Molecular Hematology

SECOND EDITION

Edited by

Drew Provan MD FRCP FRCPath
Senior Lecturer in Haematology
Department of Haematology
Barts and The Royal London School of Medicine and Dentistry
Whitechapel
London
UK

John G Gribben MD DSc FRCP FRCPath
Associate Professor
Division of Medical Oncology
Dana-Farber Cancer Institute
Department of Medicine
Brigham and Women's Hospital
Harvard Medical School
Boston
USA

Blackwell
Publishing

Molecular Hematology

Molecular Hematology

SECOND EDITION

Edited by

Drew Provan MD FRCP FRCPath
Senior Lecturer in Haematology
Department of Haematology
Barts and The Royal London School of Medicine and Dentistry
Whitechapel
London
UK

John G Gribben MD DSc FRCP FRCPath
Associate Professor
Division of Medical Oncology
Dana-Farber Cancer Institute
Department of Medicine
Brigham and Women's Hospital
Harvard Medical School
Boston
USA

Blackwell
Publishing

Blackwell Publishing, Inc., 350 Main Street, Malden, Massachusetts 02148-5020, USA
Blackwell Publishing Ltd, 9600 Garsington Road, Oxford OX4 2DQ, UK
Blackwell Publishing Asia Pty Ltd, 550 Swanston Street, Carlton, Victoria 3053, Australia

First published 2000
Second edition 2005

Library of Congress Cataloging-in-Publication Data

Molecular haematology / edited by Drew Provan, John Gribben.
-- 2nd ed.
p. ; cm.
Includes bibliographical references and index.
ISBN 1-4051-1255-7
1. Blood--Diseases--Molecular aspects. I. Provan, Andrew.
II. Gribben, John. III. Title: Molecular hematology,
[DNLM: 1. Hematologic Diseases. 2. Molecular Biology.
WH 120 M7187 2004]
RC636.M576 2004
616.1'5--dc22

2004000858

ISBN 1-4051-1255-7

A catalogue record for this title is available from the British Library

Set in 10/12 pt Minion by Sparks, Oxford – www.sparks.co.uk
Printed and bound in India by Gopsons Papers Ltd., NOIDA

Commissioning Editor: Maria Khan
Managing Editor: Rupal Malde
Production Editor: Fiona Pattison
Production Controller: Kate Charman

For further information on Blackwell Publishing, visit our website:
http://www.blackwellpublishing.com

Dedication

We would like to dedicate this book to our families, especially Val, Fraser and Peter, who provided constant encouragement and support throughout the project.

Contents

List of contributors

Nancy C Andrews MD PhD
Associate Investigator, Howard Hughes Medical Institute, Associate Professor, Harvard Medical School, and Associate in Medicine, Children's Hospital Boston, 300 Longwood Avenue, Boston, MA 02115, USA

Eyal C Attar MD
Instructor in Medicine, Harvard Medical School, Massachusetts General Hospital, Boston, MA 02129, USA

Luciano Baronciani PhD
Research Assistant, Angelo Bianchi Bonomi Hemophilia and Thrombosis Center, IRCCS Maggiore Hospital, 20122 Milan, Italy

Anthony J Bench MA PhD
Senior Scientist, Department of Haematology, Addenbrooke's Hospital, Hills Road, Cambridge CB2 2QQ, UK

P Leif Bergsagel MD FRCP(C)
Associate Professor of Medicine, Division of Hematology/Oncology, Weill Medical College of Cornell University, 1300 York Avenue, New York, NY 10021, USA

Jaqueline Boultwood
University Research Lecturer and Co-Director of Leukaemia Research Fund Molecular Haematology Unit, Nuffield Department of Clinical Laboratory Sciences, John Radcliffe Hospital, Oxford OX3 9DU, UK

James B Bussel MD
Professor of Pediatrics, Division of Pediatrics, New York Presbyterian Hospital, Weill Cornell Medical Center, 525 E. 68th Street, Payson 609, New York, NY 10021, USA

Björn Dahlbäck MD PhD
Professor of Blood Coagulation Research, Department of Clinical Chemistry, University of Lund, University Hospital, Malmo S-20502, Sweden

Riccardo Dalla-Favera MD
Director, Institute for Cancer Genetics, Columbia University, 1150 St Nicholas Avenue, New York, NY 10032, USA

Silvana Debernardi PhD
Scientist, Cancer Research UK, Medical Oncology Unit, Barts and The London, Queen Mary's School of Medicine and Dentistry, Charterhouse Square, London EC1M 6BQ, UK

Katherine A Downes MD
Assistant Professor of Pathology, Department of Pathology, Case Western Reserve University School of Medicine/University Hospitals of Cleveland, 11000 Euclid Avenue Humphrey 5611, PTH 5077, Cleveland, OH 44106, USA, and Director of Coagulation, Associate Director of Blood Banking/Transfusion Medicine, University Hospitals of Cleveland, Cleveland, OH, USA

Brian J Druker MD
Investigator, Howard Hughes Medical Institute and JELD-WEN Chair of Leukemia Research, Department of Hematology/Medical Oncology, Oregon Health Sciences University Cancer Institute, L592, 3181 SW Sam Jackson Park Road, Portland, OR 97201-3011, USA

Paul L F Giangrande BSc MD FRCP FRCPath FRCPCH
Consultant Haematologist and Director, Oxford Haemophilia Centre and Thrombosis Unit, Churchill Hospital, Oxford OX3 7LJ, UK

D Gary Gilliland MD PhD
Professor of Medicine, Howard Hughes Medical Institute, Brigham and Women's Hospital, Dana-Farber Cancer Institute, Harvard Medical School, Boston, MA 02115, USA

Anthony R Green PhD FRCP FRCPath
Wellcome Senior Fellow and Honorary Consultant Haematologist, Department of Haematology, University of Cambridge, MRC Centre, Hills Road, Cambridge CB2 2QH, UK

John G Gribben MD DSc FRCP FRCPath
Associate Professor, Division of Medical Oncology, Dana-Farber Cancer Institute, Department of Medicine, Brigham and Women's Hospital, Harvard Medical School, Boston, MA 02115, USA

Andreas Hillarp PhD
Associate Professor and Hospital Chemist, Department of Clinical Chemistry, University of Lund, University Hospital, Malmo S-20502, Sweden

Brian J P Huntley MB ChB MRCP DipRCPath
Clinical Research Fellow, Department of Haematology, University of Cambridge, MRC Centre, Hills Road, Cambridge CB2 2QH, UK

Anastasios Karadimitris PhD MRCP MRCPath
Senior Lecturer and Honorary Consultant Haematologist, Leukaemia Research Fund Bennett Senior Fellow, Department of Haematology, Imperial College London, Hammersmith Hospital Campus, Du Cane Road, London W12 0NN, UK

Ulf Klein PhD
Associate Research Scientist, Columbia University, Institute of Cancer Genetics, Russ Berrie Pavillion, 1150 St Nicholas Avenue, New York, NY 10032, USA

Anthony G Letai MD PhD
Assistant Professor in Medicine, Harvard Medical School, Smith 758, Dana-Farber Cancer Institute, 1 Jimmy Fund Way, Boston, MA 02115, USA

Debra M Lillington BSc
Head of Cytogenetics, ICRF Medical Oncology Laboratory, Barts and The London, Queen Mary's School of Medicine and Dentistry, Charterhouse Square, London EC1M 6BQ, UK

Lucio Luzzatto MD
Scientific Director, National Institute for Cancer Research, Istituto Scientifico Tumori, Largo, Rosanna Benzi, 10, 16132 Genova, Italy

Pier Mannuccio Mannucci MD
Professor and Chairman of Internal Medicine, Angelo Bianchi Bonomi Hemophilia and Thrombosis Center, IRCCS Maggiore Hospital, 20122 Milan, Italy

Keith R McCrae MD
Associate Professor of Medicine Hematology-Oncology, BRB 3, Case Western Reserve University, School of Medicine, 10900 Euclid Avenue, Cleveland, OH 44106-4937, USA

Jeffrey A Medin PhD
Associate Professor, Department of Medical Biophysics, University of Toronto, Ontario Cancer Institute, Princess Margaret Hospital, 610 University Avenue, Rm 7-411, Toronto, Ontario M5G 2M9, Canada

Graham Molineux PhD
Amgen Inc., Mailstop 15-2-A, One Amgen Center Drive, Thousand Oaks, CA 91320, USA

Ronald L Nagel MD
Irving D. Karpas Professor of Medicine, Head, Division of Hematology, Albert Einstein College of Medicine, The Bronx, New York, NY 10461, USA

Cristina Navarrete PhD
Head of Histocompatibility and Immunogenetics, National Blood Service, Colindale Avenue, London NW9 5BG, UK

Adrian C Newland MA FRCP FRCPath
Professor of Haematology, Department of Haematology, Barts and The London, Queen Mary's School of Medicine and Dentistry, Whitechapel, London E1 1BB, UK

