6 The Age of Aquarius

While the nature of gene regulation in higher organisms was being mooted, hamstrung by lack of molecular data, elsewhere a technological revolution was taking place, which would provide the toolkit to determine the repertoire of genes, their structures and products, and ultimately the composition of entire genomes.

The pace of molecular biology research and discovery was accelerating because of the expansion of the university and research sectors after World War II, especially in the 1970s when a new 'baby boomer' generation entered the workforce, eager to embrace new ways. The focus on bacterial molecular biology dissipated and there was a "mass migration in biomedicine" into "higher organisms" with the advent of gene cloning and the use of animal viruses to understand cancer.[1]

RECOMBINANT DNA AND 'GENE CLONING'

In 1972, Paul Berg, Stanley Cohen and Herbert Boyer[a] ushered in the biotechnology era by demonstrating that DNA could be cut and joined *in vitro* to make a 'recombinant' DNA molecule that could be propagated in a host cell to generate large numbers of copies by clonal amplification.

The roots of this advance lay in bacterial genetics, specifically in another strange phenomenon called 'host restriction-modification', discovered in the early 1950s by Salvador Luria, Mary Human, Giuseppe Bertani and Jean Weigle, whereby a bacteriophage infects a different strain of a bacterium (initially of *E. coli* and the closely related *Shigella dysenteriae*) orders of magnitude less efficiently than it infects the host strain from which it was derived, and reciprocally in reverse.[2–4]

A decade later, Werner Arber and Daisy Dussoix, and Mathew Meselson and colleagues, showed that this odd behavior was a manifestation of a bacterial defense system comprised of an enzyme (a 'restriction endonuclease') that cleaves foreign DNA at or near a specific nucleotide sequence, and a complementary enzyme that insulates the same sequences in the host genome, typically by methylation of one of the nucleotides. Most copies of invading viral genomes are cut by the nuclease, but a few become protected by the modification and escape cleavage. These survivors are then able to replicate efficiently, be insulated and infect the same bacterial strain with high efficiency, but not the original strain or others with different host restriction-modification systems.[5–13]

Different types of restriction endonucleases were defined:[14] Type I enzymes cleave DNA at a random distance, up to one kilobase, away from a complex recognition site;[b] Type II enzymes, of which there are several subtypes, recognize and cleave palindromic recognition sites (such as GGATCC), usually 4–8 base pairs in length, to produce either blunt or overhanging (complementary or 'sticky') ends by staggered cleavage; Type III enzymes recognize two separate non-palindromic sequences that are inversely oriented, and cleave 20–30 base pairs from the recognition site; and Type IV enzymes recognize and cleave specific modified, usually methylated, DNA sequences.[c]

The nomenclature follows a convention of an abbreviation of the name of the species and the order in which the enzyme was isolated, for example, *E. coli* strain RY13 enzyme 1, is called *Eco*R1. Thousands of different restriction enzymes are now known, including artificial enzymes produced by engineering, many of which are available commercially, fostering a molecular biology service industry to supply enzymes, cloning vectors and other tools (see below).

The enzymes studied by Arber and Meselson were Type I, which had limited utility. In 1970, however, Hamilton Smith and colleagues isolated the first Type II restriction enzyme (from *Haemophilus influenzae*, *Hin*dII), which enabled the reproducible cleavage of DNA molecules at specific sequences.[17,18] This

[a] Berg, Cohen and Boyer were all part of the south San Francisco UCSF/Stanford community, the epicenter of the early development of recombinant DNA technology.

[b] Type I restriction enzymes restrict the influx of foreign DNA via horizontal gene transfer while maintaining sequence-specific methylation of host DNA and have the ability to change sequence specificity by domain shuffling and rearrangements.[15,16]

[c] And more recently, Type V, whose cleavage sites are determined by guide RNAs (the CRISPR systems; Chapter 12).

DOI: 10.1201/9781003109242-6

was used by Kathleen Danna and Daniel Nathans to construct the first physical map of a genome, that of simian virus 40 (SV40), using size separation by gel electrophoresis of the resulting fragments, ushering in the era of 'restriction mapping'.[19]

It also brought about the era of recombinant DNA technology, as restriction fragments from different genomes could be mixed and joined together. This required another innovation – the purification and use of DNA 'ligases' capable of joining complementary or blunt DNA ends by the formation of a phosphodiester bond, isolated in 1967 by Bernard Weiss and Charles Richardson.[20] The first recombinant DNA molecule was produced by Paul Berg, Bob Symons and colleagues in 1972, who mixed *Eco*RI-cleaved SV40 DNA with a DNA segment containing lambda phage genes and the galactose operon of *E. coli*,[21] but fearing the dangers that might be created, declined to introduce these molecules into living cells.[22]

That was left to Cohen, who in 1973–1974, together with Annie Chang, Robert Helling, Boyer[d] and other colleagues, ligated *Eco*RI restriction fragments from *Staphylococcus aureus* and the frog *Xenopus laevis* with a plasmid that contained a replication origin, an antibiotic resistance gene (for selection) and was cleaved just once (i.e., linearized) by *Eco*RI. They reintroduced the recombined molecules into *E. coli* using a $CaCl_2$ ('transformation') procedure developed by Cohen.[24–26] These experiments showed that, to first approximation, genes could be successfully exchanged between species by human intervention.[e]

In 1977, Boyer's laboratory developed the first plasmid vector specifically designed for gene cloning, called pBR322, which was small, ~4 kb, and had two antibiotic resistance genes, one for selection of transformants and the other with unique restriction enzyme sites for DNA insertion to enable identification of recombinant plasmids (Figure 6.1).[29]

More sophisticated cloning vectors were developed, notably by Joachim Messing and colleagues from bacteriophage M13, containing multiple clustered sites (MCS) for restriction endonuclease cleavage, whereby double digestion prevents self-ligation and only allows re-circularization with a compatible insert.[31] Later versions contained an MCS within the *lacZ* gene, which allowed identification of colonies containing recombinant plasmids based on colorimetric detection of the encoded enzyme (beta-galactosidase) activity (blue colonies) or lack thereof (white, disrupted by an insert), and the ability to isolate single-stranded forms to aid DNA sequencing.[30,32] Cloning sites were also added into other genes that enabled direct (positive) selection of recombinant clones (e.g., [33]).

