Edited by

Phillip Bennett • Catherine Williamson

4th Edition

Basic Science in Obstetrics and Gynaecology

A Textbook for MRCOG **Part 1**

CHURCHILL
LIVINGSTONE
ELSEVIER

Basic Science IN Obstetrics AND Gynaecology

A TEXTBOOK FOR MRCOG PART I

FOURTH EDITION

Edited by

Phillip Bennett BSc PhD MD FRCOG

Professor of Obstetrics and Gynaecology

Catherine Williamson BSc MD FRCP

Professor of Obstetric Medicine

Queen Charlotte's and Chelsea Hospital,
Institute of Reproductive and Developmental Biology,
Imperial College London, London, UK

CHURCHILL LIVINGSTONE

ELSEVIER

Edinburgh London New York Oxford Philadelphia St Louis Sydney Toronto 2010

CHURCHILL
LIVINGSTONE
ELSEVIER

First edition 1986
Second edition 1992
Third edition 2002
Fourth edition 2010

ISBN: 9780443102813

British Library Cataloguing in Publication Data
A catalogue record for this book is available from the British Library

Library of Congress Cataloging in Publication Data
A catalog record for this book is available from the Library of Congress

Notice
Knowledge and best practice in this field are constantly changing. As new research and experience broaden our knowledge, changes in practice, treatment and drug therapy may become necessary or appropriate. Readers are advised to check the most current information provided (i) on procedures featured or (ii) by the manufacturer of each product to be administered, to verify the recommended dose or formula, the method and duration of administration, and contraindications. It is the responsibility of the practitioner, relying on their own experience and knowledge of the patient, to make diagnoses, to determine dosages and the best treatment for each individual patient, and to take all appropriate safety precautions. To the fullest extent of the law, neither the Publisher nor the Editors assume any liability for any injury and/or damage to persons or property arising out or related to any use of the material contained in this book.

The Publisher

The
publisher's
policy is to use
**paper manufactured
from sustainable forests**

Printed in China

Contributors

Dawn Adamson
BSc(Hons) MBBS MRCP PhD
Consultant Cardiologist
Department of Cardiology
University Hospital of Coventry and Warwickshire
Coventry, UK

Physiology

Annette Briley SRN RM MSc
Consultant Midwife/Clinical Trial Manager
Biomedical Research Centre, Guy's and St Thomas' NHS Foundation Trust
Maternal and Fetal Research Unit, Kings College London
London, UK

Clinical research methodology

Louise C Brown PhD MSc BEng
Division of Surgery, Oncology, Reproductive Biology and Anaesthetics
Imperial College London
London, UK

Statistics and evidence-based healthcare

Peter H Dixon PhD BSc
Maternal and Fetal Disease Group
Institute of Reproductive and Developmental Biology
Faculty of Medicine, Imperial College London,
Hammersmith Hospital
London, UK

Structure and function of the genome

Kate Hardy BA PhD
Professor of Reproductive Biology
Institute of Reproductive and Developmental Biology
Faculty of Medicine, Imperial College London,
Hammersmith Hospital
London, UK

Embryology

Andrew JT George MA PhD FRCPath FRSA
Professor of Molecular Immunology
Department of Immunology, Division of Medicine,
Faculty of Medicine, Imperial College London,
Hammersmith Hospital
London, UK

Immunology

Mark R Johnson PhD MRCP MRCOG
Professor of Obstetrics
Department of Maternal and Fetal Medicine
Imperial College School of Medicine
Chelsea and Westminster Hospital
London, UK

Endocrinology

Anna P Kenyon MBChB MD MRCOG
Clinical Lecturer
Institute for Women's Health
University College London
London, UK

Physiology

Sailesh Kumar
DPhil FRCS FRCOG FRANZCOG CMFM
Consultant/Senior Lecturer
Centre for Fetal Care
Queen Charlotte's and Chelsea Hospital
Imperial College London
London, UK

Fetal and placental physiology

Fiona Lyall BSc PhD FRCPath MBA
Professor of Maternal and Fetal Health
Maternal and Fetal Medicine Section
Institute of Medical Genetics
University of Glasgow
Glasgow, UK

Biochemistry

Vivek Nama MD MRCOG
Clinical Research Fellow
Maternal Medicine Department
Epsom & St Helier University Hospitals NHS Trust
Carshalton, Surrey, UK

Drugs and drug therapy

Sara Paterson-Brown FRCS FRCOG
Consultant in Obstetrics and Gynaecology
Queen Charlotte's and Chelsea Hospital
London, UK

Applied anatomy

Geoffrey L Ridgway
MD BSc FRCP FRCPath
Consultant Clinical Microbiologist and Honorary Senior Lecturer
University College London Hospitals NHS Trust
London, UK

Microbiology and virology

Neil J Sebire
MB BS BClinSci MD DRCOG FRCPath
Consultant in Paediatric Pathology
Department of Histopathology
Camelia Botnar Laboratories
Great Ormond Street Hospital
London, UK

Pathology

Hassan Shehata MRCPI MRCOG
Consultant Obstetrician & Obstetric Physician
Epsom & St Helier University Hospitals NHS Trust
Carshalton, Surrey, UK

Drugs and drug therapy

Andrew Shennan MBBS MD FRCOG
Professor of Obstetrics
Maternal and Fetal Research Unit
King's College London
St Thomas' Hospital
London, UK

