BROOK'S CLINICAL PEDIATRIC ENDOCRINOLOGY

7TH EDITION

Edited by
Mehul T. Dattani
Charles G.D. Brook

WILEY

Brook's Clinical Pediatric Endocrinology

Brook's Clinical Pediatric Endocrinology

Seventh Edition

Edited by

Mehul T. Dattani, MD FRCP FRCPCH
Professor and Head of Paediatric Endocrinology
Genetics and Genomic Medicine Programme
UCL Great Ormond Street Institute of Child Health
Great Ormond Street Hospital for Children NHS Foundation Trust
London, UK

Charles G. D. Brook, MA MD FRCP
Emeritus Professor of Paediatric Endocrinology
University College London
London, UK

WILEY

This edition first published 2020

Edition History
Wiley-Blackwell (6e, 2009)

Registered Office(s)
John Wiley & Sons, Inc., 111 River Street, Hoboken, NJ 07030, USA
John Wiley & Sons Ltd, The Atrium, Southern Gate, Chichester, West Sussex, PO19 8SQ, UK

Editorial Office
9600 Garsington Road, Oxford, OX4 2DQ, UK

For details of our global editorial offices, customer services, and more information about Wiley products visit us at www.wiley.com.

Wiley also publishes its books in a variety of electronic formats and by print-on-demand. Some content that appears in standard print versions of this book may not be available in other formats.

Library of Congress Cataloging-in-Publication Data

Names: Dattani, Mehul T., editor. | Brook, C. G. D. (Charles Groves Darville), editor.
Title: Brook's clinical pediatric endocrinology / edited by Mehul T. Dattani, Charles G.D. Brook.
Other titles: Clinical pediatric endocrinology
Description: 7th edition. | Hoboken, NJ : Wiley-Blackwell, 2020. | Preceded by Brook's clinical pediatric endocrinology / edited by Charles G.D. Brook, Peter E. Clayton, Rosalind S. Brown. 6th ed. 2009. | Includes bibliographical references and index. |
Identifiers: LCCN 2019003272 (print) | LCCN 2019003907 (ebook) | ISBN 9781119152699 (Adobe PDF) | ISBN 9781119152705 (ePub) | ISBN 9781119152682 (hardback)
Subjects: | MESH: Endocrine System Diseases | Adolescent | Child | Infant
Classification: LCC RJ418 (ebook) | LCC RJ418 (print) | NLM WS 335 | DDC 618.92/4–dc23
LC record available at https://lccn.loc.gov/2019003272

Cover Design: Wiley
Cover Image: © VikiVector/Getty Images

Set in 10/12pt Warnock by SPi Global, Pondicherry, India

10 9 8 7 6 5 4 3 2 1

Contents

List of Contributors

Mario Abinun
Honorary Clinical Senior Lecturer
Newcastle University; and
Department of Paediatric Immunology
Royal Victoria Infirmary
Newcastle upon Tyne
UK

John C. Achermann
Genetics & Genomic Medicine Programme
UCL Great Ormond Street Institute of Child Health
London, UK

Kyriaki-Sandy Alatzoglou
Consultant in Paediatrics & Endocrinology, Chelsea & Westminster Hospital NHS Foundation Trust
London, UK

Jeremy Allgrove
Consultant Paediatric Endocrinologist,
Department of Paediatric Endocrinology,
Great Ormond Street Hospital for Children Foundation Trust, London, UK

G.R. Ambler
Institute of Endocrinology and Diabetes, The Children's Hospital at Westmead and Clinical School
The University of Sydney, Sydney, Australia

Anu Bashamboo
Human Developmental Genetics, Institut Pasteur
Paris, France

Gerhard Binder
University Children's Hospital, Pediatric Endocrinology and Diabetology, Tübingen, Germany

F.J. Cameron
Department of Endocrinology and Diabetes
Royal Children's Hospital, Parkville, Victoria
Australia

Jean-Claude Carel
Assistance Publique-Hôpitaux de Paris, Robert Debré University Hospital, Paediatric Endocrinology Diabetology Department, Reference Center for Growth and Development Endocrine Diseases
Paris Diderot University, INSERM U1141
Paris, France

P.E. Clayton
Division of Developmental Biology and Medicine
School of Medical Sciences, Faculty of Biology Medicine and Health, University of Manchester; and
Department of Paediatric Endocrinology
Royal Manchester Children's Hospital
Manchester, UK

Tim D. Cheetham
University Reader and Honorary Consultant in Paediatric Endocrinology, Newcastle University; and
Department of Paediatric Endocrinology
Royal Victoria Infirmary, Newcastle upon Tyne, UK

Wassim Chemaitilly
Department of Pediatric Medicine-Division of Endocrinology, St. Jude Children's Research Hospital; and
Department of Epidemiology and Cancer Control
St. Jude Children's Research Hospital
Memphis, TN, USA

Moira Cheung
Consultant Paediatric Endocrinologist, Evelina Children's Hospital, Guy's and St Thomas' Hospital NHS Foundation Trust, London, UK

Tinh-Hai Collet
University of Cambridge Metabolic Research Laboratories, Wellcome Trust-MRC Institute of Metabolic Science, Addenbrooke's Hospital
Cambridge, UK

Martine Cools
Ghent University Hospital – Princess Elizabeth Children's Hospital, Ghent University, Ghent, Belgium

Mehul T. Dattani
Professor and Head of Paediatric Endocrinology
Genetics and Genomic Medicine Programme
UCL Great Ormond Street Institute of Child Health
Great Ormond Street Hospital for Children NHS Foundation Trust
London, UK

Leo Dunkel
Centre for Endocrinology, William Harvey Research Institute, Barts and the London School of Medicine and Dentistry, Queen Mary University of London
London, UK

I. Sadaf Farooqi
University of Cambridge Metabolic Research Laboratories, Wellcome Trust-MRC Institute of Metabolic Science, Addenbrooke's Hospital
Cambridge, UK

Emma Footit
Department of Metabolic Medicine
Great Ormond Street Children's Hospital
London, UK

Hoong-Wei Gan
Clinical Research Fellow & Paediatric Endocrinology Registrar, Genetics & Genomic Medicine Programme
UCL Great Ormond Street Institute of Child Health & Great Ormond Street Hospital for Children NHS Foundation Trust, London, UK

Evelien F. Gevers
Department of Pediatric Endocrinology, Royal London Children's Hospital, Barts Health NHS Trust; and
Centre for Endocrinology, William Harvey Research Institute, Queen Mary University of London
London, UK

Helena Gleeson
Consultant Endocrinologist
University of Birmingham Hospitals
Birmingham, UK

Sasha R. Howard
Centre for Endocrinology, William Harvey Research Institute, Barts and the London School of Medicine and Dentistry, Queen Mary University of London
London, UK

Melissa M. Hudson
Department of Epidemiology and Cancer Control
St. Jude Children's Research Hospital; and
Department of Oncology, Division of Cancer Survivorship, St. Jude Children's Research Hospital
Memphis, TN, USA

Claire R. Hughes
Centre for Endocrinology, William Harvey Research Institute, Queen Mary University of London & Royal London Children's Hospital, Barts Health NHS Trust
London, UK

Natascia Di Iorgi
Department of Pediatrics, Istituto Giannina Gaslini
University of Genova, Genova, Italy

K. Joshi
Department of Endocrinology and Diabetes, Royal Children's Hospital, Parkville, Victoria, Australia

Ritika Kapoor
Department of Paediatric Endocrinology, Kings College Hospital London, London, UK

Harshini Katugampola
Great Ormond Street Hospital for Children NHS Foundation Trust; and
Genetics & Genomic Medicine Programme, UCL GOS Institute of Child Health, London, UK

Wieland Kiess
Hospital for Children and Adolescents, Centre of Paediatric Research, Department of Women & Child Health, University of Leipzig, Leipzig, Germany

Birgit Köhler†
Department of Paediatric Endocrinology and Diabetology, Charité, Universitätsmedizin Berlin
Germany

Juliane Leger
Assistance Publique-Hôpitaux de Paris,
Robert Debré University Hospital, Paediatric Endocrinology Diabetology Department, Reference Center for Growth and Development Endocrine Diseases, Paris Diderot University, INSERM U1141
Paris, France

Mohamad Maghnie
Department of Pediatrics, Istituto Giannina Gaslini
University of Genova, Genova, Italy