Willem H Ouwehand MD PhD
Lecturer in Haematology, Division of Transfusion Medicine, Department of Haematology, University of Cambridge, Cambridge CB2 2PT, UK

John R Pawloski MD PhD
Assistant Professor of Medicine, Division of Hematology, Department of Medicine, Duke University, Durham, NC 27710, USA

Drew Provan MD FRCP FRCPath
Senior Lecturer in Haematology, Department of Haematology, Barts and The London, Queen Mary's School of Medicine and Dentistry, Whitechapel, London E1 1BB, UK

David T Scadden MD
Professor of Medicine, Harvard Medical School, AIDS Research Centre, Massachusetts General Hospital, 149, 13th Street, CNY 5212, Charleston, Boston, MA 02129-2020, USA

A Keith Stewart MD FRCP(C)
Associate Professor, Department of Medical Oncology/ Hematology, Princess Margaret Hospital, 610 University Avenue, Toronto, Ontario M5G 2M9, Canada

Marilyn J Telen MD
Wellcome Professor of Hematology and Director, Duke-UNC Comprehensive Sickle Center, Box 2615 Duke University Medical Center, Durham, NC 27710, USA

George S Vassiliou MRCP(UK) DipRCPath BSc Hons (Lond)
Leukaemia Research Fund Clinical Research Fellow/Hon. Specialist Registrar, Department of Haematology, University of Cambridge, Cambridge Institute for Medical Research, Hills Road, Cambridge CB2 2XY, UK

James S Wainscoat FRCP FRCPath
Consultant Haematologist, Department of Haematology, John Radcliffe Hospital, Marston, Oxford OX3 9DU, UK

David Weatherall FRS
Regius Professor of Medicine, University of Oxford, Institute of Molecular Medicine, John Radcliffe Hospital, Headington, Oxford OX3 9DS, UK

Bryan D Young BSc PhD
Head and Professor of ICRF Medical Oncology Laboratory, Barts and The London, Queen Mary's School of Medicine and Dentistry, Charterhouse Square, London EC1M 6BQ, UK

Foreword

In 1968, after a quest lasting 30 years, X-ray analysis of crystalline horse hemoglobin at last reached the stage when I could build a model of its atomic structure. The amino acid sequences of human globin are largely homologous to those of horse globin, which made me confident that their structures are the same. By then, the amino acid substitutions responsible for many abnormal human hemoglobins had been determined. The world authority on them was the late Hermann Lehmann, Professor of Clinical Biochemistry at the University of Cambridge, who worked in the hospital just across the road from our Laboratory of Molecular Biology. I asked him to come over to see if there was any correlation between the symptoms caused by the different amino acids substituted in the abnormal hemoglobin and their positions in the atomic model. The day we spent going through them proved one of the most exciting in our scientific lives. We found hemoglobin to be insensitive to replacements of most amino acid residues on its surface, with the notable exception of sickle cell hemoglobin. On the other hand, we found the molecule to be extremely sensitive to even quite small alterations of internal non-polar contacts, especially those near the hemes. Replacements at the contact between the α and β subunits affected respiratory function.

In sickle cell hemoglobin an external glutamate was replaced by a valine. We wrote: '*A non-polar instead at a polar residue at a surface position would suffice to make each molecule adhere to a complementary site at a neighbouring one, that site being created by the conformational change from oxy to deoxy haemoglobin*'. This was soon proved to be correct. We published our findings under the title: 'The Molecular Pathology of Human Haemoglobin'. Our paper marked a turning point because it was the first time that the symptoms of diseases could be interpreted in terms of changes in the atomic structure of the affected protein. In the years that followed, the structure of the contact between the valine of one molecule of sickle cell hemoglobin and that of the complementary site of its neighbor became known in some detail. At a meeting at Arden House near Washington in 1980, several colleagues and I decided to use this knowledge for the design of anti-sickling drugs, but after an effort lasting 10 years, we realized that we were running up against a brick wall. Luckily, the work was not entirely wasted, because we found a series of compounds that lower the oxygen affinity of hemoglobin and we realized that this might be clinically useful. One of those compounds, designed by DJ Abraham at the University of Virginia in Richmond, is now entering phase 3 clinical trials. On the other hand, our failure to find a drug against sickle cell anemia, even when its cause was known in atomic detail, made me realize the extreme difficulty of finding drugs to correct a malfunction of a protein that is caused by a single amino acid substitution. Most thalassemias are due not to amino acid substitutions, but to either complete or partial failure to synthesize α- or β-globin chains. Weatherall's chapter shows that, at the genetic level, there may be literally hundreds of different genetic lesions responsible for that failure. Correction of such lesions is now the subject of intensive work in many laboratories.

Early in the next century, the human genome will be complete. It will reveal the amino acid sequences of all the 100000 or so different proteins of which we are made. Many of these proteins are still unknown. To discover their functions, the next project now under discussion is a billion dollar effort to determine the structures of all the thousands of unknown proteins within 10 years. By then we shall know the identity of the proteins responsible for most of the several thousand different genetic diseases. Will this lead to effective treatment or will medical geneticists be in the same position as doctors were early in this century when the famous physician Sir William Osler confined their task to the establishment of diagnoses? Shall we know the cause of every genetic disease without a cure?

Our only hope lies in somatic gene therapy. AK Stewart's chapter on Molecular Therapeutics describes the many ingenious methods now under development. So far, none of these has produced lasting effects, apparently because the transferred genes are not integrated into the mammalian genome, but a large literature already grown up bears testimony to the great efforts now underway to overcome this problem.

My much-loved teacher William Lawrence Bragg used to say '*If you go on hammering away at a problem, eventually it seems to get tired, lies down and lets you catch it*.' Let us hope that somatic gene therapy will soon get tired.

M.F. Perutz
Cambridge

Perutz MF, Lehmann H. (1968) Molecular pathology of human haemoglobin. *Nature*, **219**, 902–909.

Perutz MF, Muirhead H, Cox JM, Goaman LCG. (1968) Three-dimensional Fourier synthesis of horse oxyhaemoglobin at 2.8Å resolution: the atomic model. *Nature*, **219**, 131–139.

Preface to second edition

Hematology has seen many major developments since this book was first published, as molecular techniques become more powerful and the genetics of blood disorders are unraveled. The Human Genome Project has now been completed and a huge amount of new genetic information has become available.

Of course, most funding and resources continue to be spent on understanding the molecular basis of malignant disease and hence there has been a huge increase in our knowledge base for leukemias, lymphomas and other malignancies. Molecular biology has also begun to bring with it advances in treatment, with molecules devised to reduce the tumor burden through much more subtle and selective mechanisms than have been possible with conventional chemotherapy agents. Chronic myeloid leukemia (CML), one of the best studied human malignancies, is amenable to treatment with the molecule STI571, revolutionizing our therapy of this disease. Dr Brian Druker's seminal work on this molecule is explained in detail in his chapter devoted entirely to the biology and management of CML. We have updated all the chapters on malignant hematology and have included some new chapters, such as 'Stem cells', 'Secondary myelodysplasia/acute myelogenous leukemia—assessment of risk' and 'Gene expression profiling in the study of lymphoid malignancies'.

Non-malignant disease has also enjoyed the benefits of this new genetic information and the original chapters have been updated to reflect this new knowledge. New chapters dealing with the molecular basis of blood group antigens, the molecular basis of von Willebrand disease, and platelet disorders have been added by leading clinicians and researchers in these fields.

However, despite the growing complexity of the pathogenesis, diagnosis and management of patients with blood diseases, the ethos of the book remains the same: to provide a succinct account of the molecular biology of hematological disease written at a level at which it should be of benefit to the seasoned molecular biologist and the practicing clinician alike. We have retained the original structure for the chapters, with high-quality artwork and 'Further reading' sections, in order to make the book visually appealing and relevant to modern hematology practice.

As before, we welcome any comments or suggestions from readers, which we will attempt to incorporate into the next edition.

Drew Provan (a.b.provan@qmul.ac.uk)
John Gribben (John_Gribben@dfci.harvard.edu)

Suggested general reading

Anderson KC, Ness PM (eds). (2000) *Scientific Basis of Transfusion Medicine: Implications for Clinical Practice*, 2nd edn. Philadelphia: W.B. Saunders.

Beutler E, Lichtman MA (eds). (2001) *Williams Hematology* 6th edn. New York: McGraw-Hill.

Cooper GM. (1997) *The Cell: A Molecular Approach*. Washington, DC: ASM Press.

Cox TM, Sinclair J. (1997) *Molecular Biology in Medicine*. Oxford: Blackwell Science.

Jameson JL (ed.). (1998) *Principles of Molecular Medicine*. New York: Humana Press.