Clinical research methodology

David Talbert PhD MInstP
Senior Lecturer in Biomedical Engineering
Division of Maternal and Fetal Medicine
Imperial College School of Medicine
Hammersmith Hospital
London, UK

Physics

Paul Taylor
Department of Microbiology & Virology
Royal Brompton and Harefield NHS Trust
Royal Brompton Hospital
London, UK

Microbiology and virology

Dorothy Trump MA MB BChir FRCP MD
Professor of Human Molecular Genetics
Academic Unit of Medical Genetics
University of Manchester
St Mary's Hospital
Manchester, UK

Clinical genetics

David Williams MBBS, PhD, FRCP
Consultant Obstetric Physician
Institute for Women's Health
University College London Hospital
London, UK

Physiology

Preface

The way in which junior obstetricians and gynaecologists are being trained has undergone an unprecedented evolution in the eight years since the last edition of this book. Likewise, the MRCOG Part 1 examination has evolved to reflect the exciting advances in reproductive biology, the increased emphasis upon translating basic science discoveries to the bedside, and more modern ways of assessing knowledge. A new edition of this book is therefore timely. This book has been hugely popular since it was first published under the editorship of Geoffrey Chamberlain, Michael de Swiet and the late Sir John Dewhurst, and we are pleased to continue their excellent work. We have brought in several new authors to completely revise topics that were covered in the previous editions and have introduced new chapters on molecular genetics, clinical genetics and clinical trials to reflect the growing importance of these topics in clinical practice. New multiple choice questions and extended matching questions have been devised to match the format of the examination.

We are grateful to the previous editors and authors whose work formed the foundation of the current edition. We hope that this text will continue to help future obstetricians and gynaecologists to leap one of the first hurdles in their career paths and will also be a useful source of information to facilitate their ongoing understanding of basic science as it applies to clinical practice.

Phillip Bennett and Catherine Williamson
London 2010

Acknowledgements

The editors thank the previous editors, Geoffrey Chamberlain, Michael de Swiet and the late Sir John Dewhurst, the past and present contributors and the production and editorial team at Elsevier. We are also very grateful to Mrs Ros Watts for being an efficient interface between us, the contributors and the editorial team.

Contents

Chapter **One**

1

Structure and function of the genome

Peter Dixon

CHAPTER CONTENTS

This chapter will provide a basic introduction to the human genome and some of the tools used to analyse it. Genomics and molecular biology have developed rapidly during the last few decades, and this chapter will highlight some of these advances, in particular with respect to the impact on our knowledge of the structure and function of the genome. The basic science described in this chapter is fundamental to the understanding of the field of clinical genetics, which is described in the following chapter.

Chromosomes

Inheritance is determined by genes, carried on chromosomes in the nuclei of all cells. Each adult cell contains 46 chromosomes, which exist as 23 pairs, one member of each pair having been inherited from each parent. Twenty-two pairs are homologous and are called *autosomes*. The 23rd pair is the sex chromosomes, X and Y in the male, X and X in the female.

Each cell in the body contains two pairs of autosomes plus the sex chromosomes for a total of 46, known as the diploid number (symbol N). Chromosomes are numbered sequentially with the largest first, with the X being almost as large as chromosome 1 and the Y chromosome being the smallest. This means that each cell (except gametes) has two copies of each piece of genetic information. In females, where there are two X chromosomes, one copy is silent (inactive), i.e. genes on that chromosome are not being transcribed (see below).

Each individual inherits one chromosome of each pair from their mother and one from their father following fertilization of the haploid egg (containing one of each autosome and one X chromosome) by the haploid sperm (containing one of each autosome and either an X or a Y chromosome). The sex of the individual is therefore dependent on the sex chromosome in the sperm: an X will lead to a female (with the X chromosome from the egg) and a Y chromosome will lead to a male (with an X from the egg).

Chromosomes are classified by their shape. During metaphase in cell division chromosomes are constricted and have a distinct recognizable 'H' shape with two chromatids joined by an area of constriction called the centromere. For 'metacentric' chromosomes the centromere is close to the middle of the chromosome and for 'acrocentric' chromosomes it is near to the end of the chromosome. The area or 'arm' of the chromosome above the centromere is known as the 'p arm' and the area below is the 'q arm'. For acrocentric chromosomes, the p arm is very small consisting of tiny structures called 'satellites'. Within the two arms regions are numbered from the centromere outwards to give a specific 'address' for each chromosome region (Fig. 1.1). The ends of the chromosomes are called telomeres. Chromosomes only take on the characteristic 'H' shape during a metaphase when they are undergoing division (hence giving the two chromatids).

Chromosomes are recognized by their banding patterns following staining with various compounds in the cytogenetic laboratory. The most commonly used stain is the Giemsa stain (G-banding) which gives a characteristic black and white banding pattern for each chromosome.