† Deceased

Elim Man
Paediatric Endocrinology, Queen Mary Hospital, The University of Hong Kong, HKSAR, Hong Kong; and Genetics & Genomic Medicine Programme, UCL Great Ormond Street Institute of Child Health
London, UK

Ken McElreavey
Human Developmental Genetics, Institut Pasteur
Paris, France

Giovanni Morana
Department of Pediatric Neuroradiology, Istituto Giannina Gaslini, Genova, Italy

P.G. Murray
Division of Developmental Biology and Medicine
School of Medical Sciences, Faculty of Biology, Medicine and Health, University of Manchester; and
Department of Paediatric Endocrinology
Royal Manchester Children's Hospital
Manchester, UK

Flavia Napoli
Departments of Pediatrics, Istituto Giannina Gaslini
Genova, Italy

Catherine J. Owen
Consultant in Paediatric Endocrinology
Department of Paediatric Endocrinology
Royal Victoria Infirmary
Newcastle upon Tyne
UK

Simon H.S. Pearce
Professor of Endocrinology
Institute of Genetic Medicine
Newcastle University
Newcastle upon Tyne
UK

Catherine Peters
Consultant Paediatric Endocrinologist, Great Ormond Street Hospital, London, UK

Nicolas de Roux
Biochemistry Laboratory, INSERM U1141, Robert Debré Hospital, Paris Diderot University
Paris, France

Emmanouil Saloustros
Division of Intramural Research, Eunice Kennedy Shriver National Institute of Child Health and Human Development, Bethesda, MD, USA

Nadia Schoenmakers
Wellcome Trust-Medical Research Council Institute of Metabolic Science, University of Cambridge
Cambridge, UK

Pratik Shah
Department of Paediatric Endocrinology, Great Ormond Street Children's Hospital, London, UK

Constantine A. Stratakis
Division of Intramural Research, Eunice Kennedy Shriver National Institute of Child Health and Human Development, Bethesda, MD, USA

D.K. Wherrett
Division of Endocrinology, Department of Paediatrics
Hospital for Sick Children, University of Toronto
Toronto, ON, Canada

Preface

It is nearly 40 years since I wrote the preface to the first edition of this book. There have not been fundamental changes in the clinical practice of pediatric endocrinology during those years but the era of molecular biology has changed completely the understanding of the causation of many of the disorders seen and described in this edition of the book. In years to come it will impinge on clinical practice.

Major changes have occurred in the access to information through the internet and also in the ways that a book is now assembled, prepared and printed in different parts of the world. The ready access to original literature seemed likely to make text books like this redundant but the plethora and complexity of the information available makes even more relevant the authoritative digestion of data and their presentation in a clinically useful format. This has always been the aim of the book.

One loss with the internet is the close personal relationships which used to exist between editors, authors and their publisher; so many people are now involved in the actual production of a book that it is no longer possible to identify exactly who does what. Nevertheless I thank our authors and all at Wiley for their endeavours.

Although I have claimed the right to be the sole author of the preface for this edition of *Clinical Pediatric Endocrinology* (which will be the last in which I shall be involved), it will be clear that the brains behind the book are those of my long-term colleague, now mentor and friend, Mehul T. Dattani. His time in many roles in what was my department and is now his spans 30 years. I and many others respect and admire his achievements and this edition would never have seen the light of day without him. My contribution has been trying to make the book readable, which is not always an easy task.

The number of practitioners of clinical pediatric endocrinology worldwide has increased by at least two orders of magnitude since 1981 and so no longer is this edition dedicated to just the European Society for Paediatric Endocrinology and the Lawson Wilkins Pediatric Endocrine Society but to all who strive to advance our field. Lastly, I should acknowledge with love and gratitude the way my wife Catherine has put up with this cuckoo in our nest for so many years.

Charles G. D. Brook
Hadspen Farm, Somerset, UK

Acknowledgement

I would also like to thank my children Seyan Dattani and Arushi Dattani for their help in checking the proofs.

About the Companion Website

This book is accompanied by a companion website:

www.wiley.com/go/dattani/brookcpe7

The website includes:

MCQs

1

Genetics and Genomics

Anu Bashamboo and Ken McElreavey

Human Developmental Genetics, Institut Pasteur, Paris, France

Introduction

The Human Genome Project was completed in 2003 but it is only now that we are truly in the genomic era. Next-generation sequencing (NGS), which allows genome-wide detection of variants, is transforming on an unprecedented scale our understanding of pediatric and endocrine diseases by identifying mutations that are pathogenic or confer disease risk: new genes that cause human disease are being identified at the rate of 3 per week. We all differ in our DNA sequence and medical geneticists aim to understand the significance of this genetic diversity in health and disease, which has led to the age of genomic medicine.

Understanding genetic diversity is essential to understanding the biology of diseases of various kinds, from simple *Mendelian or monogenic disorders* to more complex *multifactorial disease*, and how we respond to treatment at both population and individual levels. We have the capacity to study the human genome as an entity rather than one gene at a time and medical and clinical genetics has become part of the broader field of genomic or *precision medicine*, which seeks to apply a large-scale analysis of the human genome to provide an individual and knowledge-based approach to medical care.

Many web resources and web-based tools have been developed to help the clinicians navigate and interpret the tremendous amount of genomic data that are being generated (Table 1.1).

The term '-omics' aims at the collective characterization and quantification of pools of biological molecules that translate into the structure, function and dynamics of an organism. Genomics can be divided into *comparative genomics*, the study of the relationship of genome structure and function across different biological species or strains; *functional genomics*, which describes gene and protein functions and interactions; *metagenomics*, the study of genetic material recovered directly from environmental samples; and *epigenomics*, which is the study of the complete set of epigenetic modifications on the genetic material of a cell, known as the epigenome.

Basic Concepts in Human Genetics and Genomics

Genes and Chromosomes

Genetic information is stored in DNA in the chromosomes within the cell nucleus. DNA is a polymeric nucleic acid macromolecule composed of a five-carbon sugar (deoxyribose), a nitrogen-containing base and a phosphate group. The bases are of two types, purines and pyrimidines. In DNA, there are two purine bases, adenine (A) and guanine (G), and two pyrimidine bases, thymine (T) and cytosine (C). DNA is organized in a helical structure in which two polynucleotide chains run in opposite directions, held together by hydrogen bonds between pairs of bases, A of one chain pairing with T of the other and G with C. In the coding sequences of a gene, each set of three bases constitutes a *codon* that encodes for a particular amino acid. *Genome* refers to the totality of genetic information carried by a cell or an organism, whereas *genotype* is the genetic constitution of an individual cell or organism. With the exception of cells that develop into gametes (the germline), all cells that contribute to the body are termed somatic cells.

The human genome contained in the nucleus of the somatic cells consists of 46 chromosomes arranged in 23 pairs, 22 of which are common in both males and females and are termed autosomes, and the remaining pair being the sex chromosomes, two X chromosomes in females and an X and a Y chromosome in males. *Homologous chromosomes* refer to members of a pair of chromosomes which carry the same genes in a similar organization.

Brook's Clinical Pediatric Endocrinology, Seventh Edition. Edited by Mehul T. Dattani and Charles G. D. Brook.

Companion website: www.wiley.com/go/dattani/brookcpe7

Table 1.1 Commonly used databases in human genetic and genomic analysis.

