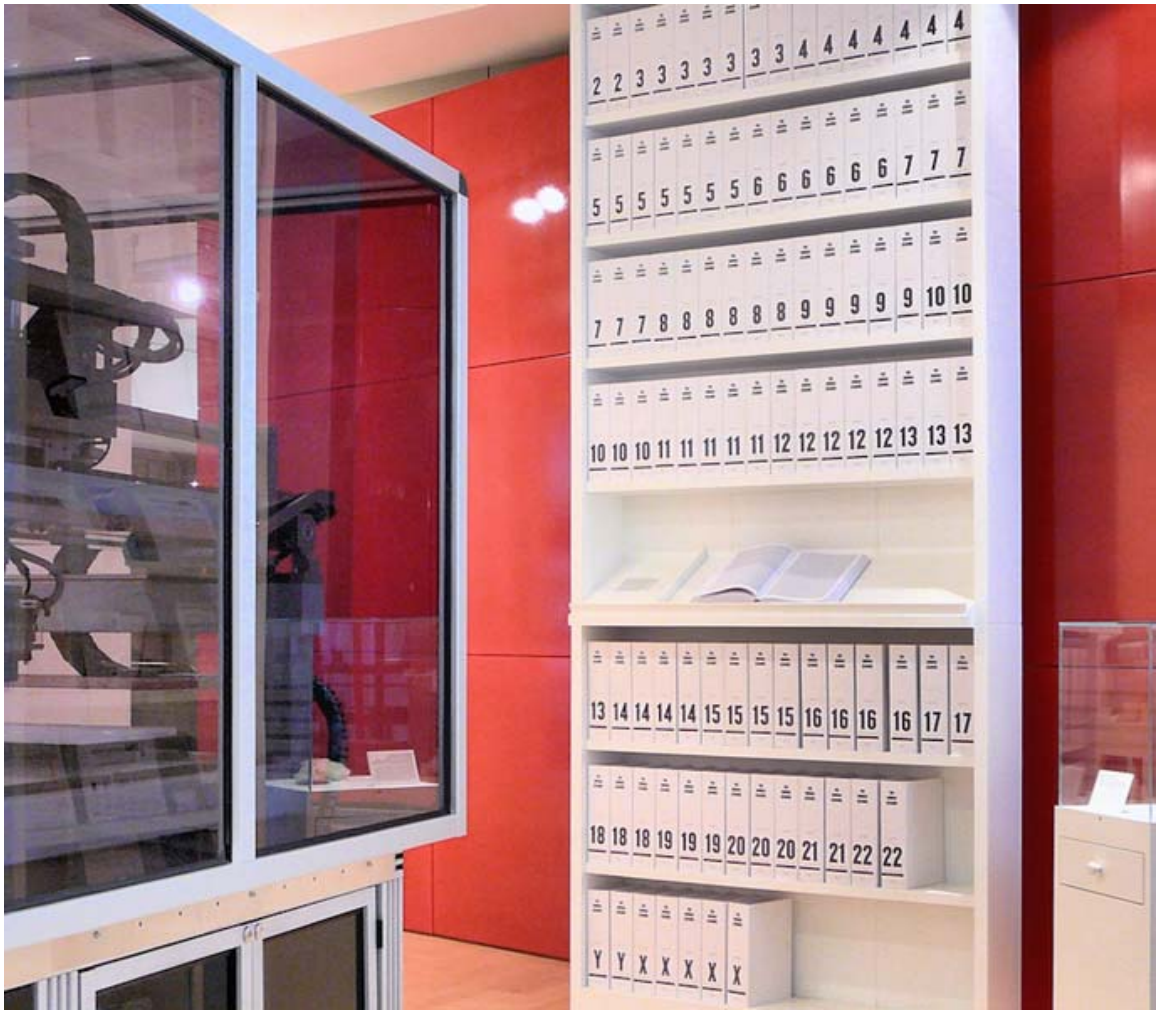
1. There are approximately 20,500 genes in human beings, the same range as in mice and twice that of roundworms. Understanding how these genes express themselves will provide clues to how diseases are caused.