Many other variations and elaborations were then, and still are, being developed on the core requirements of a 'vector' (plasmid or virus) capable of being replicated in a desired host,[f] a selectable marker (usually an antibiotic resistance gene or metabolic enzyme to complement a deficiency in the host) to discriminate transformants from non-transformants,[g] a restriction ('cloning') site (or battery thereof) to insert foreign DNA, a means of favoring[h] or discriminating recombinant clones from those containing vector alone and a means of identifying the desired insert (the target gene to be cloned)[i] among the many others that may be produced from restriction endonuclease digestion of the input DNA.

Because the production of an encoded protein was an important scientific and commercial objective, many host-vector 'expression' systems were developed in the following decades to enable the high-level transcription, translation and purification of the encoded protein (see, e.g., [34,35]), often assembled in

[d] Boyer was well aware of the potential, having written a review on DNA restriction and modification systems the year before.[23]

[e] These advances led to the famous Asilomar Conference on Recombinant DNA technology in 1975, which "placed scientific research more into the public domain, and can be seen as applying a version of the "precautionary principle' via an initial voluntary moratorium and then strict controls on recombinant DNA construction and the release of genetically modified organisms into the environment", with "one felicitous outcome [being] the increased public interest in biomedical research and molecular genetics .. [and stimulation of] knowledgeable public discussion some of the social, political, and environmental issues that are and will be emerging from genetic medicine and the use of genetically modified plants in agriculture".[27,28] The participation of the public in the implications, applications, 'ethical' considerations and prescribed limits of genetic technologies was to be revived again with the later advent of techniques for precise engineering of animal and plant genomes (Chapter 12). Many genome research programs, notably those funded by Genome Canada, have required a proportion of the funding to be allocated to the social, ethical, economic, environmental and legal aspects of the work.

[f] That is, having an origin of replication that is recognized by the host cell.

[g] DNA transformation by $CaCl_2$ treatment of cells is inefficient, ~10^{-4} at best; more efficient methods were developed later.

[h] By using two different restriction enzymes so that the vector could not be re-joined without an insert.

[i] Cloning genes that would complement a deficiency in the host cell, usually bacteria or yeast, was relatively straightforward.

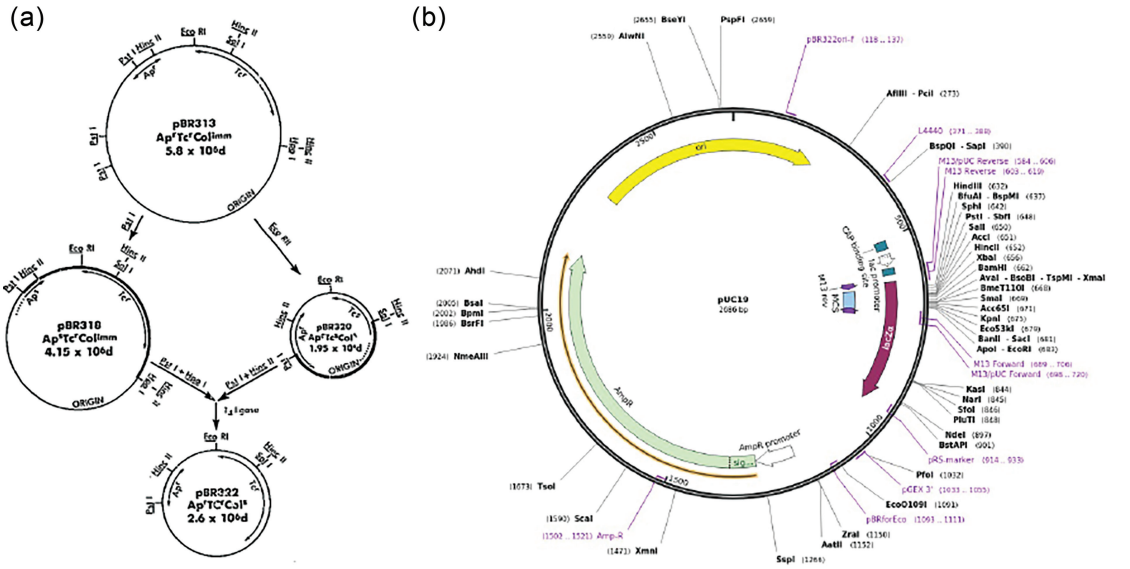

FIGURE 6.1 (a) The procedure used by Herb Boyer and colleagues to produce the plasmid cloning vector pPR322 containing unique restriction endonuclease sites with its two antibiotic resistance genes, such that recombinant plasmids may be identified by the insertional loss of one or the other. (Reproduced from Bolivar et al.[29] with permission of Elsevier.) (b) pBR322 was the precursor to more sophisticated cloning vectors such as pUC19[30] containing multiple clustered unique sites for restriction endonuclease cleavage and strategies to physically favor or genetically select for recombinant molecules. pUC19 image (https://www.addgene.org/50005/) generated by SnapGene software (snapgene.com). (Courtesy of Addgene.)

gene cloning protocol manuals that became ubiquitous in this period.[j]

ENABLING TECHNOLOGIES

The practical potential of the technology was realized immediately by Cohen and Boyer, who patented their method and started the first of the new generation of biotech companies, DNAX and Genentech, respectively, while Berg was a major proponent of strict regulation.

The initial targets were genes encoding medically important hormones, such as insulin, growth hormone and erythropoietin, among others. The difficulty was to identify the rare bacterial clone that contained the desired gene, a needle-in-a-haystack problem, especially when dealing with large genomes. This was approached via RNA, as it was reasoned that the tissues in which the proteins are highly expressed would be an enriched source of the corresponding (mRNA) coding sequences – a fortuitous approach given the (at that time) unknown problem that the protein-coding sequences of most genes are not contiguous in complex organisms (Chapter 7). To do this, additional technologies had to be developed.

The first was complementary DNA ('cDNA') synthesis – conversion of mRNAs to a DNA equivalent. This was achieved using reverse transcriptase, discovered a couple of years earlier by David Baltimore[42] and Howard Temin,[43] with an oligo dT primer that would anneal to the polyA tail of mRNAs to initiate synthesis of complementary DNA strands, then conversion of the single-stranded copies into double-stranded DNA with DNA polymerase.

