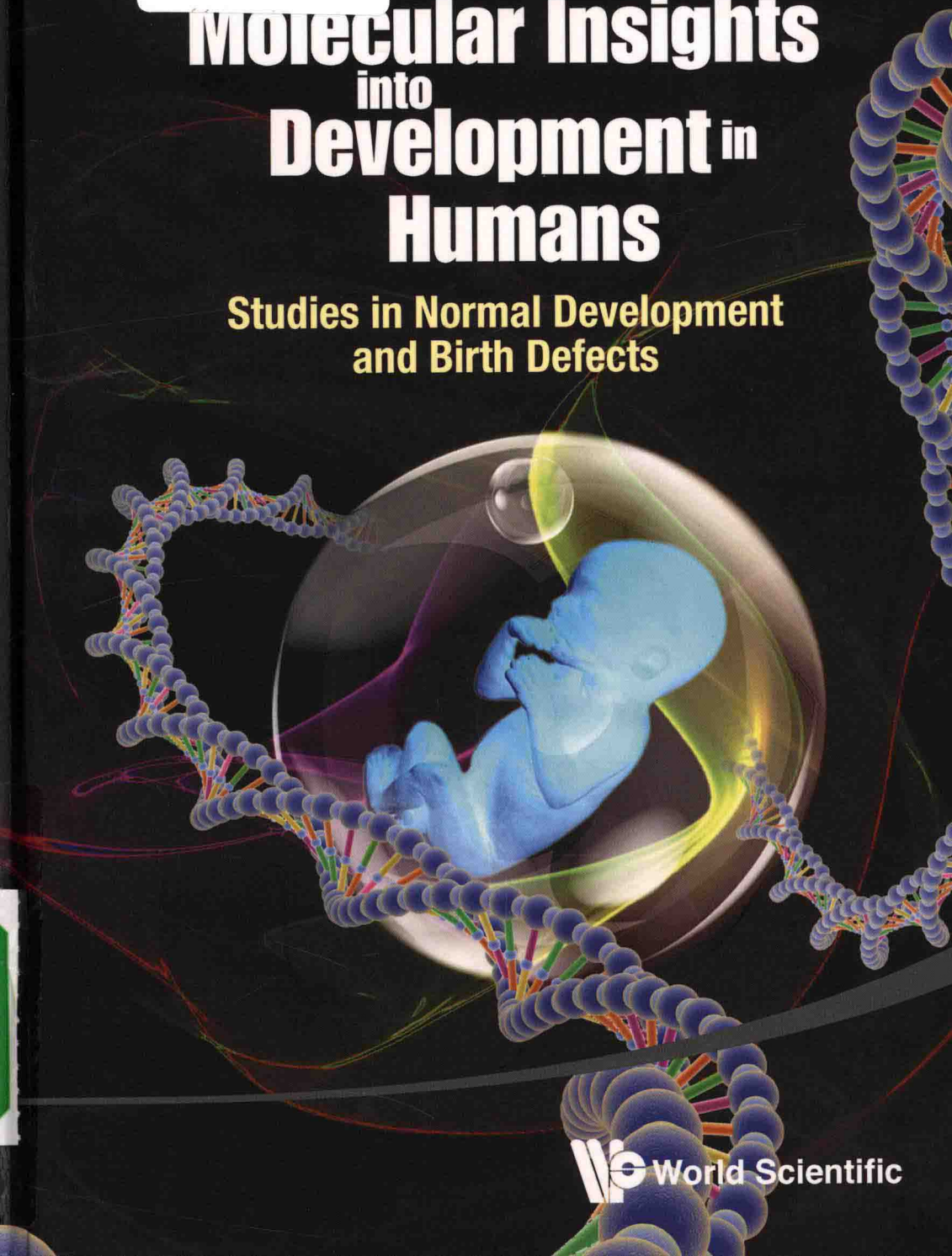
Moyra Smith

Molecular Insights into Development in Humans

Studies in Normal Development and Birth Defects



World Scientific

Molecular Insights into Development in Humans

Studies in Normal Development and Birth Defects

Moyra Smith

University of California, Irvine, USA

World Scientific

NEW JERSEY • LONDON • SINGAPORE • BEIJING • SHANGHAI • HONG KONG • TAIPEI • CHENNAI

Published by

World Scientific Publishing Co. Pte. Ltd.