Mullis KB. (1990) The unusual origin of the polymerase chain reaction. *Scientific American*, **262**, 56–65.

Roitt I. (2001) *Roitt's Essential Immunology*, 10th edn. Oxford: Blackwell Science.

Stamatoyannopoulos G, Nienhuis AW, Majerus PW, Varmus H (eds). (2000) *The Molecular Basis of Blood Diseases*, 2nd edn. Philadelphia: W.B. Saunders.

Watson JD, Gilman M, Witkowski J, Zoller M (eds). (1992) *Recombinant DNA*, 2nd edn. New York: Scientific American Books.

Acknowledgments

We would like to express thanks to Maria Khan (Commissioning Editor) and Rupal Malde (Managing Editor) for helping keep the project on track and for their continuing support throughout the project. We are very grateful to Peter Varney for his administrative assistance during this edition. As ever, David Gardner provided very high quality artwork, for which we are very grateful.

Abbreviations

AAV	adeno-associated virus
ADA	adenosine deaminase
AE1	anion exchanger protein 1
ALAS	δ-aminolaevulinate synthase
ALL	acute lymphoblastic leukemia
AML	acute myeloid leukemia (Chapters 2, 4) acute myelogenous leukemia (Chapters 5, 6, 8)
APC	activated protein C; antigen-presenting cell
APL	acute promyelocytic leukemia
ASCT	autologous stem cell transplantation
AT	antithrombin
ATRA	all-*trans*-retinoic acid
B-CLL	B-cell chronic lymphocytic leukemia
BCR	breakpoint cluster region
BM	bone marrow
BMF	bone marrow failure
BPG	2,3-biphosphoglycerate
BSS	Bernard–Soulier syndrome
CBF	core binding factor
Cbl	cobalamin
CDA	congenital dyserythropoietic anemia
CDAE1	N-terminal cytoplasmic domain of anion exchanger protein 1
CDKI	cyclin-dependent kinase inhibitor
CDR	complementarity-determining region
CFC	colony-forming cell
CFU	colony-forming unit
CFU-S	colony-forming units–spleen
CGH	comparative genomic hybridization
CH	heavy-chain constant region
CHR	complete hematologic response
CLL	chronic lymphocytic leukemia
CLM	common lymphoid progenitor
cM	centimorgan
CML	chronic myeloid leukemia
CNS	central nervous system
CSF	colony-stimulating factor
CTLA	cytotoxic T-lymphocyte antigen
DC	dyskeratosis congenita
DLBL	diffuse large B-cell lymphoma
dm	double minute (chromosome)
EGF	epidermal growth factor
EPO	erythropoietin
ERT	enzyme replacement therapy
EST	expressed sequence tag
ET	essential thrombocythemia
FA	Fanconi's anemia
FAB	French–American–British (classification of myelodysplastic syndromes)
FACS	fluorescence-activated cell sorter
FcR	Fc receptor
FG	phenylalanine-glycine
FISH	fluorescence *in situ* hybridization
G6PD	glucose-6-phosphate dehydrogenase
GAP	glycine-alanine-proline
G-CSFR	granulocyte colony-stimulating factor receptor
GDP	guanosine diphosphate
Ge	Gerbich erythrocyte antigen
GEP	gene expression profiling
GM-CSF	granulocyte macrophage colony-stimulating factor
GP	glycoprotein
GPC, D	glycophorin C, glycophorin D
GPI	glucosylphosphatidylinositol
GT	Glanzmann thrombasthenia
GTP	guanosine triphosphate
GVHD	graft-versus-host disease
HDN	hemolytic disease of the newborn
HGF	hematopoietic growth factor
HHV-8	human herpesvirus
HLA	human leukocyte antigen
HMCL	human multiple myeloma cell line
HPA	human platelet antigen
HPFF	hereditary persistence of fetal hemoglobin 1
HPP-CFC	high proliferative potential colony-forming cell
HSC	hemopoietic stem cell
HSCT	hematopoietic stem cell transplantation
hsr	homogeneously staining region
IAA	idiopathic aplastic anemia
IDDm	insulin-dependent diabetes mellitus

IFN-γ	interferon γ
Ig	immunoglobulin
IL	interleukin
IMF	idiopathic myelofibrosis
IPI	International Prognostic Index
IPPS	International Prognostic Scoring System
ISC	irreversibly sickled cell
ITD	internal tandem duplications
ITP	idiopathic thrombocytopenic purpura
IVIg	intravenous immunoglobulin
IVS	intervening sequence
kb	kilobase pair (1000 base pairs)
KS	Kaposi's sarcoma
KTLS	c-kit^{pos}Thy-1.1^{low}Lin^{neg}Sca-1^{pos} cellular phenotype
LCS	locus control region
LOH	loss of heterozygosity
LTC-IC	long-term culture-initiating cell
LTR	long terminal repeat
mAb	monoclonal antibody
M-BCR	major breakpoint cluster region
m-BCR	minor breakpoint cluster region
MBR	major breakpoint region
MCH	mean corpuscular hemoglobin
MCHC	mean corpuscular hemoglobin concentration
MCL	mantle cell lymphoma
M-CSF	macrophage colony-stimulating receptor
MCV	mean cell volume
MDS	myelodysplastic syndrome
MGDF	megakaryocyte growth and development factor
MGUS	monoclonal gammopathy of undetermined significance
MHC	major histocompatibility complex
MM	multiple myeloma
MPD	myeloproliferative disorder
MRD	minimal residual disease
mtDNA	mitochondrial DNA
NAITP	neonatal/fetal alloimmune thrombocytopenia
NHL	non-Hodgkin's lymphoma
NO	nitric oxide
NOs	nitric oxide synthase
ORF	open reading frame
PAR	protease-activated receptor
PBMC	peripheral blood mononuclear cell
PBPC	peripheral blood progenitor cell
PCR	polymerase chain reaction
PEG-MGDF	pegylated megakaryocyte growth and development factor
PETS	paraffin-embedded tissue section
PMPS	Pearson's marrow–pancreas syndrome
PNH	paroxysmal nocturnal hemoglobinuria
PUBS	peri-umbilical blood sampling
PV	polycythemia vera
RBC	red blood cell
RSCA	reference strand conformational analysis
RT-PCR	reverse transcriptase–polymerase chain reaction
SA	sideroblastic anemia
SCF	stem cell factor
SDS-PAGE	sodium dodecylsulfate–polyacrylamide gel electrophoresis
SKY	spectral karyotyping
SLE	systemic lupus erythematosus
SNO-Hb	S-nitrosohemoglobin
SNP	single-nucleotide polymorphism
TBI	total body irradiation
TCR	T-cell receptor
TF	tissue factor
TGF-β	transforming growth factor-β
THF	tetrahydrofolate
V, D, J, C	variable, diversity, joining and constant regions
VCAM-1	vascular cell adhesion molecule-1
VH	heavy-chain variable region
vWD	von Willebrand disease
vWF	von Willebrand factor
vWF:RCo	von Willebrand factor ristocetin cofactor activity
WBC	white blood cell

Chapter 1 Beginnings: the molecular pathology of hemoglobin

David Weatherall

Historical background

Linus Pauling first used the term 'molecular disease' in 1949, after the discovery that the structure of sickle cell hemoglobin differed from that of normal hemoglobin. Indeed, it was this seminal observation that led to the concept of *molecular medicine*; that is, the description of disease mechanisms at the level of cells and molecules. However, until the development of recombinant DNA technology in the mid-1970s, knowledge of events inside the cell nucleus, notably how genes function, could only be the subject of guesswork based on the structure and function of their protein products. However, as soon as it became possible to isolate human genes and to study their properties, the picture changed dramatically.

Progress over the last 20 years has been driven by technological advances in molecular biology. At first it was possible only to obtain indirect information about the structure and function of genes by DNA/DNA and DNA/RNA hybridization; that is, by probing the quantity or structure of RNA or DNA by annealing reactions with molecular probes. The next major advance was the ability to fractionate DNA into pieces of predictable size with bacterial restriction enzymes. This led to the invention of a technique that played a central role in the early development of human molecular genetics, called 'Southern blotting' after the name of its developer, Edwin Southern. This method allowed the structure and organization of genes to be studied directly for the first time and led to the definition of a number of different forms of molecular pathology.

Once it was possible to fractionate DNA, it soon became feasible to insert the pieces into vectors that are able to divide within bacteria. The steady improvement in the properties of cloning vectors made it possible to generate libraries of human DNA growing in bacterial cultures. Ingenious approaches were developed to scan the libraries to detect genes of interest; once pinpointed, the appropriate bacterial colonies could be grown to generate larger quantities of DNA carrying a particular gene. Later it became possible to sequence these genes, persuade them to synthesize their products in microorganisms, cultured cells or even other species, and hence to define their key regulatory regions.