The second was the synthesis of specific DNA (and later RNA) sequences, based on phosphonate, phosphodiester, phosphite triester and phosphoramidite anhydrous chemical synthesis methods pioneered in the 1950s, 1960s and 1970s by Alexander Todd,[44–46] Har Gobind Khorana,[47] Robert Letsinger[48,49] and Colin Reese,[50,51] among

[j] Two of the most popular have been 'Molecular Cloning: A Laboratory Manual'[36] and 'Current Protocols in Molecular Biology',[37] both regularly updated in new editions or volumes (see, e.g., [38–41]).

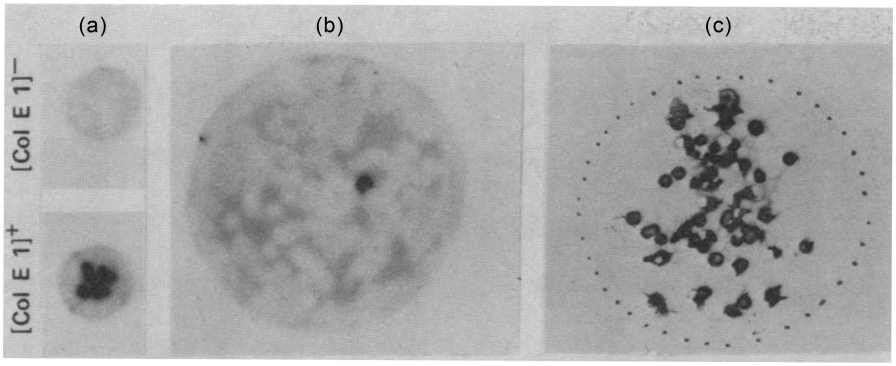

FIGURE 6.2 Autoradiographs of radiolabeled ColE1 plasmid RNA hybridized to lysed *E. coli* on nitrocellulose filters containing or lacking the plasmid at different ratios (B) 1:100 and (C) 1:1. (Reproduced from Grunstein and Hogness[70] with permission.)

others, and adapted to solid phase synthesis by Marvin Caruthers and colleagues in 1981.[52,53] These developments enabled the automation of oligonucleotide synthesis, as well as the incorporation of both natural and non-natural bases, novel linkages such as phosphorothioate or peptidyl bonds to improve biological stability or interaction strength, and other additions such as biotin for oligonucleotide capture (for reviews and recent developments see [51,54–57]), a vibrant domain of the biotechnology industry.

Synthetic designed oligonucleotides have become an indispensable part of the toolkit for molecular biology research and genetic engineering. Their uses encompass not only the detection of corresponding DNA or RNA sequences by hybridization, including highly parallel microarrays and bead arrays for target quantification and capture,[k] but also primers for DNA sequencing and amplification, introduction of restriction sites and mutations for genetic and protein engineering, construction of hybrid and other forms of artificial genes, mutagenic screening, production of antisense sequences to block gene expression, gene therapy,[58] large scale genome engineering[59] and many others. Much later enzymatic methods would be developed, using engineered terminal transferase,[60,61] which make possible the production of longer DNA sequences for synthetic biology and even the prospect of using DNA for data storage.[62]

The third technology was DNA, RNA and protein 'blotting', first developed by Ed Southern in 1975 using radioactively (and later biotin) labeled probes (usually cDNAs) to detect the location of corresponding genomic sequences in a restriction digest displayed by electrophoresis (the eponymous 'Southern blot'),[63] which played an important role in the discovery of 'genes-in-pieces' (Chapter 7). The RNA equivalent ('Northern blot') was developed by James Alwine, David Kemp and George Stark in 1977,[64] and the protein equivalent ('Western blot')[65] by Harry Towbin and colleagues in 1979 using labeled antibodies or other ligands to detect specific proteins in electrophoretic displays of cellular contents or fractions.[65–67] Subsequent variations were 'Southwestern blots'[68] and 'Northwestern blots'[69] to detect DNAs and RNAs bound by specific proteins, respectively.

These were early days, the technology was in its infancy and the cloning of specific genes was a major challenge, taking months and sometimes years. A typical strategy was to construct a cDNA 'library' from a tissue known to express the gene of interest, often involving size fractionation to enrich the desired mRNA (monitored by *in vitro* translation and Western blotting of the products), insertion of the cDNAs into a phage or plasmid vector and transformation into a bacterial host, usually *E. coli*, then screening for the desired clones among the tens of thousands of transformants by colony hybridization using radiolabeled oligonucleotide probes, developed by Michael Grunstein and Hogness in 1975[70] (Figure 6.2), commonly designed to be specific for a subsequence of the encoded protein with minimal codon redundancy. Those involved at the time can attest to the considerable celebrations that followed the successful cloning of a desired gene.

[k] Including sequence capture for 'exome' sequencing (Chapter 11) and targeted RNA sequencing (Chapter 13).

The revolutionary advance was that individual genes and genomic segments could now be isolated, amplified and characterized.

DNA SEQUENCING

The other technology was DNA sequencing, required to verify the identity and understand the details of the cloned gene. The first methods were developed in the late 1960s by George Brownlee, Fred Sanger and Bart Barrell, who used a paper fractionation method to sequence the 120nt 5S rRNA from *E. coli*,[71,72] and by Ray Wu and colleagues who used a primer extension approach (copying the sequence *in vitro* using DNA polymerase) to sequence the ends of phage lambda.[73–75] The first complete gene (encoding the MS2 RNA phage coat protein) and complete genome sequences (of phage MS2) were in fact RNA sequences, achieved by Walter Fiers and colleagues in 1972[76] and 1976,[77] respectively, using two-dimensional electrophoresis after partial nuclease digestion of the phage RNA.

In 1977, two new and more generalizable methods for DNA sequencing were published, made possible and widely applicable by the large amounts of cloned DNA. The first, by Gilbert and Allan Maxam, used terminal radiolabeling followed by base-specific partial chemical cleavage and size separation of the resulting set of fragments by (one-dimensional) electrophoresis, visualized by radiography.[78]

The second, developed by Sanger[1] and colleagues, extended Wu's primer extension method to produce the first sequence of a DNA genome (that of bacteriophage φX174) using chain terminating dideoxynucleotide analogs, terminal radiolabeling and size separation of the resulting set of fragments by electrophoresis.[79,80] Sanger sequencing (as it became known) quickly overtook the Maxam-Gilbert cleavage method, as it was easier to implement and more scalable (Figure 6.3).