5 Toh Tuck Link, Singapore 596224

USA office: 27 Warren Street, Suite 401-402, Hackensack, NJ 07601

UK office: 57 Shelton Street, Covent Garden, London WC2H 9HE

Library of Congress Cataloging-in-Publication Data

Smith, Moyra, author.

Molecular insights into development in humans : studies in normal development and birth defects / by Moyra Smith.

p. ; cm.

Includes bibliographical references and index.

ISBN 978-9814630580 (hardcover : alk. paper)

I. Title.

[DNLM: 1. Human Development. 2. Congenital Abnormalities--genetics. WS 103]

RB155.5

616'.042--dc23

2014041575

British Library Cataloguing-in-Publication Data

A catalogue record for this book is available from the British Library.

Typeset by Stallion Press

Email: enquiries@stallionpress.com

Printed in Singapore

Molecular Insights into Development in Humans

Studies in Normal Development and Birth Defects

This work is dedicated to four friends and colleagues
on three continents who have encouraged and inspired me,
Dr. Susan Dyson and Dr. Simon Prinloo in South Africa,
Dr. David Hopkinson in England, and Pamela Flodman in the USA.
Simon, thank you for being a catalyst.

PREFACE

> "And meanwhile the most incredible miracles are happening all around us, stones when we lift them and let them go fall to the ground, the sun shines, bees visit flowers, seeds grow into plants, a cell in nine months multiplies its weight a few thousands and thousands of times and is a child; and men think creating the world they live in".
>
> Aldous Huxley, 1924

And the miracle extends beyond multiplication of cells to their differentiation to multiple different tissues and specialized organs. It is the potential for pluripotency and mechanisms of differentiation that particularly inspire further study.

Our concepts of pluripotency and reversal of differentiation were greatly expanded with reports in 2006 that the introduction into the cell of only four transcription factors could promote differentiated cells to become pluripotent. Intense studies have been carried out over the past decade to determine the factors required to promote differentiation of pluripotent cells to differentiated cells. These studies have relevance not only to regenerative medicine but also enhance our concepts of normal development.

In this book, along with descriptions of development of particular organ systems, I explore aspects of relevant studies on pluripotent stem cells.

A second quotation relevant to the goals of this book and to studies on congenital malformations comes from Dr. James Paget in 1882.

> "We ought not to set them aside with idle thoughts or idle words about "curiosities" or "chances". Not one of them is without meaning; not one that might not become the beginning of excellent knowledge, if only we could answer the question-why is it rare? Or being rare why did in this instance happen".

Dr. David Smith used this quotation in 1970 in his seminal publication, *Recognizable patterns of human malformation*. He inspired generations of clinicians and researchers to carry out careful examinations, to document patterns of human malformation, to document family history and to track the long-term consequences of malformations.

Availability of techniques for detailed chromosome analyses in the last three decades of the 20th century revealed that a number of recognizable syndromic forms of human malformation were due to trisomy or monosomy of specific chromosomes or dosage changes of specific chromosome segments.

The availability of human DNA sequence data for each of the human chromosomes through efforts of the Human Genome Project and enhanced capabilities of sequencing DNA from individuals has led to identification of specific gene mutations that give rise to structural or functional defects. In addition scientists are now able to explore the epigenome, secondary modifications of DNA, histone and chromatin and nucleosome positioning that play critical roles in regulation of gene expression.

In this book, I present examples of instances where insights into steps in embryonic development have been gained through discovery of specific gene alterations or mutations in cases with developmental defects.

