From Genes to Genomics to Proteomics

4

Michael R. Speicher

Abstract In human genetics many initial research initiatives focused on single genes or were performed on a gene-by-gene basis. However, recent findings, especially those about the extensive transcriptional activity of the genome, changed the concept of what a gene is supposed to be. In addition, novel high-throughput approaches and numerous innovative technologies, such as gene and expression microarrays, mass spectrometry, new sequencing methods, and many more, now enable us to address complex diseases and to unravel underlying involved regulatory patterns. These high-throughput assays resulted in a shift from studying Mendelian disorders towards multifactorial diseases, although monogenic diseases still provide a unique opportunity for elucidating gene function. This chapter describes current concepts about the definition of a gene, possible consequences of mutations and the latest developments in the areas of genomics, transcriptomics, and proteomics and their potential to add to a better understanding of factors contributing to phenotypic features.

Contents

M.R. Speicher (✉)
Institute of Human Genetics, Medical University of Graz,
Harrachgasse 21/8, 8010 Graz, Austria
e-mail: michael.speicher@medunigraz.at

M.R. Speicher et al. (eds.), *Vogel and Motulsky's Human Genetics: Problems and Approaches,*
DOI 10.1007/978-3-540-37654-5_4, © Springer-Verlag Berlin Heidelberg 2010

4.1 Single-Gene Approaches

Prior to the era of high-throughput analyses, typical research initiatives focused on single genes or were performed on a gene-by-gene basis. However, even research focusing on a single gene may already represent a very complex challenge. Some principles of working with single genes are described below. A particular focus will be on the limitations which have propelled the development of numerous innovative technologies, such as gene and expression microarrays, mass spectrometry, and proteomics, and many more, which now allow investigators to reveal underlying complex regulatory patterns.

4.1.1 *What Is a Gene?*

Before discussing the steps from genes to proteomics we should reflect on what a "gene" is actually supposed to be. In 1909 the term "gene" was used for the first time by Wilhelm Johannsen. Ever since, the concept of a gene has been under constant development, and numerous gene definitions have been proposed and adjusted as our knowledge of genes has evolved over the past decades. A somewhat surprising result is that although the term "gene" is one of the most commonly used expressions in genetics and although genes are constantly being characterized and more and more mutations in genes are being linked to diseases, the term itself in fact remains poorly defined. An excellent history of operational definitions of a gene over the past decades together with an attempt at an updated definition was recently provided by Gerstein et al. [25]. The authors rightfully argue that the provocative findings of the ENCODE Project [17], which elucidated the complexity of the RNA transcripts produced by the genome, have to change previous definitions of a gene. The preceding views of a gene were centered on protein coding (Fig. 4.1) and did not take the extensive transcriptional activity of the genome into account, most likely because the full extent of transcriptional activity was unknown prior to the ENCODE Project.

Based on the knowledge derived from the ENCODE Project, Gerstein et al. [25] proposed the following, updated definition for a gene: "The gene is a union of genomic sequences encoding a coherent set of potentially overlapping functional products." An illustration of how to apply this definition is provided in Fig. 4.2.

Another implication of this definition is that 5′ and 3′ untranslated regions (UTRs), despite their importance for translation, regulation, stability, and/or localization of mRNAs, would not be part of a gene because they do not participate in encoding the final product of a protein-coding gene. In order to compensate for this, Gerstein et al. [25] suggested a new "category" for regulatory and untranslated regions playing an important part in gene expression, by naming these regions "gene-associated." This terminology may help to acknowledge that additional DNA sequences outside

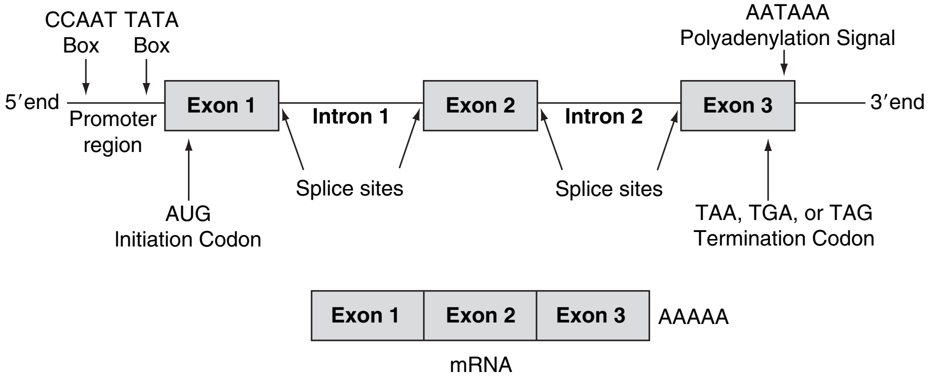


Fig. 4.1 Representative classic view of a gene. Transcription may be initiated from the promoter region located at the 5′ side of a gene. The promoter region often contains a TATA or a CCAAT box and is enriched for the paired nucleotides cytosine and guanine (CG islands). Genes consist of translated (exons) and noncoding (introns) portions. The open reading frame (*ORF*) is situated between the initiation codon (*AUG*) and the termination codon (*TAA*, *TGA*, or *TAG*). Sequences encoding the polyA tail of the protein are located at the end of a gene. The precursor RNA is spliced so that intronic sequences are removed and messenger RNA (mRNA) is formed

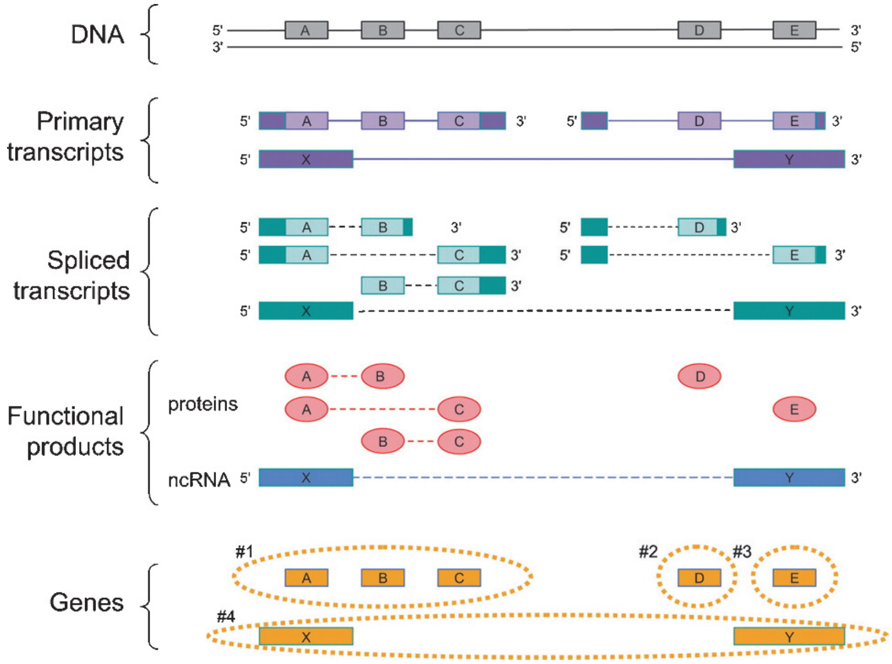


Fig. 4.2 Gerstein et al. [25] proposed a new definition for genes, and this figure illustrates how this definition should be applied. In this region the *gray rectangles* correspond to exonic/protein-coding sequences. Three primary transcripts originate from this genomic region. Two of these transcripts consist, in addition to the 5′ and 3′ un-translated regions, of some of the exons (*A*, *B*, *C* or *D*, and *E*); intronic sequences are represented by *solid lines*. The third transcript (*X* and *Y*) does not encode a protein but is a noncoding RNA (ncRNA) product. Therefore, such a transcript may share its genomic sequences with protein-coding segments; however, it usually does not exactly correspond to exons. The result of alternative splicing in this example is that the first two transcripts encode five protein products (*A-B*, *A-C*, *B-C*, *D*, and *E*; the *dashed lines* illustrate connectivity between the RNA sequences). Thus, exons A, B, and C generate transcripts, each derived from two of these DNA segments. In contrast, the products originating from D and E share a 5′ untranslated region, but their translated regions do not overlap after alternative splicing. The noncoding RNA product is not a coproduct of the protein-coding genes. The functional products are 5 different proteins shown in *ovals* (connected by *dashed lines*) and one RNA product (*rectangles*, also connected by a *dashed line*). As a consequence this region harbors four genes indicated at the bottom within the *orange dashed lines*. Sequence segments A, B, and C comprise gene 1, whereas gene 2 contains D, gene 3 E and gene 4 X and Y. From [25]

of the respective gene themselves have important roles in contributing to gene function.

From this new definition it follows that only continuous DNA sequences coding for a protein or RNA product without overlapping products correspond to the classic and most commonly used view of gene. In fact, the vast majority of our knowledge about "genes" and their functions centers on this subclass of genes. Thus, with these new evolving concepts it is obvious that even "monogenic" disorders are at present incompletely explored and a lot remains to be discovered.

As the updated definition emphasizes the final products of a gene, it disregards intermediate products originating from a genomic region that may happen to overlap. This implies that the number of genes in the human genome is going to increase significantly when the survey of the human transcriptome has been completed.

4.1.2 Mutations

The aforementioned summary of the complexity of a gene and its possible transcripts also suggests that the distinction between pathogenic and nonpathogenic mutations is often very difficult. In general, there are three different types of mutations. *Deletions* involve the loss of at least one nucleotide, whereas *insertions* represent the addition of at least one nucleotide. Both deletions and insertions cause a shift of the reading frame and are therefore also referred to as frameshift mutations. Usually the resulting sequences no longer code for a functional gene product and are thus dubbed "nonsense mutations." Since insertions and deletions usually disturb the gene function significantly, they are often associated with diseases and are therefore frequently pathogenic.

In contrast, a contribution to specific phenotypic features of the *substitution* or *exchange* of a single nucleotide is often very difficult to establish. An exchange of one purine for another purine or of one pyrimidine for another is called transition, whereas an exchange of a purine for a pyrimidine or vice versa is a transversion. A nucleotide substitution does not result in a shift of the reading frame, and possible consequences depend on how a codon has been altered. For example, a substitution may alter a codon so that a wrong amino acid will be present at this site, which is referred to as a "missense mutation." Such missense mutations may have consequences ranging from no changes to severe functional changes, and it is often very difficult to establish the outcome of such mutations. A nucleotide substitution is called a "silent mutation" if the resulting codon still corresponds to the same amino acid. This is possible because of the redundancy of the genetic code, as different nucleotide sequences may code for the same amino acid sequence. For example, the four nucleotide base pairs GCC, GCG, GCT and GCA all code for the amino acid alanine. If GCC represented a codon within an open reading frame a substitution at the third position from C to G or from C to A would still represent a codon with the nucleotide sequence for alanine. Much has been learnt about the phenotypic consequences of mutations, but there are many examples of missense mutations, variants in DNA elements of unknown function, and silent changes in coding regions for which pathogenicity is questionable. Thus, another difficult challenge is to prove that an altered allele is causal to the disease in question.

For example, silent mutations frequently have no consequences for the phenotype. However, in order to illustrate the often enormous difficulties in determining the significance of mutations, two striking examples demonstrating that even "silent" mutations may have severe consequences for a phenotype will be discussed below.

4.1.3 Silent Mutations and Phenotypic Consequences

Two particularly fascinating "silent" mutations with significant consequences for the phenotype are described here.

Hutchinson–Gilford progeria syndrome (HGPS) is a rare genetic disorder. Affected individuals show very early signs of aging, such as loss of hair, lipodystrophy, scleroderma, decreased joint mobility, osteolysis, and facial features resembling those of aged persons, and they die at an average age of 13. In the vast majority (90%), progressive atherosclerosis of the coronary and cerebrovascular arteries is the cause of death [30]. HGPS belongs to a group of conditions called laminopathies, which affect nuclear lamins. The lamins belong to the multiprotein family of intermediate filaments and can be regarded as the main determinants of the nuclear architecture. HGPS is caused by mutations in *LMNA*, resulting in an abnormally formed lamin A. In the majority of progeria patients a classic p.G608G (c.1824C>T) mutation in exon 11 can be found. It is predicted that this mutation is a silent mutation, as it does not cause any change at the amino acid level. However, this change improves the match to a consensus splice donor, activating a cryptic splice site [14, 18]. Owing to this activation of a cryptic splice site, 150 nucleotides, up to the start codon of exon 12, are removed [14, 18]. The last step in the posttranslational processing of prelamin A cannot occur without these nucleotides, so that the mutant prelamin A persists. The mutant prelamin A is called progerin, and it is the presence of progerin, and not the lack of normal lamin A, that causes the phenotype [55].

The second example is the identification of a synonymous *single-nucleotide polymorphism* (SNP), which did not produce altered coding sequences in the *Multidrug Resistance* 1 (*MDR*1) gene [35]. The *MDR*1 gene product is a P-glycoprotein multiple-transmembrane protein pump contributing to the pharmacokinetics of drugs, which is associated with the multidrug resistance of cancer cells. Although *MDR1* harbors many SNPs, some SNPs have been associated with reduced functionality of the pump. This was observed for two SNPs (e.g., C1236T and C3435T) even though neither changes the amino acid sequence of P-glycoprotein. For example, the C1236T polymorphism changes at amino acid position 412a GGC codon to GGT and both encode glycine, whereas the C3435T polymorphism changes at position 1145 ATC to ATT, which in each case encodes isoleucine. However, Kimchi-Sarfaty et al. [35] were able to demonstrate that both polymorphisms resulted in changes from frequent to infrequent codons. As a consequence, ribosome trafficking is slowed down at the corresponding mRNA regions. These alterations likely affect the cotranslational folding pathway of P-glycoprotein, resulting in a different final conformation and eventually in altered substrate specificity. Thus, silent mutations of synonymous codons (changing from frequent to infrequent) in certain genes may alter translation kinetics of mRNA, which might in turn affect final protein conformation.