Site	Content	URL
National Center for Biotechnology Information	A portal that provides access to a wealth of biomedical and genomic information. Includes PubMed, OMIM, dbSNP, Clinvar, expression data sets. Suite of tools for data and sequence analysis (e.g. BLAST)	http://www.ncbi.nlm.nih.gov
Mendelian Inheritance in Man (MIM)	A comprehensive database of human genes and genetic disorders	http://www.ncbi.nlm.nih.gov/omim
ClinGen	Authoritative central resource that defines the clinical relevance of genes and variants for use in precision medicine and research	https://www.clinicalgenome.org
Ensembl	Genome browser for vertebrate genomes that supports research in comparative genomics, evolution, sequence variation and transcriptional regulation. Annotates genes, computes multiple alignments, predicts regulatory function and collects disease data	http://www.ensembl.org
University California, Santa Cruz (UCSC), genome browser	Genome browser offering access to genome sequence data from vertebrate and invertebrate species and major model organisms. Integrated with a large collection of analysis tools	https://genome.ucsc.edu
GeneCards	Provides comprehensive information on all human genes. It integrates gene data from ~125 web sources, including genomic, transcriptomic, proteomic, genetic, clinical and functional information	http://www.genecards.org
Human Gene Mutation Database (HGMD)	Collates published gene lesions responsible for human inherited disease	www.hgmd.cf.ac.uk/ac
Mouse Genome Informatics at the Jackson Laboratories	International database resource for the laboratory mouse, providing integrated genetic, genomic, and biological data to facilitate the study of human health and disease	http://www.informatics.jax.org
DECIPHER database	Collects clinical information about rare genomic variants and displays this information on the human genome map	https://decipher.sanger.ac.uk
Database of Genomic Variants (DGV)	A curated catalogue of human genomic structural variation	http://dgv.tcag.ca/dgv/app/home
Exome Aggregation Consortium (ExAC) browser	Exome data set >60,000 unrelated individuals. Provides both a reference set of allele frequencies and constraint metrics giving information on whether a gene is tolerant or intolerant to variation	http://exac.broadinstitute.org
ClinVar	Aggregates information about genomic variation and its relationship to human health.	http://www.ncbi.nlm.nih.gov/clinvar
Sequence Variant Nomenclature	Provides guidelines for sequence variation nomenclature	http://varnomen.hgvs.org
dbSNP	Genetic variation within and across different species. Not limited to SNPs, it contains a range of molecular variation	http://www.ncbi.nlm.nih.gov/SNP
F-SNP	Provides integrated information about the functional effects of SNPs obtained from 16 bioinformatics tools and databases. Helps identify and focus on SNPs with potential pathological effect to human health	http://compbio.cs.queensu.ca/F-SNP
Biological General Repository for Interaction Datasets (BioGRID)	Database of protein–protein interactions, genetic interactions, chemical interactions, and post-translational modifications	http://thebiogrid.org

Table 1.1 (Continued)

Site	Content	URL
PhenomicDB	A multi-organism phenotype–genotype database including human, mouse, fruit fly, *C. elegans*, and other model organisms	http://www.phenomicdb.de
Phencode	Connects human phenotype and clinical data in various locus-specific mutation databases with data on genome sequences, evolutionary history and function in the UCSC Genome Browser	http://phencode.bx.psu.edu
Human Epigenome Atlas	Includes human reference epigenomes and the results of their integrative and comparative analyses. Provides details of locus-specific epigenomic states like histone marks and DNA methylation across tissues and cell types, developmental stages, physiological conditions, genotypes and disease states	http://www.genboree.org/epigenomeatlas
Encyclopedia of DNA Elements (ENCODE)	Catalogue of functional elements in the human genome, including elements that act at the protein and RNA levels and regulatory elements that control cells and circumstances in which a gene is active	https://www.encodeproject.org
Genomics England 100,000 Genomes Project	The project will sequence 100,000 genomes from around 70,000 people. Participants are National Health Service (UK) patients with a rare disease, plus their families, and patients with cancer	www.genomicsengland.co.uk/the-100000-genomes-project

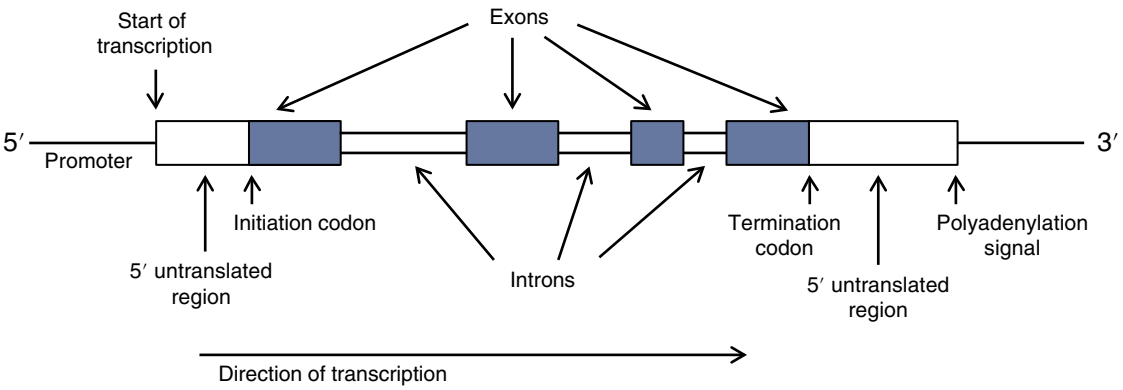

Figure 1.1 Example of a typical mammalian gene structure. A typical gene has regulatory regions preceding the coding exons interspersed by non-coding introns. The individual labelled features are discussed in detail in the text.

A *gene* is a sequence of DNA in the genome required for the expression of a functional product, including a polypeptide or RNA molecule (Figure 1.1). The majority of human genes are organized as coding regions called *exons* interrupted by one or more non-coding regions termed *introns*. Introns are initially transcribed into RNA in the nucleus but are not present in mature mRNA, which also has flanking 5′ and 3′ *untranslated regions* (UTRs). The latter contains a signal for the addition of adenosine residues (the polyA tail) to the end of the mature mRNA.

Other sequences within the 3′ UTR are important for translation efficiency, localization and stability, whereas the 5′ UTR is important in the regulation of RNA translation. It is important when discussing genes to define what is meant by the terms *trans* and *cis*. *trans*-Acting usually means 'acting from a different molecule', whereas *cis*-acting means 'acting from the same molecule'. In genetics and genomics, *cis*-acting elements refer to DNA sequences in the vicinity of a gene that are required for gene expression; trans-acting factors, either proteins or some classes of RNA molecules, bind to the *cis*-acting sequences to control gene expression.

Many genes produce not just one but multiple proteins, which is achieved either by *alternative splicing* of the coding segments of genes or by numerous types of

biochemical modifications of the resulting proteins so that the 19,000 genes in the human genome are estimated to generate over a million different proteins. The other point to remember is that individual proteins rarely work by themselves. The cell is composed of modular supramolecular complexes and each complex performs an independent, discrete biological function that could not be achieved by the independent components of the complex. The transfer of information from the DNA strand to the protein is mediated by RNA, which directs the synthesis and sequence of polypeptides.

Genetic information is stored in genes in the form of a genetic code in which the sequence of adjacent bases determines the sequence of amino acids in the polypeptide. RNA is synthesized from DNA by *transcription* and the RNA carrying the coded information is termed *messenger RNA* (mRNA), which is transported from the nucleus to the cytoplasm where it is translated to synthesize the protein. This constitutes the central dogma of molecular biology.

Regulation of Gene Expression

Gene expression is the production of correct RNA, which is a complex process where the RNA must be expressed in the appropriate cell type in the correct amount and, in some cases, at a precise developmental time. Nucleic acid sequences flanking the coding sequences and in some cases within the coding sequences provide the molecular signals for gene transcription. A *promoter* region that contains sequences necessary for the initiation of transcription lies at the 5′ end of most genes. An *enhancer* is a short (50–1500 bp) region of DNA that can be bound by proteins (transcription factors) to increase the likelihood that transcription of a particular gene will occur. Enhancers are generally located up to 1 Mbp away from the gene and can be upstream or downstream of the gene it regulates. The orientation of an enhancer may even be inverted without having an effect on its function.

Genes that are necessary for complex and multiple developmental processes usually have a number of enhancers with overlapping functions. A good example is the *SOX9* locus: the developmental timing and tissue-specific transcriptional regulation of *SOX9* are highly complex and involve multiple elements located in flanking regions of at least 1 Mb upstream and 1.6 Mb downstream and these show strikingly different phenotypes when mutated. Upstream rearrangements are associated with campomelic dysplasia and fall within two clusters located about 400 kb apart. Large (>1 Mb) duplications 5′ to *SOX9* (i.e. downstream) are associated with brachydactyly anonychia (symmetric brachydactyly of the hands/feet, hyponychia or anonychia). Pierre Robin sequence (micrognathia, cleft palate and macroglossia) is caused by either a deletion located 1.38 Mb upstream or a deletion located 1.56 Mb downstream of *SOX9*.

Another regulatory element, termed *RevSex*, is located 600 kb upstream of *SOX9* gene. Three copies of *RevSex* are associated with testicular or ovotesticular disorders of sex development (DSD), whereas deletions (one copy) of *RevSex* are associated with 46,XY gonadal dysgenesis.