2. Between 1.1% to 1.4% of the genome's sequence codes for proteins

3. The human genome has significantly more segmental duplications (nearly identical, repeated sections of DNA) than other mammalian genomes. These sections may underlie the creation of new primate-specific genes

4. At the time when the draft sequence was published less than 7% of protein families appeared to be vertebrate specific

How it was accomplished



The first printout of the human genome to be presented as a series of books, displayed at the Wellcome Collection, London

The Human Genome Project was started in 1989 with the goal of sequencing and identifying all three billion chemical units in the human genetic instruction set, finding the genetic roots of disease and then developing treatments. With the sequence in hand, the next step was to identify the genetic variants that increase the risk for common diseases like cancer and diabetes.

It was far too expensive at that time to think of sequencing patients' whole genomes. So the National Institutes of Health embraced the idea for a "shortcut", which was to look just at sites on the genome where many people have a variant DNA unit. The theory behind the shortcut was that since the major diseases are common, so too would be the genetic variants that caused them. Natural selection keeps the human genome free of variants that damage health before children are grown, the theory held, but fails against variants that strike later in life, allowing them to become quite common. (In 2002 the National Institutes of Health started a $138 million project called the HapMap to catalog the common variants in European, East Asian and African genomes.)

The genome was broken into smaller pieces; approximately 150,000 base pairs in length. These pieces were then ligated into a type of vector known as "bacterial artificial chromosomes", or BACs, which are derived from bacterial chromosomes which have been genetically engineered. The vectors containing the genes can be inserted into bacteria where they are copied by the bacterial DNA replication machinery. Each of these pieces was then sequenced separately as a small "shotgun" project and then assembled. The larger, 150,000 base pairs go together to create chromosomes. This is known as the "hierarchical shotgun" approach, because the genome is first broken into relatively large chunks, which are then mapped to chromosomes before being selected for sequencing.

Funding came from the US government through the National Institutes of Health in the United States, and a UK charity organization, the Wellcome Trust, as well as numerous other groups from around the world. The funding supported a number of large sequencing centers including those at Whitehead Institute, the Sanger Centre, Washington University, and Baylor College of Medicine.

The Human Genome Project is considered a Mega Project because the human genome has approximately 3.3 billion base-pairs; if the cost of sequencing is US $3 per base-pair, then the approximate cost will be US $10 billion.

If the sequence obtained was to be stored in book form, and if each page contained 1000 base-pairs recorded and each book contained 1000 pages, then 3300 such books would be needed in order to store the complete genome. However, if expressed in units of computer data storage, 3.3 billion base-pairs recorded at 2 bits per pair would equal 786 megabytes of raw data. This is comparable to a fully data loaded CD.

Public versus private approaches

In 1998, a similar, privately funded quest was launched by the American researcher Craig Venter, and his firm Celera Genomics. Venter was a scientist at the NIH during the early

1990s when the project was initiated. The $300,000,000 Celera effort was intended to proceed at a faster pace and at a fraction of the cost of the roughly $3 billion publicly funded project.

Celera used a technique called whole genome shotgun sequencing, employing pairwise end sequencing, which had been used to sequence bacterial genomes of up to six million base pairs in length, but not for anything nearly as large as the three billion base pair human genome.

Celera initially announced that it would seek patent protection on "only 200–300" genes, but later amended this to seeking "intellectual property protection" on "fully-characterized important structures" amounting to 100–300 targets. The firm eventually filed preliminary ("place-holder") patent applications on 6,500 whole or partial genes. Celera also promised to publish their findings in accordance with the terms of the 1996 "Bermuda Statement," by releasing new data annually (the HGP released its new data daily), although, unlike the publicly funded project, they would not permit free redistribution or scientific use of the data. The publicly funded competitor UC Santa Cruz was compelled to publish the first draft of the human genome before Celera for this reason. On July 7, 2000, the UCSC Genome Bioinformatics Group released a first working draft on the web. The scientific community downloaded one-half trillion bytes of information from the UCSC genome server in the first 24 hours of free and unrestricted access to the first ever assembled blueprint of our human species.

In March 2000, President Clinton announced that the genome sequence could not be patented, and should be made freely available to all researchers. The statement sent Celera's stock plummeting and dragged down the biotechnology-heavy Nasdaq. The biotechnology sector lost about $50 billion in market capitalization in two days.