4.1.4 Mutation Detection by Sanger Sequencing

In 1977 Fred Sanger published three seminal method papers on the rapid determination of DNA sequences [53, 54, 60], for which he received his second Nobel prize in Chemistry in 1980. This technology, which besides Sanger sequencing is also referred to as dideoxynucleotide sequencing, provided a tool for deciphering complete genes and later entire genomes. In fact, Sanger sequencing evolved into the only DNA sequencing method used for three decades after it was first described. DNA can be prepared for sequencing by two approaches: for targeted resequencing, which is done in most diagnostic routine applications: primers flanking the target regions are used to amplify the respective region. In contrast, for shotgun de novo sequencing, DNA is randomly fragmented and cloned into a plasmid, which is subsequently used to transform *Escherichia coli* (Fig. 4.3a). The latter approach played a pivotal role in deciphering the human genome. The

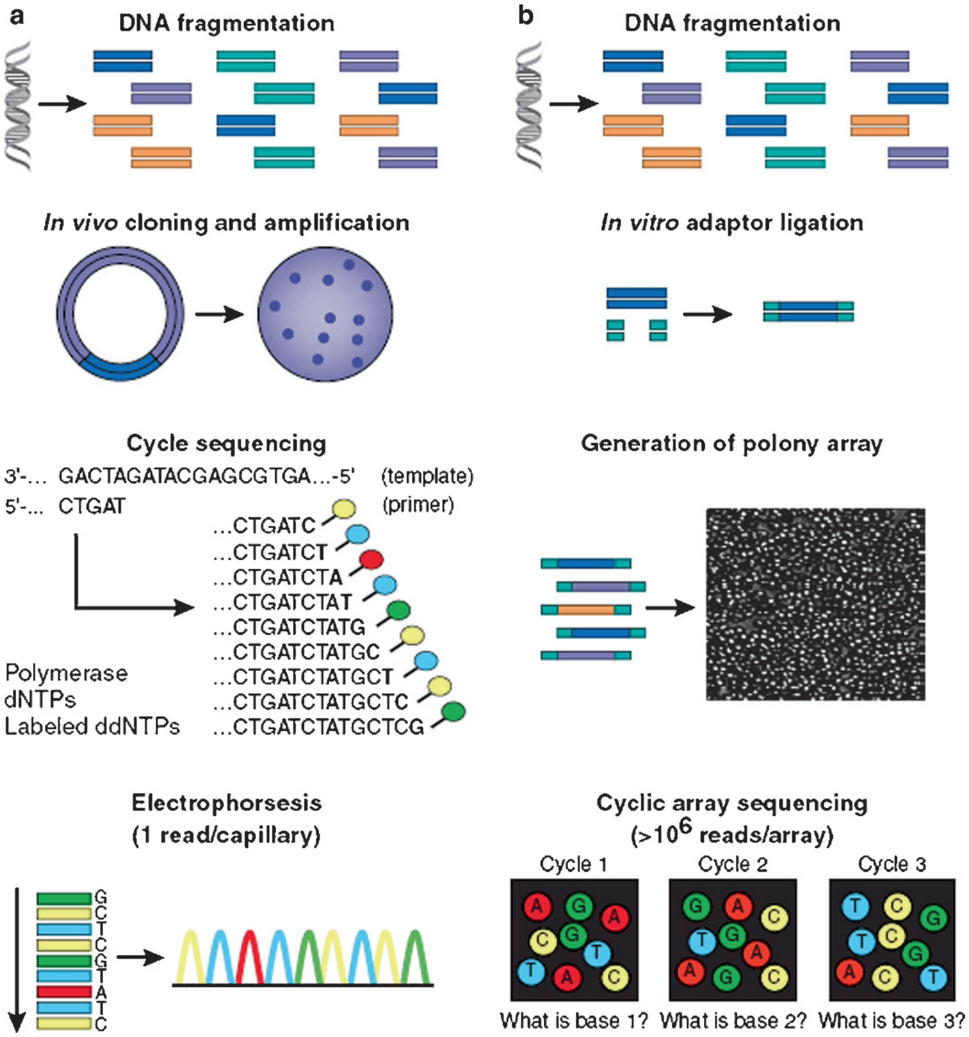


Fig. 4.3 Comparison between (**a**) Sanger and (**b**) next-generation sequencing. (Reprinted by permission from Macmillan Publishers Ltd: Nature Biotechnology [58], copyright 2008)

results of both approaches are multiple templates, which are then subjected to the sequencing reactions, consisting of repeated rounds of template denaturation, primer annealing, and primer extension. In each cycle of the sequencing reaction the primer extension is stochastically terminated by the integration of dideoxynucleotides (ddNTPs), which are labeled with a fluorochrome. This results in a mixture of extension products of different lengths, and the label of the respective terminating ddNTPs reflects the nucleotide identity of its terminal position. Subsequently the sequence can be determined by high-resolution electrophoretic separation of the single-stranded, end-labeled extension products in a capillary-based polymer gel. The DNA sequence is deciphered by analysis of the fluorescent labels at the end of the fragments. The discrete lengths of the fragments determine the nucleotide position, and the nucleotide itself is encoded by laser excitation of the fluorescent labels and a four-color detection of the emission spectra, which are then translated into DNA sequence by appropriate software (Fig. 4.3a).

The application of Sanger sequencing for deciphering the entire human genome represented particularly large-scale sequencing efforts, which were conducted in factory-like environments called sequencing centers. These centers had a specialized and dedicated infrastructure consisting of hundreds of DNA-sequencing instruments, robotics, bioinformatics, computer databases, instrumentation, and a large number of personnel. The aim of deciphering the entire human genome dramatically changed the throughput requirements of DNA sequencing and propelled developments such as automated capillary electrophoresis. Many capillary-based sequencing systems have 96 or more capillaries, meaning that 96 sequence reads can be processed in parallel. However, a simple increase in the number of capillaries was not sufficient for the new enduring tasks in genomics, which required the development of entirely new technologies, as summarized in the next paragraph.

4.1.5 Next-Generation Sequencing

The next-generation sequencing revolution started in 2005 with two seminal papers describing a sequence-by-synthesis technology [47] and a multiplex polony-sequencing protocol [59]. The parallel sequencing throughput capacity is perhaps the most important feature setting next-generation sequencers apart from conventional capillary-based sequencing. In fact, instead of running 96 capillaries or samples at a time, next-generation sequencing allows the processing of millions of sequence reads simultaneously (Fig. 4.3b). This massive parallel sequencing requires only one or two instruments instead of several hundred Sanger-type DNA capillary sequencers and naturally involves significantly fewer personnel operating the machines. Another important difference is that next-generation sequence reads do not depend on vector-based cloning, but are instead derived from fragment libraries. This alone allows a significant speeding up of sequencing (Fig. 4.4). Another difference is that read lengths are shorter (35–250 bp for next-generation sequencing, as against 650–800 bp for capillary sequencers). Next-generation sequencing, often also referred to as second-generation sequencing, and the evolving third-generation sequencing will be discussed in greater detail in Sect. 4.4.

4.1.6 The Importance of Monogenic Mendelian Disorders

The quest for high-throughput assays is also accomanied by a shift away from the Mendelian disorders towards multifactorial diseases. This neglects the fact that linking naturally occurring pathogenic mutations with monogenic disorders provides a unique opportunity for elucidating gene function [1]. Studies on Mendelian traits reveal irreplaceable insights into mutation processes and their associated molecular pathophysiology. Furthermore, it was the investigations of Mendelian disorders that disclosed the existence of genetic phenomena, such as uniparental disomy or parental imprinting.

Only in-depth analysis of monogenic disorders can unravel the consequences of different mutations within the same gene that can give rise to distinct phenotypes. For example, among the most striking examples are mutations in the aforementioned *LMNA* gene, which can cause not only the Hutchinson–Gilford progeria syndrome but also several other, different phenotypes, which are often summarized as primary laminopathies. Phenotypic consequences of mutations in *LMNA* can be further subdivided into laminopathies with striated muscular atrophy [including

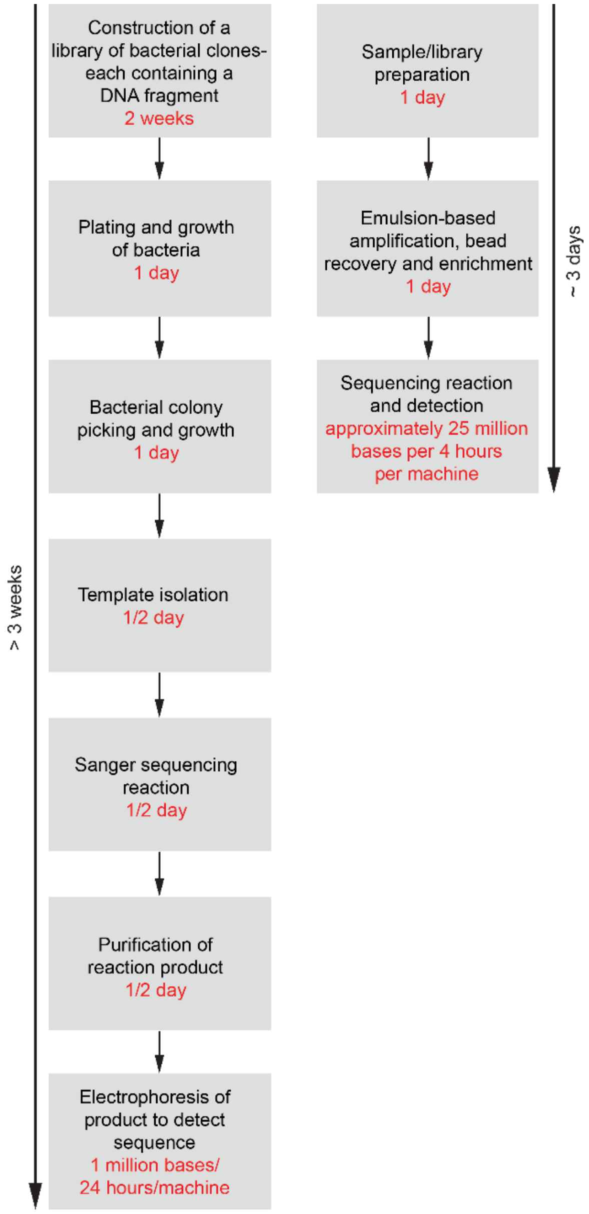

Construction of a library of bacterial clones-each containing a DNA fragment
2 weeks
Plating and growth of bacteria
1 day
Bacterial colony picking and growth
1 day
Template isolation
1/2 day
Sanger sequencing reaction
1/2 day
Purification of reaction product
1/2 day
Electrophoresis of product to detect sequence
1 million bases/ 24 hours/machine
> 3 weeks
Sample/library preparation
1 day
Emulsion-based amplification, bead recovery and enrichment
1 day
Sequencing reaction and detection
approximately 25 million bases per 4 hours per machine
~ 3 days

Fig. 4.4 Flow diagrams comparing the time-lapse needed for traditional Sanger sequencing (*left*) and massively parallel sequencing as used for the 454 system (*right*). (Reprinted by permission from Macmillan Publishers Ltd: Nature [52], copyright 2005)

Emery-Dreifuss muscular dystrophy (EDMD2; OMIM 181350), autosomal dominant limb girdle muscular dystrophy 1B (LGMD1B; OMIM 159001) and dilated cardiomyopathy, 1A (DCM1A; OMIM 115200)], laminopathies affecting peripheral nerves [(Charcot-Marie-Tooth disease type 2B1 (CMT2B1; OMIM 605588)] and laminopathies with loss of or reduced adipose tissues [familial partial lipodystrophy, Dunnigan type (FPLD2; OMIM 151660) and congenital generalized lipodystrophy, type 2 (CGL2, OMIM 269700)] [41]. A probably new *LMNA*-associated disease entity that may be classified as a congenital muscular dystrophy (LMNA-related congenital muscular dystrophy, or L-CMD) has recently been described [51] and suggests that even more phenotypes may be caused by mutations in this gene.

Even monogenic diseases have considerable phenotypic complexity, often depending on the genetic background and the status of modifier genes, which may modulate the consequences of specific mutations. In addition, such epigenetic changes as the genomic distribution of 5-methylcytosine DNA and histone acetylation may change the outcome of a mutation, and to make these issues even more complicated, such epigenetic modifications may change as we age [22]. Thus, monogenic disorders are in fact examples of oligogenic inheritance and vary along a continuum from simple to complex disorders [1]. Allelic variation in genes or other functional DNA sequences that modify the phenotypic severity of a monogenic disorder or control variation in gene expression provide links to additional genomic causes related to phenotypic variability.

4.2 Gene Regulation

Genes can be regulated by various means (Fig. 4.5). Obviously there is a "many-to-many" relationship between regulatory regions, epigenetic mechanisms, small RNAs, and genes. In fact, gene expression is a multilevel process, which is controlled by regulatory proteins and DNA sequences (genetic regulation) and by chromatin remodeling and the position of chromosomes in the nucleus (epigenetic regulation). In addition, gene regulation may be affected by complex sets of RNAs that do not produce proteins.

4.2.1 Genetic Regulation

At the beginning of transcription the base sequences of genes are transcribed into RNA by RNA polymerase II. Multiple accessory factors determine the transcriptional start and end points for RNA polymerase II. An important component is the promoter, typically located close to the gene it regulates, which facilitates the transcription of a gene.

Promoters comprise two interacting parts, i.e., the basal promoter elements and the enhancer elements. Basal promoter elements bind accessory transcription initiation factors that position RNA polymerase II in the right place and direction. These basal elements are composed of short, low-complexity sequences (such as the TATA element). Enhancer elements bind regulatory factors that specify the physiological conditions or cell types where the gene will be expressed. The enhancer and basal promoter complexes interact at both functional and physical levels to determine how often an RNA transcript is produced. Enhancers can work over large distances of DNA in both directions.

4.2.2 Epigenetic Regulation

In addition, there are several epigenetic components influencing gene expression, such as histone modifications, DNA methylation, and position effects (Fig. 4.5a).