CONTENTS

Preface ..... **vii**

Chapter 1: Genomes, Genes, Structure and Function 1
Chapter 2: Epigenetics 24
Chapter 3: Signaling Systems 41
Chapter 4: Pluripotency to Differentiation 60
Chapter 5: Development and Growth Abnormalities 71
Chpater 6: Metabolism and Organelles 88
Chapter 7: Brain Development 111
Chapter 8: Cranio–Facial Development and Defects 143
Chapter 9: Molecular Aspects of Heart Development 150
Chapter 10: Vasculogenesis Malformations and Hematopoiesis 157
Chapter 11: Abdominal Wall and Gastro-Intestinal Tract 164
Chapter 12: Lung and Diaphragm Development 168
Chapter 13: Liver and Pancreas Development 173
Chapter 14: Bone and Extra-Cellular Matrix 184
Chapter 15: Kidney and Urinary Tract Development, Malformations in Human Genes 192
Chapter 16: Sex Determination Differentiation and Endocrine Glands 195

Conclusion **208**
References **211**
Index **249**

CHAPTER 1

GENOMES, GENES, STRUCTURE AND FUNCTION

Molecular embryology involves analysis of gene expression and delineation of the key gene products involved in determining differentiation at the cellular and tissue level.

Overlapping Layers within Genome Architecture

Overlapping layers within the genome architecture need to be taken into account in analyzing gene expression, these include:

(a) The linear sequence of DNA and linear gene structure;
(b) The embedding of DNA in histone rich chromatin that undergoes modifications that impact gene expression;
(c) The arrangement of DNA and chromatin within the nucleus at different stages of the cell cycle.

During metaphase distinct chromosomes can be identified. More precise identification of each of the human chromosomes was initially achieved through development of specific histological staining techniques that revealed banding patterns on chromosomes (Caspersson *et al.*, 1970; Seabright, 1972). Molecular cytogenetics began in 1977 when labeled DNA probes were hybridized to chromosomes.

Analysis of folding and looping of DNA and chromatin in interphase nuclei and of three-dimensional structure is currently being actively analyzed. There is evidence that gene regulation is influenced by looping of chromatin and DNA and by dynamic changes in the positioning of loops on which specific genes and regulators are positioned. Fluorescence *in-situ* hybridization (FISH) techniques provided means to examine the location of specific gene targets within the nucleus. FISH studies initially provided evidence that specific chromosomes have preferred positions in the nucleus. The arrangement differed in different cell types (Bickmore, 2013).

Molecular Organization of Eukaryotic Genes

The molecular organization of eukaryotic genes, including human genes has been actively ongoing since the late 1970s when it became clear that these genes contain coding segments (exons) interspersed with non-coding segments introns (Leder, 1978). Analyses revealed that transcription is initiated from the 5 prime (5′) end of the gene from a site adjacent to promoter sequences that are located further 5′ (upstream) of the transcription initiation site. The promoter sequences are not transcribed but comprise elements that facilitate transcription: specific regions within the promoter bind polymerases essential for transcription.

Detailed molecular studies revealed that the primary mRNA transcript undergoes cleavage at specific points and subsequent rejoining and that this splicing process removes introns. The precise location of splicing sometimes varied depending on the nucleotide sequences present at specific positions. Specific nucleotide sequences at the 3′ end of the transcript directs binding of an endonuclease that lead to cleavage at the 3′ end and this cleavage site then polyadenylated.

Developmental Control of Gene Expression

Development is characterized by temporal and spatial differences in gene expression. Development requires the integrative action of gene promoters and cis-regulatory elements including those that lie close to promoters and those that lie at some distance from transcription start sites. Important cis-acting elements that impact gene expression include enhancers, silencers, insulators and transcription factor binding sites.

In this section, I review information on promoters, on cis-regulatory elements including enhancers and transcription factors and their binding sites. Following this aspects of transcription, alternative splicing, alternate generation of 3′ end sequences and aspects of translation are then reviewed.

Pal *et al.* (2011) emphasized that analysis of transcription and its regulation is essential for deciphering cell and tissue specific gene functions. Alternative transcription initiation and alternative transcription termination give rise to alternative mRNA transcripts and these may undergo alternative splicing and then give rise to alternate protein isoforms.