RNA Editing

RNA editing is a molecular process through which cells can make discrete changes to specific nucleotide sequences within a RNA molecule. RNA editing includes nucleotide additions and insertions as well as nucleobase modifications such as cytidine (C) to uridine (U) and adenosine (A) to inosine (I), which is termed deamination. RNA editing in mRNAs alters the amino acid sequence of the encoded protein so that it differs from that predicted by the genomic DNA sequence.

Various post-translational modifications after protein biosynthesis can extend the chemical repertoire of the 20 amino acids by introducing new functional groups such as phosphate, acetate, amide groups or methyl groups. This can occur on the amino acid side chains or at the protein's carboxy (C-) or amino (N-) termini. The most common post-translational modifications in descending order are phosphorylation, acetylation, N-linked glycosylation, palmitoylation and O-linked glycosylation.

Classes of RNA Molecules and their Functions

In recent years there has been a substantial shift in understanding the importance of different classes of RNA molecules in biological processes. Ninety percent of the human genome is transcribed, but many of the resulting transcripts and the factors regulating their transcription remain uncharacterized. The vast majority of the transcribed genome comprises diverse classes of *non-coding RNAs* (ncRNAs) that may play key roles in different biochemical and cellular processes with profound implications for human health and disease. The cellular repertoire of ncRNAs consists of small housekeeping RNAs, such as *ribosomal RNAs* (rRNAs) and *transfer RNAs*, *microRNAs*, and *long ncRNAs* (lncRNAs). Assigning molecular, cellular, and physiological functions to well-annotated ncRNAs is the current challenge in this field. ncRNAs are emerging as key players in pediatric and endocrine diseases.

MicroRNAs, Small Interfering RNAs, and Piwi-interacting RNAs

Small RNAs usually refers to any class of non-coding RNA of 19–32 nt (Table 1.2). The role of small RNAs is continuing to be explored but their main function is *gene silencing* through RNA-mediated mechanisms. *RNA silencing* is an umbrella term for all small RNA-mediated inhibition of transcription, translation and deactivation of transposable elements. RNA silencing is widely regarded as a master controller of gene regulation but small RNAs have important roles in an increasing variety of eukaryotic biological processes. For example, small RNAs may play a role in transgenerational inheritance and epigenetic memory.

In eukaryotes, there are three major classes of small RNAs involved in post-transcriptional regulation: microRNAs (miRNAs), small interfering RNAs (siRNAs) and Piwi-interacting RNAs (piRNA). All find target RNAs by base pairing between complementary sequences, causing target RNA degradation and/or translational repression. Each type has its own preferred class of RNA targets, reflecting their biological functions. miRNAs and siRNAs bind to proteins in the Argonaute subfamily, whereas piRNAs bind to the Piwi subfamily of proteins.

MicroRNAs (miRNAs) are an abundant class of small evolutionarily conserved regulatory RNAs about 19–22 nucleotides (nt) in length. They are thought to play fundamental roles in most biological processes including disease. Over 1500 miRNAs have been identified that regulate the expression of up to 60% of mammalian genes. The canonical miRNA pathway starts with the transcription of miRNA genes by RNA polymerase II, which results in the production of the primary miRNA (pri-miRNA). The pri-miRNA transcript is cleaved by a protein complex consisting of Drosha/DGCR8 to generate an ~80 base-pair precursor miRNA (pre-miRNA) with a characteristic hairpin secondary structure vital for enabling export from the nucleus.

Dicer, an RNase III/helicase multi-domain enzyme, processes the pre-miRNA into a ~22 bp miRNA. The gene encoding DICER is termed *DICER1*. *DICER1* syndrome is an inherited disorder that increases the risk of a variety of cancerous and benign tumours, including pleuropulmonary blastoma, cystic nephroma, multinodular goiter and Sertoli–Leydig cell tumours of the ovary, which typically develop in affected women in their teens or twenties. Some Sertoli–Leydig cell tumours release testosterone resulting in virilization.

Incorporated into one member of the Argonaute (Ago) protein family is the RNA-induced silencing complex (RISC), a mature miRNA that binds typically to the 3′-UTR of the mRNA and inhibits its translation via various mechanisms including mRNA degradation. The key determinant of target recognition is a short sequence complementarity between the miRNA seed sequence (the second–eighth nucleotides of the miRNA) and the target mRNA. The maturation and function of miRNAs are highly dependent on the coordinated action of several RNA-binding proteins.

The miRNA genes are mainly clustered in the genome and are transcribed as *polycistronic* primary transcripts. Forty percent of miRNA genes lie in the introns of protein and non-protein coding genes. These are usually,

Table 1.2 Characteristics of the three major classes of small RNAs involved in post-transcriptional regulation.

Properties	miRNA	piRNA	siRNA
Size (nt)	20–24 (usually 22)	26–31	20–25
Origin	Endogenous and ubiquitous	Endogenous to germ cell lineages	Exogenous or endogenous
Evolutionary conservation	Eukaryotes	Vertebrates and invertebrates	Eukaryotes
Precursor	Single-stranded RNA	Single-stranded RNA	Double-stranded RNA
Biogenesis	Dicer dependent	Dicer independent	Dicer dependent
Base-pair match to target	Imperfect	Perfect	Perfect
Distribution	Cytoplasmic and nuclear	Cytoplasmic and nuclear	Cytoplasmic
Ago dependence	Ago subfamily	Piwi subfamily	Ago subfamily
Target nucleic acid	3′UTR, 5′-UTR, promoters, coding regions, pseudogenes	Transposons	mRNA, promoters
Main functions	Translation inhibition, mRNA degradation, transcriptional and post-transcriptional silencing	Transposon silencing, transcriptional and post-transcriptional silencing	mRNA degradation, transcriptional and post-transcriptional silencing

though not exclusively, found in a sense orientation and thus are regulated together with their host genes. Pre-miRNAs that are spliced directly out of introns are termed *Mirtrons*. Approximately 16% of pre-miRNAs are modified by nuclear RNA editing, leading to changes in biological function. This is mainly mediated by adenosine deaminases acting on RNA to catalyse adenosine to inosine (A to I) transitions. RNA editing can result in disruption of nuclear processing as well as alter downstream processes including cytoplasmic miRNA processing and target specificity.

Gene regulation by miRNAs is of key importance in many fundamental biological processes such as cellular differentiation, proliferation, migration and apoptosis. In disease, some circulating blood miRNA levels are proportional to the degree of severity of the pathology, such as drug-induced liver injury, cardiovascular infection, cancer, Alzheimer's disease, inflammation and metabolic diseases (obesity). Altered expression of miRNAs in diabetes causes malfunction of insulin release and insulin resistance. The use of miRNAs as biomarkers for type 1 diabetes (T1D) risk is attractive as these markers could be used to identify individuals at risk for developing T1D before symptoms appear. Twelve miRNAs were found to be more concentrated in sera from children and adolescents with newly diagnosed T1D compared with sera from age-matched controls. Among them, miR-25 was associated with improved glycaemic control and better residual β-cell function, suggesting that this miRNA could be used during early and intensive management of newly diagnosed diabetes to improve blood glucose control and reduce microvascular complications.

piRNA is the largest class of small non-coding RNA molecules expressed in animal cells. piRNAs are generated from various portions of long single-stranded precursor RNAs transcribed from genomic loci termed piRNA clusters, which are often >100 kb in size. They are distinguished from miRNA by their size (26–31 nt), lack of sequence conservation and increased sequence complexity. The majority are antisense to transposon sequences, indicating that transposons are the piRNA target. piRNAs direct the Piwi proteins to their transposon targets for gene silencing. The piRNA-mediated repression of transposons is best characterized in the germline. piRNAs are necessary for spermatogenesis in humans.

siRNAs are derived from long double-stranded precursor RNAs (dsRNAs). Endogenously formed dsRNAs are exported to the cytoplasm where they are cleaved into 20–25-nt duplexes by Dicer. One strand of these fragments, usually the antisense strand, is incorporated into multiprotein RISCs composed of one of a family of Argonaute proteins together with auxiliary proteins that extend or modify the function. In contrast to miRNAs, siRNAs have a sequence fully complementary to their target mRNA and usually have a single target mRNA. Depending on the source of dsRNA precursor, siRNAs can be further divided into exogenous and endogenous siRNAs (exo- and endo-siRNAs, respectively).