The early work in the field of human molecular genetics focussed on diseases in which there was some knowledge of the genetic defect at the protein or biochemical level. However, once linkage maps of the human genome became available, following the identification of highly polymorphic regions of DNA, it was possible to search for any gene for a disease, even where the cause was completely unknown. This approach, first called 'reverse genetics' and later rechristened 'positional cloning', led to the discovery of genes for many important diseases.

As even more DNA markers became available and as methods for sequencing were improved and automated, thoughts turned to the next major goal in this field, which was to determine the complete sequence of the bases that constitute our 30 000 or so genes and all that lies between them: the Human Genome Project. This remarkable endeavor was partially completed recently and should be finished within the next few years. The further understanding of the functions and regulation of our genes will require multidisciplinary research encompassing many different fields. The next stage in the Human Genome Project, called 'genome annotation', entails analyzing the raw DNA sequence in order to determine its biological significance. One of the main ventures in the era of functional genomics will be in what is termed 'proteomics', the large-scale analysis of the protein products of genes. The ultimate goal will be to try to define the protein complement, or proteome, of cells and how the many different proteins interact with one another. To this end, large-scale facilities are being established for isolating and purifying the protein

products of genes that have been expressed in bacteria. Their structure can then be studied by a variety of different techniques, notably X-ray crystallography and nuclear magnetic resonance spectroscopy. The crystallographic analysis of proteins is being greatly facilitated by the use of X-ray beams from a synchrotron radiation source.

During this remarkable period of technical advance, considerable progress has been made towards an understanding of the pathology of disease at the molecular level. This has had a particular impact on hematology, leading to advances in the understanding of gene function and disease mechanisms in almost every aspect of the field.

The inherited disorders of hemoglobin, the thalassemias and structural hemoglobin variants, the commonest human monogenic diseases, were the first to be studied systematically at the molecular level and a great deal is known about their genotype–phenotype relationships. This field led the way to molecular hematology and, indeed, to the development of molecular medicine. Thus, even though the genetics of hemoglobin is complicated by the fact that different varieties are produced at particular stages of human development, the molecular pathology of the hemoglobinopathies provides an excellent model system for understanding any monogenic disease and the complex interactions between genotype and environment that underlie many multigenic disorders.

In this chapter we will consider the structure, synthesis and genetic control of the human hemoglobins, describe the molecular pathology of the hemoglobin disorders individually, and discuss briefly how the complex interactions of their different genotypes produce a remarkably diverse family of clinical phenotypes. Readers who wish to learn more about the methods of molecular genetics, particularly as applied to the study of hemoglobin disorders, are referred to the reviews cited at the end of this chapter.

The structure, genetic control and synthesis of normal hemoglobin

Structure and function

The varying oxygen requirements during embryonic, fetal and adult life are reflected in the synthesis of different structural hemoglobins at each stage of human development. They all have the same general tetrameric structure, however, consisting of two different pairs of globin chains, each attached to one heme molecule. Adult and fetal hemoglobins have α chains combined with β chains (Hb A, $\alpha_2\beta_2$), δ chains (Hb A_2, $\alpha_2\delta_2$) and γ chains (Hb F, $\alpha_2\gamma_2$). In embryos, α-like chains called ζ chains combine with γ chains to produce Hb Portland ($\zeta_2\gamma_2$), or with ε chains to make Hb Gower 1 ($\zeta_2\varepsilon_2$), while α and ε chains form Hb Gower 2 ($\alpha_2\varepsilon_2$). Fetal hemoglobin is heterogeneous; there are two varieties of γ chain that differ only in their amino acid composition at position 136, which may be occupied by either glycine or alanine; γ chains containing glycine at this position are called G_γ chains, those with alanine, A_γ chains (Figure 1.1).

The synthesis of hemoglobin tetramers consisting of two unlike pairs of globin chains is absolutely essential for the effective function of hemoglobin as an oxygen carrier. The classical sigmoid shape of the oxygen dissociation curve, which reflects the allosteric properties of the hemoglobin molecule, ensures that, at high oxygen tensions in the lungs, oxygen is readily taken up and later released effectively at the lower tensions encountered in the tissues. The shape of the curve is quite different to that of myoglobin, a molecule which consists of a single globin chain with heme attached to it, which, like abnormal hemoglobins that consist of homotetramers of like-chains, has a hyperbolic oxygen dissociation curve.

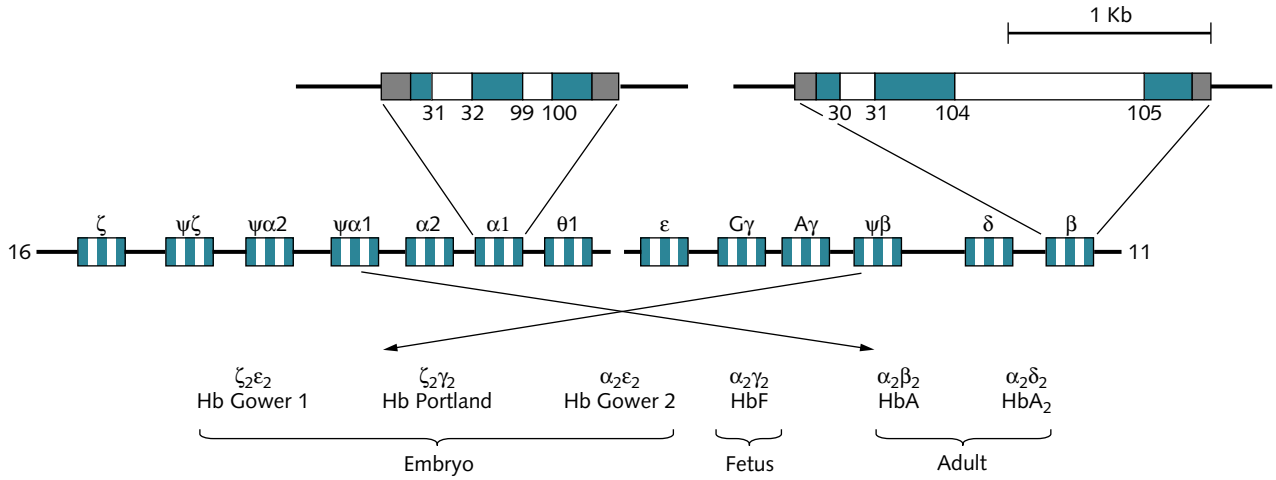

Fig. 1.1 The genetic control of human hemoglobin production in embryonic, fetal and adult life

The transition from a hyperbolic to a sigmoid oxygen dissociation curve, which is absolutely critical for normal oxygen delivery, reflects cooperativity between the four heme molecules and their globin subunits. When one of them takes on oxygen, the affinity of the remaining three increases markedly; this happens because hemoglobin can exist in two configurations, deoxy(T) and oxy(R), where T and R represent the tight and relaxed states, respectively. The T configuration has a lower affinity than the R for ligands such as oxygen. At some point during the addition of oxygen to the hemes, the transition from the T to the R configuration occurs and the oxygen affinity of the partially liganded molecule increases dramatically. These allosteric changes result from interactions between the iron of the heme groups and various bonds within the hemoglobin tetramer, which lead to subtle spatial changes as oxygen is taken on or given up.

The precise tetrameric structures of the different human hemoglobins, which reflect the primary amino acid sequences of their individual globin chains, are also vital for the various adaptive changes that are required to ensure adequate tissue oxygenation. The position of the oxygen dissociation curve can be modified in several ways. For example, oxygen affinity decreases with increasing CO_2 tension (the Bohr effect). This facilitates oxygen loading to the tissues, where a drop in pH due to CO_2 influx lowers oxygen affinity; the opposite effect occurs in the lungs. Oxygen affinity is also modified by the level of 2,3-biphosphoglycerate (2,3-BPG) in the red cell. Increasing concentrations shift the oxygen dissociation curve to the right, that is, they reduce oxygen affinity, while diminishing concentrations have the opposite effect. 2,3-BPG fits into the gap between the two β chains when it widens during deoxygenation, and interacts with several specific binding sites in the central cavity of the molecule. In the deoxy configuration the gap between the two β chains narrows and the molecule cannot be accommodated. With increasing concentrations of 2,3-BPG, which are found in various hypoxic and anemic states, more hemoglobin molecules tend to be held in the deoxy configuration and the oxygen dissociation curve is therefore shifted to the right, with more effective release of oxygen.

Fetal red cells have greater oxygen affinity than adult red cells, although, interestingly, purified fetal hemoglobin has an oxygen dissociation curve similar to that of adult hemoglobin. These differences, which are adapted to the oxygen requirements of fetal life, reflect the relative inability of Hb F to interact with 2,3-BPG compared with Hb A. This is because the γ chains of Hb F lack specific binding sites for 2,3-BPG.

In short, oxygen transport can be modified by a variety of adaptive features in the red cell that include interactions between the different heme molecules, the effects of CO_2 and differential affinities for 2,3-BPG. These changes, together with more general mechanisms involving the cardiorespiratory system, provide the main basis for physiological adaptation to anemia.