Promoters

Promoters are not transcribed; they contain sequence elements that enhance transcription capability. Promoters include core domains and regulatory domains. Key sequence elements within the promoters include the polymerase II binding site and transcription binding sites. Protein coding genes are transcribed by RNA polymerase II and core elements included in these genes include TATA boxes (sequence rich in thymine and adenine), CpG islands (repeats of cytosine and guanine) are often present in the promoter regions. Chromatin modifications are key in determining transcription initiation and are discussed in the subsequent chapter on epigenetics. TATA boxes are present in promoters of approximately 24% of human genes. Key to transcription of TATA box-containing promoters is a TATA box binding protein (TBP). This protein binds to the TATA box and acts to properly position the RNA polymerase II. It also serves as a scaffold for the assembly of transcription factors.

Alternate promoter use

Different stages of development and different cellular and tissue conditions may require use of different promoters that are differentially regulated. Furthermore sequence differences in promoters lead to binding of different transcription factors.

Factors that determine the use of alternate promoters include DNA sequences in regulatory regions and histone modifications. Pal *et al.* (2011) reported that 50% of the multi-transcript genes they studied used multiple promoters. They also determined that genes which used multiple promoters also exhibited alternative transcription termination and alternative splicing. They carried out mRNA sequencing and epigenetic analysis to develop an inventory of transcript variants that occur in development of the cerebellum in the mouse. Their studies revealed extensive changes in the expression of transcript variants of a number of different genes during differentiation of the granular cells of the cerebellum.

Examples of genes with multiple promoters

The uridine diphosphate glucuronosyl transferase 1A (*UGT1A*) gene on human chromosome 2q37 has 13 alternative promoters. The *BDNF* gene

that encodes "brain derived neurotrophic factor" is expressed from nine different promoters. The gene gives rise to multiple different transcripts that result from use of different promoters, different transcription start sites and from alternative splicing of exons. Neuronal activity impacts *BDNF* transcription and splicing (Autry and Monteggia, 2012).

Sakata *et al.* (2009) reported that *BDNF* promoter IV plays a particularly important role in *BDNF* transcription. They reported that absence of that promoter in mice led to deficits in specific neurotransmitter functions in gamma amino butyric acid (GABA) ergic interneurons in the prefrontal cortex.

Methods to analyze alternate promoters of a specific gene: CAP sequencing

Messenger RNA contains a 5′ CAP sequence in which 7-methylguanosine is linked to the first transcribed nucleotide in a phosphodiesterase bond. The CAP synthesizing complex is associated with RNA polymerase. CAGE sequencing is a sequencing method that involves biotinylation of this 5′ 7-methylguanosine and then selection through streptavidin binding. This capture enables sequencing of 5′ mRNA adjacent to the 5′ CAP. CAGE sequencing has revealed that most genes have multiple promoters and there is frequent tissue specific or developmental stage specific promoter usage. Faulkner *et al.* (2009) reported that repetitive DNA and retro-transposons (ancient sequences that can amplify themselves and are mobile in the genome) are often associated with the 5′ region of protein coding genes and may serve as alternate promoters. They proposed that retro-transposon transcription has a key influence on the transcriptional output of the mammalian genome.

Promoters have a 3′ splice donor but lack a 5′ splice sequence. Each different promoter can then bind to the downstream exon of a gene through binding to the splice acceptor adjacent to that exon.

Transcription

Transcription factors are key constituents of gene expression regulatory systems and are central elements in determination of differentiation and development. Two distinct domains that occur within transcription factor

proteins include a DNA binding domain that permits binding to sequence elements in DNA and secondly an activator domain that directly impacts transcription. In some cases the transcription factor interacts with a co-activator domain. The repertoire of transcription factors includes general or basal transcription factors that facilitate transcription and specific transcription factors that bind only to specific DNA sequences.