LncRNAs

LncRNAs, non-protein-coding RNA transcripts longer than 200 nucleotides (nt), are emerging as key regulators of diverse cellular processes. The definition of lncRNAs continues to evolve. The first reported example of a long non-coding RNA (lncRNA) was the *H19* transcript, which lacked large open reading frames and was not translated into protein. Later work revealed the existence of thousands of lncRNAs in the human genome. The expression of lncRNAs is usually low but they are transcribed in a highly regulated manner, either from their own promotor sequence or as a by-product of other transcriptional processes. Although some lncRNAs are located within intergenic sequences, the majority are transcribed as complex, interlaced networks of overlapping sense and antisense transcripts that often include protein-coding genes. They are generally, but not exclusively, spliced, 5′-capped and 3′-polyadenylated, and transcribed by RNA polymerase II.

Approximately one-third to one-half of lncRNAs overlap protein-coding genes. Genic lncRNAs can be further divided into those that overlap protein-coding loci in the sense or antisense direction and those that overlap exonic or intronic regions of the protein-coding gene. A universal classification does not exist. The 200-nt cut-off to define their size is arbitrary and does not represent a biological distinction. A lncRNA may code for a polypeptide but it must have coding-independent functions, as shown for the steroid receptor RNA activator (SRA), a well-characterized bifunctional lncRNA involved in the nuclear receptor-mediated regulation of gene expression. The SRA1 gene expresses both SRA RNA and the SRA protein (SRAP). This gene is involved in the regulation of many NR and non-NR activities, including metabolism, adipogenesis and chromatin organization. The encoded protein, SRAP, acts as a transcriptional repressor by binding to the non-coding RNA. SRA coactivates a range of nuclear receptors including ERα and ERβ in a ligand-dependent manner by direct interaction with other co-regulatory proteins. lncRNAs are becoming increasingly important in oncology. The dysregulation of lncRNAs expression is highly specific to the cancer type as compared to the protein-coding genes. lncRNAs are being identified as drivers of cancer with their potential functions being predicted. This is providing a framework for the development of new cancer diagnostics, stratification and precision treatments.

Other Small ncRNA Classes

Small nucleolar RNAs (snoRNAs) are a class of small RNA molecules that primarily guide chemical modifications of other RNAs, mainly ribosomal RNAs, transfer RNAs and small nuclear RNAs. There are two main classes, the C/D box snoRNAs, which are associated with methylation, and the H/ACA box snoRNAs, which are associated with pseudouridylation.

Y non-coding RNAs were initially found in the cytoplasm of mammalian cells. There are four non-coding Y RNAs in humans: hY1, hY3, hY4 and hY5 RNA. Y RNA fragments are not involved in the miRNA pathway but they are essential factors for the initiation step of chromosomal DNA replication in human cell nuclei. The mechanism is poorly understood but is thought to be mediated by interactions with chromatin and transcription initiation proteins.

Small nuclear ribonucleic acid (snRNA) are found within the splicing speckles and Cajal bodies of the cell nucleus in eukaryotic cells. The length of an average snRNA is ~150 nucleotides. Their primary function is the processing of pre-messenger RNA in the nucleus. They also aid the regulation of transcription factors (7SK RNA) or RNA polymerase II (B2 RNA) and telomere maintenance.

Circular RNAs (CircRNAs) are a family of naturally occurring endogenous ncRNAs with widespread distribution and diverse functions. These ~100 nucleotides long, single-stranded RNA molecules form a circle through covalent binding. CircRNAs mainly arise from the exons of protein-coding genes but they can also be derived from introns, untranslated regions, intergenic loci and antisense sequences of known transcripts. CircRNAs are common in the eukaryotic transcriptome and abundant in exosomes. CircRNAs show a high sequence conservation with specific expression in various tissues during different developmental stages. Some circRNAs can interact with miRNAs and can function as miRNA sponges in mammalian cells. CircRNAs are becoming increasingly important in medicine by serving as biomarkers for non-invasive diagnosis of atherosclerosis, neurodegenerative diseases and cancers.

Gene Mutations and Inheritance

Any permanent heritable change in the sequence of genomic DNA is termed a *mutation*.

Classes of Gene Mutations

Fifty percent of all disease-causing mutations are *missense,* which are caused by a single nucleotide substitution (*point mutation*) in the DNA coding sequence of a gene that results in the replacement of one amino acid by another in the final protein product. Nucleotide changes that involve the substitution of one purine for the other (A for G or G for A) or one pyrimidine for the other (C for T or T for C) are termed *transitions.* The replacement of a purine for a pyrimidine (or vice versa) is a *transversion*. Missense mutations are often referred to as *nonsynonymous* mutations, whereas point mutations that do not alter the amino acid in the final protein product are referred to as *synonymous* mutations. Although the latter were largely ignored since they were considered to have no functional consequences, there is a growing realization that they can be associated with disease by affecting the stability of the mRNA, mRNA folding, translation fidelity and miRNA–mRNA interaction or by creating novel RNA splice sites. More than 50 human diseases are associated with synonymous mutations including Crohn's disease, Treacher Collins syndrome and Crouzon syndrome.

A point mutation in a DNA coding sequence that results in the replacement of an amino acid codon to one of the three termination codons is termed a *nonsense mutation*. Depending on its position, the resulting transcript is predicted to be recognized by the nonsense-mediated decay surveillance complexes and degraded. If this does not occur, the resulting truncated protein is usually unstable and degraded. Ten percent of all disease-causing mutations are nonsense mutations, which may affect the processing of RNA. For introns to be excised from unprocessed RNA and the exons spliced together to form a mature mRNA requires a specific nucleotide sequence located at the exon–intron (5′ donor site) or the intron–exon (3′ acceptor site) junctions. Mutations that affect these required bases at either the splice donor or acceptor site interfere with (and in some cases abolish) normal RNA splicing at that site. A second class of splicing mutations involves intron base substitutions that do not affect the donor or acceptor site sequences themselves. This class of mutations creates alternative donor or acceptor sites that compete with the normal sites during RNA processing. Thus, at least a proportion of the mature mRNA in such cases may contain improperly spliced intron sequences.

Mutations can also be caused by the insertion, inversion, translocation or deletion of DNA sequences. This may involve only a single base pair or up to several million base pairs. *Frameshift mutations* occur when small deletions or insertions occur within coding sequences and involve a number of bases that are not a multiple of 3. These generate a different sequence of codons from the point of the insertion or deletion and usually generate a downstream termination codon.