In the cell, the chromosomes are folded many hundreds of times around histone proteins and are usually only visible under a microscope during mitosis and meiosis. DNA is composed of a deoxyribose backbone, the 3-position (3′) of each deoxyribose being linked to the 5-position (5′) of the next by a phosphodiester bond. At the 2-position each deoxyribose is linked to one of four nucleic acids, the purines (adenine or guanine) or the pyrimidines (thymine or cytosine). Each DNA molecule is made up of two such strands in a double helix with the nucleic acid bases on the inside. This is the famous double helix structure that was first proposed by Watson and Crick in 1953. The bases pair by hydrogen bonding, adenine (A) with thymine (T) and cytosine (C) with guanine (G). DNA is replicated by separation of the two strands and synthesis by DNA polymerases of new complementary strands. With one notable exception, the reverse transcriptase produced by viruses, DNA polymerases always add new bases at the 3′ end of the molecule. RNA has a structure similar to that of DNA but is single stranded. The backbone consists of ribose, and uracil (U) is used in place of thymine (Fig. 1.2).

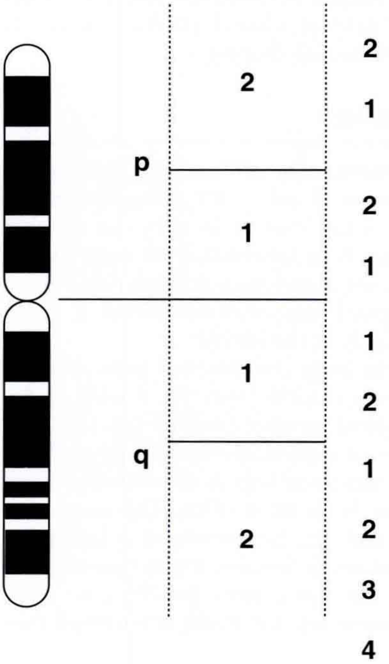

Figure 1.1 • Diagrammatic representation of the X chromosome. Note that the short arm (referred to as p) and the long arm (referred to as q) are each divided into two main segments labelled 1 and 2, within which the individual bands are also labelled 1, 2, 3, etc. (Courtesy of Dorothy Trump.)

Gene structure and function

DNA is organized into discrete functional units known as genes. Genes contain the information for the assembly of every protein in an organism via the translation of the DNA code into a chain of amino acids to form proteins. DNA that encodes a single amino acid consists of three bases, or letters. With four letters and three positions in each 'word', there are 64 possible

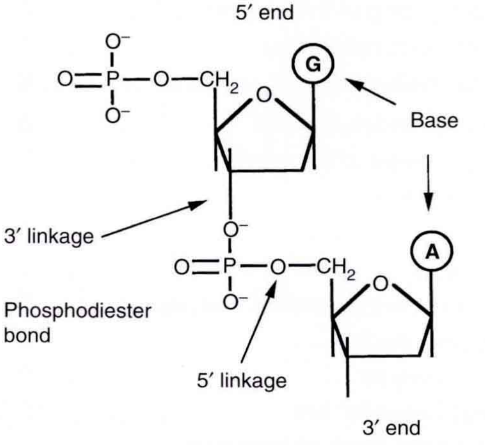

Figure 1.2 • The sugar phosphate backbone of DNA.

combinations of DNA, but in fact only 20 amino acids are coded for (Table 1.1). Therefore, the third base of a codon is often not crucial to determining the amino acid – a phenomenon known as wobble.

A diagram of a typical gene structure is shown (Fig. 1.3). Each gene gives rise to a message (messenger RNA), which can be interpreted by the cellular machinery to make the protein that the gene encodes.

Genes are split into exons, which contain the coding information, and introns, which are between the coding regions and may contain regulatory sequences that control when and where a gene is expressed. Promoters (which control basal and inducible activity) are usually upstream of the gene, whereas enhancers (which usually regulate inducible activity only) can be found throughout the genomic sequence of a gene. The two base pair sequences at the boundary of introns and exons (the splice acceptor and donor sites), identical in over 99% of genes, are known as the splice junction (Fig. 1.3); they signal cellular splicing machinery to cut and paste exonic sequences together at this point. The first residue of each gene is almost always methionine, encoded by the codon ATG.

Recent estimates based on the genome sequence put the number of genes at <30 000, a huge reduction from earlier estimates. This means that the vast majority of

Table 1.1 The genetic code

1st position	2nd position				3rd position
	T	C	A	G	
T	Phe	Ser	Tyr	Cys	T
	Phe	Ser	Tyr	Cys	C
	Leu	Ser	STOP	STOP	A
	Leu	Ser	STOP	Tyr	G
C	Leu	Pro	His	Arg	T
	Leu	Pro	His	Arg	C
	Leu	Pro	Gln	Arg	A
	Leu	Pro	Gln	Arg	G
A	Ile	Thr	Asn	Ser	T
	Ile	Thr	Asn	Ser	C
	Ile	Thr	Lys	Arg	A
	Met	Thr	Lys	Arg	G
G	Val	Ala	Asp	Gly	T
	Val	Ala	Asp	Gly	C
	Val	Ala	Glu	Gly	A
	Val	Ala	Glu	Gly	G

Note that in RNA thymidine (T) is replaced by uracil (U).