Transcription initiation

Control of transcription initiation is essential to development. Distal and proximal enhancers and promoters are involved in this process. Transcription start sites can be identified through analysis of sequence at the 5′ end of full-length cDNAs (complementary DNA segments sequenced from messenger RNA). Kawaji *et al.* (2006) reported that for a specific gene, transcription start site selection varied in different tissues and that most genes do not have a specific single transcription start site. They determined that for a specific gene, alternative start sites were present and these were spread across the 5′ gene region. Furthermore for a specific gene transcription, start site selection varied in different tissues.

Recent studies have revealed that transcription factors may bind at the transcription start site or downstream or upstream of that site. The transcription machinery that assembles at the promoter includes 27 polypeptides including general transcription factors.

Enhancers contain specific DNA sequence elements to which transcription factors bind. However the low sequence specificity of the 6–12 nucleotide elements that bind transcription factors results in the potential for a specific sequence element to bind different transcription factors. It is therefore possible that at different stages of development different transcription factors may bind to a specific enhancer element. Furthermore it seems likely that different combinations of transcription factors may bind to a specific enhancer site.

Pioneer transcription factors

Spitz and Furlong (2012) reviewed evidence for the existence and function of pioneer transcription factors. These bind to specific genomic sequence sites and enhancer sequences and lead to alterations in nucleosome

positioning in adjacent genomic regions. The pioneer transcription factors do not necessarily activate the enhancers to which they bind. The pioneer transcription factors may subsequently be replaced by other transcription factors. Pioneer transcription factors may also serve to protect enhancer sites from methylation. Examples of transcription factors that may have pioneer function include MyoD and PAX5.

Spitz and Furlong (2012) emphasized that expression of a specific gene is dependent upon enhancers, available promoter elements and the three-dimensional arrangement of genome segments.

An important consideration is whether structural chromosome changes; e.g., duplication, deletions, translocations and inversions alter the three-dimensional chromosome organization and the potential for promoter enhancer interactions.

Transcription factors and pluripotency of cells

The field of transcription factor research took a great leap forward with the discovery of the key role of transcription factors in converting somatic cells, such as fibroblasts, to pluripotent stem cells. Stem cells and induction of pluripotency and differentiation will be discussed in a subsequent chapter.

Function and expression of human transcription factors

In a review of human transcription factors, Vaquerizas *et al.* (2009) presented data on 1,391 manually curated sequence specific transcription factors. Complete genome sequence analysis of DNA elements that bind to transcription factors have led to the creation of a number of different databases with inventories of transcription factors. Classifications of transcription factors are based on DNA binding characteristics, on protein domains and structural homologies and on the basis of the biological processes in which they participate. Vaquerizas *et al.* (2009) noted that in some cases the structural characteristics of a transcription factor provide some insight into its biological function, e.g., homeodomain transcription factors are often involved in developmental processes. Classifications are sometimes based on the tissue or tissue in which the transcription

factor is expressed. It is important to note that transcription factors undergo extensive protein–protein interactions and in some cases combinations of transcription factors determine regulation.

Vaquerizas *et al.* (2009) classified transcription factors into 23 families and added a 24 "other" category for undefined transcription factors. Zinc finger transcription factors constituted the most predominant class, approximately 680 of 700, followed by homeodomain factors approximately 250 of 700 and the helix-loop-helix transcription factors 80 of 700. They also classified transcription factors according to biological function based on literature reports. Of 741 factors thus analyzed 263 were involved in developmental processes, 221 in cellular processes, 109 in metabolic processes, 66 in responses to stimuli, 30 in immune processes, 28 in reproductive processes and 24 in localization.

Vaquerizas *et al.* (2009) reported that results of expression studies revealed that approximately one third of transcription factors were expressed primarily in one tissue. These included the heart specific transcription factor NKX2-1, and the fetal brain expressed transcription factor MYCN. They reported that the central nervous system compartment, including whole brain, spinal cord and fetal brains had seven transcription factors in common. These included the thyroid hormone alpha-receptor (THRA) and the aryl-hydrocarbon receptor ARNT2.