Genetic control of hemoglobin

The α- and β-like globin chains are the products of two different gene families which are found on different chromosomes (Figure 1.1). The β-like globin genes form a linked cluster on chromosome 11, spread over approximately 60 kb (kb = kilobase or 1000 nucleotide bases). The different genes that form this cluster are arranged in the order 5′–ε–^{G}γ–^{A}γ–ψβ–δ–β–3′. The α-like genes also form a linked cluster, in this case on chromosome 16, in the order 5′–ζ–ψζ–ψα1–α2–α1–3′. The ψβ, ψζ and ψα genes are pseudogenes; that is, they have strong sequence homology with the β, ζ and α genes but contain a number of differences that prevent them from directing the synthesis of any products. They may reflect remnants of genes that were functional at an earlier stage of human evolution.

The structure of the human globin genes is, in essence, similar to that of all mammalian genes. They consist of long strings of nucleotides that are divided into coding regions, or exons, and non-coding inserts called 'intervening sequences' (IVS), or introns. The β-like globin genes contain two introns, one of 122–130 between codons 30 and 31 and one of 850–900 base pairs between codons 104 and 105 (the exon codons are numbered sequentially from the 5′ to the 3′ end of the gene; that is, from left to right). Similar, though smaller, introns are found in the α and ζ globin genes. These introns and exons, together with short non-coding sequences at the 5′ and 3′ ends of the genes, represent the major functional regions of the particular genes. However, there are also extremely important regulatory sequences that subserve these functions, which lie outside the genes themselves.

At the 5′ non-coding (flanking) regions of the globin genes, as in all mammalian genes, there are blocks of nucleotide homology. The first, the ATA box, is about 30 bases upstream (to the left) of the initiation codon; that is, the start word for the beginning of protein synthesis (*see below*). The second, the CCAAT box, is about 70 base pairs upstream from the 5′ end of the genes. About 80–100 bases further upstream there is the sequence GGGGTG, or CACCC, which may be inverted or duplicated. These three highly conserved DNA sequences, called 'promoter elements', are involved in the initiation of transcription of the individual genes. Finally, in the 3′ non-coding region of all the globin genes there is the sequence AATAAA, which is the signal for cleavage and polyA addition to RNA transcripts (*see below: Gene action and globin synthesis*).

The globin gene clusters also contain several sequences that constitute regulatory elements, which interact to promote erythroid-specific gene expression and coordination of the changes in globin gene activity during development. These

include the globin genes themselves and their promoter elements—enhancers (regulatory sequences that increase gene expression despite being located at a considerable distance from the genes) and 'master' regulatory sequences called, in the case of the β globin gene cluster, the 'locus control region' (LCR); and, in the case of the α genes, HS40 (a nuclease-hypersensitive site in DNA 40 kb from the α globin genes). Each of these sequences has a modular structure made up of an array of short motifs that represent the binding sites for transcriptional activators or repressors.

Gene action and globin synthesis

The flow of information between DNA and protein is summarized in Figure 1.2. When a globin gene is transcribed, messenger RNA (mRNA) is synthesized from one of its strands, a process which begins with the formation of a transcription complex consisting of a variety of regulatory proteins together with an enzyme called RNA polymerase (*see below*). The primary transcript is a large mRNA precursor which contains both intron and exon sequences. While in the nucleus, this molecule undergoes a variety of modifications. First, the introns are removed and the exons are spliced together. The intron/exon junctions always have the same sequence: GT at their 5′ end, and AG at their 3′ end. This appears to be essential for accurate splicing; if there is a mutation at these sites this process does not occur. Splicing reflects a complex series of intermediary stages and the interaction of a number of different nuclear proteins. After the exons are joined, the mRNAs are modified and stabilized; at their 5′ end a complex CAP structure is formed, while at their 3′ end a string of adenylic acid residues (polyA) is added. The mRNA processed in this way moves into the cytoplasm, where it acts as a template for globin chain production. Because of the rules of base pairing—that is, cytosine always pairs with thymine, and guanine with adenine—the structure of the mRNA reflects a faithful copy of the DNA codons from which it is synthesized; the only difference is that, in RNA, uracil (U) replaces thymine (T).

Amino acids are transported to the mRNA template on carriers called transfer RNAs (tRNAs); there are specific tRNAs for each amino acid. Furthermore, because the genetic code is redundant (that is, more than one codon can encode a particular amino acid), for some of the amino acids there are several different individual tRNAs. Their order in the globin chain is determined by the order of codons in the mRNA. The tRNAs contain three bases, which together constitute an anticodon; these anticodons are complementary to mRNA codons for particular amino acids. They carry amino acids to the template, where they find the appropriate positioning by codon–anticodon base-pairing. When the first tRNA is in po-

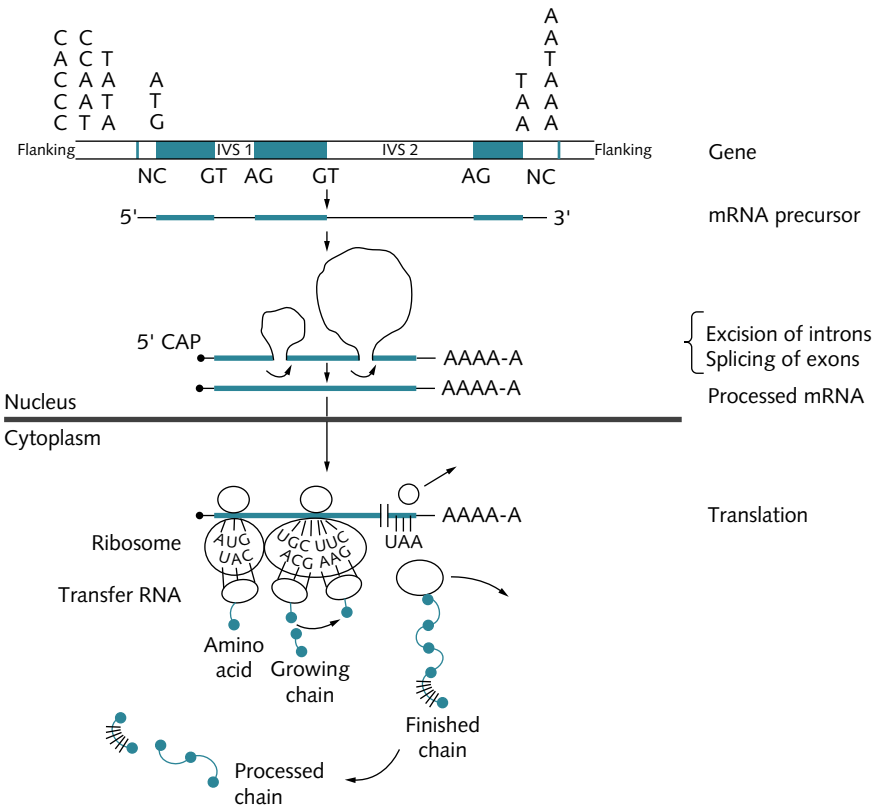

Fig. 1.2 The mechanisms of globin gene transcription and translation

sition, an initiation complex is formed between several protein initiation factors together with the two subunits which constitute the ribosomes. A second tRNA moves in alongside and the two amino acids that they are carrying form a peptide bond between them; the globin chain is now two amino acid residues long. This process is continued along the mRNA from left to right, and the growing peptide chain is transferred from one incoming tRNA to the next; that is, the mRNA is translated from 5′ to 3′. During this time the tRNAs are held in appropriate steric configuration with the mRNA by the two ribosomal subunits. There are specific initiation (AUG) and termination (UAA, UAG and UGA) codons. When the ribosomes reach the termination codon, translation ceases, the completed globin chains are released, and the ribosomal subunits are recycled. Individual globin chains combine with heme, which has been synthesized through a separate pathway, and then interact with one like chain and two unlike chains to form a complete hemoglobin tetramer.

Regulation of hemoglobin synthesis

The regulation of globin gene expression is mediated mainly at the transcriptional level, with some fine tuning during translation and post-translational modification of the gene products. DNA that is not involved in transcription is held tightly packaged in a compact, chemically modified form that is inaccessible to transcription factors and polymerases and which is heavily methylated. Activation of a particular gene is reflected by changes in the structure of the surrounding chromatin, which can be identified by enhanced sensitivity to nucleases. Erythroid lineage-specific nuclease-hypersensitive sites are found at several locations in the β globin gene cluster. Four are distributed over 20 kb upstream from the ε globin gene in the region of the β globin LCR (Figure 1.3). This vital regulatory region is able to establish a transcriptionally active domain spanning the entire β globin gene cluster. Several enhancer sequences have been identified in this cluster. A variety of regulatory proteins bind to the LCR, and to the promoter regions of the globin genes and to the enhancer sequences. It is thought that the LCR and other enhancer regions become opposed to the promoters to increase the rate of transcription of the genes to which they are related.