The major mechanism for suppressing widespread transcription is probably sequestration of potential transcription start sites by wrapping most of the genome in nucleosomes (see Sect. 3.2.2). Typical transcription start sites are found in nucleosome-free regions generated by DNA sequences that are intrinsically resistant to nucleosome wrapping. Another mechanism is the targeted modification and removal of nucleosomes in order to expose the underlying promoter sequences (see Sect. 3.2.2). Thus, functional eukaryotic promoters must not only attract RNA

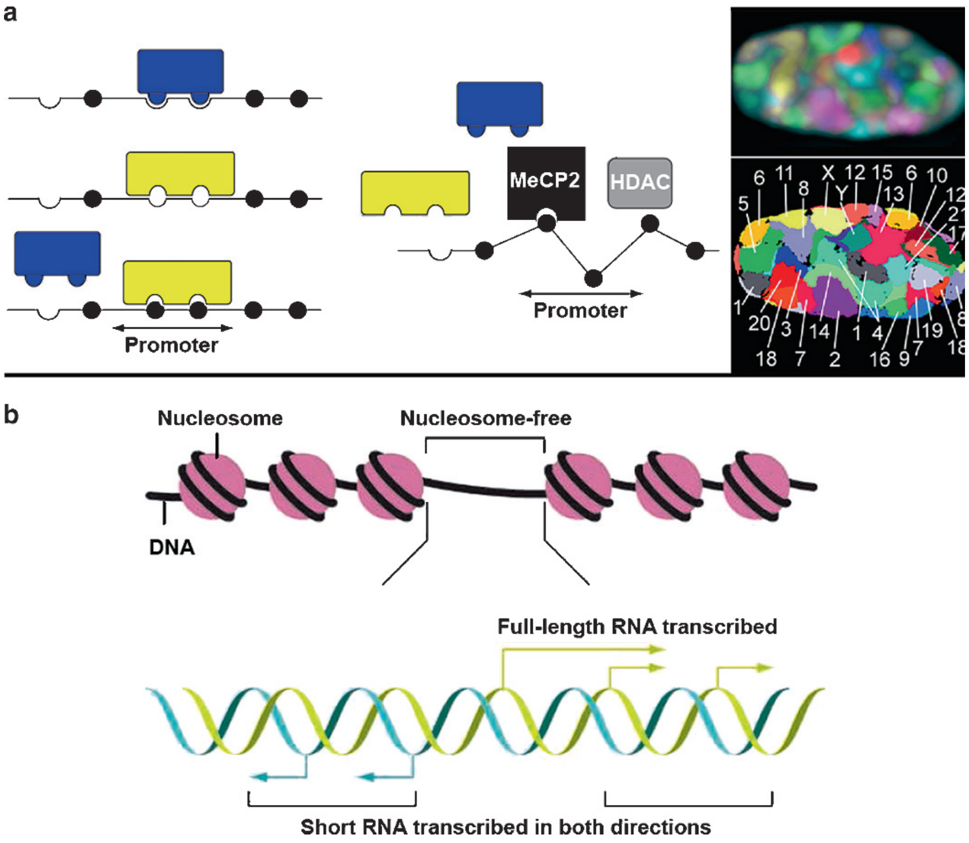


Fig. 4.5 (**a**, **b**) Different means of gene regulation. (**a**) *Left panel*: Consequences of epigenetic regulation by histone methylation: the configuration of the promoter region changes so that transcription factors cannot bind and expression of the respective gene is suppressed; *right panel*: Chromosomes occupy nonrandom positions in cell nuclei, these position effects influence genes expression. From [4]. (**b**) Example for possible gene regulation by small RNAs: promotor-associated transcripts (transcription start sites and transcripts are represented as *bent arrows*) within nucleosome-free DNA close to the promoters may influence gene expression. (From [6]. Reprinted with permission from AAAS)

polymerase II, but also evade nucleosomal repression. The epigenetic modes of gene regulation by histone modifications, DNA methylation, and position effects are discussed in detail in Sects. 3.2.2, 3.4.4.2, and 3.6.2.1.

4.2.3 Regulatory Transcripts of Small RNAs

More recently it has become clear that gene regulation may be affected by complex sets of small (20–30 nucleotides) RNAs that do not produce proteins, i.e., noncoding RNAs (ncRNAs; Fig. 4.5b). In general, effects of small RNAs on gene expression are inhibitory, as small RNAs may bind effector proteins to target nucleic acid molecules through base-pairing interactions. Therefore, activities of small RNAs are frequently summed up as "RNA silencing."

In humans the two main categories of small RNAs – among several classes – are short, interfering RNAs (siRNAs) and microRNAs (miRNAs) [2, 8]. These small RNAs are important regulators of gene expression that control both physiological and pathologic processes (e.g., development and cancer) miRNAs are regulators of endogenous genes, whereas siRNAs are defenders of genome integrity in response to foreign or invasive nucleic acids such as transposons and transgenes. An important distinction between miRNAs and siRNAs is whether or not they silence their own expression. Almost all siRNAs silence the same locus as they were derived from, and they only sometimes have the ability to silence other loci as well. In contrast, most miRNAs do not silence their own loci but do silence other genes. Both RNAs have double-stranded precursors and depend upon the same two families of proteins: Dicer enzymes to excise them from their precursors and Ago proteins to support their silencing effector functions [2, 8]. Single-stranded forms of both miRNAs and siRNAs associate with effector assemblies, which have been dubbed RNA-induced silencing complexes (RISCs).

The genes to be silenced are determined by the small RNA component, which identifies the respective complementary nucleotide sequence. The silencing can be monitored by increased expression of small RNAs or, conversely, by dilution or removal of old ones.

Furthermore, a new class of short RNA transcripts begins near the expected transcription start sites upstream of protein-encoding sequences (Fig. 4.5b). These RNAs often occur in the direction opposite to that of the protein-coding region [12, 29, 50, 57]. Although the function of these RNAs is presently not well defined, they may have an impact on how promoters delineate transcription start sites. These new RNAs are largely derived from DNA in nucleosome-free regions and may therefore arise from random, weak basal promoter elements that escape suppression [12, 29, 50, 57]. Hence, these short promoter-associated RNAs may simply result from incomplete suppression of cryptic initiation which, however, does not exclude an associated function by affecting the expression of the nearby gene.

4.3 "-omics" Sciences

Single-biomarker analysis is increasingly being replaced by multiparametric analysis of genes, transcripts, or proteins, now subsumed under the term "omics" sciences. The current nomenclature of omics sciences includes genomics for DNA variants, transcriptomics for mRNA, proteomics for proteins, and metabolomics for intermediate products of metabolism. The omics sciences use high-throughput techniques often allowing simultaneous examination of changes in the genome (DNA), transcriptome (messenger RNA [mRNA]), proteome (proteins), or metabolome (metabolites) in a biological sample, with the goal of understanding the physiology or mechanisms of disease. Insights derived from the complementary fields of omics sciences are expected to assist the development of new diagnostic, prognostic, and therapeutic tools. The omics sciences have in common that they require the development of novel informatic applications and sophisticated dimensionality reduction strategies. They have an enormous potential to unravel disease and physiological mechanisms and can identify clinically exploitable biomarkers from huge experimental datasets and offer insights into the molecular mechanisms of diseases.

The characteristics of the individual omics sciences and their integration to systems biology can be summarized as follows:

Genomics: Genomics seeks to define our genetic substrate and describes the study of the genomes of organisms.

Transcriptomics: Transcriptomics refers to the detailed analysis of the entire transcriptome, i.e., of all expressed sequences.

Proteomics: Proteomics explores the structure and function of proteins, which are the end-effectors of our genes. Proteomics has been revolutionized in the past decade by the application of techniques such as protein arrays, two-dimensional gel electrophoresis, and mass spectrometry. These techniques have tremendous potential for biomarker development, target validation, diagnosis, prognosis, and an optimization of treatment in medical care, especially in the field of clinical oncology.

Systems Biology: The integration of omic techniques is called "systems biology." This discipline aims at defining the interrelationships of several, or ideally all, of the elements in a system, rather than studying each element independently. Thus, systems biology will capture information from genomics, transcriptomics, proteomics, metabolomics, etc. and combine it with theoretical models in order to predict the behavior of a cell or organism.

4.4 Genomics

Genomics is the systematic study of the genomes of organisms. The field includes intensive efforts to determine the entire DNA sequence of organisms and fine-scale genetic mapping efforts. The investigation of single genes does not usually fit the definition of genomics. However, as the function of a single gene may affect many other genes, the border between single-gene analysis and genomics is often blurred.

4.4.1 Genomes of Organisms

A major branch of genomics is still concerned with the sequencing of the genomes of various organisms. The genome of the first free-living organism that was completely sequenced (*Haemophilus influenzae* in 1995) had a size of 1.8 Mb [21]. This was followed by complete sequences for *Mycoplasma genitalium* [23] and *Mycobacterium tuberculosis* [11], and subsequently by many other archeal, bacterial, and eukaryotic genomes. A rough draft of the human genome was presented in 2001 [39, 67], followed by an auspiciously completed version in 2004 [32]. Today sequencing efforts for other genomes continue. However, especially the resequencing of genomes, e.g., of human genomes to establish the variability between the genomes of different individuals, was propelled by the new possibilities of next-generation sequencing, which have added the sequences of other human individuals or of the first entire tumor genomes.

4.4.2 Array and Other Technologies

Genomics has certainly benefited from various array technologies that allowed the systematic analysis of entire genomes with various resolutions. These array technologies and other currently frequently employed important diagnostic tools, such as ChIP on chip and MLPA, are described and discussed in detail in Chap. 3 (Sect. 3.4.4.4). However, perhaps the most important recent development in genomics stems from next-generation (also referred to as second-generation) sequencing and the evolving third-generation sequencing (also referred to as single-molecule DNA sequencing).

4.4.3 Next-Generation Sequencing

Next-generation sequencing has already been introduced briefly in Sect. 4.1. An important issue of the new sequencing technologies is a significant reduction in costs: the public Human Genome Project spent US $\$3 \times 10^9$ to sequence the human genome, and the National Human Genome Research Institute at the National Institutes of Health aimed at a reduction of these costs to US $\$10^3$ by 2014 (www.genome.gov/12513210). The new DNA-sequencing platforms now available do indeed have the potential to achieve the same sequencing results of the Human Genome Project at perhaps 1% of the cost. However, the data obtained with next-generation sequencing depends heavily on the high-quality reference sequence produced by the Human Genome Project. The key to the increased efficiency of the new methods lies in massive parallelization of the biochemical and measurement steps. The second important issue is a significant increase in DNA sequencing speed.

So far there are several commercial next-generation DNA sequencing systems, such as Roche's (454) Genome Sequencer 20/FLX Genome Analyzer, Illumina's Solexa 1G sequencer, Applied Biosystem's SOLiD system, and the Polonator G.007 (Dover Systems/Harvard).

4.4.3.1 Roche's (454) GS FLX Genome Analyzer

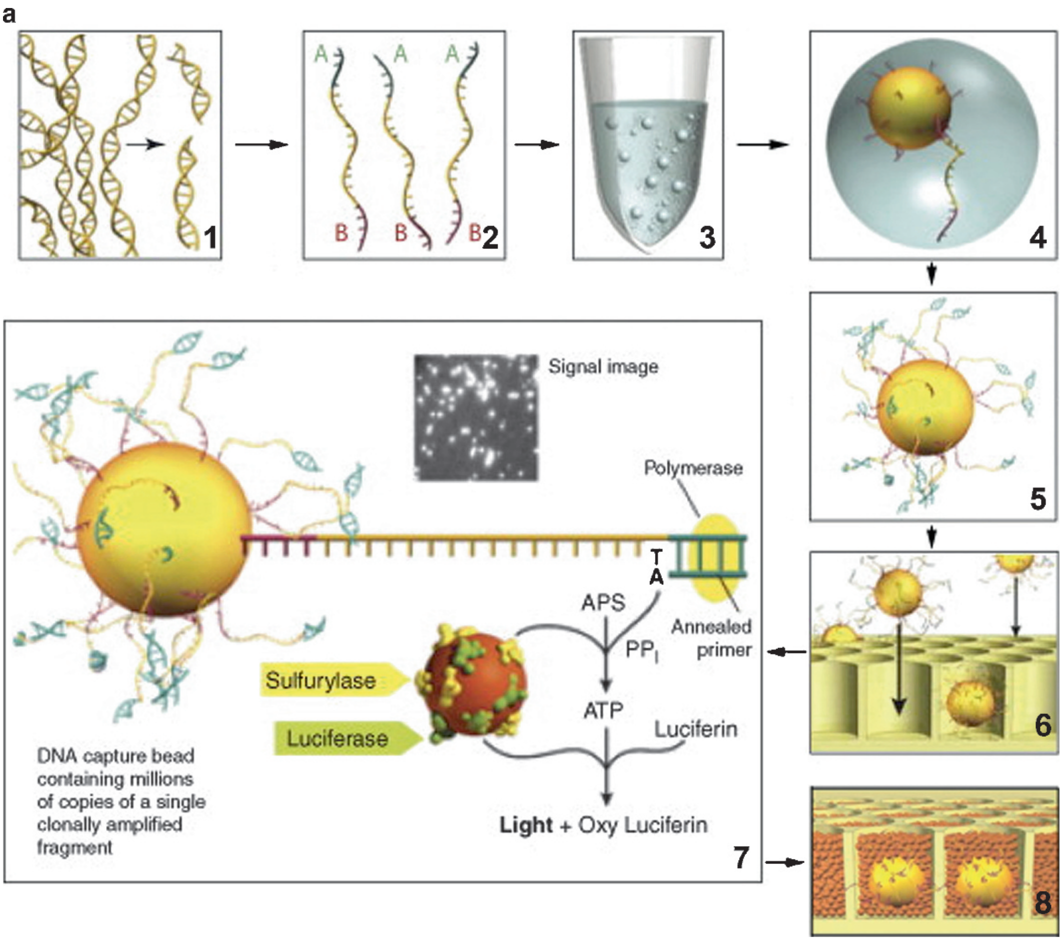
This system was commercially introduced in 2004 [47] and is based on pyrosequencing [49]. The sample preparation starts with fragmentation of the genomic DNA (Fig. 4.6a, b). In a next step, adapter sequences are attached to the ends of the DNA pieces to allow the DNA fragments to bind to beads, which have millions of oligomers attached to their surfaces, each with a complementary sequence to the adapter sequences. This is done under conditions allowing only one DNA fragment to bind to each bead. Subsequently, the DNA strands of the library are amplified by emulsion PCR: the beads, each with a single unique DNA fragment, are encased in droplets of oil, which isolate individual agarose beads and keep them apart from their neighbors to ensure that the amplification is uncontaminated. Each droplet contains all reactants needed to amplify the DNA, so that after some hours each agarose bead surface contains more than 1,000,000 copies of the original annealed DNA fragment. This number of DNA strands is needed to produce a detectable signal in the subsequent sequencing reaction. For this sequencing reaction the DNA template-carrying beads are loaded into picoliter reactor wells, each of which just has space for one bead. In these wells pyrosequencing [49], a sequencing-by-synthesis method, takes place, because DNA complementary to each template strand is synthesized. The pyrosequencing reactions flow through each well, and nucleotide and reagent solutions are delivered into it in a sequential fashion. The nucleotide bases used for sequencing release a chemical group as the base forms a bond with the growing DNA chain. This group drives a light-emitting reaction in the presence of specific enzymes and luciferin. The light from the luciferase activity reflects which templates are adding that particular nucleotide, and the emitted light is directly proportional to the amount of the particular nucleotide incorporated. Average read length per sample (or per bead) is about 250 bp.