Incremental technical improvements were made, which increased the length of the sequence reads. The next big leap forward was the introduction of fluorescently labeled primers by Leroy Hood and colleagues in 1986[81] and chain terminators by James Prober and colleagues in 1987,[82] which led to the development of the first automated DNA sequencer by Lloyd Smith in the same year,[83] using a repertoire of labels that allowed all four base-specific chain termination events to be identified in single reaction and read by continuous electrophoresis past a photodetector, with the data directly analyzed by a linked computer (Figure 6.4). The later development of 'sequencing by synthesis' (SBS)[84] using reversible chain terminators in high density on solid phase surfaces, resulted in another step change in the volume of data and a reciprocal decrease in cost and enabled the industrialization and massive parallelization of DNA sequencing that led to the genome projects at the turn of the century and ultimately to the feasibility of personal genome sequencing for precision healthcare[85] (Chapters 10 and 11).

THE GOLD RUSHES

These new technologies led to a stampede in the late 1970s and following years to clone and sequence genes or cDNAs encoding proteins of interest from bacteria, archaea, fungi, plants and animals, the ease of which depended on the availability of suitable genetic complementation (for genes in microorganisms) or tissues in multicellular organisms where the gene was highly expressed. For the latter reason, the first vertebrate cDNAs to be isolated were those encoding hemoglobin,[87] immunoglobulins, the chicken egg protein ovalbumin, and highly expressed muscle and milk proteins. They were followed by many others as the technology developed and became adopted across a wider spectrum of biological and biomedical disciplines, not just biochemistry and genetics, but also botany, zoology, microbiology, developmental biology, physiology, pharmacology, cell biology, pathology, anthropology, evolutionary biology, cancer biology, etc.

Importantly, molecular biology connected plant and animal developmental and behavioral genetics with biochemistry. Many of the genes that affect phenotype in model organisms and others in other species were cloned and sequenced, leading to an explosion of discovery and characterization of whole new families of proteins involved in body plan specification, cell differentiation and cell biology. Many of these genes turned out to be similar from yeast and invertebrates to humans.

[1] Sanger was one of the few people to be awarded two science Nobel Prizes in the same category (Chemistry), for protein sequencing and DNA sequencing, the other being John Bardeen (Physics) for developing the transistor and superconductor theory. The only other dual winners were Marie Skłodowska-Curie (Physics and Chemistry) and Linus Pauling (Chemistry and Peace).

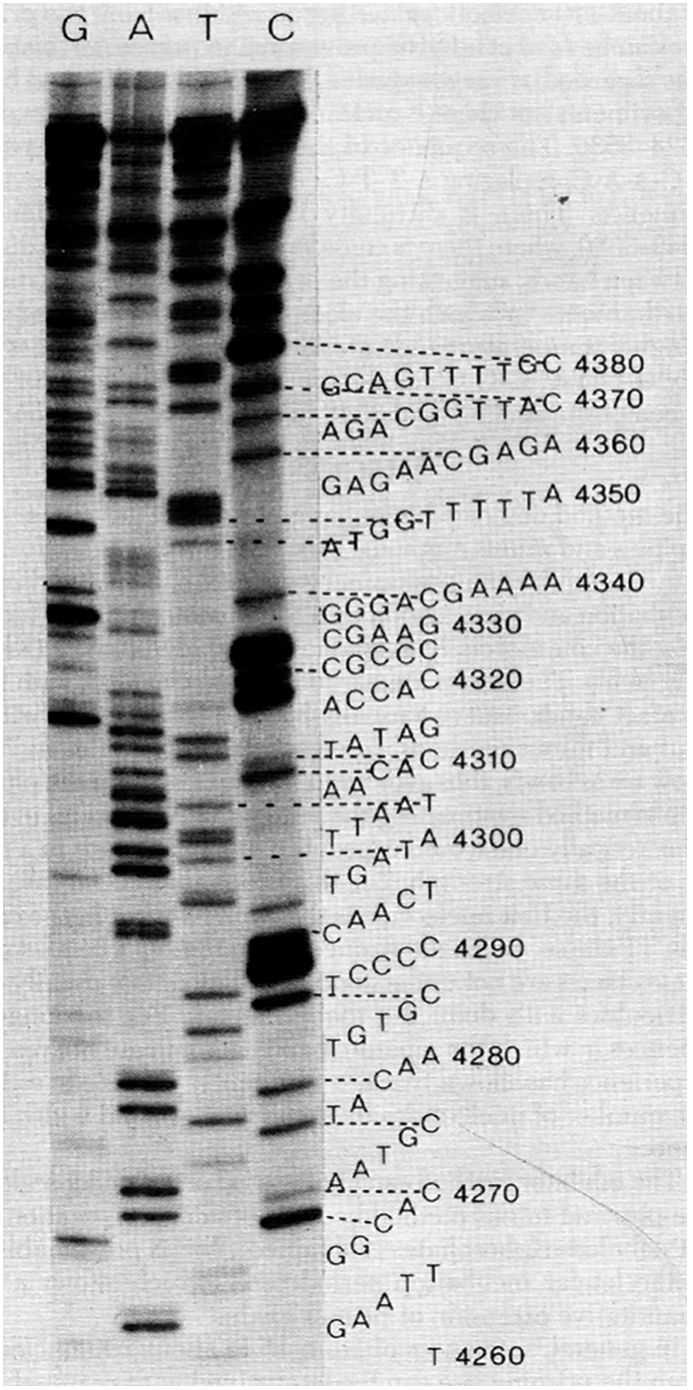

FIGURE 6.3 One of the autoradiographs presented by Sanger and colleagues in their 1977 paper on DNA sequencing,[80] showing electrophoretic size separation of X174 DNA sequences copied *in vitro* from specific primers using radiolabeled nucleotide triphosphates, with each of the four separate reactions containing specific A, G, C or T chain terminating nucleotide analogs (ddATP, ddGTP, ddCTP and ddTTP). The nucleotide sequence is read bottom to top (5'>3') from the ascending fragments in the different tracks. Later refinements optimized the reaction conditions, including the ratios of ddNTPs to dNTPs and the use of radiolabeled primers to yield even labeling. (Reproduced with author permission from Sanger et al.[80])