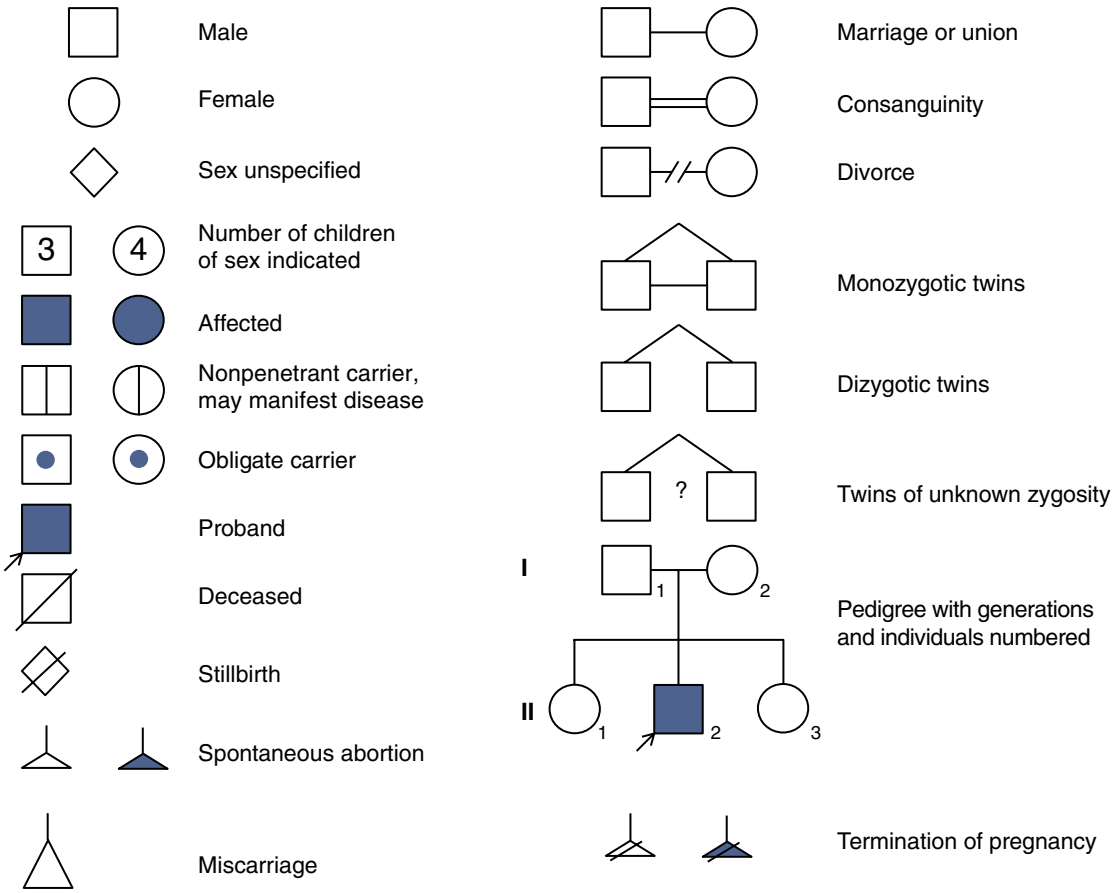

Figure 1.2 Symbols commonly used for creating pedigree charts.

Another form of mutation is *gene conversion*, a process by which one DNA sequence replaces a *homologous* sequence so that the sequences become identical after the conversion event. Gene conversion can be allelic, meaning that one allele of the same gene replaces another allele, or ectopic, meaning that one *paralogous* DNA sequence is converted to another. Gene congential plays an important role in congenital adrenal hypoplasia due to mutations involving the *CYP21A2* gene, which is located on chromosome 6p21.3, 30 kb from the *CYP21A1P* pseudogene. Both *CYP21A2* and *CYP21A1P* show sequence identity of 98% between exons and 96% between introns. The high sequence identity and close proximity of *CYP21A2* and *CYP21A1P* can generate gene conversion that results in the transfer of deleterious mutations from the pseudogene to *CYP21A2*. It is estimated that 25% of all disease-causing mutations are due to deletions or insertions.

Patient and Family History

A detailed patient and family history and thorough clinical and biochemical investigation are essential to understand if the disorder follows one of five basic modes of inheritance for single-gene diseases or if the inheritance pattern follows a more complex inheritance with incomplete penetrance or variable expressivity. The family history should include the drawing of the pedigree where individuals are represented by symbols (usually circles for female and squares for male), solid to indicate someone affected by a trait and unfilled for unaffected. Figure 1.2 illustrates some of the conventions followed in constructing pedigrees. A detailed pedigree analysis can reveal the inheritance patterns in a family. Pedigree analysis is also useful for analysing populations with limited progeny data from multiple generations.

Mendelian Inheritance Patterns

The basic laws of inheritance are important to understand patterns of disease transmission. Single-gene or monogenic diseases are usually inherited in one of several patterns depending on the location of the gene in the genome or whether one or two copies of the gene are needed for normal biological activity. The five modes of inheritance for single-gene diseases are *autosomal*

Table 1.3 Types of inheritance and their associated family history.

Type of inheritance	Family history pattern
Autosomal dominant	Individuals carrying one mutated copy of the gene will be affected by the disease. Each affected person usually has one affected parent, although *de novo* mutations occur. Usually occurs in every generation of the family
Autosomal recessive	Individuals carrying two mutated copies of the gene will be affected. Parents are usually unaffected and each must carry a copy of the mutated gene (carriers). Usually not seen in every generation
Mitochondrial	Maternally inherited. Both males and females can be affected. Can appear in every generation of a family
X-linked dominant	Females are more frequently affected than males. Fathers cannot pass on X-linked traits to their sons
X-linked recessive	Males are more frequently affected than females. Affected males are often seen in each generation. Both parents of an affected daughter must be carriers. Only the mother must be a carrier of an affected son

dominant, *autosomal recessive*, *X-linked dominant*, *X-linked recessive* and *mitochondrial inheritance* (Table 1.3). Single-gene disorders affect 2% of the population sometime during a lifetime.

Single-gene disorders are dominant or recessive. A dominant phenotype is expressed in both homozygotes and heterozygotes for a mutant allele, whereas a recessive phenotype is expressed only in homozygotes for the mutant allele. Most dominant disorders are rare and usually seen in the heterozygous state. Thus approximately one-half of offsprings will inherit a dominant trait. Homozygotes for dominant traits have usually a more severe phenotype or fail to survive. Many autosomal dominant mutations result in *haploinsufficiency*, which occurs when the single functional copy of the gene does not produce enough gene product to produce a wild-type phenotype.

A *dominant negative mutation* dominantly affects the phenotype by means of a defective protein or RNA molecule that interferes with the function of the normal gene product in the same cell. Most recessive disorders are due to mutations that result in the reduction or elimination of the function of the gene product and are termed *loss-of-function* mutations. Examples include 5-alpha reductase deficiency, due to autosomal recessive mutations in the *SRD5A2* gene, and congenital adrenal hyperplasia, due to homozygous mutations in the *CYP21A2*, *CYP11B1* or *CYP11A1* genes.

The majority of loci on the X chromosome show *X-linked inheritance* because they participate in meiotic recombination only during female gametogenesis, when there are two X chromosomes; they cannot recombine with the Y during male gametogenesis. Males have a single X and are therefore *hemizygous* with respect to X-linked genes. 46,XY males are never heterozygous for alleles at X-linked loci, whereas females can be heterozygous or homozygous at X-linked loci. To compensate for the double complement of X-linked genes in females, alleles for most X-linked genes are expressed from only one of the two X chromosomes in any given cell.

A *sex-limited trait* is a phenotype expressed in only one sex, although the gene that determines the trait is carried by both sexes and therefore autosomal. This is an on or off phenomenon. Sex-limited genes cause the two sexes to show different phenotypes, despite having the same genotype. They are responsible for sexual dimorphism, a phenotypic (directly observable) difference between males and females of the same species (e.g. lactation). This is not to be confused with *sex-linked traits*, which is the phenotypic expression of an allele present on the sex chromosome of the individual. A classic example is male-limited precocious puberty, an autosomal dominant disorder in which affected boys develop secondary sexual characteristics and undergo an adolescent growth spurt at about 4 years of age. In some families, the phenotype is due to mutations in the gene that encodes the receptor for luteinizing hormone (*LCGHR*). A *sex-influenced trait* refers when the expressivity of the phenotype is influenced by the sex, for example, body and facial hair.

For many phenotypes the mode of inheritance may depend on the gene involved. Non-syndromic disorders of testis determination, 46,XY complete or partial gonadal dysgenesis, can be inherited in a number of different ways that include sex-limited autosomal recessive (e.g. *DHH*), sex-limited autosomal dominant with variable expressivity and incomplete penetrance (e.g. *NR5A1*), Y-linked (*SRY*) or X-linked (hemizygous duplication of *NR0B1*).

Non-Mendelian Inheritance Patterns

Mitochondrial DNA (mtDNA) molecules, are present in tens to thousands of copies per cell. If a cell contains mitochondria that contain only a pure population of mutant mtDNA, it is termed *homoplasmy*. Alternatively, if the cell has mitochondria, some with and some without mutation, it is termed *heteroplasmy*. Disorders involving mtDNA mutations are characterized by maternal inheritance. Sperm mitochondria are eliminated from the embryo, so that mtDNA is always inherited from the mother. Thus, all the children of a female who is

homoplasmic for a mtDNA mutation will inherit the mutation, whereas none of the offspring of a male carrying the same mutation will do so. mtDNA mutations cause human diseases that often involve the central nervous and musculoskeletal systems. The proportions of normal and mutant mtDNA in the cells making up different tissues often result in incomplete penetrance of the phenotype, variable expression and *pleiotropy*.

Amplification of a repeat sequence is observed in disorders such as Huntington disease and fragile X syndrome. In the former, a simple trinucleotide repeat is located in the coding region and in a transcribed but untranslated region of the *FRM1* gene in the latter. *Trinucleotide repeat* disorders often show a *parent-of-origin* effect. Large expansions of the CAG repeat that cause juvenile Huntington disease are generally of paternal origin, whereas large expansions of the CGG repeat in fragile X syndrome are often of maternal origin.