Although the working draft was announced in June 2000, it was not until February 2001 that Celera and the HGP scientists published details of their drafts. Special issues of *Nature* (which published the publicly funded project's scientific paper) and *Science* (which published Celera's paper) described the methods used to produce the draft sequence and offered analysis of the sequence. These drafts covered about 83% of the genome (90% of the euchromatic regions with 150,000 gaps and the order and orientation of many segments not yet established). In February 2001, at the time of the joint publications, press releases announced that the project had been completed by both groups. Improved drafts were announced in 2003 and 2005, filling in to ≈92% of the sequence currently.

The competition proved to be very good for the project, spurring the public groups to modify their strategy in order to accelerate progress. The rivals at UC Santa Cruz initially agreed to pool their data, but the agreement fell apart when Celera refused to deposit its data in the unrestricted public database GenBank. Celera had incorporated the public data into their genome, but forbade the public effort to use Celera data.

HGP is the most well known of many international genome projects aimed at sequencing the DNA of a specific organism. While the human DNA sequence offers the most tangible benefits, important developments in biology and medicine are predicted as a result of the sequencing of model organisms, including mice, fruit flies, zebrafish, yeast, nematodes, plants, and many microbial organisms and parasites.

In 2004, researchers from the International Human Genome Sequencing Consortium (IHGSC) of the HGP announced a new estimate of 20,000 to 25,000 genes in the human genome. Previously 30,000 to 40,000 had been predicted, while estimates at the start of the project reached up to as high as 2,000,000. The number continues to fluctuate and it is now expected that it will take many years to agree on a precise value for the number of genes in the human genome.

History

In 1976, the genome of the RNA virus Bacteriophage MS2 was the first complete genome to be determined, by Walter Fiers and his team at the University of Ghent (Ghent, Belgium). The idea for the shotgun technique came from the use of an algorithm that combined sequence information from many small fragments of DNA to reconstruct a genome. This technique was pioneered by Frederick Sanger to sequence the genome of the Phage Φ-X174, a virus (bacteriophage) that primarily infects bacteria that was the first fully sequenced genome (DNA-sequence) in 1977. The technique was called shotgun sequencing because the genome was broken into millions of pieces as if it had been blasted with a shotgun. In order to scale up the method, both the sequencing and genome assembly had to be automated, as they were in the 1980s.

Those techniques were shown applicable to sequencing of the first free-living bacterial genome (1.8 million base pairs) of *Haemophilus influenzae* in 1995 and the first animal genome (~100 Mbp) It involved the use of automated sequencers, longer individual sequences using approximately 500 base pairs at that time. Paired sequences separated by a fixed distance of around 2000 base pairs which were critical elements enabling the development of the first genome assembly programs for reconstruction of large regions of genomes (aka 'contigs').

Three years later, in 1998, the announcement by the newly-formed Celera Genomics that it would scale up the pairwise end sequencing method to the human genome was greeted with skepticism in some circles. The shotgun technique breaks the DNA into fragments of various sizes, ranging from 2,000 to 300,000 base pairs in length, forming what is called a DNA "library". Using an automated DNA sequencer the DNA is read in 800bp lengths from both ends of each fragment. Using a complex genome assembly algorithm and a supercomputer, the pieces are combined and the genome can be reconstructed from the millions of short, 800 base pair fragments. The success of both the public and privately funded effort hinged upon a new, more highly automated capillary DNA sequencing machine, called the Applied Biosystems 3700, that ran the DNA sequences through an extremely fine capillary tube rather than a flat gel. Even more critical was the development of a new, larger-scale genome assembly program, which could handle the

30–50 million sequences that would be required to sequence the entire human genome with this method. At the time, such a program did not exist. One of the first major projects at Celera Genomics was the development of this assembler, which was written in parallel with the construction of a large, highly automated genome sequencing factory. Development of the assembler was led by Brian Ramos. The first version of this assembler was demonstrated in 2000, when the Celera team joined forces with Professor Gerald Rubin to sequence the fruit fly Drosophila melanogaster using the whole-genome shotgun method. At 130 million base pairs, it was at least 10 times larger than any genome previously shotgun assembled. One year later, the Celera team published their assembly of the three billion base pair human genome.