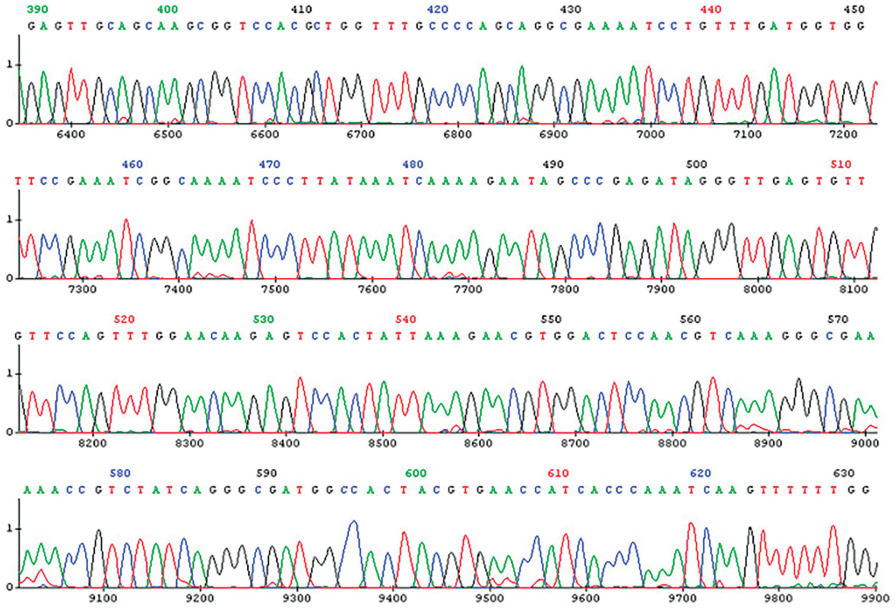

FIGURE 6.4 An example of the output of an automated fluorescent DNA sequencer. (Reproduced from Foret et al.[86] with permission of Elsevier.)

There was scientific gold to be unearthed everywhere. Investigators built their careers on the discovery of an important gene and study of its associated biology, all the better if the encoded protein may have medical significance.[m] This in turn reinforced the orthodox view of genetic information and fostered a generation of molecular biologists occupied with the "brutal reductionism" of identifying and characterizing genes encoding proteins.[92]

Nonetheless, there were wonderful discoveries in the decades that led up to the genome sequencing projects at the turn of the century. One could – as many have – fill a book on these alone: genes controlling cell division or enabling host colonization by bacteria; genes controlling flowering in plants; genes encoding molecular machines, all the way from ribosomal proteins to chloroplasts to flagella and muscle fibers; genes forming the cytoskeleton, and those encoding histones and histone modifiers, etc.

The avalanche of protein sequences (deduced from cloned genes) also led to the recognition of similar functional modules in different proteins, such as protease, phosphorylation, methylation, nucleotide binding and DNA binding domains, nuclear and mitochondrial localization signals, secretion signals, etc., information about which is now housed in databases such as Pfam.[93,94]

HOX GENES

The cloning in the 1980s by Walter Gehring and others of the genes in the Drosophila *bithorax* complex - studied by Lewis, Hogness[n] and others - identified the 'homeotic' proteins and the core 'homeobox' domain,[98–100] as well as many others identified by genetic screens, which enabled their expression patterns during development to be monitored.

[m] This is a source of unconscious bias by investigators and may explain in part why many biomedical studies have proven difficult to reproduce.[88–91]

[n] Anticipating the first successful cloning of eukaryotic DNA, in 1972 Hogness proposed using large insert clones to enable the detailed study of chromosome structure. His laboratory generated the first random clones from any organism in 1974, mapped a cloned DNA segment to a specific chromosomal location a few months later, and by early 1975, had generated clone libraries encompassing the entire *Drosophila* genome. In the late 1970s and early 1980s, Hogness, Lewis, Wellcome Bender and colleagues achieved the first 'positional cloning' of a gene, *Utrabithorax* (*Ubx*), and then others, using chromosomal 'walking' and 'jumping' aided by inversions. Many of the mutant alleles in the loci studied turned out to be the result of chromosomal breakage or transposon insertions rather than alterations to protein-coding sequences,[95–97] contrasting with the spectrum of chemical mutagen-induced single base changes used widely in mammalian genetic studies.

Unexpectedly, work from Mike Akam and colleagues revealed that the regulatory loci in the *bithorax* complex did not encode proteins, but rather expressed non-protein-coding RNAs,[101–103] but these were overlooked in favor of the homeotic proteins and the preconception that regulatory regions functioned in *cis* by binding regulatory proteins.

There was also a strong emphasis on finding equivalents of genes identified in model organisms (including, for example, neurological and transporter proteins) in other species by sequence homology, using initially a Southern blot variation dubbed 'zoo blots' and later sequence similarity.[99,104–106] Such approaches led to the discovery that not only do homeotic gene clusters occur in vertebrates in multiple copies but also that their introns (Chapter 7), relative orientation and temporal expression patterns (including antisense transcripts) are conserved between *Drosophila* and mammals[107–111](Figure 6.5).

Many other genes involved in *Drosophila* development were found to have human orthologs, a great surprise at the time, including that encoding the homeobox-containing protein Pax6, which is required for eye morphogenesis in both insects and mammals, indicating a common evolutionary origin[o] despite the differences between compound and camera-type eyes.[106] They also included many mutated in human diseases, like the homolog of the fly gene *patched* in the etiology of the skin cancer basal cell carcinoma[p] and the brain tumor medulloblastoma.[112,113]

ONCOGENES AND TUMOR SUPPRESSORS

Many genes that play a role in the etiology of cancer[q] – 'oncogenes' (in which mutations and activation drive cancer) and 'tumor suppressors' (whose inactivation facilitates cancer) – were unearthed in those years and are still being identified today.