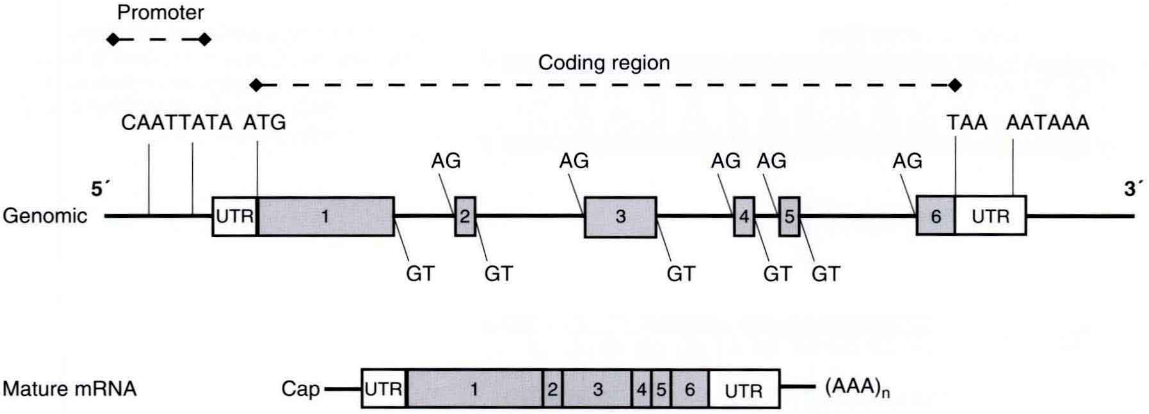

Figure 1.3 • Schematic representation of generalized gene structure. The upper panel shows the genomic organization of a typical gene (with a variety of key features indicated) and the lower panel the mRNA resulting from the transcription of this gene. Key features indicated include the consensus splice sites GT (donor) and AG (acceptor), the initiation codon (ATG), the stop codon (TAA) and polyadenylation signal (AATAAA). Typical promoter motifs are indicated (CAAT and TATA) together with 5′ and 3′ untranslated regions (UTR). Mature mRNAs have a protective 5′ cap (a guanosine nucleotide connected to the mRNA by means of a 5′ to 5′ triphosphate linkage).

human DNA does not contain a coding sequence (i.e. exons), but is rather an intronic sequence: structural motifs and regulatory regions. This is distinct from lower organisms, e.g. bacteria, where >95% of the DNA is a coding sequence. Just exactly why this 'unused' DNA is present remains somewhat enigmatic. The other implication of this finding is that the huge complexity of humans compared to other organisms with similar numbers of genes must arise from more subtle regulation of gene expression, rather than greater numbers of different genes.

The central dogma of molecular biology

The central dogma of molecular biology concerns the information flow pathway in cells and can be simply summarized as: 'DNA makes RNA makes protein, which in turn can facilitate the two prior steps'. These steps are now explained in more detail.

Transcription

'Transcription' is the process of the information encoded in DNA being transferred into a strand of messenger RNA (mRNA). During transcription the RNA polymerase, which constructs the complementary mRNA, reads from the DNA strand complementary to the RNA molecule. This is known as the anti-sense strand while the opposite strand, which has the same base pair composition as the RNA molecule (with thymidine (T) in place of uracil (U) as mentioned previously), is the sense strand. Gene sequences are expressed as the sequence of the sense strand of DNA, although it is in fact the anti-sense strand which is read (Fig. 1.4). The vast majority of genes consist of a 5′ untranslated region (UTR) containing response elements to which proteins may bind that influence transcription. The 5′ regions of genes are frequently characterized by elements such as the TATA and CAAT boxes (Fig. 1.3) and are often richer in GC pairs than elsewhere in the genome. This is frequently the case around the 5′ ends of 'housekeeping' genes that are constitutively expressed in the majority of tissues. There then follows the transcribed sequence. The expressed coding parts of the gene are known as the exons, while the intervening sequences are known as introns. The coding portion of the gene is often interrupted by one or more non-coding intervening sequences, although numerous examples of single exon genes exist. Initially, the RNA molecule transcribes both introns and exons and is known as heavy nuclear RNA (hnRNA). The exons are perfectly spliced out (as marked by the splice boundary sequences) and a protective cap added before the now mature mRNA exits the nucleus. Hence, cytoplasmic mRNA consists only of coding regions flanked by untranslated regions at the two ends. A polyadenine (poly A) tail is added to most mRNA molecules at their 3′ end, facilitated by the polyadenylation signal found past the stop codon in the coding sequence. This tail, found on the great majority of expressed mRNAs, serves to protect the RNA from degradation prior to translation by the ribosome (see below).

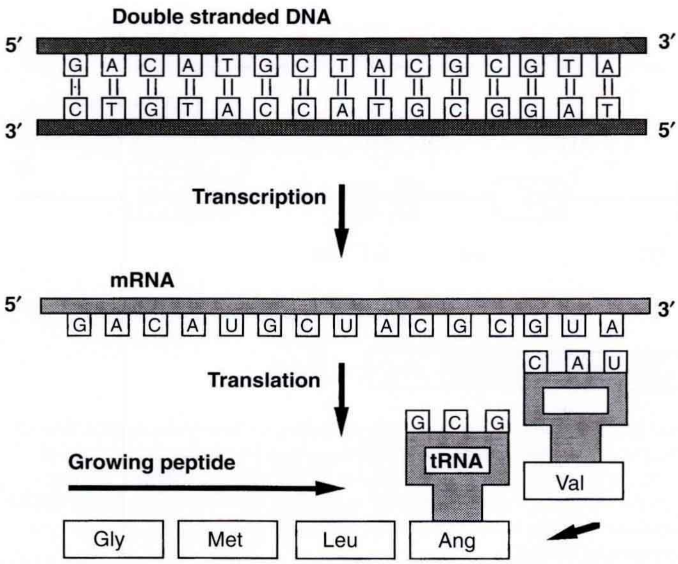

Figure 1.4 • Transcription and translation. Double-stranded DNA is transcribed forming a complementary single-stranded molecule of RNA. The mRNA is translated by tRNA (transfer RNA) to form the peptide chain.