These regulatory regions contain sequence motifs for various ubiquitous and erythroid-restricted transcription factors. Binding sites for these factors have been identified in each of the globin gene promoters and at the hypersensitive-site regions of the various regulatory elements. A number of the factors which bind to these areas are found in all cell types. They include Sp1, Yy1 and Usf. In contrast, a number of transcription factors have been identified, including GATA-1, EKLF and NF-E2, which are restricted in their distribution to erythroid cells and, in some cases, megakaryocytes and mast cells. The overlapping of erythroid-specific and ubiquitous-factor binding sites in several cases suggests that competitive binding may

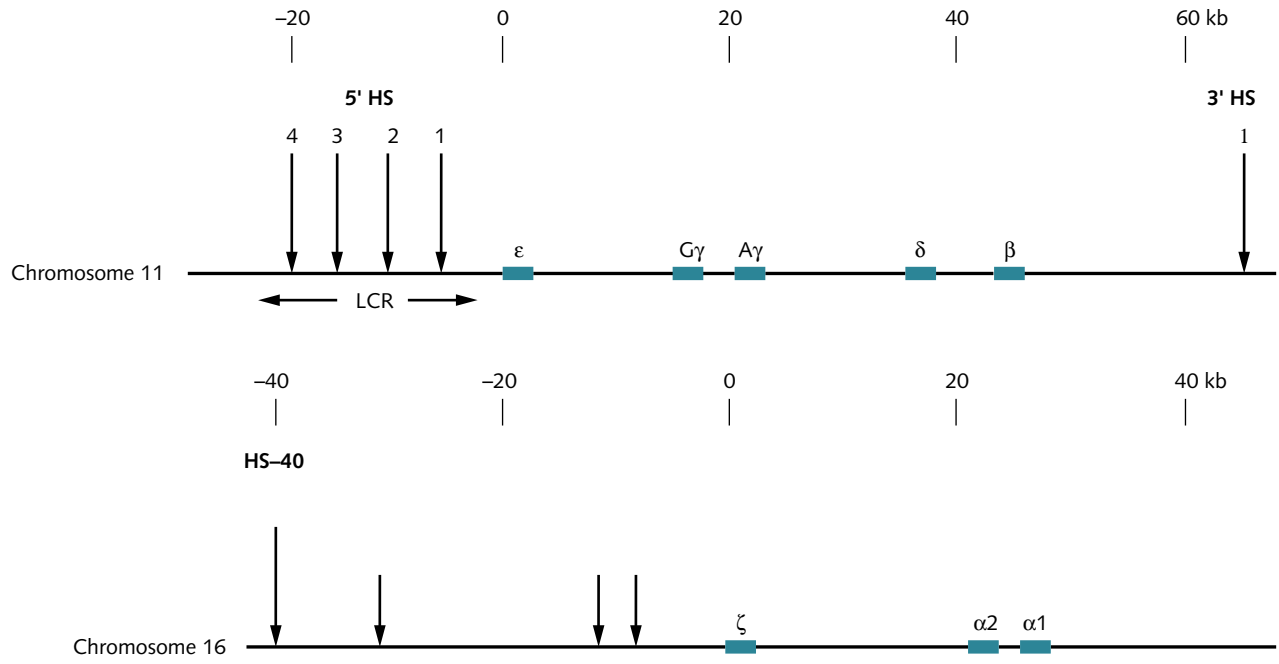

Fig. 1.3 The positions of the major regulatory regions in the β and α globin gene clusters
The arrows indicate the position of the erythroid lineage-specific nuclease-hypersensitive sites. HS = hypersensitive.

play an important part in the regulation of erythroid-specific genes. Another binding factor, SSP, the stage selector protein, appears to interact specifically with ε and γ genes.

The binding of hematopoietic-specific factors activates the LCR, which renders the entire β globin gene cluster transcriptionally active. These factors also bind to the enhancer and promoter sequences, which work in tandem to regulate the expression of the individual genes in the clusters. It is likely that some of the transcriptional factors are developmental-stage-specific, and hence may be responsible for the differential expression of the embryonic, fetal and adult globin genes. The α globin gene cluster also contains an element, HS40, which has some structural features in common with the β LCR, although it is different in aspects of its structure. A number of enhancer-like sequences have also been identified, although it is becoming clear that there are fundamental differences in the pattern of regulation of the two globin gene clusters.

In addition to the different regulatory sequences outlined above, there are also sequences which may be involved specifically with 'silencing' of genes, notably those for the embryonic hemoglobins, during development.

Some degree of regulation is also mediated by differences in the rates of initiation and translation of the different mRNAs, and at the post-transcriptional level by differential affinity for different protein subunits. However, this kind of post-transcriptional fine tuning probably plays a relatively small role in determining the overall output of the globin gene products.

Regulation of developmental changes in globin gene expression

During development, the site of red cell production moves from the yolk sac to the fetal liver and spleen, and thence to bone marrow in the adult. Embryonic, fetal and adult hemoglobin synthesis is approximately related in time to these changes in the site of erythropoiesis, although it is quite clear that the various switches, between embryonic and fetal and between fetal and adult hemoglobin synthesis, are beautifully synchronized throughout these different sites. Fetal hemoglobin synthesis declines during the later months of gestation and Hb F is replaced by Hbs A and A_2 by the end of the first year of life.

Despite a great deal of research, very little is known about the regulation of these different switches from one globin gene to another during development. Work from a variety of different sources suggests that there may be specific regions in the α and β globin gene clusters that are responsive to the action of transcription factors, some of which may be developmental-stage-specific. However, proteins of this type have not yet been isolated, and nothing is known about their regulation and how it is mediated during development.

The molecular pathology of hemoglobin

As is the case for most monogenic diseases, the inherited disorders of hemoglobin fall into two major classes. First, there are those that result from a reduced output of one or other globin genes, the *thalassemias*. Second, there is a wide range of conditions that result from the production of *structurally abnormal globin chains*; the type of disease depends on how the particular alteration in protein structure interferes with its stability or function. Of course, no biological classification is entirely satisfactory; those which attempt to define the hemoglobin disorders are no exception. There are some structural hemoglobin variants which happen to be synthesized at a reduced rate and hence are associated with a clinical picture similar to thalassemia. And there are other classes of mutations which simply interfere with the normal transition from fetal to adult hemoglobin synthesis, a family of conditions that is given the general title 'hereditary persistence of fetal hemoglobin'. Furthermore, because these diseases are all so common and occur together in particular populations, it is not uncommon for an individual to inherit a gene for one or other form of thalassemia and a structural hemoglobin variant. The rather heterogeneous group of conditions that results from all these different mutations and interactions is summarized in Table 1.1.

Table 1.1 The thalassemias and related disorders.

α **Thalassemia**
α^0
α^+
Deletion ($-\alpha$)
Non-deletion (α^T)
β **Thalassemia**
β^0
β^+
Normal Hb A_2
'Silent'
δβ **Thalassemia**
$(\delta\beta)^+$
$(\delta\beta)^0$
$(^A\gamma\delta\beta)^0$
γ **Thalassemia**
δ **Thalassemia**
εγδβ **Thalassemia**
Hereditary persistence of fetal hemoglobin
Deletion
$(\delta\beta)^0$
Non-deletion
Linked to β globin genes
$^G\gamma\beta^+$
$^A\gamma\beta^+$
Unlinked to β globin genes

Over recent years, the determination of the molecular pathology of the two common forms of thalassemia, α and β, has provided a remarkable picture of the repertoire of mutations that can underlie human monogenic disease. Similarly, studies of the relationship between structure and function in the structurally abnormal hemoglobins have provided a great deal of information about normal human hemoglobin function.

In the sections that follow we will describe, in outline, the different forms of molecular pathology that underlie these conditions.

The β thalassemias

There are two main classes of β thalassemia, β^0 thalassemia, in which there is an absence of β globin chain production, and β^+ thalassemia, in which there is a variable reduction in the output of β globin chains. As shown in Figure 1.4, mutations of the β globin genes may cause a reduced output of gene product at the level of transcription or mRNA processing, translation, or through the stability of the globin gene product.

Defective β globin gene transcription

There are a variety of mechanisms that interfere with the normal transcription of the β globin genes. First, the genes may be either completely or partially deleted. Overall, deletions of the β globin genes are not commonly found in patients with β thalassemia, with one exception: a 619 bp deletion involving the 3′ end of the gene is found frequently in the Sind populations of India and Pakistan, where it constitutes about 30% of the β thalassemia alleles. Other deletions are extremely rare.