The first oncogene was identified by a brilliant experiment which sought to understand how the Rous sarcoma[r] virus (RSV) transforms avian cells into a cancerous state. RSV was discovered in 1911 by Francis Peyton Rous, who showed that this retrovirus, from which Baltimore and Temin later isolated reverse transcriptase, was the infectious agent present in cell-free extracts of chicken tumors that could transmit cancer to other birds,[116] consistent with observations of others in leukemia[117] and sarcoma.[118] The surprising finding that cancer could be caused by a virus was, as so often the case, not believed and was "met with reactions ranging from indifference to skepticism and outright hostility",[119] although Rous ultimately received the Nobel Prize (55 years later) for his discovery. Analysis of viral mutants that could replicate but not 'transform' cells in culture led to the isolation of the *v-src* gene, encoding a protein tyrosine kinase (which phosphorylates other proteins). Just as importantly, *v-src* was shown to be a constitutively activated version of a normal human gene, the 'proto-oncogene' *c-SRC*, which is mutated in many cancers.[119,120] Subsequent studies, initially of Burkitt's Lymphoma in the early 1980s, showed that somatic chromosomal translocations involving the *c-myc* gene could create oncogenic hybrids,[121] which also occurs in the *bcr-abl* fusion characteristic of chronic myeloid leukemia.[122]

In 1969, Henry Harris showed that fusion of normal cells with tumor cells suppressed their tumorigenicity, indicating that cells express genes that control cell growth, which are lost in cancers. In 1971, Alfred Knudson and others studying rare cases of familial retinoblastoma hypothesized that the heritability was due to a loss-of-function mutation in one copy of a germline gene, followed by a later *de novo* (somatic) mutation in the other allele: the 'two-hit hypothesis'.[123] Nearly 15 years later this led to the identification of the first tumor suppressor gene, encoding the retinoblastoma tumor suppressor protein, *RB1*.[124,125] The two-hit hypothesis explained the relationship between inherited and acquired mutations in cancer predisposition genes, including *TP53*, also referred to as 'the guardian of the genome', which was discovered in 1979 by several groups and is mutated in about 50% of all cancers.[126,127]

[o] The evolution of the eye has been a popular and controversial topic in evolutionary biology and often cited as an example of 'intelligent design'.

[p] Tracked down by what is termed 'positional cloning' (also referred to as 'forward genetics'), whereby genetic and physical mapping techniques are combined to home in on the chromosomal locations of mutations causing serious genetic disorders (Chapter 11).

[q] Cancer is fundamentally a life-threatening disease of metazoans, resulting from the reversion of individual cells in complex differentiated organisms to a primitive, atavistic state,[114] wherein mutations disrupt the interactions between ancestral genes that promote cellular growth and those that control cell division and differentiation in multicellular development[115] (Chapter 15).

[r] Sarcoma is the collective name given to cancers of connective tissue.

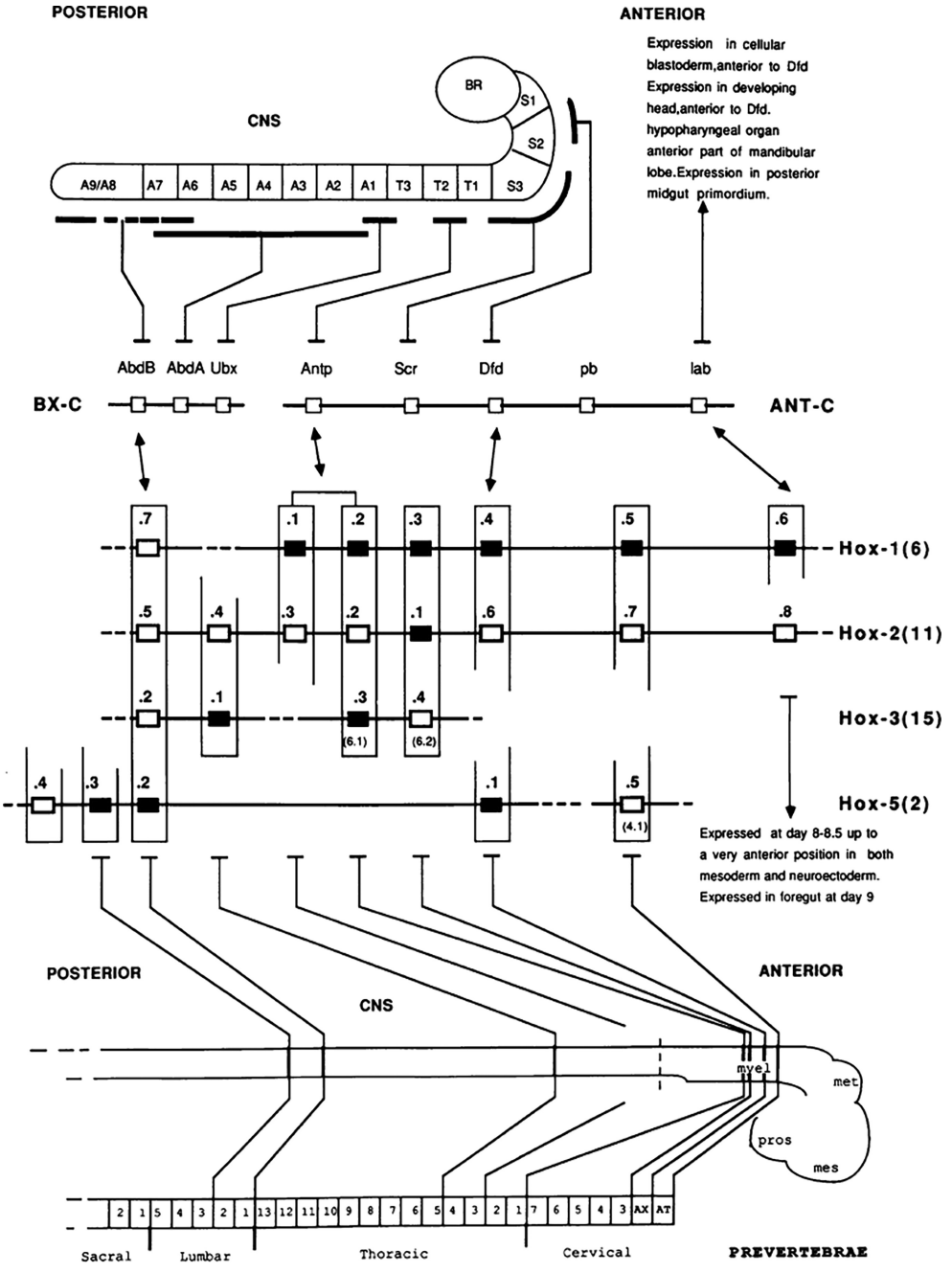

FIGURE 6.5 Schematic representation of the correlation between the *Drosophila* homeotic gene complexes and the murine (vertebrate) *Hox* gene network. (Reproduced from Duboule and Dollé[107] with permission from John Wiley and Sons.) The upper part represents the domains of expression of *Drosophila* homeotic genes in the embryonic central nervous system (CNS). In the central part all the genes belonging to the same subfamily are indicated by the vertical open or closed rectangles, the latter being the *Hox* loci that had been studied by comparative *in situ* hybridization experiments and whose expression domains had been defined at that time. The bottom part schematically represents the antero-posterior boundaries of expression of these genes along the fetal CNS and pre-vertebral column.