Translation

The term 'translation' describes the process whereby the cellular machinery reads the mRNA code and creates a chain of polypeptides (i.e. a protein). Once in the cytoplasm, the mRNA message is translated into protein by a ribosome. Ribosomes, consisting of a complex bundle of proteins and ribosomal RNA, attach to mRNA at the 5′ end. Protein synthesis begins at the amino terminal and amino acids are sequentially added at the freshly made carboxyl end. Amino acids are brought into the reaction by specific transfer RNA (tRNA) molecules. Each tRNA is a single-stranded molecule which folds in a way that allows complementary base pairing between parts of the same strand. The specific configuration allows the tRNA molecule to bind to its specific amino acid. There remains, unpaired, at one end of the molecule, three bases which are complementary to the codon coding for the amino acid. This anticodon binds to the codon of the mRNA and places the amino acid in the correct sequence of the protein (Fig. 1.4). Usually, several ribosomes translate a single mRNA molecule at any one time.

Replication

'Replication' is the process whereby DNA is copied or replicated to permit transmission of genetic information to offspring. DNA replication is performed prior to cell division, when an identical copy must be made for each daughter cell resulting from division. Replication occurs before mitosis, the normal form of cellular division where resulting cells have identical DNA to the original. Meiosis, the second form of cellular division, occurs during gametogenesis, and results in haploid cells, i.e. cells with half the usual complement of DNA. In meiosis the resulting cells (gametes) are haploid, i.e. carry only a single copy of the genomic sequence.

It is important to note that since this dogma was first established in 1958 by Crick, a number of exceptions have been identified. For example retroviruses (e.g. HIV-1) can cause information to flow from RNA to DNA by integrating their genome (carried as RNA) into that of the host. A second example is ribozymes, which are functional enzymes composed solely of RNA and hence have no need to be translated into protein.

Regulation of gene expression

When a gene is actively being transcribed into mRNA and then translated into a protein, it is said to be 'expressed'. Gene expression can be controlled at several levels. Transcription of DNA into mRNA is generally regulated by the binding of specific proteins, known as transcription factors, to the region of DNA just upstream, or 5′, of the coding sequence itself. Other proteins can bind enhancer sequences that may be within the gene or a long way upstream or downstream.

The promoter contains specific DNA sequence motifs which bind transcription factors. In general, transcription factors become active when the cells receive some form of signal and then translocate to the nucleus, where they bind to specific sequences in the promoters of specific genes and activate transcription. Other genes, often known as housekeeping genes, have a constant level of expression and are not induced in this way.

Many different types of transcription factor exist with different modes of action. Typical examples of two types will be considered here, namely intracellular nuclear hormone receptors (which are transcription factors) and cell surface receptors, which are capable of activating transcription factors.

Members of the nuclear hormone receptor superfamily, such as the progesterone receptor and the thyroid hormone receptor, are present mainly in the cytoplasm of the cell. When a steroid hormone crosses the lipid bilayer of the cell membrane, it binds to the receptor which is usually dimerized to form pairs of receptor molecules. The receptor/hormone dimer complex then translocates to the nucleus and binds to response elements in the promoters of target genes, where it activates (or indeed represses) transcription. This process also involves the recruitment of many other co-factors to the dimer complex which are also involved in regulation of the expression of the target gene.

Cell surface receptors, subsequent to binding of ligands, can activate pathways leading to the formation of active transcription factors. For example activation of tyrosine kinase-linked receptors on the cell surface may lead to a series of phosphorylation events within the cell, culminating in the phosphorylation of the protein Jun. Jun will then combine with the protein Fos to form a dimer transcription factor called AP-1, which can bind to specific AP-1 binding sites in the promoters of responsive genes.

In another example of cell surface receptor action, the 'inflammatory' transcription factor NFκB exists in the cytoplasm of cells as dimers bound to an inhibitory protein IκB. Mediators of inflammation, such as the inflammatory cytokine interleukin 1β, bind to cell surface receptors and activate a chain of biochemical events that result in the phosphorylation and subsequent breakdown of IκB. Uninhibited NFκB dimers then translocate to the nucleus to activate genes whose promoters contain NFκB DNA binding motifs.

Gene expression can also be controlled by regulation of the stability of the transcript. Most mRNA molecules are protected from degradation by the presence of their poly-A tail. Degradation of mRNA is controlled by specific destabilizing elements within the sequence of the molecule. One type of destabilizing element has been well characterized. The Shaw–Kayman or AU-rich sequence (ARE) is a region of RNA, usually within the 3′ untranslated region, in which the motif AUUUA is repeated several times. Rapid response genes, whose expression is rapidly switched on and then off again in response to some signal, often contain an ARE within their 3′ untranslated region. Binding of specific proteins to the ARE leads to removal of the mRNA's poly-A tails and then to degradation of the molecule.