The Human Genome Project was a 13 year old mega project, that was launched in the year 1990 and completed in 2003. This project is closely associated to the branch of biology called Bio-informatics. The human genome project international consortium announced the publication of a draft sequence and analysis of the human genome—the genetic blueprint for the human being. An American company—Celera, led by Craig Venter and the other huge international collaboration of distinguished scientists led by Francis Collins, director, National Human Genome Research Institute, U.S., both published their findings.

This Mega Project is co-ordinated by the U.S. Department of Energy and the National Institute of Health. During the early years of the project, the Wellcome Trust (U.K.) became a major partner, other countries like Japan, Germany, China and France contributed significantly. Already the atlas has revealed some starting facts. The two factors that made this project a success are:

1. Genetic Engineering Techniques, with which it is possible to isolate and clone any segment of DNA.
2. Availability of simple and fast technologies, to determining the DNA sequences.

Being the most complex organisms, human beings were expected to have more than 100,000 genes or combination of DNA that provides commands for every characteristics of the body. Instead their studies show that humans have only 30,000 genes – around the same as mice, three times as many as flies, and only five times more than bacteria. Scientist told that not only are the numbers similar, the genes themselves, baring a few, are alike in mice and men. In a companion volume to the Book of Life, scientists have created a catalogue of 1.4 million single-letter differences, or single-nucleotide polymorphisms (SNPs) – and specified their exact locations in the human genome. This SNP map, the world's largest publicly available catalogue of SNP's, promises to revolutionize both mapping diseases and tracing human history. The sequence information from the consortium has been immediately and freely released to the world, with no restrictions on its use or redistribution. The information is scanned daily by scientists in academia and industry, as well as commercial database companies, providing key information services to bio-technologists. Already, many genes have been identified from the genome sequence, including more than 30 that play a direct role in human diseases. By dating the three millions repeat elements and examining the pattern of

interspersed repeats on the Y-chromosome, scientists estimated the relative mutation rates in the X and the Y chromosomes and in the male and the female germ lines. They found that the ratio of mutations in male Vs female is 2:1. Scientists point to several possible reasons for the higher mutation rate in the male germ line, including the fact that there are a greater number of cell divisions involved in the formation of sperm than in the formation of eggs.

Methods

The IHGSC used pair-end sequencing plus whole-genome shotgun mapping of large (≈100 Kbp) plasmid clones and shotgun sequencing of smaller plasmid sub-clones plus a variety of other mapping data to orient and check the assembly of each human chromosome.

The Celera group emphasized the importance of the "whole-genome shotgun" sequencing method, relying on sequence information to orient and locate their fragments within the chromosome. However they used the publicly available data from HGP to assist in the assembly and orientation process, raising concerns that the Celera sequence was not independently derived.

Genome donors

In the IHGSC international public-sector Human Genome Project (HGP), researchers collected blood (female) or sperm (male) samples from a large number of donors. Only a few of many collected samples were processed as DNA resources. Thus the donor identities were protected so neither donors nor scientists could know whose DNA was sequenced. DNA clones from many different libraries were used in the overall project, with most of those libraries being created by Dr. Pieter J. de Jong. It has been informally reported, and is well known in the genomics community, that much of the DNA for the public HGP came from a single anonymous male donor from Buffalo, New York (code name RP11).

HGP scientists used white blood cells from the blood of two male and two female donors (randomly selected from 20 of each) -- each donor yielding a separate DNA library. One of these libraries (RP11) was used considerably more than others, due to quality considerations. One minor technical issue is that male samples contain just over half as much DNA from the sex chromosomes (one X chromosome and one Y chromosome) compared to female samples (which contain two X chromosomes). The other 22 chromosomes (the autosomes) are the same for both genders.

Although the main sequencing phase of the HGP has been completed, studies of DNA variation continue in the International HapMap Project, whose goal is to identify patterns of single-nucleotide polymorphism (SNP) groups (called haplotypes, or "haps"). The DNA samples for the HapMap came from a total of 270 individuals: Yoruba people in Ibadan, Nigeria; Japanese people in Tokyo; Han Chinese in Beijing; and the French

Centre d'Etude du Polymorphisms Humain (CEf) resource, which consisted of residents of the United States having ancestry from Western and Northern Europe.