Such experiments resolved the controversy about the common causes of cancer – it is due to mutations that result in the ectopic activation of genes that promote cell division[121] and the loss of function of other genes that constrain cell division and migration.[128] These mutations can be inherited, occur spontaneously or be induced by DNA-damaging carcinogens and radiation, a complex landscape that is still being mapped by sequencing of tumor genomes in thousands of human cancers (Chapter 11).

Driven by medical need, the most intensely studied genes in the human genome[s] are those involved in cancer and other diseases.[134] Many altered cancer-causing genes, such as the breast cancer predisposition gene*s* *BRCA1* and *BRCA2*, maintain genome integrity,[135,136] with mutations leading to the chaotic genomes seen in many cancers. Other genes, such as the mismatch repair genes associated with Lynch syndrome,[t] cause a high tumor mutation burden, creating 'neo-antigens'[138] that can be recognized by the immune system.[138–142]

IMMUNOLOGY AND MONOCLONAL ANTIBODIES

Gene cloning also provided molecular insight into the previously arcane world of immunology, showing that antibody genes undergo rearrangement and hypermutation to generate a wide arsenal of antigen-recognition molecules. It was found that cells that express antibodies to foreign antigens undergo secondary changes to improve the binding of the antibody, as well as clonal selection,[143,144] as had been predicted by Macfarlane Burnett.[145] It also led to the identification of inflammatory molecules ('cytokines') that excite immune responses and drive autoimmune disorders,[146] as well as tangentially to the development and production of mouse monoclonal antibodies in culture, pioneered by Georges Köhler and César Milstein in 1975[147] (Figure 6.6) and later humanized by Greg Winter and colleagues,[148] which have proved so efficacious as therapeutics for cancer and autoimmune diseases, among other conditions and applications.

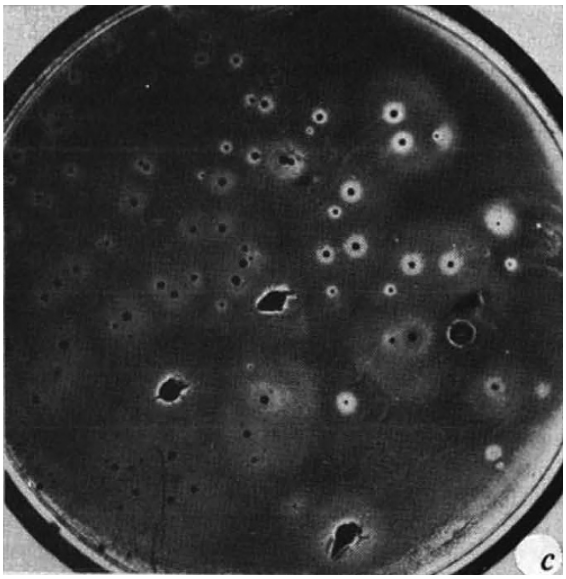

FIGURE 6.6 Photograph of immortalized cells created by a fusion between a myeloma cell line and spleen cells from a mouse immunized with sheep red blood cells, showing individual clones that secrete 'monoclonal' antibodies that lyse sheep red cells, indicated by the halos. (Reproduced from Köhler and Milstein[147] with permission from Springer Nature.)

BIOTECHNOLOGICAL EXPLOITATION

Not only did the gene cloning revolution create a vibrant technology support industry, it also transformed the pharmaceutical industry. The cloning, engineering and high-level expression of genes encoding human hormones such as insulin (which had previously been isolated from pig and cattle pancreas), erythropoietin (used to stimulate red cell production after bone marrow transplantation[u]) and growth hormone, among others, spawned multibillion-dollar products and companies, such as Genentech and Amgen.

Valuable tools were also developed by gene cloning and manipulation, notably the green fluorescent protein from jellyfish and its variants with different emission wavelengths by Osamu Shimomura, Douglas Prasher, Martin Chalfie and Roger Tsien,[149]

[s] Second only to TP53, and the most popular non-human gene, is the mouse *Rosa26*, which was identified in 1991 by Philippe Soriano and Glenn Friedrich as a locus that is ubiquitously active in mammals,[129] and subsequently widely used for the construction of transgenic mice and other species, as well as transgenic human cells.[130–132] The *Rosa* locus encodes two overlapping non-protein-coding RNAs,[133] whose functions are presently unknown.

[t] Which result in dinucleotide repeat instability.[137]

[u] And used illegally by athletes, as is growth hormone, to boost oxygen carrying capacity and muscle mass, respectively.