Epigenetics

The field of epigenetics is concerned with modifications of DNA and chromatin that do not affect the underlying DNA sequence. In recent years, the importance of these modifications has come to light and this is now a very active area of research.

Epigenetic modification of DNA

The principal epigenetic modification of DNA is methylation, whereby a methyl group ($-CH_3$) is added to a cytosine, converting it to 5-methylcytosine. This can only occur where a cytosine is next to a guanine, i.e. joined by a phosphate linkage, and is usually described as CpG to distinguish it from a cytosine base-paired to a guanine via hydrogen bonds across the double helix.

Methylation, particularly in the 5′ promoter regions of genes that are often GC-rich, is associated with silencing. Humans have at least three DNA methyl transferases, and the process is critical to imprinting (parent of origin-dependent gene expression) and X inactivation. Abnormal DNA methylation is being increasingly recognized as playing a role in cancer cell development.

Epigenetic modification of histones

Histone proteins are associated with DNA to form nucleosomes, which make up chromatin. Two of each histone protein (2A, 2B, 3 and 4) form the octameric core of the nucleosome, with H1 histone attached and linking nucleosomes to form the 'beads on a string' structure. Chromatin structure plays an important role in regulation of gene expression, and this structure is heavily influenced by modifications of the histone proteins. These modifications usually occur on the tail region of the protein, and include methylation, acetylation, phosphorylation and ubiquitination. Combinations of modifications are considered to constitute a code (the so-called histone code), which it is hypothesized, control DNA–chromatin interaction. A comprehensive understanding of these mechanisms has not yet been elucidated; however some functions have been worked out in detail. For example, deacetylation allows for tight bunching of chromatin, preventing gene expression.

Mitochondrial DNA

In addition to the genomic DNA present within cells, another type of DNA is present – mitochondrial DNA. The mitochondria are small organelles within cells that have a unique double-layered membrane and are the energy source for cellular activity and metabolism via production of adenosine triphosphate (ATP). They have their own genome (mtDNA), consisting of a single circular piece of DNA of 16 568 base pairs and encoding 37 genes. Mitochondria are only ever inherited maternally because all the mitochondria in a zygote come from the ovum and none from the sperm. Mitochondrial DNA can be used for confirming family relatedness through analysis of the maternal lineage. In addition, mitochondrial DNA has been successfully and reproducibly extracted from ancient DNA samples, largely due to the high copy number compared with nuclear DNA. Mutations in mitochondrial DNA are responsible for a number of human diseases (see Ch. 2).

Studying DNA

The vast majority of DNA samples used for genetic analysis originate from a peripheral blood sample, usually collected in a 10 mL tube containing an anticoagulant, e.g. EDTA. From this sample, large quantities of DNA are easily extracted from the leucocytes using one of the many commercial kits available. This has replaced the older method of phenol/chloroform extraction. Alternatively, if only a small amount of DNA is required, buccal swabs can be used to collect DNA. As this is non-invasive, it has considerable advantage, for example where patients are needle-phobic, or where DNA is required from small children. It is also possible to extract usable quantities of DNA from very small amounts of tissue or blood from archive samples such as formalin-fixed paraffin-embedded sections.

Mendelian genetics and linkage studies

The majority of advances in recent years in disease gene identification have come from the field of Mendelian disease. This refers to diseases (e.g. cystic fibrosis or muscular dystrophy) where the inheritance pattern follows classical Mendelian principles, i.e. those established by Gregor Mendel at the end of the nineteenth century. His work, long before the existence of DNA was known, established simple rules for inheritance of characteristics (phenotypes). That is, a disease can be dominant (requiring only one mutant allele to have the disease), recessive (requires two) or X-linked (one mutant allele on the X chromosome and hence much more common in males). Since the first gene was identified by linkage/positional cloning in 1986, well over 1000 Mendelian disease genes have been identified, initially by the use of linkage studies.

Linkage studies rely on the use of large, phenotypically well-characterized families. Typically, 12 or more affected family members are required for tracing autosomal dominant diseases, but far smaller families with as few as three affected individuals can be used for recessive diseases. Family members are typed for polymorphic markers throughout the genome in order to detect which regions the affected individuals share, and hence are more likely to contain the disease gene. The marker of choice for these studies is usually short tandem repeats (STRs) which are more commonly known as microsatellites. These markers are repeat sequences that most commonly consist of dinucleotide base repeats, e.g. $(CA)_n$, but they may also comprise tri- or tetranucleotide repeats. These markers exhibit length polymorphism, such that they are different lengths in different individuals, and can be heterozygous. For example an individual may carry at one marker position one repeat of five units and one of seven. These different repeat lengths are easily detectable by common molecular biology techniques. If a disease gene is close to a particular marker, i.e. linked, it will almost always be inherited with it. Thus, if affected individuals all show the same length repeat at a particular marker, the disease gene may be close by. Statistical analysis is used to formalize the results and give likelihood ratios, the LOD score, or the location of a disease locus.