They noted that the serum response factor (SRF) is an example of a universally expressed transcription factor that is involved in multiple processes and that frequently combines with other factors to exert its effects. Protein–protein interactions are key to co-operative interactions between transcription factors.

Chromosomal distribution of transcription factor loci

Vaquerizas *et al.* (2009) analyzed the chromosomal distribution of human transcription factor loci. They reported that 20% of these loci mapped to 23 high-density clusters. The gene loci that encode HOX transcription factors (homeodomain containing proteins) map in specific clusters and other clusters contain zinc finger transcription factors. The short and long arms of human chromosome 19 are locations of particularly striking clusters of transcription factor loci. These authors proposed that repetitive

tandem duplication events led to generation of the series of paralogous, i.e., related genes. A second type of cluster occurred in the region of centromeres and telomeres. Genes in these clusters are non-paralogous. Vacquerizas *et al.* (2009) proposed that these clusters resulted from the intense genomic shuffling that occurs in these regions and genes in these clusters.

In a report on a specific transcription factor database TFClass, Wingender *et al.* (2012) provided classification of 1,558 human transcription factors and noted that if different isoforms are taken into account there are at least 2,900 different transcription factors. They noted that these factors regulate gene expression through binding to sequence elements in promoters, enhancers, silencers and other regulatory elements and emphasized that co-occurring cis-regulatory elements form comprehensive regulatory modules.

Although specific protein DNA recognition codes have been discovered for several classes, there remain classes of unclassified transcription factors. Taxonomies of DNA binding characteristics have evolved over the years and expression analysis, co-precipitation assays and bioinformatics capacities have greatly expanded. Wingender *et al.* (2012) ranked transcription factors into super-classes, classes, families and sub-families; they also included gene and peptide information. The super-class designation referred to the general topology of the DNA binding region, e.g., zinc finger transcription factors. Class designation included structural characteristics, e.g., presence of four zinc fingers. Family information included factors with structural and functional similarities.

The five most abundant sub-class categories of transcription factors defined in the classification of Wingender *et al.* (2012) include 52% with a zinc co-coordinating domain, 27% with a helix-turn-helix domain, 11% with a basic domain, 4% with an immunoglobulin type fold and 3% with other alpha-helical DNA binding domains.

Identification of transcription factor binding sites in DNA

One method for identifying DNA sites where transcription factors are bound is the CHIP Seq method. In this procedure formaldehyde is used to facilitate formation of bonds between DNA and bound protein. Following

fragmentation of chromatin, antibodies against specific transcription factors are used to isolate chromatin fragments that have the specific transcription factor bound and therefore react with the corresponding antibody. Following this the antibody bound fragments are heated at 65°C or they are treated with proteinase to release the protein. The DNA fragments released may be captured on to microarrays or they may be bead captured and sequenced.

Nuclear receptors

These proteins act as transcription factors after they bind ligands such as steroid or thyroid hormones or other specific fat-soluble ligands. Ligand binding induces specific structural changes in the nuclear receptor proteins and these structural changes are necessary to activate the transcription factor activity (Mangelsdorf *et al.*, 1995).

Nuclear receptor proteins contain DNA binding domains and ligand binding domains. Ligand bound nuclear receptor proteins interact with other co-regulatory proteins. The co-regulatory proteins may be involved in chromatin remodeling and histone modifications.

Nuclear receptors play important roles during embryonic development and in later life.

Cis-regulatory elements that impact transcription

Cis-regulatory elements include enhancers, silencers, promoters and insulators.

Enhancers

Buecker and Wysocka (2012) reviewed the role of enhancers in gene regulation. They proposed that enhancers act as information integration hubs that facilitate precise spatiotemporal gene expression during embryogenesis.

Enhancers may be active or poised. The key activity of enhancers is the ability to drive gene expression at a distance. They are uncoupled from promoters and from transcription start sites. Enhancers respond to signaling cascades and can drive gene expression in multiple tissues.