A much more common group of mutations, which results in a moderate decrease in the rate of transcription of the β globin genes, involves single nucleotide substitutions in or near the TATA box at about −30 nucleotides (nt) from the transcription start site, or in the proximal or distal promoter elements at −90 nt and −105 nt. These mutations result in decreased β globin mRNA production, ranging from 10 to 25% of the normal output. Thus, they are usually associated with the mild forms of β^+ thalassemia. They are particularly common in African populations, an observation which explains the unusual mildness of β thalassemia in this racial group. One particular mutation, C→T at position −101 nt to the β globin gene, causes an extremely mild deficit of β globin mRNA. Indeed, this allele is so mild that it is completely silent in carriers and can only be identified by its interaction with more severe β thalassemia alleles in compound heterozygotes.

Mutations that cause abnormal processing of mRNA

As mentioned earlier, the boundaries between exons and introns are marked by the invariant dinucleotides GT at the donor (5′) site and AG at the acceptor (3′) site. Mutations (base changes) that affect either of these sites completely abolish normal splicing and produce the phenotype of β^0 thalassemia. The transcription of genes carrying these mutations appears to be normal, but there is complete inactivation of splicing at the altered junction.

Another family of mutations involves what are called 'splice site consensus sequences'. Although only the GT dinucleotide is invariant at the donor splice site, there is conservation of adjacent nucleotides and a common, or consensus, sequence of these regions can be identified. Mutations within this sequence can reduce the efficiency of splicing to varying degrees because they lead to alternate splicing at the surrounding cryptic sites. For example, mutations of the nucleotide at

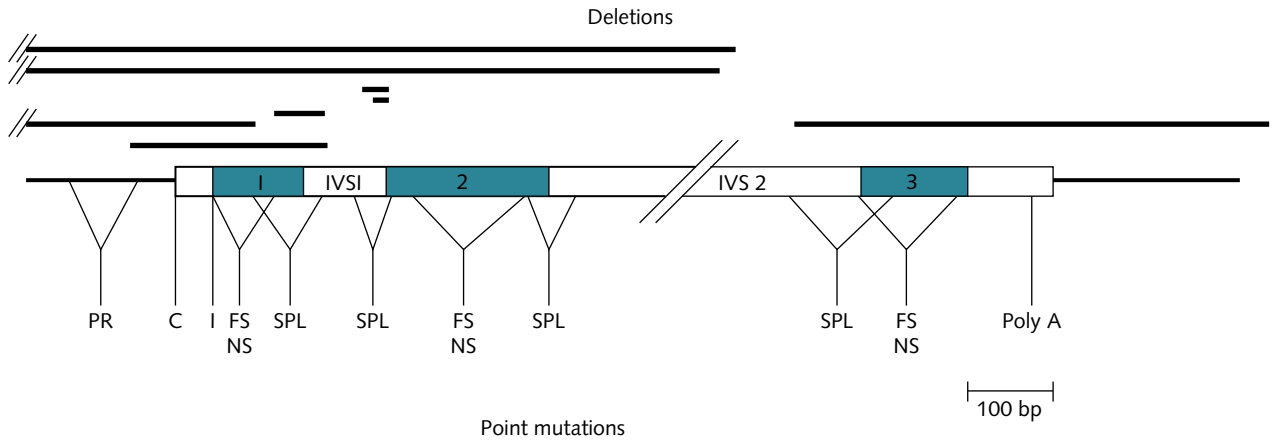

Fig. 1.4 The mutations of the β globin gene that underlie β thalassemia
The heavy black lines indicate the length of the deletions. The point mutations are designated as follows: PR, promoter; C, CAP site; I, initiation codon; FS, NS, frameshift and nonsense mutations; SPL, splice mutations; Poly A, poly A addition site mutations.

position 5 of IVS-1 (the first intervening sequence), G→C or T, result in a marked reduction of β chain production and in the phenotype of severe β^+ thalassemia. On the other hand, the substitution of C for T at position 6 in IVS-1 leads to only a mild reduction in the output of β chains.

Another mechanism that leads to abnormal splicing involves 'cryptic splice sites'. These are regions of DNA which, if mutated, assume the function of a splice site at an inappropriate region of the mRNA precursor. For example, a variety of mutations activate a cryptic site which spans codons 24–27 of exon 1 of the β globin gene. This site contains a GT dinucleotide, and adjacent substitutions that alter it so that it more closely resembles the consensus donor splice site result in its activation, even though the normal splice site is intact. A mutation at codon 24 GGT→GGA, though it does not alter the amino acid which is normally found in this position in the β globin chain (glycine), allows some splicing to occur at this site instead of the exon–intron boundary. This results in the production of both normal and abnormally spliced β globin mRNA and hence in the clinical phenotype of severe β thalassemia. Interestingly, mutations at codons 19, 26 and 27 result in both reduced production of normal mRNA (due to abnormal splicing) and an amino acid substitution when the mRNA which is spliced normally is translated into protein. The abnormal hemoglobins produced are hemoglobins Malay, E and Knossos, respectively. All these variants are associated with a mild β^+ thalassemia-like phenotype. These mutations illustrate how sequence changes in coding rather than intervening sequences influence RNA processing, and underline the importance of competition between potential splice site sequences in generating both normal and abnormal varieties of β globin mRNA.

Cryptic splice sites in introns may also carry mutations that activate them even though the normal splice sites remain intact. A common mutation of this kind in Mediterranean populations involves a base substitution at position 110 in IVS-1. This region contains a sequence similar to a 3′ acceptor site, though it lacks the invariant AG dinucleotide. The change of the G to A at position 110 creates this dinucleotide. The result is that about 90% of the RNA transcript splices to this particular site and only 10% to the normal site, again producing the phenotype of severe β^+ thalassemia (Figure 1.5). Several other β thalassemia mutations have been described which generate new donor sites within IVS-2 of the β globin gene.

Another family of mutations that interferes with β globin gene processing involves the sequence AAUAAA in the 3′ untranslated regions, which is the signal for cleavage and polyadenylation of the β globin gene transcript. Somehow, these mutations destabilize the transcript. For example, a T→C substitution in this sequence leads to only one-tenth of the normal amount of β globin mRNA transcript and hence to the phenotype of a moderately severe β^+ thalassemia. Another example of a mutation which probably leads to defective processing of function of β globin mRNA is the single base substitution, A→C, in the CAP site. It is not yet understood how this mutation causes a reduced rate of transcription of the β globin gene.

There is another small subset of rare mutations which involve the 3′ untranslated region of the β globin gene and are associated with relatively mild forms of β thalassemia. It is

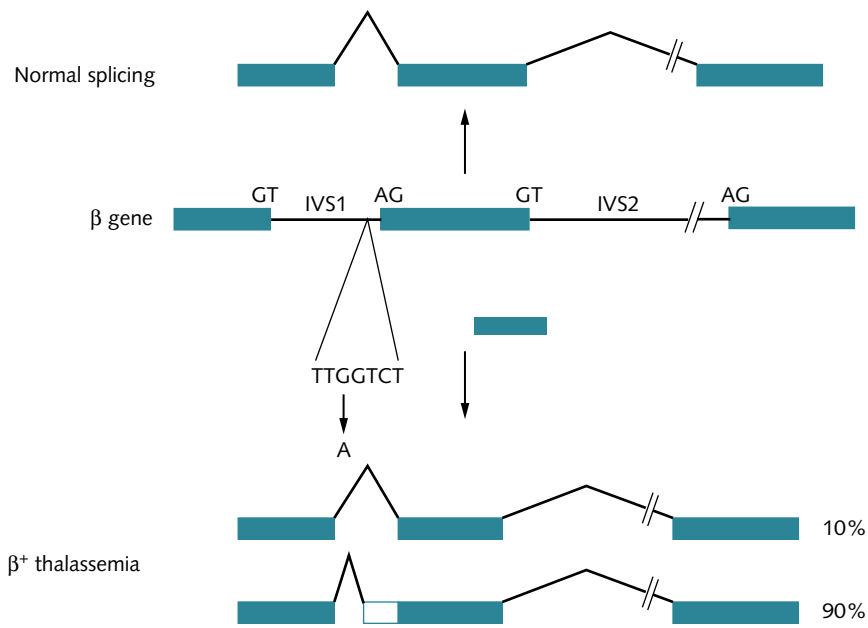

Fig. 1.5 The generation of a new splice site in an intron as the mechanism for a form of β^+ thalassemia
For details see text.

thought that these interfere in some way with transcription but the mechanism is unknown.

Mutations that result in abnormal translation of β globin mRNA

There are three main classes of mutations of this kind. Base substitutions that change an amino acid codon to a chain termination codon prevent the translation of β globin mRNA and result in the phenotype of β^0 thalassemia. Several mutations of this kind have been described; the commonest, involving codon 17, occurs widely throughout Southeast Asia. Similarly, a codon 39 mutation is encountered frequently in the Mediterranean region.