In the recent past, linkage studies were followed by positional cloning to identify a disease gene. This method of gene identification is so called because genes are identified primarily on the basis of their position in the genome, with no underlying assumptions about the protein they encode. After the linkage of a disease had been achieved, a physical map of the linked region was constructed. This was done using large-scale cloning vectors such as YACs (yeast artificial chromosomes) or BACs (bacterial artificial chromosomes), which contain inserts of up to a megabase (1 000 000 base pairs) of the human genome. Libraries of the whole genome were screened with the microsatellite markers used that had been linked to the disease and a series of overlapping clones, or contig, of the linked region constructed. Once this had been established, these clones would be searched for genes which when identified would be screened for mutations in affected patients. This search would have utilized a variety of methods such as direct library hybridization or exon trapping to identify genes within the contig. Much of this work however is now unnecessary due to the greatest advance in the field of human genetics in the last few years – the completion of the sequence of the human genome.

The sequencing of the genome

The completion of the human genome sequencing project has transformed the field of genetics. In brief, BAC (see above) libraries were constructed from the DNA of a handful of anonymous donors, and arranged in order around the genome using genetic markers with established positions. Each BAC was then sequenced and, by the use of high-powered computers, the sequence was assembled, first into the original BAC and then, by matching overlaps, to build up a sequence for the entire genome. The genome centres involved in this project utilized vast numbers of sequencing machines and a production-line environment to achieve the throughput required. In addition to the publicly funded consortium, a private company also produced a complete human genome sequence using a slightly different methodology.

Individual labs and researchers now have access to the entire genome dataset from the publicly funded project freely available on the internet. This information is an invaluable resource and has greatly accelerated research into the molecular aetiology of genetic disease. Once the position of a disease gene has been confirmed (linkage), scientists can now employ an *in-silico* (i.e. computer-based) approach to identifying the disease gene. Practically, this involves searching databases for all the identified genes in a region and then sequencing them in affected individuals to look for mutations. These 'positional candidates' are often prioritized using other sources of information such as tissue expression pattern or predicted function. Once mutations have been identified, functional studies of mutant forms of the protein to determine the exact nature of the molecular aetiology of the disease in question are often pursued.

Completion of this project has enabled genome centres to focus on two other areas: that of whole-genome sequencing of other organisms for comparative purposes, and so-called 'deep resequencing' to identify the spectrum of genetic variation in human populations.

Analysis of complex traits

The vast majority of so-called 'genetic' disease does not fall into the category of Mendelian disease. Rather, it is caused by so-called complex genetic disease or traits, where a number of genetic factors interacting with the environment result in a disease phenotype. It is this area of genetics that current research is most focused upon.

An example of such a disease in obstetrics is pre-eclampsia (see later chapters). It is important to note that in this type of genetic disease the mutant gene may only be having a small effect on disease susceptibility, and for each disease a large number of genes together with environmental influences may be playing a role.

Methods of analysis of complex traits can be broadly divided into two areas: family-based studies and case–control studies. Family-based studies are usually based upon microsatellite typing approaches (see above), whereas association studies (otherwise known as case–control studies) generally employ another kind of genetic marker, single nucleotide polymorphisms (SNPs). SNPs are much more frequent throughout the genome (every 1000 bases or so) and although they have a lower information content than microsatellites can be used for much finer mapping studies, thanks to their more frequent occurrence.

Family-based studies rely on large collections of nuclear families, parent–offspring trios and/or affected or discordant sibling (sib) pairs. The term discordant refers to disease status, i.e. a discordant sib pair comprises one affected and one unaffected individual. Unaffected family members act as controls.

The dissection of complex traits using these approaches has been problematic for many years for a variety of reasons. These include insufficient sample size (i.e. underpowered studies), inappropriate controls (in association studies) and a lack of knowledge about the underlying structure of the genome (i.e. the patterns of linkage disequilibrium, or the underlying non-random association of markers). In addition, very little was known on a genome-wide scale about the pattern of naturally occurring human variation. However, with a more complete understanding of the structure of the genome, and ever-larger sample resources, significant and reproducible associations of genetic variation with common human disease are emerging. Technology has played a role too, with it now being possible to type many thousands of SNPs in a single experiment using DNA array technologies.

Molecular biology techniques

The manipulation of DNA, RNA and proteins at a molecular level is collectively referred to as molecular biology. This term encompasses a huge range of techniques some of which are outlined here. All of these techniques are in routine use in clinical and research labs around the world.

Restriction endonucleases

One of the key tools used to manipulate DNA is restriction endonucleases. These enzymes, which have been isolated from a wide range of bacteria, cut or restrict DNA at a certain site determined by the base sequence. The reaction occurs under certain conditions, i.e. at the correct temperature and in the correct buffer (usually supplied by the manufacturer). These known recognition sites can be used to manipulate DNA for cloning, blotting, etc. The enzymes have usually been isolated from microorganisms, and their name reflects the organism from which they have been isolated. For example, the common restriction enzyme EcoRI, which cuts or restricts DNA at the sequence GAATTC, was isolated from *Escherichia coli* RY13. *Note*: the recognition of the restriction site depends upon double-stranded DNA, and the cleavage can result in an overhang of a few bases ('sticky ends') or a straight cut across both strands ('blunt ends').