4.4.3.2 Illumina's Solexa IG Sequencer

Illumina's Genome Analyzer, also commonly referred to as the "Solexa," was the second system commercially launched, in 2006. It is based on "sequencing by synthesis" [3] and is the only next-generation sequencing system that employs bridge-PCR [19] rather than emulsion-PCR (Fig. 4.7). The system applies high-density clonal single-molecule arrays consisting of genomic DNA fragments immobilized to the surface of a reaction chamber. In a first step, DNA fragments are generated by random shearing, and these are then ligated to a pair of oligonucleotides in a forked adapter configuration (Fig. 4.7a). These products can be amplified with two different oligonucleotide primers, which result in double-stranded DNA fragments with different adapter sequences at either end (Fig. 4.7a). In a next step these DNA fragments are denatured and a microfluid cluster station is used to anneal the single strands to the respective complementary oligonucleotides, which are covalently attached to the surface of a glass flow cell (Fig. 4.7b). A new strand is generated using the original strand as template in an extension reaction with an isothermal polymerase. Accordingly, the original strand is removed by denaturation. The adapter sequence of each newly generated strand is annealed to another surface-bound complementary oligonucleotide. This leads to formation of a bridge, and a new site for synthesis of a second strand is generated (Fig. 4.7b). Repeated cycles of annealing, extension, and denaturation result in growth of clusters, each apparently about 1 μm in diameter (Fig. 4.7c). Approximately 50×10^6 separate clusters can be generated per flow cell. For sequencing, each cluster is supplied with polymerase and four differently labeled fluorescent nucleotides (Fig. 4.7d). Based on the concept of "sequencing-by-synthesis," each base incorporation is followed by an imaging step to identify the incorporated nucleotide at each cluster. This iterative process needs about 2.5 days to generate read lengths of 36 bases. As each flow cell has 50×10^6 clusters, each analytical run generates more than 1 billion base pairs (Gb).

4.4.3.3 Applied Biosystem's SOLiD System

The SOLiD (sequencing by *o*ligo *l*igation and *d*etection) system was commercially released in 2007 and represents a development of work published in 2005 [59]. It



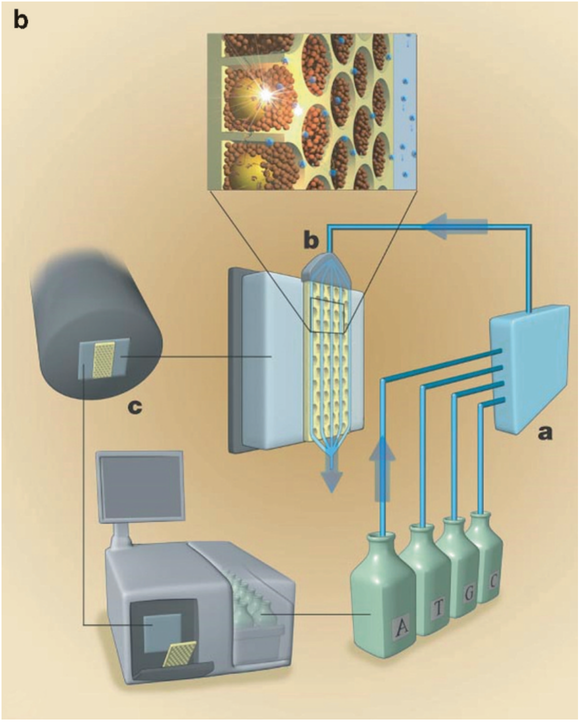


Fig. 4.6 (**a**, **b**) Steps involved in sequencing with the 454 system. (**a**) After isolation genomic DNA is fragmented (*1*) and ligated to adapters (*2*). The DNA is denatured to prepare them for emulsion PCR (*3*). Fragments are bound to beads under conditions that usually allow only one fragment per bead (*4*). The beads are captured in the droplets of a PCR reaction mixture-in-oil emulsion so that a PCR-amplification can be performed within each droplet (*5*). As a result, each bead carries 10 million copies of a unique DNA template. After breaking the emulsion the DNA strands are denatured, and beads carrying single-stranded DNA clones are placed into picotiter plates, i.e., wells of a fiberoptic slide (*6*). In these wells the pyrosequencing reaction takes place (*7* and *8*). (A composite from figures in [46] and [47]) (**b**) Major subsystems of the 454 sequencing instrument: (**b***a*) fluidic assembly; (**b***b*) flow chamber including the well-containing fibre-optic slide; (**b***c*) CCD camera, which captures the light emitted during the pyrosequencing reaction and a computer for instrument control. From [47], reprinted by permission from Macmillan Publishers Ltd: *Nature*, copyright 2005

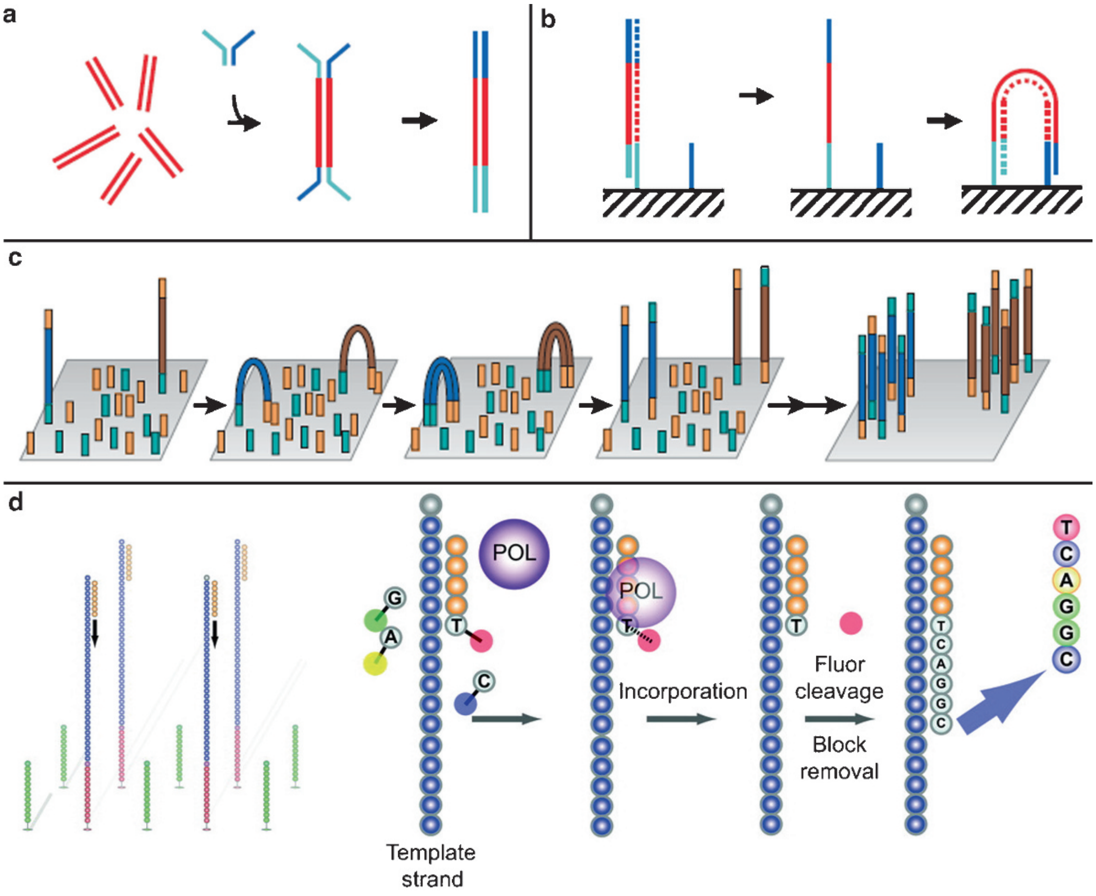


Fig. 4.7 (**a–d**) Steps involved in sequencing with the Illumina system. (**a**) DNA is fragmented by random shearing, and the fragments are then ligated to a pair of oligonucleotides. (**b**) The DNA fragments are denatured and annealed to the respective complementary oligonucleotides, which are covalently attached to the surface of a glass flow cell. A new strand is generated using the original strand as template. The "bridge" amplification relies on captured DNA strands arching over to that they can hybridize to an adjacent anchor oligonucleotide. By this means a bridge is formed and a new site for synthesis of a second strand is generated. (**c**) The adapter sequence of each newly generated strand is annealed to another surface-bound complementary oligonucleotide. Repeated cycles of annealing, extension, and denaturation result in growth of clusters, each appearing about 1 μm in diameter (**c**). (**d**) For sequencing the clusters are denatured, and after a chemical cleavage reaction and wash only forward strands remain for single-end sequencing. Each cluster is supplied with polymerase and four differently labeled fluorescent nucleotides, and each base incorporation is followed by an imaging step to identify the incorporated nucleotide at each cluster. (Reprinted by permission from Macmillan Publishers Ltd: (a,b) Nature [3], (c) Nature Biotechnology [58], copyright 2008)

also employs emulsion-PCR. After amplification the emulsion is broken and beads are covalently attached to the surface of a solid planar substrate, resulting in a dense, disordered array. The ligation-based sequencing process starts with the annealing of a universal primer complementary to the specific adapters on the library fragments. In each sequencing cycle a partially degenerate population of 8mer fluorescently labeled octamers is added (Fig. 4.8). These semi-degenerate oligos are structured in such a way that the label correlates with the identity of the central 2 bp in the octamer ("XX" in Fig. 4.8; the correlation with 2 bp, rather than 1 bp, is the basis of two-base encoding). When an 8mer oligo matches, it can hybridize adjacent to the universal primer 3′ end. The DNA ligase can then seal the phosphate backbone. After oligo-ligation a fluorescent readout consisting of imaging in four channels identifies the fixed base with the fluorescence label (the fifth position in Fig. 4.8). Subsequently, a chemical cleavage step removes the sixth through eighth bases ("zzz" in Fig. 4.8) via a modified linkage between bases 5 and 6, which deletes the fluorescent

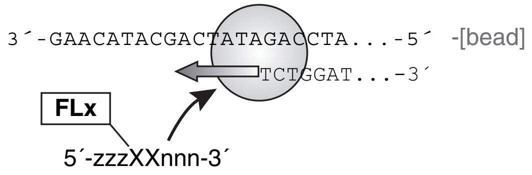


Fig. 4.8 Steps involved in sequencing with the Abi SOLiD system. Sample preparation is similar to that used in 454 technology, because DNA fragments are also ligated to oligonucleotide adapters linked to beads and clonally amplified by emulsion PCR. The ligation-based sequencing process starts with the annealing of a universal primer (5′-zzzXXnnn-3′) complementary to the specific adapters on the library fragments. These semidegenerate oligos are structured in such a way that the label correlates with the identity of the central 2 bp in the octamer ("XX"). Matching 8mer oligos can hybridize adjacent to the universal primer 3′ end and DNA ligase can then seal the phosphate backbone. After oligoligation a fluorescent readout consisting of imaging in four channels identifies the fixed base with the fluorescence label (here the fifth position). Subsequently, a chemical cleavage step removes the sixth through eighth bases (*zzz*), which leaves a free end for another cycle of ligation. For more details see text. (Reprinted by permission from Macmillan Publishers Ltd: Nature Biotechnology [58], copyright 2008)

group and leaves a free end for another cycle of ligation. Several cycles of that kind will iteratively interrogate an evenly spaced, discontinuous set of bases, in this example the sequence of each fragment at five nucleotide intervals. The system is then reset (by denaturation of the extended primer), and the process is repeated with a different offset (e.g., a primer set back from the original position by one or several bases) so that a different set of discontinuous bases is interrogated on the next round of serial ligations. A 6-day instrument run generates sequence read lengths of 35 bases. Placing two flow-cell slides in the instrument per analytical run can produce a combined output of more than 4 Gb of sequence.

A system related to the SOLiD system is the Polonator, which was also developed by the group of George M. Church at Havard [59].

4.4.4 Third-Generation Sequencing

Unlike many of the aforementioned high-speed sequencing technologies currently in use, third-generation sequencing is still under development. This technology is also often referred to as "single-molecule" sequencing, and it reads from individual DNA fragments without the need for amplification, or the risk of introducing errors, or the use of expensive reagents, such as fluorescent tags. As a consequence, third-generation sequencing has the potential to be even faster and cheaper than next-generation sequencing.

There are several different third-generation sequencing approaches, such as exonuclease sequencing, sequencing by synthesis, nanopore sequencing, and transmission electron microscopy [27]. For example, the principle of nanopore sequencing is that DNA can be detected as it passes through a pore by the interruption in the flow of ions through the aperture. The pores, made from a ring of seven α-hemolysin membrane proteins, are the same as those pushed into the membranes of other cells by the infectious bacterium *Staphylococcus aureus* in order to create damaging holes. The identity of each of the four bases traversing the hole might be revealed by distinctive changes in ion flow, which can be read as an electrical signal.

Companies which will likely offer commercial products within the near future include Helicos Bioscience, Complete Genomics, Pacific Biosciences, and Oxford Nanopore.

4.4.5 Personalized Genomics

In April 2008, 454 Life Science sequenced the entire genome of James Watson within 2 months for less than US$1 million [70]. In November 2008, Illumina reported the sequence of the human genome of a person of West African descent [3] and of a person of Han Chinese descent [69], each obtained for about US$250,000 within 8 weeks. At the same time, using the same technology the first complete DNA sequencing of a cytogenetically normal acute myeloid leukemia genome was reported [40]. Thus, the cheap sequencing enabled by next-generation sequencing heralds an era of "personal genomics." In fact, the routine use of whole-genome sequencing as a research tool in human genetics is now possible. At present it actually seems impossible to imagine the potential of third-generation sequencing. For example, Pacific Biosciences, which uses a single-molecule technology with DNA polymerase, aims at producing entire human genomes in less than 3 min by 2013. If these ambitious goals can be realized, the sequencing of an entire genome for about US$1,000 becomes reality and may introduce personalized genomics to the routine work-up in human genetics.

The rapid progress in genetic screening assays and DNA sequencing techniques promises to increase our understanding of the complex relationship between the human genetic make-up (the genotype) and its associated traits (the phenotype). However, what can we expect to learn from the sequences of individual genomes? The first complete genomes demonstrated that it will be extremely difficult to extract medically, or even biologically, reliable inferences from individual sequences. Without any doubt, whole-genome sequencing allows the identification of SNPs, as well as insertion/deletion polymorphisms and structural variations. However, at present they do not accurately define copy-number variants (CNVs, Sect. 3.4.4.4) at the nucleotide level. Thus, next-generation sequencing will improve the catalogue of variants existing in human genomes – SNPs by the million, insertion/deletion polymorphisms by the hundred thousand and structural variants by the thousand. The numbers of these variants will not directly provide information about how such polymorphisms contribute to the wide spectrum of human traits, yet they do provide a necessary step toward accurately defining genomic loci that are likely to be implicated in those traits. Therefore, association studies using complete individual genomes may become the approach of choice for understanding the complexity of human biology and disease.

4.4.6 Gene Function

Regardless of what definition of gene is being used, there is no question that genotype determines phenotype, often together with some environmental factors. At the molecular level, DNA sequences determine the sequences of functional molecules. Thus, an important consequence of the new gene definition as discussed in Sect. 4.1 is that the protein or RNA products must be functional for the purpose of assigning them to a particular gene [25]. This of course results in the important question of, "What is a function?". Many genes remain functionally uncharacterized in the physiological context of disease development. Importantly, the same pathologic mutation may – depending on the genetic background in which it occurs – have different consequences on the phenotype, which is often referred to as expressivity or penetrance. Therefore, high-throughput biochemical and mutational assays, molecular profiling, and interaction studies are needed to define function on a large scale. This is one of the purposes of the -omics sciences.