The second class involves the insertion or deletion of one, two or four nucleotides in the coding region of the β globin gene. These disrupt the normal reading frame, cause a frameshift, and hence interfere with the translation of β globin mRNA. The end result is the insertion of anomalous amino acids after the frameshift until a termination codon is reached in the new reading frame. This type of mutation always leads to the phenotype of β^0 thalassemia.

Finally, there are several mutations which involve the β globin gene initiation codon and which, presumably, reduce the efficiency of translation.

Unstable β globin chain variants

Some forms of β thalassemia result from the synthesis of highly unstable β globin chains which are incapable of forming hemoglobin tetramers, and which are rapidly degraded, leading to the phenotype of β^0 thalassemia. Indeed, in many of these conditions no abnormal globin chain product can be demonstrated by protein analysis and the molecular pathology has to be interpreted simply on the basis of a derived sequence of the variant β chain obtained by DNA analysis.

Recent studies have provided some interesting insights into how complex clinical phenotypes may result from the synthesis of unstable β globin products. For example, there is a spectrum of disorders that result from mutations in exon 3 which give rise to a moderately severe form of β thalassemia in heterozygotes. It has been found that nonsense or frameshift mutations in exons I and II are associated with the absence of messenger RNA from the cytoplasm of red cell precursors. This appears to be an adaptive mechanism, called 'nonsense-mediated decay', whereby abnormal messenger RNA of this type is not transported to the cytoplasm, where it would act as a template for the production of truncated gene products. However, in the case of exon III mutations, apparently because this process requires the presence of an intact upstream exon, the abnormal messenger RNA is transported into the cytoplasm and hence can act as a template for the production of unstable β globin chains. The latter precipitate in the red cell precursors together with excess α chains to form large inclusion bodies, and hence there is enough globin chain imbalance in heterozygotes to produce a moderately severe degree of anemia.

The molecular pathology of the α thalassemias

The molecular pathology of the α thalassemias is more complicated than that of the β thalassemias, simply because there are two α globin genes per haploid genome. Thus, the normal α globin genotype can be written αα/αα. As in the case of β thalassemia, there are two major varieties of α thalassemia, α^+ and α^0 thalassemia. In α^+ thalassemia one of the linked α globin genes is lost, either by deletion (–) or mutation (T); the heterozygous genotype can be written –α/αα or α^Tα/αα. In α^0 thalassemia the loss of both α globin genes nearly always results from a deletion; the heterozygous genotype is therefore written – –/αα. In populations where specific deletions are particularly common—Southeast Asia (SEA) or the Mediterranean region (MED)—it is useful to add the appropriate superscript, as follows: $--^{SEA}$/αα or $--^{MED}$/αα. It follows that when we speak of an 'α thalassemia gene' what we are really referring to is a haplotype; that is, the state and function of both of the linked α globin genes.

α^0 Thalassemia

Three main molecular pathologies, all involving deletions, have been found to underlie the α^0 thalassemia phenotype. The majority of cases result from deletions that remove both α globin genes and a varying length of the α globin gene cluster (Figure 1.6). Occasionally, however, the α globin gene cluster is intact but is inactivated by a deletion which involves the major regulatory region HS40, 40 kb upstream from the α globin genes. Finally, the α globin genes may be lost as part of a truncation of the tip of the short arm of chromosome 16.

As well as providing us with an understanding of the molecular basis for α^0 thalassemia, detailed studies of these deletions have yielded more general information about the mechanisms that underlie this form of molecular pathology. For example, it has been found that the 5′ breakpoints of a number of deletions of the α globin gene cluster are located approximately the same distance apart and in the same order along the chromosome as their respective 3′ breakpoints; similar findings have been observed in deletions of the β globin gene cluster. These deletions seem to have resulted from illegitimate recombination events which have led to the deletion of an integral number of chromatin loops as they pass through their nuclear attachment points during chromosomal replication. Another long deletion has been characterized

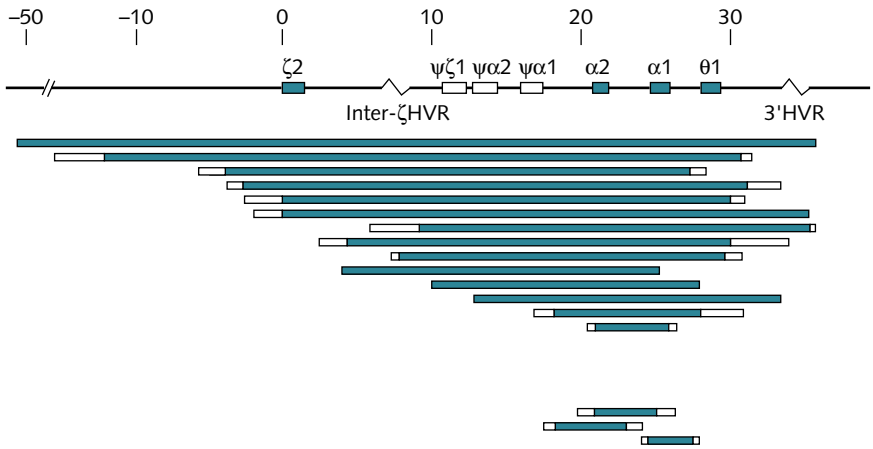

Fig. 1.6 Some of the deletions that underlie α^0 and α^+ thalassemia
The heavy red lines indicate the lengths of the deletions. The unshaded regions indicate uncertainty about the precise breakpoints. The three small deletions at the bottom of the figure represent the common α^+ thalassemia deletions.

in which a new piece of DNA bridges the two breakpoints in the α globin gene cluster. The inserted sequence originates upstream from the α globin gene cluster, where it normally is found in an inverted orientation with respect to that found between the breakpoints of the deletion. Thus it appears to have been incorporated into the junction in a way that reflects its close proximity to the deletion breakpoint region during replication. Other deletions seem to be related to the family of Alu-repeats, simple repeat sequences that are widely dispersed throughout the genome; one deletion appears to have resulted from a simple homologous recombination between two repeats of this kind that are usually 62 kb apart.

A number of forms of α^0 thalassemia result from terminal truncations of the short arm of chromosome 16 to a site about 50 kb distal to the α globin genes. The telomeric consensus sequence TTAGGGn has been added directly to the site of the break. Since these mutations are stably inherited, it appears that telomeric DNA alone is sufficient to stabilize the ends of broken chromosomes.

The molecular pathology of α^+ thalassemia

As mentioned earlier, the α^+ thalassemias result from the inactivation of one of the duplicated α globin genes, either by deletion or point mutation.

α^+ Thalassemia due to gene deletions

There are two common forms of α^+ thalassemia that are due to loss of one or other of the duplicated α globin genes, $-\alpha^{3.7}$ and $-\alpha^{4.2}$, where 3.7 and 4.2 indicate the sizes of the deletions. The way in which these deletions have been generated reflects the underlying structure of the α globin gene complex (Figure 1.7). Each α gene lies within a boundary of homology, approximately 4 kb long, probably generated by an ancient duplication event. The homologous regions, which are divided by small inserts, are designated X, Y and Z. The duplicated Z boxes are 3.7 kb apart and the X boxes are 4.2 kb apart. As the result of misalignment and reciprocal crossover between these segments at meiosis, a chromosome is produced with either a single ($-\alpha$) or triplicated ($\alpha\alpha\alpha$) α globin gene. As shown in Figure 1.7, if a crossover occurs between homologous Z boxes 3.7 kb of DNA are lost, an event which is described as a rightward deletion, $-\alpha^{3.7}$. A similar crossover between the two X boxes deletes 4.2 kb, the leftward deletion $-\alpha^{4.2}$. The corresponding triplicated α gene arrangements are called $\alpha\alpha\alpha^{anti\ 3.7}$ and $\alpha\alpha\alpha^{anti\ 4.2}$. A variety of different points of crossing over within the Z boxes give rise to different length deletions, still involving 3.7 kb.

Non-deletion types of α^+ thalassemia

These disorders result from single or oligonucleotide mutations of the particular α globin gene. Most of them involve the α2 gene but, since the output from this locus is two to three times greater than that from the α1 gene, this may simply reflect ascertainment bias due to the greater phenotypic effect and, possibly, a greater selective advantage.

Overall, these mutations interfere with α globin gene function in a similar way to those that affect the β globin genes. They affect the transcription, translation or post-translational stability of the gene product. Since the principles are the same as for β thalassemia, we do not need to describe them in detail with one exception, a mutation which has not been observed in the β globin gene cluster. It turns out that there is a family of mutations that involves the α2 globin gene termination codon, TAA. Each specifically changes this codon so that an amino acid is inserted instead of the chain terminating. This is followed by 'read-through' of α globin mRNA, which is not normally translated until another in-phase termination codon is reached. The result is an elongated α chain with 31 additional residues at the C terminal end. Five hemoglobin variants of this type have been identified. The commonest,