The polymerase chain reaction

The polymerase chain reaction (PCR) is the bedrock of molecular biology and refers to a procedure whereby a known sequence of DNA (the target sequence) can be amplified many millions of times to generate enough copies to visualize, clone, sequence or manipulate in many other ways. A known DNA sequence is amplified first by using a uniquely designed pair of primers at the start (5′) and end (3′; on the reverse strand) of the sequence to be amplified. The primers are thus small pieces of DNA, known as oligonucleotides (oligos), and are usually synthesized by commercial companies for relatively minimal cost. The primers are used in combination with a buffer, a source of deoxyribose nucleotide triphosphate (dNTP) building blocks, the target DNA and Taq polymerase. This polymerase, first isolated from *Thermophilus aquaticus*, is able to replicate DNA at high temperatures. Once prepared, the reaction is placed into a thermal cycler. The reaction proceeds through a number of repeated cycles where the DNA template is denatured, the primers anneal and the polymerase extends the products. Cycling of these three temperatures (one for each of the above steps) results in an exponential amplification of the target sequence. Following amplification, products can be visualized by agarose gel electrophoresis (see below).

Many other commonly used applications are based around the principles of PCR. For example, reverse transcription PCR (RT-PCR), which can be applied to RNA analysis. This technique uses reverse transcriptase enzymes isolated from retroviruses to generate DNA copies of template RNA to detect expression of a particular gene. This approach is further enhanced by quantitative RT-PCR, where relative or absolute expression levels of a particular message can be measured.

Another development of PCR is whole genome amplification, which relies on the use of specialist polymerases to amplify the entire genome in a single reaction, a very useful tool when the amount of sample available is limited.

Electrophoresis

DNA molecules are slightly negatively charged and hence, under the right conditions, will migrate towards a positive charge. This phenomenon can be exploited to visualize DNA. For example the results of a PCR reaction (see above) can be assessed in this way, or a sample of genomic DNA digested with a restriction enzyme can be separated. DNA samples are loaded onto an agarose gel (a sieving mixture of seaweed extract) in the range of 0.5–4% (depending on the size range of DNA to be separated) in a tank containing running buffer (commonly Tris/borate/EDTA). Under an electric current the DNA will migrate at a rate proportional to its size. The samples can then be visualized under a UV light box after the addition of ethidium bromide, or one of the newer less toxic alternatives (e.g. Sybersafe). Larger DNA molecules and RNA samples can also be visualized by electrophoresis. Slightly different conditions are used to protect the RNA, which is inherently more unstable than DNA, and specialized running equipment is need to separate DNA molecules >10 kb in size.

Blotting

DNA (in the case of Southern blotting), RNA (northern) and protein (western) can be fixed to nylon membranes for further analysis, e.g. for screening with a radioactively labelled probe (DNA/RNA) or with an antibody raised to an epitope of interest (proteins). This is a fairly straightforward and routine procedure, which enables a range of downstream experiments to be carried out. For example, a genomic DNA digest can be screened with a radiolabelled or biotinylated probe for a gene sequence of interest, or an antibody raised against a particular protein can be used to screen for that protein in cellular extracts.

Sequencing

DNA sequencing is now a rapid and straightforward process. The sequence of an amplified fragment of DNA is determined using a variation of the PCR method incorporating fluorescently labelled bases which can be read by a laser detection system. In this application, a PCR cycle is performed using only one primer, either forward or reverse, and the labelled nucleotides. This results in linear amplification of product with consecutive lengths of sequence with a fluorescent tag corresponding to the final base of the fragment. When run on a slab gel or capillary and read by a laser, the sequence is determined by the sequential reading of each base. Recent advances in the use of capillary-based machines with multiple channels have resulted in a huge increase in throughput and capacity, and facilitated the rapid acceleration in efforts to sequence the entire human genome.

Cloning vectors and cDNA analysis

As outlined above, the human genome sequence now makes it unnecessary to clone genes from a candidate region before mutation analysis. However, cloning is still a critical part of the analysis of gene function subsequent to mutation detection. For example, using some of the techniques outlined above in the molecular biology section, the expression pattern of a gene can be studied, factors that induce transcription can be identified, and so on. Many of these techniques rely on the use of cDNA clones. These are vectors of much smaller size than YACs and are carried and propagated in bacteria as plasmids or phage. They may also be introduced into cell lines by transfection. The vectors contain an insert of DNA, which corresponds to the full-length mRNA of the gene in question; this is known as copy DNA (cDNA) and contains only the exonic material of the gene. Clones may be screened from libraries or in many cases purchased from commercial sources. Isolation and propagation of these clones in a suitable host strain of bacteria allows detailed analysis of gene function.

Expression studies

A detailed explanation of protein analysis is beyond the scope of this chapter. Key concepts to understand are that proteins can be expressed in mammalian and bacterial systems, their interactions studied and function analysed. A recent approach gaining popularity is to use short interfering RNA (siRNA) to 'knock-down' genes of interest in both *in-vitro* and *in-vivo* systems. In this approach, a vector is introduced which expresses short pieces of carefully designed RNA. These RNA molecules interact with cellular machinery and interfere with endogenously expressed mRNA by targeting it for degradation. This results in the reduction, or knocking down, of the expression of the target gene by up to 80% of the original expression level.

In-silico analysis

The free availability of the human genome sequence via the internet has greatly enhanced the use of computer analysis for molecular biology. This has led to an enormous rise in the discipline of 'bioinformatics', which can be simply defined as deriving knowledge from computer analysis of biological data.