Gene functions must be clearly defined. This is a tremendous task, considering that biological function has many facets owing to the diversity of cellular activities. Defining the function of a gene is difficult, and it may be influenced by a membership in a specific pathway or a complex network in which the gene product interacts. Depending on this, the function of a gene may have effects across a wide range of spatial and temporal scales.

The gene ontology (GO) database uses a clearly defined and computationally friendly vocabulary for representing the cellular, biochemical, and physiological roles of gene products in a systematic fashion [28]. GO provides a standardized way to assess whether a given number of genes have similar functions. GO terms are organized in a tree-like structure, starting from more general at the root to the most specific at the leaves distributed across three main semantic domains – molecular function, biological process, and cellular location. However, GO describes many, but not all specific biological properties of known genes. In addition to GO, there are many other publicly available data sources, which can be used to get some information about possible gene-product functions (e.g., [31]; Chaps. 29.1–29.3).

Furthermore, there are multiple computational and statistical methods which can be used to deduce the functions of poorly characterized genes from genomic and proteomic datasets via association networks [31].

Many efforts have been made to assign functions to genes computationally. These gene-function predictions are based on parameters such as sequence similarity, the co-occurrence of the protein products in the same macromolecular complex, similarity in mRNA, and protein-expression patterns [71].

A particular challenge in the postgenome era is the deciphering of the biological function of individual genes and gene networks that drive disease. Therefore, at present, alternatives to traditional forward genetics approaches are sought. Such alternatives could consist in the construction of molecular networks defining the molecular states of a system underlying disease. Unlike classic genetics approaches aiming at the indentification of genes underlying genetic loci associated with disease, such approaches seek to identify whole gene networks responding *in trans* to genetic loci driving disease, and in turn leading to variations in the disease traits. The promise of these studies is that investigating how a network of gene interactions affects disease will come to complement more strongly the classic focus of how a single protein or RNA affects disease. Thus, a more detailed picture of the particular network states driving disease may be derived. This in turn may pave the way for more progressive treatments of disease, which may ultimately involve targeting whole networks, as opposed to current therapeutic strategies focused on targeting one or two genes [9].

4.5 Transcriptomics

4.5.1 Capturing the Cellular Transcriptome, Expression Arrays, and SAGE

A detailed analysis of the entire transcriptome requires sophisticated high-throughput approaches. Quantitative real-time PCR (qRT-PCR) represents a very effective technology for gene expression analysis, as it is indeed very quantitative and has a high sensitivity, enabling very accurate measurements of low-abundance transcripts. However, qRT-PCR provides less throughput than the technologies listed in the following paragraphs. Still, many see qRT-PCR as the "gold standard" against which other methods are validated.

Microarray chips have evolved to the most successful and most commonly-used technology for gene expression profiling [13, 56]. Numerous commercially available high-density microarray platforms are accessible, allowing the analysis of more or less entire transcriptomes of complex organisms with relative technical simplicity at low cost.

In parallel with the development of microarrays, computational methods for the analysis of the resulting large data sets were improved and standardized reporting and interpretation guidelines were developed [5]. In principle, two approaches are used for microarray analysis: First, as with CGH (Sect. 3.4.3.5.3), the two differently labeled RNAs are hybridized to the same array and the different fluorescence intensities are compared with one another. In the second approach only one RNA is hybridized to an oligonucleotide platform and stored reference data are being used to derive a comparison.

Microarray-based experiments are performed with RNA isolated from a specific tissue source, which is labeled with a detectable marker. This labeled RNA is then hybridized to arrays comprised of gene-specific probes representing thousands of individual genes.

Each experiment creates a massive amount of data requiring analysis by elaborate computational tools. There are two principle forms of data analysis, i.e., unsupervised and supervised hierarchical clustering analysis. The latter approach detects gene-expression patterns that discriminate tumors on the basis of predefined clinical information [16, 26].

Microarray-based gene expression has propelled our knowledge about transcriptome changes in disease and in physiological conditions. For example, the transcriptome of a normal cell type can be compared with the transcriptome of the same cell type with a specific disease, e.g., after malignant transformation, to elucidate disease-specific alterations. Another frequent application is the analysis of physiological changes, e.g., the comparison of the transcriptome of young versus old cell donors to decipher aging-related changes in the transcriptome [24, 42].

Serial analysis of gene expression (SAGE) [65] represents another approach for gene expression analysis (Fig. 4.9). SAGE is an RNA library-based technology which requires the sequencing of millions of cDNA tags from each library. These tags are then assigned to their genomic location by bioinformatics tools. The main advantage of SAGE is that the transcriptome analysis does not depend on the sequences represented on an array platform. However, SAGE involves significant sequencing efforts, making cost an important issue, so that this technology has not been affordable for many laboratories. Still, the aforementioned new next-generation or third-generation sequencing technologies should significantly decrease costs and may make SAGE even more attractive.

Other, more recent transcriptome analysis approaches are cap analysis of gene expression (CAGE) [7, 36] and polony multiplex analysis of gene expression (PMAGE) [34].

4.5.2 Regulatory Networks

The particular challenges in transcriptomics are to identify every transcript of each cell type and the analysis

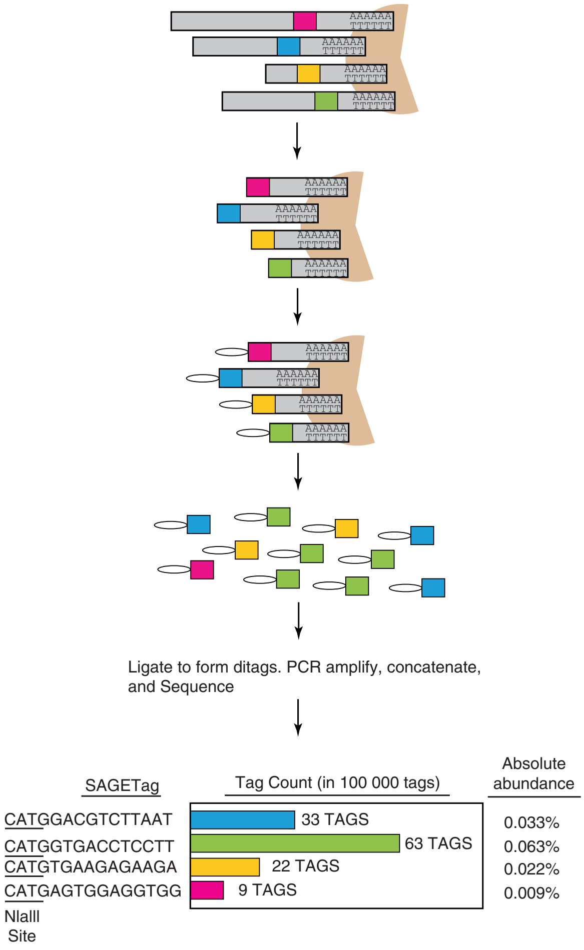


Fig. 4.9 Outline of serial analysis of gene expression (SAGE). In a first step poly-A RNA is captured on oligo-dT-coated beads and subjected to double-stranded cDNA synthesis. The poly-A RNA is cut at defined positions within each transcript by cleavage with an anchoring enzyme (usually NlaIII). Subsequently, linkers are ligated to the immobilized cDNA fragments. These linkers harbor a restriction enzyme type IIs site so that a "tagging enzyme" (usually BsmFI) cuts a short (15-bp) tag from the cDNA. These tags are ligated to form ditags, which can be amplified by PCR. The amplification products are then concatemerized and cloned. Individual tags are then identified by sequencing of concatemere clones. Absolute abundances of tags are calculated by dividing the observed abundance of any tag by the total number of tags analyzed. (Reprinted from [66], with permission from Elsevier)

of how transcription changes during development, with time and space, and especially according to environmental alterations. An integral part of these research efforts is to unravel the control mechanisms which regulate the transcriptome. One aim is the identification of regulatory networks for each cell type under different conditions, which may be an important prerequisite for the development of new therapeutic options. As a consequence, system approaches have been developed over the past years to elucidate transcriptional regulatory networks from high-throughput data [61].

The definition of networks includes the identification of all expressed transcripts under any developmental and growth condition. Furthermore, all possible physical interactions between transcriptional regulators and regulatory elements have to be delineated.

As complete transcriptomes of cells are cataloged at increasingly finer levels of detail, we may be able to discern the rules that determine where RNAs are made and how they are processed. However, such rules may change under certain conditions. For example, a cryptic transcription start site upstream of the "correct" initiation site might produce an RNA with additional protein-coding sequence or altered translation efficiency. A minor transcription start site within a gene could produce a truncated protein variant targeted at a different subcellular location. If any of these events provide some selective advantage the cryptic transcription start site could, over the course of time, become an alternative one and eventually the real transcription start site. Such evolutions can only be addressed if entire networks are being analyzed.

4.5.3 Outlier Profile Analysis

A particular challenge in transcriptome analysis could be the inability to extract the essence of recurring specific characteristics that may only be present on a subset of cases within a group. This may be especially true in RNA that has been extracted from cancer samples and which may show heterogeneous patterns of gene amplification, fusion, mutation, or deletion. To overcome these problems, a novel bioinformatics approach dubbed "cancer outlier profile analysis" has been developed as a means of identifying recurring patterns of gene overexpression that may characterize distinct subsets of known cancer types, but may not be detectable with traditional analysis methods (such as *t*-tests or signal-to-noise ratios) [62]. By using cancer outlier profile analysis, two members of the ETS family of transcription factors, ETV1 and ERG, were identified as outliers in prostate cancer. Additional analysis of cDNA transcripts of ERG and ETV1 in prostate cancer cell lines indicated fusion of the 5′ untranslated region of *TMPRSS2* (a prostate-specific, strongly androgen-regulated gene) to either ERG or ETV1. Indeed, cytogenetic analyses performed subsequently confirmed the presence of translocations involving the *TMPRSS2* locus on chromosome 21q22.3 and the corresponding chromosomes harboring one of the ETS family genes. Thus, purely computational manipulation and meta-analysis of existing high-throughput gene expression datasets has eventually led to discovery of a novel group of recurring chromosomal translocations in prostate cancer, which had been neglected by all previously performed cytogenetic or molecular cytogenetic technologies [62].

4.5.4 High-Throughput Long- and Short-Read Transcriptome Sequencing

The same group as initiated outlier profile analysis developed an integrative analysis of high-throughput long- and short-read transcriptome sequencing of cancer cells to discover novel gene fusions [45]. This strategy may represent a powerful tool for the discovery of novel gene chimeras using high-throughput sequencing, opening up an important class of cancer-related mutations for comprehensive characterization [45]. At the same time it becomes obvious that the new sequencing technologies can also be applied to the transcriptome and that they will have a tremendous impact on transcriptomics.

4.5.5 Disease Classification

Interestingly, it has been shown that cancer types can be subclassified based on their *gene* expression patterns. Therefore, gene expression data are often referred to as “signatures” or “molecular portraits,” because most tumors show unique expression patterns [10]. Together with appropriate statistical analysis, new or improved classifications have been developed based on expression microarrays for a variety of tumors, such as breast, ovary, prostate, colon, gastric, lung, kidney, brain, leukemia, and lymphoma (reviewed in [10]). These analyses demonstrated that some gene pathways, especially those involved in cell-cycle control, adhesion and motility, apoptosis, and angiogenesis, are frequently affected. Furthermore, these analyses point to pathways, which may represent especially promising targets for therapeutic interventions.

4.5.6 Tools for Prognosis Estimation

Gene expression data have also been used to establish prognostic categories, e.g., in leukemias, breast cancers, and other tumor types [38]. For example, several studies suggest that a panel of 70 genes is sufficient to classify breast cancer into prognostic categories [63, 64]. These analyses resulted in the first multigene panel test approved by the FDA for predicting breast cancer relapse [63].

However, a meta-analysis of seven of the most prominent studies on cancer prognosis based on microarray-expression profiling failed to reproduce the original data in five of these studies [48]. The other two studies yielded much weaker prognostic information than the original data. This suggests that larger sample sizes and careful validation are needed before definite statements about the clinical usefulness of such prognosis predictors can be made. Thus, at present the use of these gene arrays as diagnostic markers cannot yet be recommended [38].

4.6 Proteomics

The proteome is the entire set of proteins encoded by the genome, whereas proteomics is the discipline which studies the global set of proteins and their expression, function, and structure. Proteomics is – after genomics – often considered as a next step in the study of biological systems. Whereas an organism’s genome is relatively stable, and therefore more or less constant, the proteome differs from cell to cell and from time to time, making the analysis of the proteome more complicated. Even within a particular cell type, cells may make different sets of proteins at different times or under different conditions. Furthermore, any protein can undergo a wide range of posttranslational modifications, such as phosphorylation, ubiquitination, methylation, acetylation, and so on. As a particular gene can generate multiple distinct proteins, the number of proteins exceeds the number of genes in the corresponding genome by far.

As neither DNA nor mRNA reflects the function of proteins, a number of sophisticated technologies are needed to study individual proteins or the proteome.

4.6.1 From Low-Throughput to High-Throughput Techniques

There are a number of low-throughput techniques which allow testing for the presence of proteins and which can quantify them accurately. These analyses are often performed under certain conditions, e.g., to measure any protein changes during a particular physiological setting or during defined disease stages. Such techniques include Western blot, immunohistochemical staining, and enzyme-linked immunosorbent assay (ELISA). However, in a similar way to DNA or RNA analyses, the study of a protein can quickly become very complex. A frequent aim of proteomics is the identification of biomarkers. This usually requires a detailed understanding of multiple proteins and the complexities of protein-protein interactions. With such an amount of complexity, high-throughput approaches are needed.

At the beginning of proteomics, protein composition studies were performed on two-dimensional gel electrophoresis, which separates proteins in one dimension by molecular weight and in the second dimension by isoelectric point. Spots in the polyacrylamide gel can be cut, and proteins are identified using trypsin digestion and mass spectrometry (MS; Fig. 4.10). The MS tracing provides information on the mass/charge ratio (*m/z* ratio) of ions. These ratio values can be used to search protein databases. Such a two-dimensional polyacrylamide gel electrophoresis is suitable for high-throughput protein profiling. Basically, procedures to identify biomarkers from clinical specimens can be classified into two principle methodologies: mass spectrometer-based methods and antibody array-based methods, which are similar to DNA microarrays. Mass spectrometry-based approaches are more suitable in cases where the nature of the biomarkers or biosignatures is unknown. In contrast, targeted antibody arrays, which appear to be more cost effective, are more popular for testing proteins for known key pathways.