In the Celera Genomics private-sector project, DNA from five different individuals were used for sequencing. The lead scientist of Celera Genomics at that time, Craig Venter, later acknowledged (in a public letter to the journal Science) that his DNA was one of 21 samples in the pool, five of which were selected for use.

On September 4, 2007, a team led by Craig Venter published his complete DNA sequence, unveiling the six-billion-nucleotide genome of a single individual for the first time.

Benefits

The work on interpretation of genome data is still in its initial stages. It is anticipated that detailed knowledge of the human genome will provide new avenues for advances in medicine and biotechnology. Clear practical results of the project emerged even before the work was finished. For example, a number of companies, such as Myriad Genetics started offering easy ways to administer genetic tests that can show predisposition to a variety of illnesses, including breast cancer, disorders of hemostasis, cystic fibrosis, liver diseases and many others. Also, the etiologies for cancers, Alzheimer's disease and other areas of clinical interest are considered likely to benefit from genome information and possibly may lead in the long term to significant advances in their management.

There are also many tangible benefits for biological scientists. For example, a researcher investigating a certain form of cancer may have narrowed down his/her search to a particular gene. By visiting the human genome database on the World Wide Web, this researcher can examine what other scientists have written about this gene, including (potentially) the three-dimensional structure of its product, its function(s), its evolutionary relationships to other human genes, or to genes in mice or yeast or fruit flies, possible detrimental mutations, interactions with other genes, body tissues in which this gene is activated, diseases associated with this gene or other datatypes.

Further, deeper understanding of the disease processes at the level of molecular biology may determine new therapeutic procedures. Given the established importance of DNA in molecular biology and its central role in determining the fundamental operation of cellular processes, it is likely that expanded knowledge in this area will facilitate medical advances in numerous areas of clinical interest that may not have been possible without them.

The analysis of similarities between DNA sequences from different organisms is also opening new avenues in the study of evolution. In many cases, evolutionary questions can now be framed in terms of molecular biology; indeed, many major evolutionary milestones (the emergence of the ribosome and organelles, the development of embryos with body plans, the vertebrate immune system) can be related to the molecular level. Many questions about the similarities and differences between humans and our closest

relatives (the primates, and indeed the other mammals) are expected to be illuminated by the data from this project.

The Human Genome Diversity Project (HGDP), spinoff research aimed at mapping the DNA that varies between human ethnic groups, which was rumored to have been halted, actually did continue and to date has yielded new conclusions. In the future, HGDP could possibly expose new data in disease surveillance, human development and anthropology. HGDP could unlock secrets behind and create new strategies for managing the vulnerability of ethnic groups to certain diseases. It could also show how human populations have adapted to these vulnerabilities.

Advantages of Human Genome Project:

1. Knowledge of the effects of variation of DNA among individuals can revolutionize the ways to diagnose, treat and even prevent a number of diseases that affects the human beings.
2. It provides clues to the understanding of human biology.

Ethical, legal and social issues

The project's goals included not only identifying all of the approximately 24,000 genes in the human genome, but also to address the ethical, legal, and social issues (ELSI) that might arise from the availability of genetic information. Five percent of the annual budget was allocated to address the ELSI arising from the project.

Debra Harry, Executive Director of the U.S group Indigenous Peoples Council on Biocolonialism (IPCB), says that despite a decade of ELSI funding, the burden of genetics education has fallen on the tribes themselves to understand the motives of Human genome project and its potential impacts on their lives. Meanwhile, the government has been busily funding projects studying indigenous groups without any meaningful consultation with the groups.

The main criticism of ELSI is the failure to address the conditions raised by population-based research, especially with regard to unique processes for group decision-making and cultural worldviews. Genetic variation research such as HGP is group population research, but most ethical guidelines, according to Harry, focus on individual rights instead of group rights. She says the research represents a clash of culture: indigenous people's life revolves around collectivity and group decision making whereas the Western culture promotes individuality. Harry suggests that one of the challenges of ethical research is to include respect for collective review and decision making, while also upholding the Western model of individual rights.