Mosaicism is the presence of at least two cell lines that differ genetically but are derived from a single zygote in an individual or a tissue. Mosaicism can be categorized as somatic and/or germline. Mosaicism for numerical or structural abnormalities of chromosomes is clinically important and somatic mutations are recognized as a major contributor to many types of cancer. Somatic mosaicism refers to population of cells that carry a mutation in some tissues of the body but not in the gametes, whereas the cells carrying the mutation may be restricted to the gamete lineage in germline mosaicism. In some individuals both somatic lineages and the germline may be affected. 45,X/46,XY is a disorder of sex development associated with sex chromosome aneuploidy and mosaicism of the Y chromosome associated with highly variable clinical phenotypes, ranging from partial virilisation and ambiguous genitalia at birth to individuals with completely male or female gonads.

Common Disorders with Complex Inheritance Patterns

Many common disorders such as myocardial infarction, Alzheimer disease and diabetes do not follow Mendelian inheritance patterns seen in single gene disorders. They result from complex interactions between a number of genetic and environmental factors and follow a multifactorial or complex inheritance pattern.

Complex phenotypes can be divided into *qualitative* and *quantitative traits*. A qualitative trait is the presence or absence of the disorder, whereas a quantitative trait is a measurable physiological or biochemical aspect of the disorder such as the body mass index in obesity.

Familial aggregation of a common phenotype does not always mean that the cause must be genetic. Family members may develop a disorder by chance since, as well as genes, family members often share a common environment, diet, socio-economic status and culture. The familial aggregation of a disorder can be measured by comparing the frequency of the disorder in the relatives of an affected proband with its frequency in the general population. The more common a disorder in the general population, the more likely it is that the familial aggregation may be a coincidence.

Another approach to determine familial aggregation is case–control studies. Patients with the disorder are compared with carefully chosen control individuals who do not have the disorder. Often this is a spouse since they usually match the case in terms of age, ancestry (previously referred to as ethnicity) and environment. These types of studies are subject to errors that include ascertainment biases and failure to correctly match case–control subjects. Control individuals should differ from cases only in their disease status. All other factors should be matched. If they are not matched, a case–control study may find significant associations that are due to differences in, for example, ancestry rather than any relationship to the presence or absence of the disorder.

The genetic contribution to a complex disorder can be dissected by measuring allele sharing between affected and unaffected relatives. The concept is simple: when a genetic contribution is important in a disorder, the frequency of disease concordance increases as the degree of relatedness increases. Monozygotic twins (MZ) are the most extreme example as they have all their alleles in common. A first-degree relative shares ½ of alleles, a second-degree relative ¼, and a third-degree relative ⅛ and so on. One of the most common ways of separating the genetic contribution from the environment is to study MZ and dizygotic twins (DZ). DZ reared together allow the measurement of disease concordance in a similar environment, whereas MZ twins provide the opportunity to study genotypically identical individuals reared in similar or different environments. Greater disease concordance in MZ versus DZ twins is strong evidence for a genetic contribution to the disease.

Heritability was developed to quantify the role of genetic differences in determining variability of quantitative traits. It is the fraction of the total phenotypic variance of a quantitative trait caused by genes. Due to genetic differences between individuals, the higher the heritability, the greater the variability of the phenotype. Heritability measures the fraction of phenotype variability that can be attributed to genetic variation but it does not indicate the degree of genetic influence on the development of a trait of an individual.

Good examples of complex or multifactorial disorders include hypospadias and T1D. Hypospadias is one of the most common congenital disorders in males occurring in 1:200–1:300 male births. Anterior (glandular) and

middle (penile) forms comprise 70–80% and 15–20% of cases, respectively. Hypospadias shows familial clustering, with 7% of cases having affected first-, second- or third-degree relatives. The chance that a brother of an affected boy will also have hypospadias is 9–17%. Studies of family as well as twins of known zygosity have estimated the heritability of hypospadias to be 57–77%, meaning that 57–77% of the phenotypic variability can be attributed to genetic variability. Like many other common disorders, such as infertility, current data indicate that hypospadias might be monogenic in a small proportion of the families (e.g. the mutations in genes *NR5A1, AR, FGFR2* and *MAMLD1*) but that there is a multifactorial cause in the majority of cases.

T1D is a result of interplay between genetic predisposition, environmental factors and reprograming of the immune system. The destruction of pancreatic β-cells affects the level of insulin secretion leading to disease development. Destruction of pancreatic β-cells is mediated by an altered immune response due to genetic anomalies resulting in increase of pro-inflammatory cytokines and autoreactive T and B lymphocytes. Twin studies have estimated that 88% of phenotypic variance of T1D in Finland is due to genetic factors and the remaining due to unshared environmental factors. Genome-wide association studies have identified more than 50 variants associated with increased risk for T1D. The HLA class II region has the strongest impact on T1D risk. However, more than 40 non-HLA loci that impact upon the risk of developing T1D have been identified. Many of these genes are associated with immune function including *interleukin (IL)-2Ra*, *PTN22*, *IL-10*, *CCR5* and *IL-2*.

Uniparental Disomy

Uniparental disomy (UPD) occurs when a person receives two copies of a chromosome or of part of a chromosome from one parent and no copy from the other. UPD usually arises from the failure of the two members of a chromosome pair to separate properly into two daughter cells during meiosis in the parent's germline (nondisjunction). The resulting gametes contain either two copies of a chromosome (disomic) or no copy of that chromosome (nullisomic). This leads to a conception with either three copies of one chromosome (trisomy) or a single copy of a chromosome (monosomy).

If a second event occurs by either the loss of one of the extra chromosomes in a trisomy or the duplication of the single chromosome in a monosomy, the karyotypically normal cell may have a growth advantage compared to the aneuploid cells.

A postfertilization error can also lead to UPD, by either somatic recombination or gene conversion. Two types of UPD can be defined – uniparental heterodisomy (UPhD), where the two different alleles of the same parent are transmitted, and uniparental isodisomy (UPiD), where two identical copies of one allele of the contributing parent are present.

UPD may have clinical relevance for several reasons. For example, either *isodisomy* or *heterodisomy* can disrupt parent-specific genomic imprinting, resulting in imprinting disorders. Additionally, isodisomy leads to large blocks of homozygosity, which may lead to the uncovering of recessive genes. Uniparental inheritance of imprinted genes can result in phenotypic anomalies. Examples include Prader–Willi, Angelman and Silver–Russell syndromes. Prader–Willi syndrome, characterized by hypothalamic–pituitary abnormalities, is caused by deletion or inactivation of genes on the paternally inherited chromosome 15, while the maternal copy, which may be of normal sequence, is imprinted and therefore silenced. Angelman syndrome is a neurodevelopmental disorder caused by the loss of maternally inherited genes on chromosome 15 and paternal imprinting. Silver–Russell syndrome is a clinically heterogeneous disorder characterized by severe *in utero* growth restriction and poor postnatal growth, body asymmetry, irregular craniofacial features and several additional minor malformations. The etiology is complex and current evidence strongly implicates imprinted genes. Approximately half of all patients exhibit DNA hypomethylation at the *H19/IGF2* imprinted domain; around 10% have maternal UPD of chromosome 7.

Penetrance and Expressivity

Some disorders are not expressed in an individual, even though the individual carries the mutation causing the phenotype in other members of the family. This is termed *penetrance*, defined as the probability that a gene mutation will have a phenotypic expression. Penetrance is an all or nothing concept. If only a proportion of people carrying the genotype display the phenotype, the trait is said to show incomplete penetrance. If all carriers show the phenotype, then the trait is said to have complete or full penetrance. For example, familial cases of central precocious puberty show reduced or incomplete penetrance.

Expressivity refers to the severity of the phenotype in different individuals carrying the same disease-causing genotype. If the severity of the phenotype differs in people with the same disease-causing genotype, the phenotype shows *variable expressivity*. Disease expressivity includes age of onset, rate of progression, severity and the manifestation of other comorbidities. There are numerous examples of studies of non-identical twins who share the same environment and carry the same disease-causing genotype yet display distinct phenotypes.

This suggests that genetic factors are acting as modifiers of the phenotype. The effect of one gene or allele on the phenotypic outcome of a second gene or locus is termed epistasis, genetic interaction, digenic inheritance, oligogenic inheritance or genetic modifier. Although these are essentially synonyms, there are important distinctions that can be drawn.