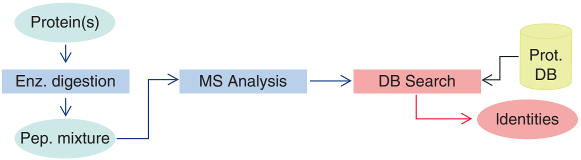


Fig. 4.10 Outline of an experiment in which proteins from 2D gel electrophoresis are identified after enzymatic digestion to create a protein mixture and mass spectrometry of the resulting peptides. The MS tracing provides information on the mass/charge ratio (*m/z* ratio) of ions, which can be used to search protein databases. (Adapted from [15]. Reprinted with permission from AAAS)

4.6.2 Mass Spectrometer-Based Methods

The central analytical technique for protein research and for the study of biomolecules is mass spectrometry (MS) [15]. MS is the method most commonly used for the investigation and identification of proteins. MS operates to create ions from neutral proteins, peptides, or metabolites. Therefore, MS depends on effective technologies to softly ionize and to transfer the ionized molecules from the condensed phase into the gas phase without excessive fragmentation. Thus, an MS consists of two main components – an ionization source and a mass analyzer. There are two commonly used techniques to transfer molecules into the gas phase and ionize them prior to mass separation, i.e., electrospray ionization (ESI) [20] and matrix-assisted laser desorption/ionization (MALDI) [33]. After ionization the mass analyzer utilizes the electric charge of the particulates for their separation by speed and/or direction, dependent on the intrinsic m/z of the ion. The types of ion mass separation may include, for example, time-of-flight (ToF), quadrupole electric fields (Q), ion trap (IT), Fourier transform ion cyclotron resonance (FT-ICR) and the Orbitrap [15]. The mass spectrum is characteristic of the molecular mass and/or structure of the metabolite.

Single-stage mass spectrometers are used to evaluate the molecular mass of a polypeptide. However, MS can also provide information about additional structural features, such as amino acid sequence or types of posttranslational modifications. Such analyses are performed after the initial mass determination. Specific ions are selected and fragmented, and structural features of the respective peptides can be deduced from the analysis of these fragments' masses. As two MS analyses are sequentially performed these analyses are usually referred to as tandem MS (MS/MS) [15].

However, like all other approaches, the promising proteomic profiling technologies via MS also have some shortcomings. These include potential artifacts attributable to sample collection and storage, the inherent qualitative nature of mass spectrometers defined by instrument sensitivity, resolution, mass accuracy, dynamic range and throughput, and finally potential artifacts introduced by high-abundance proteins in the serum [38].

4.6.3 Antibody Array-Based Methods

Alternative proteomic strategies include protein microarrays, which depend on immobilization of proteins on a solid support in a way that preserves their folded conformations [44]. For example, antibodies are spotted on the solid surface onto which unmodified proteins are applied. After binding of the proteins to their respective antibodies, a second antibody, which recognizes the same protein and which is labeled for detection by fluorescence, is applied. Such an approach has been referred to as a "sandwich ELISA assay" [37].

Rather like DNA arrays, the direct chemical modification of proteins provides a direct assay mode. Proteins can be labeled with different fluorescent dyes, e.g., as Cy3 and Cy5, and can then be applied to the antibody-spotted slide. This allows the simultaneous analyses of hundreds of target proteins on the same slide. Such an assay is semiquantitative and makes the comparison of two samples applied on the same array, e.g., control versus treated, or normal versus cancer, as in CGH experiments, possible. As in DNA arrays, false-positive or false-negative results have to be excluded, making further validation with other methodologies necessary.

The use of antibody arrays is mainly intended for initial screening of large numbers of proteins to identify candidates for further research. Additional applications include the analysis of posttranslational modifications (such as phosphorylation, acetylation, glycosylation, among others) in complex mixtures of proteins and the analysis of protein/protein interactions [43].

4.6.4 Proteomic Strategies

Several strategies for the analysis of proteins or the proteome have evolved. *MS analysis of substantially purified proteins* corresponds to the aforementioned, classic approach: two-dimensional (2D) gel electrophoresis followed by the mass-spectrometric identification of the protein(s) in a single gel spot. The targeted proteins are digested and identified by mass spectrometry.

In contrast, for *MS analysis of complex peptide mixtures,* also referred to as shotgun proteomics, complex protein samples are digested. The resulting peptide samples are extensively fractionated and analyzed by automated MS/MS. Such an approach allows the analysis of protein samples derived from complete cell lysates or tissue extracts, subcellular fractions, isolated organelles, or other subproteomes.

Furthermore, the establishment of comparative peptide patterns is an important issue. Beside the aforementioned antibody arrays to which two differently labeled protein samples are applied, such a comparison can also be made by 2D gel electrophoresis. For each sample to be analyzed, 2D patterns are generated and the patterns are compared to identify quantitative or qualitative changes. Observed differences can then be further characterized, for example, by sequencing or by determining their posttranslationally modified state.

Future strategies aim at more efficient approaches than those available at present. Such strategies may avoid the situation where the proteome is rediscovered in every experiment. Instead, it would be desirable to use the information from prior proteomic experiments as a guideline for new experiments. This requires the generation of extensive (complete) databases with information to both known and theoretical peptides and their respective proteins to facilitate the targeted, nonredundant analysis of information-rich peptides [15].

4.6.5 Proteomics for Screening and Diagnosis of Disease (Diagnostic and Prognostic Biomarkers)

MS and antibody arrays have evolved into popular platforms for protein screening. It is of special importance that they offer the advantage of multiplexing, can be performed with low sample requirement, and they have the potential for up-scaling using automation.

The availability of methods for measuring the abundance of proteins simultaneously in multiplexed assay formats has opened up opportunities in basic and disease-related research. These technologies can be applied to studies requiring large surveys of changes in protein abundance, to biomarker identification and validation, and to clinical diagnostics using selected targets.

The technologies have matured and can now be used not only for broad protein expression analysis, but also for defining signal transduction pathways, for molecular classification of diseases, for compound profiling and toxicology studies, and for the analysis of patients' individual sensitivities to drugs.

4.7 Conclusions

In human biology the elucidation of gene-product function and regulation is a fundamental objective. In most scenarios a focused single-gene approach is insufficient, making omics sciences indispensible. Owing to its relative stability, the in-depth analysis of the human genome now represents, especially because of new sequencing technologies, an amenable task, although the real extent of genomic variability is so far unknown. The recent completion of the genomic sequences of human and other mammalian species provides researchers with access to a wealth of relevant sequence information necessary for the functional characterization of gene products in a systematic and comprehensive manner. However, analyses of both transcriptome and proteome appear to be significantly more complex than the analysis of the genome. At transcriptome level, the functional characterization of noncoding RNAs represents what will presumably be the greatest challenge. Furthermore, proper biological activity and cellular homeostasis depend on spatially and temporally restricted partitioning of functionally related sets of gene products. A basic and conserved mode of biological control is the organ- and organelle-selective protein accumulation. Therefore, the fundamental biological information encrypted in the human genome can only be understood by the study of the global patterns of protein synthesis and subcellular localization across the major mammalian organ systems. However, at present, much of the human proteome remains poorly annotated in terms of tissue- and organelle-selective expression.

One of the outstanding questions in expression profiling is how well mRNA levels indeed reflect protein abundance and may represent the biological basis for any measurable differences. Although protein synthesis is dependent on mRNA, in many studies often only a modest relationship between mRNA and protein levels was reported. There may be numerous causes for incomplete proteome/transcriptome coverage, such as sample complexity, unknown protein modifications, poor recovery and detection of lower abundance and membrane-associated proteins, and the fact that certain proteins may also be transported between tissues, particularly those associated with circulatory or endocrine functions. This hampers a rigorous definition of the expressed proteome. Hence, the biological significance of differences in mRNA abundance detected among tissues remains to be elaborated at the protein level.

Furthermore, another limitation of transcriptional profiling is that little information is gleaned with respect to the subcellular localization of the translated gene products. Therefore, proteomic methods of examining protein expression and subcellular localization on a genome-wide scale should provide additional insight into the biological context of uncharacterized gene products that can naturally lead to testable hypotheses regarding function.

References

1. Antonarakis SE, Beckmann JS (2006) Mendelian disorders deserve more attention. Nat Rev Genet 7:277–282
2. Bartel DP (2004) MicroRNAs: genomics, biogenesis, mechanism, and function. Cell 116:281–297
3. Bentley DR, Balasubramanian S, Swerdlow HP, Smith GP, Milton J, Brown CG, Hall KP, Evers DJ, Barnes CL, Bignell HR, Boutell JM, Bryant J, Carter RJ, Keira Cheetham R, Cox AJ, Ellis DJ, Flatbush MR, Gormley NA, Humphray SJ, Irving LJ, Karbelashvili MS, Kirk SM, Li H, Liu X, Maisinger KS, Murray LJ, Obradovic B, Ost T, Parkinson ML, Pratt MR, Rasolonjatovo IM, Reed MT, Rigatti R, Rodighiero C, Ross MT, Sabot A, Sankar SV, Scally A, Schroth GP, Smith ME, Smith VP, Spiridou A, Torrance PE, Tzonev SS, Vermaas EH, Walter K, Wu X, Zhang L, Alam MD, Anastasi C, Aniebo IC, Bailey DM, Bancarz IR, Banerjee S, Barbour SG, Baybayan PA, Benoit VA, Benson KF, Bevis C, Black PJ, Boodhun A, Brennan JS, Bridgham JA, Brown RC, Brown AA, Buermann DH, Bundu AA, Burrows JC, Carter NP, Castillo N, Chiara E, Catenazzi M, Chang S, Neil Cooley R, Crake NR, Dada OO, Diakoumakos KD, Dominguez-Fernandez B, Earnshaw DJ, Egbujor UC, Elmore DW, Etchin SS, Ewan MR, Fedurco M, Fraser LJ, Fuentes Fajardo KV, Scott Furey W, George D, Gietzen KJ, Goddard CP, Golda GS, Granieri PA, Green DE, Gustafson DL, Hansen NF, Harnish K, Haudenschild CD, Heyer NI, Hims MM, Ho JT, Horgan AM, Hoschler K, Hurwitz S, Ivanov DV, Johnson MQ, James T, Huw Jones TA, Kang GD,

Kerelska TH, Kersey AD, Khrebtukova I, Kindwall AP, Kingsbury Z, Kokko-Gonzales PI, Kumar A, Laurent MA, Lawley CT, Lee SE, Lee X, Liao AK, Loch JA, Lok M, Luo S, Mammen RM, Martin JW, McCauley PG, McNitt P, Mehta P, Moon KW, Mullens JW, Newington T, Ning Z, Ling Ng B, Novo SM, O'Neill MJ, Osborne MA, Osnowski A, Ostadan O, Paraschos LL, Pickering L, Pike AC, Pike AC, Chris Pinkard D, Pliskin DP, Podhasky J, Quijano VJ, Raczy C, Rae VH, Rawlings SR, Chiva Rodriguez A, Roe PM, Rogers J, Rogert Bacigalupo MC, Romanov N, Romieu A, Roth RK, Rourke NJ, Ruediger ST, Rusman E, Sanches-Kuiper RM, Schenker MR, Seoane JM, Shaw RJ, Shiver MK, Short SW, Sizto NL, Sluis JP, Smith MA, Ernest Sohna Sohna J, Spence EJ, Stevens K, Sutton N, Szajkowski L, Tregidgo CL, Turcatti G, Vandevondele S, Verhovsky Y, Virk SM, Wakelin S, Walcott GC, Wang J, Worsley GJ, Yan J, Yau L, Zuerlein M, Rogers J, Mullikin JC, Hurles ME, McCooke NJ, West JS, Oaks FL, Lundberg PL, Klenerman D, Durbin R, Smith AJ (2008) Accurate whole human genome sequencing using reversible terminator chemistry. Nature 456:53–59