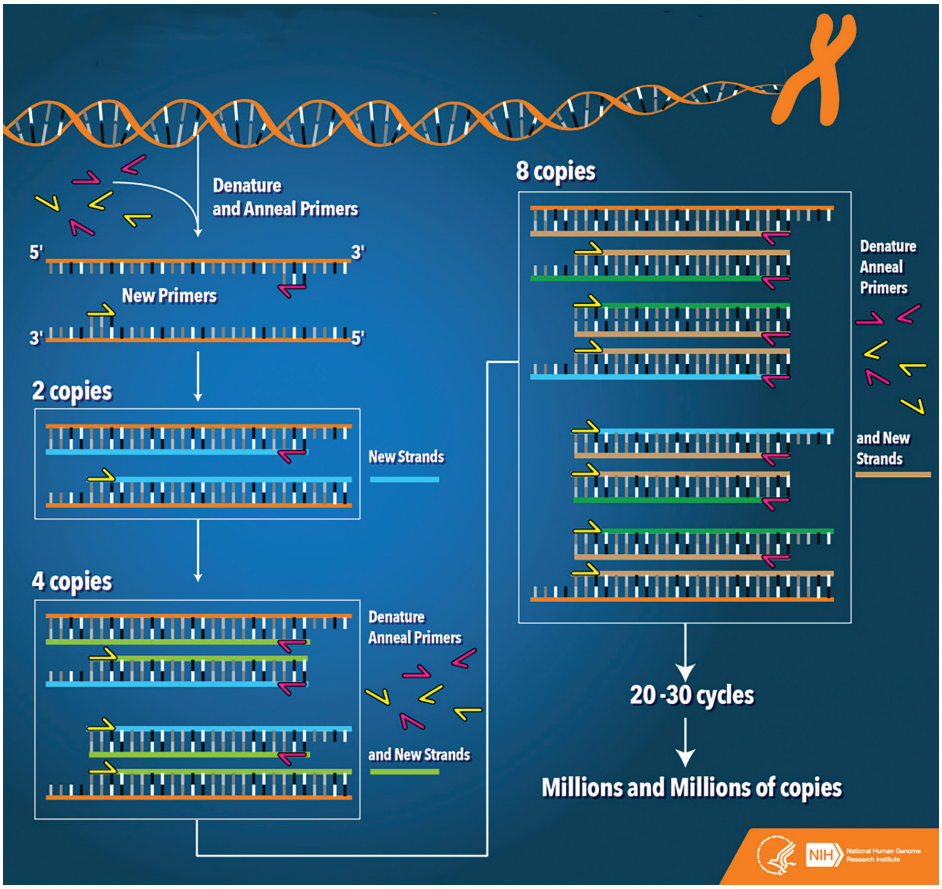

FIGURE 6.7 The principle of exponential amplification of targeted DNA sequences by PCR. Image modified from the US National Human Genome Research Institute fact sheet.[157]

and firefly luciferase by Marlene DeLuca and colleagues,[150,151] which have been widely used in cell and developmental biology to track the expression of genes fused to these visual 'reporters'.

Many of the genes discovered and characterized during this period were patented for medical or industrial use, largely on the basis of the inventiveness of the technology and the novel uses claimed for the products of these genes, a practice that was later circumscribed.[152] However, and despite criticism of gene patenting, it had the beneficial effect of allowing the development of new pharmaceuticals, which require (limited) monopoly rights to protect and recover the required massive investments in clinical safety and efficacy trials, following the tragic teratogenic effects of the anti-nausea drug thalidomide on limb development in embryos.[153]

CELL-FREE DNA AMPLIFICATION AND SHOTGUN CLONING

In 1983, Kary Mullis conceived a brilliantly simple strategy to amplify defined segments of DNA *in vitro*, using flanking oligonucleotide primers and DNA polymerases for cyclic, exponential replication of the targeted sequence, termed 'Polymerase Chain Reaction' or PCR[154] (Figure 6.7). The crucial technical advance was the use of thermostable DNA polymerases isolated from thermophilic archaea,[155] originally identified by Thomas Brock in hot springs in Yellowstone National Park in 1964.[156] PCR allowed ultra-sensitive detection and amplification of known DNA segments and transformed gene cloning, genetic engineering and diagnostic assays for mutations and infectious agents, especially viruses.

However, most genes that had been cloned in those days encoded proteins that had been identified biochemically or genetically. There were many more, as Craig Venter, Mark Adams and colleagues demonstrated in 1992 when they introduced the concept of 'shotgun' mRNA cloning and sequencing ('expressed sequence tags') to double the number of known human proteins in a single publication.[158] There was also great controversy when the US National Institutes of Health attempted to patent these genes *en masse*.[159] Similar agnostic approaches identified thousands of new genes in other organisms including plants.[160] Importantly, the shotgun strategy along with advances in the technology for high-throughput DNA sequencing allowed gene discovery on an industrial scale[161–164] and set the foundations for the genome projects (Chapter 10).

A WORLD OF PROTEINS

Throughout this period gene cloning and characterization was almost exclusively focused on protein-coding sequences,[v] due to a number of intrinsic and mutually reinforcing biases: expectational bias that most genes encode proteins; perceptual bias due to the strong phenotypes of disabling protein-coding mutations that are readily observed and genomically mapped; a sampling bias, as protein-coding genes are generally highly expressed;[w] technical bias due to the use of oligo(dT) priming of cDNA synthesis, which favors mRNAs; the difficulty of sequencing vast tracts of non-protein-coding DNA, and reticence to do so; and the problem of identifying causative mutations among the many variations in introns and 'intergenic' sequences.

The concept of a gene became synonymous with 'open reading frames', reinforcing the presumed equivalence of gene and protein, which in turn had a major influence on the interpretation of the discoveries of the mosaic structure of eukaryotic genes and the vast tracts of non-protein-coding sequences in animal and plant genomes (Chapter 7).

As observed by Ed Rubin and Lewis:

> Ironically, the success in cloning and studying individual genes dampened enthusiasm for an organized genome project, which was seen as unnecessary. Over 1300 genetically characterized genes—nearly 10% of all the genes in Drosophila—have been cloned and sequenced by individual labs. This is over twice the percentage of genes in any other animal for which both the loss-of-function phenotype and sequence have been determined. Nevertheless, for flies as well as other animals, less than a third of genes have obvious phenotypes when mutated, emphasizing the critical importance of genome sequencing as a gene discovery method.[95]

A very large fraction of discovered proteins in all kingdoms of life have no known function.[172]

On the other hand, not only did the genetic and biochemical approaches used to identify and characterize proteins reveal surprising cases of regulatory RNAs (Chapters 8, 9, 12 and 13), genome sequencing and high-throughput assays later showed that the largest class of proteins in the human genome is RNA binding proteins.[173,174]

FURTHER READING

Cohen S.N. (2013) DNA cloning: A personal view after 40 years. *Proceedings of the National Academy of Sciences* 110: 15521.

Roberts R.J. (2005) How restriction enzymes became the workhorses of molecular biology. *Proceedings of the National Academy of Sciences* 102: 5905.

Rubin G.M. and Lewis E.B. (2000) A brief history of *Drosophila*'s contributions to genome research. *Science* 287: 2216–8.

Russo E. (2003) Special report: The birth of biotechnology. *Nature* 421: 456–7.

Watson J.D. and Tooze J. (1981) *The DNA Story. A Documentary History of Gene Cloning.* (W.H. Freeman and Company, New York).

[v] Some early work indicated that polymorphisms in the 5' region of human protein-coding genes are associated with variations in gene expression and disorders, such as hemoglobinopathies and hypertriglyceridemia.[165–168]

[w] In general, protein-coding genes are more highly and broadly expressed than genes that express regulatory RNAs, which show high cell specificity,[169–171] although there are exceptions (Chapter 13).