If a mutation in the primary gene is both necessary and sufficient to cause disease, the presence of an allele at a second gene has a purely modifying role on the severity of the phenotype. Digenic or oligogenic inheritance is termed when alleles at two (di-) or more (oligo-) genes are required to manifest the pathology. In practice this distinction is often blurred. It is important to note that variability in the expression of the phenotype could also be due to environmental factors, perhaps interacting with genetic variants. Examples of genetic modifiers include variation in an interacting protein partner or, if the protein is a DNA-binding transcription factor, variation in target binding sequences. An example of variable expressivity is the phenotype associated with mutations in the *NR5A1* gene. Here, mutations involving the same amino acid change are associated with 46,XY complete gonadal dysgenesis in some individuals and infertility in others. Digenic and, more rarely, oligogenic inheritance has been reported in individuals with central hypogonadotropic hypogonadism (CHH). Where the causal mutation has been identified, over 80% of patients with CHH have a monogenic cause but ~12% have digenic and 2.5% oligogenic inheritance.

Human Populations and Genetic Variation

Overview of Human Genetic Variation

Mendelian phenotypes result from mutations that alter the function, localization and/or the presence of a protein. Even though protein-coding sequences comprise only around 2% of the human genome, linkage analyses on pedigrees with various disorders have shown that the vast majority of disease-causing mutations are variants that directly impact protein expression or function. This excludes ascertainment bias. Overall, clinically recognized Mendelian phenotypes occur in ~0.4% of all live births and 8% of live births have a genetic disorder recognizable by early adulthood. The Human Genome Project and subsequent *annotation* efforts have established that there are around 19,000 predicted protein-coding genes in humans. The consequences of germline mutations (single nucleotide variants [SNVs] and copy number variants [CNVs]) are known for more than 2300 of these genes. Around 3300 genes have been implicated in Mendelian disorders and this figure is growing at around 300 per year.

Humans are 99.9% identical with respect to their DNA sequence. A typical human genome from an apparently healthy individual differs from the reference genome at 4.1–5 million sites (>99.9% SNVs or indels) and carries 300–600 non-synonymous mutations that are found in <1% of the general population (minor allele frequency, MAF < 0.01). This includes around 150 mutations that are not (yet) present on any of the public variation databases and that are a combination of *de novo* or family or community-specific DNA variants. All of us inherit about 100 likely loss-of-function or nonsense variants from our parents and around 25–30 variants per genome that have been reported to be associated with rare diseases (ClinVar: http://www.ncbi.nlm.nih.gov/clinvar).

One of the surprises from the large amount of genomic data generated from healthy control populations in recent years, such as the 1000 genomes project, is the relatively high prevalence of mutations that have previously been reported as causing severe disease. This suggests that a combination of incomplete penetrance, a false assignment of pathogenicity or a wide range in the expressivity of the phenotype may be more common features of disease mutations than is generally appreciated.

Genetic variation in the general human population can be interrogated using dbSNP (http://www.ncbi.nlm.nih.gov/SNP) or the Exome Aggregation Consortium (ExAC; http://exac.broadinstitute.org). The Exome Aggregation Consortium (ExAC) data set contains exome sequence data from more than 60,000 individuals with an assigned geographic ancestry. Approximately 60.9% of the samples in the ExAC reference cohort are of European ancestry, compared with 13.7% of South Asian ancestry, 9.6% of Latino ethnicity, 8.6% of African (African American) ancestry and 7.2% of East Asian ancestry.

Allele Frequencies Differ in Different Populations

Although most variants are common across human populations, rare gene variants can show markedly different patterns across different human communities. The 1000 genomes project established that there are several hundred thousand SNVs that show considerable differences in allelic frequencies in geographically and ancestry distinct populations. There are several explanations for this. Local populations may have adapted to their specific environments and genetic variants that facilitated this adaptation were selected by evolution (positive selection), which could explain the high frequency of mutation in the cystic fibrosis transmembrane conductance regulator (*CFTR*) gene in individuals of European ancestry. Carriers of *CFTR*

mutations may have had more resistance to cholera and other dehydrating intestinal disorders or are more resistant to contracting tuberculosis.

The demographic history of a population can also have a dramatic impact on allele frequencies in modern populations. Migration can change allele frequency by the process of gene flow, defined as the slow diffusion of genes across a barrier. This usually involves a large population and a gradual change in gene frequencies. The genes of migrant populations with their own characteristic allelic frequencies are gradually merged into the gene pool of the population into which they have migrated. Historically small and/or isolated populations or populations that experienced a population bottleneck can also effect allelic frequencies. Tay–Sachs disease in Ashkenazi Jews is an example, where a Tay–Sachs mutation arose by chance in a small breeding population and led to a *founder effect*.

Chance events can have a much greater effect on allele frequencies in a small population than in a large one. If the population is small, random effects, such as increased fertility or survival of the carriers of a mutation that occurred for reasons *unrelated* to carrying the mutant allele, may cause the allele frequency to change from one generation to the next. This is termed *genetic drift*. Whatever the mechanism, large-scale sequencing projects are showing that disease-causing alleles at relatively high frequencies in specific populations and communities as well as rare variants may be an important contributor to common diseases. This has an important impact on the clinical work-up of a patient, where it is essential to determine the ancestry of the affected individual.

Copy Number Variation (CNV)

Human populations also show extensive structural polymorphism, both deletions and duplications of chromosomal segments and, consequently, in the number of genes in these segments. Approximately two-thirds of the human genome is composed of repeats and 4.8–9.5% of the genome contributes to CNVs. Indeed, CNVs are thought to account for ~1% of the variation between two individuals. In contrast, SNVs are thought to account for ~0.1% of the variation. CNVs can arise both meiotically and somatically and can therefore contribute to variation between identical twins as well as variation between different organs and tissues of the same individual.

Smaller deletions and insertions (typically >50 kb) can be detected by comparative genomic hybridization (CGH) or multiplex ligation-dependent probe amplification (MLPA) analysis. MLPA is a variation of the multiplex polymerase chain reaction. For a short sequence of target DNA, two adjacent probes are designed to contain the forward and reverse primer sequence, respectively. In addition, one or both probes contain a stuffer sequence of which the length can be varied during the experiment. The probes are hybridized against the target DNA and subsequently ligated. Only if ligation happened does a functional PCR strand appear, so that amplification only happens if target DNA is present in the sample. The amount of PCR product is proportional to the amount of target DNA present in the sample, making the technique suitable for quantitative measurements.

Comparative genomic hybridization is a molecular cytogenetic method for analysing CNVs relative to ploidy level in the DNA of a test sample compared to a reference sample. Classically this was performed by differentially labelling a reference and test genome and hybridizing to an immobilized substrate such as a microarray. The fluorescence ratios provide a representation of the relative DNA CNV. Today, this is performed by a combination of hybridization of unlabelled DNA to target oligomers and enzymatic single-base extension to incorporate a labelled nucleotide for assay read-out but this is likely to be superseded by whole genome sequencing (WGS) in the near future.

There are still many hurdles in the clinical interpretation of CNVs. They are relatively common and there are many examples of known pathogenic CNVs exhibiting reduced penetrance and/or variable expressivity. This can result in a more severely affected child, who has inherited a CNV from a seemingly normal parent. The 22q11.2 deletion (del) syndrome is a classic example of this. In general, large, rare recurrent deletions and duplications are straightforward to interpret because of considerable genetic and phenotypic evidence. Typically, in a large multi-centre clinic, 15–20% of cases with developmental delay are associated with diagnostic findings from whole genome chromosomal microarray (CMA) analysis. Among these diagnostic cases, many rare CNVs are detected for which the potential functional significance is unknown and they are referred to as variants of uncertain (or unknown) significance (VUS). It is important to bear in mind that ~100 genes can be completely deleted from the human genome without phenotypic consequences.

The major challenge in this field is the detection and interpretation of small (>1–5 kb) rearrangements. These are generally too small to be detected by conventional microarrays but can be detected by WGS. Current data indicate that each individual carries many thousands of these small CNVs. Because of the absence of information in public databases on these small rearrangements, the interpretation of these small variants in challenging. CNVs are of considerable importance in pediatric endocrinology. A considerable proportion of individuals with