4. Bolzer A, Kreth G, Solovei I, Koehler D, Saracoglu K, Fauth C, Müller S, Eils R, Cremer C, Speicher MR, Cremer T (2005) Three-dimensional maps of all chromosome positions indicate a probabilistic order in human male fibroblast nuclei and prometaphase rosettes. PLoS Biol 3:e157
5. Brazma A, Hingamp P, Quackenbush J, Sherlock G, Spellman P, Stoeckert C, Aach J, Ansorge W, Ball CA, Causton HC, Gaasterland T, Glenisson P, Holstege FC, Kim IF, Markowitz V, Matese JC, Parkinson H, Robinson A, Sarkans U, Schulze-Kremer S, Stewart J, Taylor R, Vilo J, Vingron M (2001) Minimum information about a microarray experiment (MIAME)-toward standards for microarray data. Nat Genet 29:365–371
6. Buratowski S (2008) Transcription. Gene expression--where to start? Science 322:1804–1805
7. Carninci P, Sandelin A, Lenhard B, Katayama S, Shimokawa K, Ponjavic J, Semple CA, Taylor MS, Engström PG, Frith MC, Forrest AR, Alkema WB, Tan SL, Plessy C, Kodzius R, Ravasi T, Kasukawa T, Fukuda S, Kanamori-Katayama M, Kitazume Y, Kawaji H, Kai C, Nakamura M, Konno H, Nakano K, Mottagui-Tabar S, Arner P, Chesi A, Gustincich S, Persichetti F, Suzuki H, Grimmond SM, Wells CA, Orlando V, Wahlestedt C, Liu ET, Harbers M, Kawai J, Bajic VB, Hume DA, Hayashizaki Y (2006) Genome-wide analysis of mammalian promoter architecture and evolution. Nat Genet 38:626–635
8. Carthew RW, Sontheimer EJ (2009) Origins and Mechanisms of miRNAs and siRNAs. Cell 136:642–655
9. Chen Y, Zhu J, Lum PY, Yang X, Pinto S, MacNeil DJ, Zhang C, Lamb J, Edwards S, Sieberts SK, Leonardson A, Castellini LW, Wang S, Champy MF, Zhang B, Emilsson V, Doss S, Ghazalpour A, Horvath S, Drake TA, Lusis AJ, Schadt EE (2008) Variations in DNA elucidate molecular networks that cause disease. Nature 452:429–435
10. Chung CH, Bernard PS, Perou CM (2002) Molecular portraits and the family tree of cancer. Nat Genet 32:533–540
11. Cole ST, Brosch R, Parkhill J, Garnier T, Churcher C, Harris D, Gordon SV, Eiglmeier K, Gas S, Barry CE 3 rd, Tekaia F, Badcock K, Basham D, Brown D, Chillingworth T, Connor R, Davies R, Devlin K, Feltwell T, Gentles S, Hamlin N, Holroyd S, Hornsby T, Jagels K, Krogh A, McLean J, Moule S, Murphy L, Oliver K, Osborne J, Quail MA, Rajandream MA, Rogers J, Rutter S, Seeger K, Skelton J, Squares R, Squares S, Sulston JE, Taylor K, Whitehead S, Barrell BG (1998) Deciphering the biology of Mycobacterium tuberculosis from the complete genome sequence. Nature 393:537–544
12. Core LJ, Waterfall JJ, Lis JT (2008) Nascent RNA sequencing reveals widespread pausing and divergent initiation at human promoters. Science 322:1845–8
13. DeRisi J, Penland L, Brown PO, Bittner ML, Meltzer PS, Ray M, Chen Y, Su YA, Trent JM (1996) Use of a cDNA microarray to analyse gene expression patterns in human cancer. Nat Genet 14:457–460
14. De Sandre-Giovannoli A, Bernard R, Cau P, Navarro C, Amiel J, Boccaccio I, Lyonnet S, Stewart CL, Munnich A, Le Merrer M, Levy N (2003) Lamin A truncation in Hutchinson-Gilford Progeria. Science 300:2055
15. Domon B, Aebersold R (2006) Mass spectrometry and protein analysis. Science 312:212–217
16. Eisen MB, Spellman PT, Brown PO, Botstein D (1998) Cluster analysis and display of genome-wide expression patterns. Proc Natl Acad Sci USA 95:14863–14868
17. ENCODE Project Consortium, Birney E, Stamatoyannopoulos JA, Dutta A, Guigó R, Gingeras TR, Margulies EH, Weng Z, Snyder M, Dermitzakis ET, Thurman RE, Kuehn MS, Taylor CM, Neph S, Koch CM, Asthana S, Malhotra A, Adzhubei I, Greenbaum JA, Andrews RM, Flicek P, Boyle PJ, Cao H, Carter NP, Clelland GK, Davis S, Day N, Dhami P, Dillon SC, Dorschner MO, Fiegler H, Giresi PG, Goldy J, Hawrylycz M, Haydock A, Humbert R, James KD, Johnson BE, Johnson EM, Frum TT, Rosenzweig ER, Karnani N, Lee K, Lefebvre GC, Navas PA, Neri F, Parker SC, Sabo PJ, Sandstrom R, Shafer A, Vetrie D, Weaver M, Wilcox S, Yu M, Collins FS, Dekker J, Lieb JD, Tullius TD, Crawford GE, Sunyaev S, Noble WS, Dunham I, Denoeud F, Reymond A, Kapranov P, Rozowsky J, Zheng D, Castelo R, Frankish A, Harrow J, Ghosh S, Sandelin A, Hofacker IL, Baertsch R, Keefe D, Dike S, Cheng J, Hirsch HA, Sekinger EA, Lagarde J, Abril JF, Shahab A, Flamm C, Fried C, Hackermüller J, Hertel J, Lindemeyer M, Missal K, Tanzer A, Washietl S, Korbel J, Emanuelsson O, Pedersen JS, Holroyd N, Taylor R, Swarbreck D, Matthews N, Dickson MC, Thomas DJ, Weirauch MT, Gilbert J, Drenkow J, Bell I, Zhao X, Srinivasan KG, Sung WK, Ooi HS, Chiu KP, Foissac S, Alioto T, Brent M, Pachter L, Tress ML, Valencia A, Choo SW, Choo CY, Ucla C, Manzano C, Wyss C, Cheung E, Clark TG, Brown JB, Ganesh M, Patel S, Tammana H, Chrast J, Henrichsen CN, Kai C, Kawai J, Nagalakshmi U, Wu J, Lian Z, Lian J, Newburger P, Zhang X, Bickel P, Mattick JS, Carninci P, Hayashizaki Y, Weissman S, Hubbard T, Myers RM, Rogers J, Stadler PF, Lowe TM, Wei CL, Ruan Y, Struhl K, Gerstein M, Antonarakis SE, Fu Y, Green ED, Karaöz U, Siepel A, Taylor J, Liefer LA, Wetterstrand KA, Good PJ, Feingold EA, Guyer MS, Cooper GM, Asimenos G, Dewey CN, Hou M, Nikolaev S, Montoya-Burgos JI, Löytynoja A, Whelan S, Pardi F, Massingham T, Huang H, Zhang NR, Holmes I, Mullikin JC, Ureta-Vidal A, Paten B, Seringhaus M, Church D, Rosenbloom K, Kent WJ, Stone EA, NISC Comparative Sequencing Program; Baylor College of

Medicine Human Genome Sequencing Center; Washington University Genome Sequencing Center; Broad Institute; Children's Hospital Oakland Research Institute, Batzoglou S, Goldman N, Hardison RC, Haussler D, Miller W, Sidow A, Trinklein ND, Zhang ZD, Barrera L, Stuart R, King DC, Ameur A, Enroth S, Bieda MC, Kim J, Bhinge AA, Jiang N, Liu J, Yao F, Vega VB, Lee CW, Ng P, Shahab A, Yang A, Moqtaderi Z, Zhu Z, Xu X, Squazzo S, Oberley MJ, Inman D, Singer MA, Richmond TA, Munn KJ, Rada-Iglesias A, Wallerman O, Komorowski J, Fowler JC, Couttet P, Bruce AW, Dovey OM, Ellis PD, Langford CF, Nix DA, Euskirchen G, Hartman S, Urban AE, Kraus P, Van Calcar S, Heintzman N, Kim TH, Wang K, Qu C, Hon G, Luna R, Glass CK, Rosenfeld MG, Aldred SF, Cooper SJ, Halees A, Lin JM, Shulha HP, Zhang X, Xu M, Haidar JN, Yu Y, Ruan Y, Iyer VR, Green RD, Wadelius C, Farnham PJ, Ren B, Harte RA, Hinrichs AS, Trumbower H, Clawson H, Hillman-Jackson J, Zweig AS, Smith K, Thakkapallayil A, Barber G, Kuhn RM, Karolchik D, Armengol L, Bird CP, de Bakker PI, Kern AD, Lopez-Bigas N, Martin JD, Stranger BE, Woodroffe A, Davydov E, Dimas A, Eyras E, Hallgrímsdóttir IB, Huppert J, Zody MC, Abecasis GR, Estivill X, Bouffard GG, Guan X, Hansen NF, Idol JR, Maduro VV, Maskeri B, McDowell JC, Park M, Thomas PJ, Young AC, Blakesley RW, Muzny DM, Sodergren E, Wheeler DA, Worley KC, Jiang H, Weinstock GM, Gibbs RA, Graves T, Fulton R, Mardis ER, Wilson RK, Clamp M, Cuff J, Gnerre S, Jaffe DB, Chang JL, Lindblad-Toh K, Lander ES, Koriabine M, Nefedov M, Osoegawa K, Yoshinaga Y, Zhu B, de Jong PJ (2007) Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. Nature 447:799–816

18. Eriksson M, Brown WT, Gordon LB, Glynn MW, Singer J, Scott L, Erdos MR, Robbins CM, Moses TY, Berglund P, Dutra A, Pak E, Durkin S, Csoka AB, Boehnke M, Glover TW, Collins FS (2003) Recurrent de novo point mutations in lamin a cause Hutchinson-Gilford progeria syndrome. Nature 423:293–298
19. Fedurco M, Romieu A, Williams S, Lawrence I, Turcatti G (2006) BTA, a novel reagent for DNA attachment on glass and efficient generation of solid-phase amplified DNA colonies. Nucleic Acids Res 34:e22
20. Fenn J, Mann M, Meng C, Wong S, Whitehouse C (1988) Electrospray ionisation for mass spectrometry of large biomolecules. Science 246:64–71
21. Fleischmann RD, Adams MD, White O, Clayton RA, Kirkness EF, Kerlavage AR, Bult CJ, Tomb JF, Dougherty BA, Merrick JM et al (1995) Whole-genome random sequencing and assembly of Haemophilus influenzae Rd. Science 269:496–512
22. Fraga MF, Ballestar E, Paz MF, Ropero S, Setien F, Ballestar ML, Heine-Suñer D, Cigudosa JC, Urioste M, Benitez J, Boix-Chornet M, Sanchez-Aguilera A, Ling C, Carlsson E, Poulsen P, Vaag A, Stephan Z, Spector TD, Wu YZ, Plass C, Esteller M (2005) Epigenetic differences arise during the lifetime of monozygotic twins. Proc Natl Acad Sci USA 102:10604–10609
23. Fraser CM, Gocayne JD, White O, Adams MD, Clayton RA, Fleischmann RD, Bult CJ, Kerlavage AR, Sutton G, Kelley JM, Fritchman RD, Weidman JF, Small KV, Sandusky M, Fuhrmann J, Nguyen D, Utterback TR, Saudek DM, Phillips CA, Merrick JM, Tomb JF, Dougherty BA, Bott KF, Hu PC, Lucier TS, Peterson SN, Smith HO, Hutchison CA 3 rd, Venter JC (1995) The minimal gene complement of Mycoplasma genitalium. Science 270: 397–403
24. Geigl JB, Langer S, Barwisch S, Pfleghaar K, Lederer G, Speicher MR (2004) Analysis of gene expression patterns and chromosomal changes associated with aging. Cancer Res 64:8550–8557
25. Gerstein MB, Bruce C, Rozowsky JS, Zheng D, Du J, Korbel JO, Emanuelsson O, Zhang ZD, Weissman S, Snyder M (2007) What is a gene, post-ENCODE? History and updated definition. Genome Res 17:669–681
26. Golub TR, Slonim DK, Tamayo P, Huard C, Gaasenbeek M, Mesirov JP, Coller H, Loh ML, Downing JR, Caligiuri MA, Bloomfield CD, Lander ES (1999) Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. Science 286:531–537
27. Gupta PK (2008) Single-molecule DNA sequencing technologies for future genomics research. Trends Biotechnol 26:602–611
28. Harris MA, Clark J, Ireland A, Lomax J, Ashburner M, Foulger R, Eilbeck K, Lewis S, Marshall B, Mungall C, Richter J, Rubin GM, Blake JA, Bult C, Dolan M, Drabkin H, Eppig JT, Hill DP, Ni L, Ringwald M, Balakrishnan R, Cherry JM, Christie KR, Costanzo MC, Dwight SS, Engel S, Fisk DG, Hirschman JE, Hong EL, Nash RS, Sethuraman A, Theesfeld CL, Botstein D, Dolinski K, Feierbach B, Berardini T, Mundodi S, Rhee SY, Apweiler R, Barrell D, Camon E, Dimmer E, Lee V, Chisholm R, Gaudet P, Kibbe W, Kishore R, Schwarz EM, Sternberg P, Gwinn M, Hannick L, Wortman J, Berriman M, Wood V, de la Cruz N, Tonellato P, Jaiswal P, Seigfried T, White R, Gene Ontology Consortium (2004) The gene ontology (GO) database and informatics resource. Nucleic Acids Res 32(Database issue): D258–D261
29. He Y, Vogelstein B, Velculescu VE, Papadopoulos N, Kinzler KW (2008) The antisense transcriptomes of human cells. Science 322:1855–1857
30. Hennekam RC (2006) Hutchinson-Gilford progeria syndrome: review of the phenotype. Am J Med Genet 140: 2603–2624
31. Hu P, Bader G, Wigle DA, Emili A (2007) Computational prediction of cancer-gene function. Nat Rev Cancer 7:23–34
32. International Human Genome Sequencing Consortium (2004) Finishing the euchromatic sequence of the human genome. Nature 431:931–945
33. Karas M, Hillenkamp F (1988) Laser desorption ionization of proteins with molecular masses exceeding 10,000 daltons. Anal Chem 60:2299–2301
34. Kim JB, Porreca GJ, Song L, Greenway SC, Gorham JM, Church GM, Seidman CE, Seidman JG (2007) Polony multiplex analysis of gene expression (PMAGE) in mouse hypertrophic cardiomyopathy. Science 316:1481–1484
35. Kimchi-Sarfaty C, Oh JM, Kim IW, Sauna ZE, Calcagno AM, Ambudkar SV, Gottesman MM (2007) A "silent" polymorphism in the MDR1 gene changes substrate specificity. Science 315:525–528
36. Kodzius R, Kojima M, Nishiyori H, Nakamura M, Fukuda S, Tagami M, Sasaki D, Imamura K, Kai C, Harbers M, Hayashizaki Y, Carninci P (2006) CAGE: cap analysis of gene expression. Nat Methods 3:211–222

37. Kopf E, Zharhary D (2007) Antibody arrays–an emerging tool in cancer proteomics. Int J Biochem Cell Biol 39: 1305–1317
38. Kulasingam V, Diamandis EP (2008) Strategies for discovering novel cancer biomarkers through utilization of emerging technologies. Nat Clin Pract Oncol 5:588–599
39. Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, Devon K, Dewar K, Doyle M, FitzHugh W, Funke R, Gage D, Harris K, Heaford A, Howland J, Kann L, Lehoczky J, LeVine R, McEwan P, McKernan K, Meldrim J, Mesirov JP, Miranda C, Morris W, Naylor J, Raymond C, Rosetti M, Santos R, Sheridan A, Sougnez C, Stange-Thomann N, Stojanovic N, Subramanian A, Wyman D, Rogers J, Sulston J, Ainscough R, Beck S, Bentley D, Burton J, Clee C, Carter N, Coulson A, Deadman R, Deloukas P, Dunham A, Dunham I, Durbin R, French L, Grafham D, Gregory S, Hubbard T, Humphray S, Hunt A, Jones M, Lloyd C, McMurray A, Matthews L, Mercer S, Milne S, Mullikin JC, Mungall A, Plumb R, Ross M, Shownkeen R, Sims S, Waterston RH, Wilson RK, Hillier LW, McPherson JD, Marra MA, Mardis ER, Fulton LA, Chinwalla AT, Pepin KH, Gish WR, Chissoe SL, Wendl MC, Delehaunty KD, Miner TL, Delehaunty A, Kramer JB, Cook LL, Fulton RS, Johnson DL, Minx PJ, Clifton SW, Hawkins T, Branscomb E, Predki P, Richardson P, Wenning S, Slezak T, Doggett N, Cheng JF, Olsen A, Lucas S, Elkin C, Uberbacher E, Frazier M, Gibbs RA, Muzny DM, Scherer SE, Bouck JB, Sodergren EJ, Worley KC, Rives CM, Gorrell JH, Metzker ML, Naylor SL, Kucherlapati RS, Nelson DL, Weinstock GM, Sakaki Y, Fujiyama A, Hattori M, Yada T, Toyoda A, Itoh T, Kawagoe C, Watanabe H, Totoki Y, Taylor T, Weissenbach J, Heilig R, Saurin W, Artiguenave F, Brottier P, Bruls T, Pelletier E, Robert C, Wincker P, Smith DR, Doucette-Stamm L, Rubenfield M, Weinstock K, Lee HM, Dubois J, Rosenthal A, Platzer M, Nyakatura G, Taudien S, Rump A, Yang H, Yu J, Wang J, Huang G, Gu J, Hood L, Rowen L, Madan A, Qin S, Davis RW, Federspiel NA, Abola AP, Proctor MJ, Myers RM, Schmutz J, Dickson M, Grimwood J, Cox DR, Olson MV, Kaul R, Raymond C, Shimizu N, Kawasaki K, Minoshima S, Evans GA, Athanasiou M, Schultz R, Roe BA, Chen F, Pan H, Ramser J, Lehrach H, Reinhardt R, McCombie WR, de la Bastide M, Dedhia N, Blöcker H, Hornischer K, Nordsiek G, Agarwala R, Aravind L, Bailey JA, Bateman A, Batzoglou S, Birney E, Bork P, Brown DG, Burge CB, Cerutti L, Chen HC, Church D, Clamp M, Copley RR, Doerks T, Eddy SR, Eichler EE, Furey TS, Galagan J, Gilbert JG, Harmon C, Hayashizaki Y, Haussler D, Hermjakob H, Hokamp K, Jang W, Johnson LS, Jones TA, Kasif S, Kaspryzk A, Kennedy S, Kent WJ, Kitts P, Koonin EV, Korf I, Kulp D, Lancet D, Lowe TM, McLysaght A, Mikkelsen T, Moran JV, Mulder N, Pollara VJ, Ponting CP, Schuler G, Schultz J, Slater G, Smit AF, Stupka E, Szustakowski J, Thierry-Mieg D, Thierry-Mieg J, Wagner L, Wallis J, Wheeler R, Williams A, Wolf YI, Wolfe KH, Yang SP, Yeh RF, Collins F, Guyer MS, Peterson J, Felsenfeld A, Wetterstrand KA, Patrinos A, Morgan MJ, de Jong P, Catanese JJ, Osoegawa K, Shizuya H, Choi S, Chen YJ, International Human Genome Sequencing Consortium (2001) Initial sequencing and analysis of the human genome. Nature 409:860–921
40. Ley TJ, Mardis ER, Ding L, Fulton B, McLellan MD, Chen K, Dooling D, Dunford-Shore BH, McGrath S, Hickenbotham M, Cook L, Abbott R, Larson DE, Koboldt DC, Pohl C, Smith S, Hawkins A, Abbott S, Locke D, Hillier LW, Miner T, Fulton L, Magrini V, Wylie T, Glasscock J, Conyers J, Sander N, Shi X, Osborne JR, Minx P, Gordon D, Chinwalla A, Zhao Y, Ries RE, Payton JE, Westervelt P, Tomasson MH, Watson M, Baty J, Ivanovich J, Heath S, Shannon WD, Nagarajan R, Walter MJ, Link DC, Graubert TA, DiPersio JF, Wilson RK (2008) DNA sequencing of a cytoenetically normal acute myeloid leukaemia genome. Nature 456:66–72
41. Liu B, Zhou Z (2008) Lamin A/C, laminopathies and premature ageing. Histol Histopathol 23:747–763
42. Ly DH, Lockhart DJ, Lerner RA, Schultz PG (2000) Mitotic misregulation and human aging. Science 287:2486–2492
43. MacBeath G (2002) Protein microarrays and proteomics. Nat Genet 32:526–532
44. MacBeath G, Schreiber SL (2000) Printing proteins as microarrays for high-throughput function determination. Science 289:1760–1763
45. Maher CA, Kumar-Sinha C, Cao X, Kalyana-Sundaram S, Han B, Jing X, Sam L, Barrette T, Palanisamy N, Chinnaiyan AM (2009) Transcriptome sequencing to detect gene fusions in cancer. Nature 458:97–101
46. Mardis ER (2008) The impact of next-generation sequencing technology on genetics. Trends Genet 24:133–141
47. Margulies M, Egholm M, Altman WE, Attiya S, Bader JS, Bemben LA, Berka J, Braverman MS, Chen YJ, Chen Z, Dewell SB, Du L, Fierro JM, Gomes XV, Godwin BC, He W, Helgesen S, Ho CH, Irzyk GP, Jando SC, Alenquer ML, Jarvie TP, Jirage KB, Kim JB, Knight JR, Lanza JR, Leamon JH, Lefkowitz SM, Lei M, Li J, Lohman KL, Lu H, Makhijani VB, McDade KE, McKenna MP, Myers EW, Nickerson E, Nobile JR, Plant R, Puc BP, Ronan MT, Roth GT, Sarkis GJ, Simons JF, Simpson JW, Srinivasan M, Tartaro KR, Tomasz A, Vogt KA, Volkmer GA, Wang SH, Wang Y, Weiner MP, Yu P, Begley RF, Rothberg JM (2005) Genome sequencing in microfabricated high-density picolitre reactors. Nature 437:376–380
48. Michiels S, Koscielny S, Hill C (2005) Prediction of cancer outcome with microarrays: a multiple random validation strategy. Lancet 365:488–492
49. Nyrén P, Pettersson B, Uhlén M (1993) Solid phase DNA minisequencing by an enzymatic luminometric inorganic pyrophosphate detection assay. Anal Biochem 208: 171–175
50. Preker P, Nielsen J, Kammler S, Lykke-Andersen S, Christensen MS, Mapendano CK, Schierup MH, Jensen TH (2008) RNA exosome depletion reveals transcription upstream of active human promoters. Science 322:1851–1854
51. Quijano-Roy S, Mbieleu B, Bönnemann CG, Jeannet PY, Colomer J, Clarke NF, Cuisset JM, Roper H, De Meirleir L, D'Amico A, Ben Yaou R, Nascimento A, Barois A, Demay L, Bertini E, Ferreiro A, Sewry CA, Romero NB, Ryan M, Muntoni F, Guicheney P, Richard P, Bonne G, Estournet B (2008) De novo LMNA mutations cause a new form of congenital muscular dystrophy. Ann Neurol 64:177–186
52. Rogers YH, Venter JC (2005) Genomics: massively parallel sequencing. Nature 437:326–327
53. Sanger F, Air GM, Barrell BG, Brown NL, Coulson AR, Fiddes CA, Hutchison CA, Slocombe PM, Smith M (1977) Nucleotide sequence of bacteriophage phi X174 DNA. Nature 265:687–695

54. Sanger F, Nicklen S, Coulson AR (1977) DNA sequencing with chain-terminating inhibitors. Proc Natl Acad Sci USA 74:5463–5467
55. Scaffidi P, Misteli T (2005) Reversal of the cellular phenotype in the premature aging disease Hutchinson-Gilford Progeria syndrome. Nat Med 11:440–445
56. Schena M, Shalon D, Davis RW, Brown PO (1995) Quantitative monitoring of gene expression patterns with a complementary DNA microarray. Science 270:467–470
57. Seila AC, Calabrese JM, Levine SS, Yeo GW, Rahl PB, Flynn RA, Young RA, Sharp PA (2008) Divergent transcription from active promoters. Science 322:1849–1851
58. Shendure J, Ji H (2008) Next-generation DNA sequencing. Nat Biotechnol 26:1135–1145
59. Shendure J, Porreca GJ, Reppas NB, Lin X, McCutcheon JP, Rosenbaum AM, Wang MD, Zhang K, Mitra RD, Church GM (2005) Accurate multiplex polony sequencing of an evolved bacterial genome. Science 309:1728–1732
60. Smith M, Brown NL, Air GM, Barrell BG, Coulson AR, Hutchison CA 3 rd, Sanger F (1977) DNA sequence at the C termini of the overlapping genes A and B in bacteriophage phi X174. Nature 265:702–705
61. Tegner J, Bjorkegen J (2007) Perturbations to uncover gene networks. Trends Genet 23:34–41
62. Tomlins SA, Rhodes DR, Perner S, Dhanasekaran SM, Mehra R, Sun XW, Varambally S, Cao X, Tchinda J, Kuefer R, Lee C, Montie JE, Shah RB, Pienta KJ, Rubin MA, Chinnaiyan AM (2005) Recurrent fusion of TMPRSS2 and ETS transcription factor genes in prostate cancer. Science 310:644–648
63. van de Vijver MJ, He YD, LJ van't Veer, Dai H, Hart AA, Voskuil DW, Schreiber GJ, Peterse JL, Roberts C, Marton MJ, Parrish M, Atsma D, Witteveen A, Glas A, Delahaye L, van der Velde T, Bartelink H, Rodenhuis S, Rutgers ET, Friend SH, Bernards R (2002) A gene-expression signature as a predictor of survival in breast cancer. N Engl J Med 347:1999–2009
64. van't Veer LJ, Dai H, van de Vijver MJ (2002) Gene expression profiling predicts clinical outcome of breast cancer. Nature 415:530–536
65. Velculescu VE, Zhang L, Vogelstein B, Kinzler KW (1995) Serial analysis of gene expression. Science 270:484–487
66. Velculescu VE, Vogelstein B, Kinzler KW (2000) Analysing uncharted transcriptomes with SAGE. Trends Genet 16:423–425
67. Venter JC, Adams MD, Myers EW, Li PW, Mural RJ, Sutton GG, Smith HO, Yandell M, Evans CA, Holt RA, Gocayne JD, Amanatides P, Ballew RM, Huson DH, Wortman JR, Zhang Q, Kodira CD, Zheng XH, Chen L, Skupski M, Subramanian G, Thomas PD, Zhang J, Gabor Miklos GL, Nelson C, Broder S, Clark AG, Nadeau J, McKusick VA, Zinder N, Levine AJ, Roberts RJ, Simon M, Slayman C, Hunkapiller M, Bolanos R, Delcher A, Dew I, Fasulo D, Flanigan M, Florea L, Halpern A, Hannenhalli S, Kravitz S, Levy S, Mobarry C, Reinert K, Remington K, Abu-Threideh J, Beasley E, Biddick K, Bonazzi V, Brandon R, Cargill M, Chandramouliswaran I, Charlab R, Chaturvedi K, Deng Z, Di Francesco V, Dunn P, Eilbeck K, Evangelista C, Gabrielian AE, Gan W, Ge W, Gong F, Gu Z, Guan P, Heiman TJ, Higgins ME, Ji RR, Ke Z, Ketchum KA, Lai Z, Lei Y, Li Z, Li J, Liang Y, Lin X, Lu F, Merkulov GV, Milshina N, Moore HM, Naik AK, Narayan VA, Neelam B, Nusskern D, Rusch DB, Salzberg S, Shao W, Shue B, Sun J, Wang Z, Wang A, Wang X, Wang J, Wei M, Wides R, Xiao C, Yan C, Yao A, Ye J, Zhan M, Zhang W, Zhang H, Zhao Q, Zheng L, Zhong F, Zhong W, Zhu S, Zhao S, Gilbert D, Baumhueter S, Spier G, Carter C, Cravchik A, Woodage T, Ali F, An H, Awe A, Baldwin D, Baden H, Barnstead M, Barrow I, Beeson K, Busam D, Carver A, Center A, Cheng ML, Curry L, Danaher S, Davenport L, Desilets R, Dietz S, Dodson K, Doup L, Ferriera S, Garg N, Gluecksmann A, Hart B, Haynes J, Haynes C, Heiner C, Hladun S, Hostin D, Houck J, Howland T, Ibegwam C, Johnson J, Kalush F, Kline L, Koduru S, Love A, Mann F, May D, McCawley S, McIntosh T, McMullen I, Moy M, Moy L, Murphy B, Nelson K, Pfannkoch C, Pratts E, Puri V, Qureshi H, Reardon M, Rodriguez R, Rogers YH, Romblad D, Ruhfel B, Scott R, Sitter C, Smallwood M, Stewart E, Strong R, Suh E, Thomas R, Tint NN, Tse S, Vech C, Wang G, Wetter J, Williams S, Williams M, Windsor S, Winn-Deen E, Wolfe K, Zaveri J, Zaveri K, Abril JF, Guigó R, Campbell MJ, Sjolander KV, Karlak B, Kejariwal A, Mi H, Lazareva B, Hatton T, Narechania A, Diemer K, Muruganujan A, Guo N, Sato S, Bafna V, Istrail S, Lippert R, Schwartz R, Walenz B, Yooseph S, Allen D, Basu A, Baxendale J, Blick L, Caminha M, Carnes-Stine J, Caulk P, Chiang YH, Coyne M, Dahlke C, Mays A, Dombroski M, Donnelly M, Ely D, Esparham S, Fosler C, Gire H, Glanowski S, Glasser K, Glodek A, Gorokhov M, Graham K, Gropman B, Harris M, Heil J, Henderson S, Hoover J, Jennings D, Jordan C, Jordan J, Kasha J, Kagan L, Kraft C, Levitsky A, Lewis M, Liu X, Lopez J, Ma D, Majoros W, McDaniel J, Murphy S, Newman M, Nguyen T, Nguyen N, Nodell M, Pan S, Peck J, Peterson M, Rowe W, Sanders R, Scott J, Simpson M, Smith T, Sprague A, Stockwell T, Turner R, Venter E, Wang M, Wen M, Wu D, Wu M, Xia A, Zandieh A, Zhu X (2001) The sequence of the human genome. Science 291:1304–1351
68. Voelkerding KV, Dames SA, Durtschi JD (2009) Next-generation sequencing: from basic research to diagnostics. Clin Chem 55:641–658
69. Wang J, Wang W, Li R, Li Y, Tian G, Goodman L, Fan W, Zhang J, Li J, Zhang J, Guo Y, Feng B, Li H, Lu Y, Fang X, Liang H, Du Z, Li D, Zhao Y, Hu Y, Yang Z, Zheng H, Hellmann I, Inouye M, Pool J, Yi X, Zhao J, Duan J, Zhou Y, Qin J, Ma L, Li G, Yang Z, Zhang G, Yang B, Yu C, Liang F, Li W, Li S, Li D, Ni P, Ruan J, Li Q, Zhu H, Liu D, Lu Z, Li N, Guo G, Zhang J, Ye J, Fang L, Hao Q, Chen Q, Liang Y, Su Y, San A, Ping C, Yang S, Chen F, Li L, Zhou K, Zheng H, Ren Y, Yang L, Gao Y, Yang G, Li Z, Feng X, Kristiansen K, Wong GK, Nielsen R, Durbin R, Bolund L, Zhang X, Li S, Yang H, Wang J (2008) The diploid genome sequence of an Asian individual. Nature 456:60–65
70. Wheeler DA, Srinivasan M, Egholm M, Shen Y, Chen L, McGuire A, He W, Chen YJ, Makhijani V, Roth GT, Gomes X, Tartaro K, Niazi F, Turcotte CL, Irzyk GP, Lupski JR, Chinault C, Song XZ, Liu Y, Yuan Y, Nazareth L, Qin X, Muzny DM, Margulies M, Weinstock GM, Gibbs RA, Rothberg JM (2008) The complete genome of an individual by massively parallel DNA sequencing. Nature 452:872–876
71. Zhang B, Horvath S (2005) A general framework for weighted gene co-expression network analysis. Stat Appl Genet Mol Biol 4:Article17