The Decoding of Eukaryotic Genomes

P.M. Selzer et al., *Applied Bioinformatics*, https://doi.org/10.1007/978-3-319-68301-0_4

4.1 The Sequencing of Complete Genomes

A new era in genome research started in 1995 with the publication of the first completely sequenced bacterial genome from the human pathogen *Haemophilus influenzae*. For the first time one could analyze a complete genome, including both genes and their regulatory regions. Three years later, the sequencing of the first multicellular eukaryotic genome, from the nematode *Caenorhabditis elegans,* was completed. Eukaryotic genomes are larger and far more complex than those from bacteria (▶ Chap. 7). A comparison of eukaryotic and prokaryotic genomes demonstrated that genes encoding proteins constitute a much smaller proportion of the eukaryotic genome. Thus, in humans and mice just 1.4% of the genome actually encodes genes, and only 5% of both genomes are highly conserved even though both share approximately 80% gene orthology. In addition to protein-encoding genes, conserved regions contain important regulatory elements, non-protein-encoding genes, and regions important for chromosome structure. For the greater proportion of the genome, however, there are few data regarding function (Mouse Genome Sequencing Consortium 2002).

The relatively low number of genes identified in the human genome was at first surprising. At the beginning of the human genome sequencing project, it had been estimated that the number of genes would be on the order of 100,000–150,000. To date, however, only 19,000–20,000 genes have been demonstrated (Ezkurdia et al. 2014). A similar number of genes were also estimated for the mouse genome. Interestingly, humans possess only about 3,000 genes more than the nematode *C. elegans*. In view of the fact that a human body contains several billion cells, whereas *C. elegans* has just 959 somatic cells, this small difference in the number of genes is remarkable.

4.2 Characterization of Genomes Using STS and EST Sequences

4.2.1 Sequence-Tagged Sites are Landmarks in the Human Genome

Sequencing the entire human genome was a huge achievement. More than three billion nucleotides had to be sequenced and assembled in the right order. In a sense, the project could be compared to assembling a large jigsaw puzzle. It was first necessary to establish landmarks in the genome to allow for the correct placement of sequence regions. The most important landmarks in the genome are sequence-tagged sites (STSs), short DNA sequences 200–500 nucleotides long that are present only once in the genome of an organism. STSs are generated by the polymerase chain reaction (PCR), a method for the amplification of specific nucleotide sequences. Because STSs are unique, they can always be specifically amplified by PCR from genomic DNA.

DNA clones are examined by database searches for the existence of matching STS regions and then positioned on chromosomes or in genomes. Using this approach, a precise physical map of the humane genome could be generated.

A database dedicated to STSs has existed since 1994; this is the dbSTS [dbsts], which was transferred to a division of GenBank in 2013. Here, one can find all the information

available for individual STSs, including the STS name, sequences of the oligonucleotides necessary for PCR amplification, size of the PCR product, conditions for the PCR, and the nucleotide sequence of the STS.

Shortly following publication of the concept of STS-based mapping in 1989, it was recognized that STSs could also be generated from complementary DNA (cDNA) clones. Such cDNA clones originate from cellular mRNA and, thus, correspond to the expressed genes of a cell. In addition to genome mapping, STSs derived from cDNA can also be used to localize genes within a genome. Indeed, by 1996, a genetic map of the human genome had been assembled.

4.2.2 Expressed Sequence Tags

It was soon realized that partial sequences of cDNA clones could also be used in the discovery of new genes (Adams et al. 1991). Because cDNA clones are derived from expressed genes, the sequences were called expressed sequence tags (ESTs). ESTs are generated by the end-sequencing of cDNAs (Fig. 4.1). ESTs are easy to produce at a reasonable price, and many EST projects have resulted in the identification of new genes. However, the concept of EST sequencing also met with opposition. Critics noted that sequencing just cDNA would miss important and nonexpressed gene regulatory regions. Second, some ESTs are just too short to assign a gene function, and, finally,

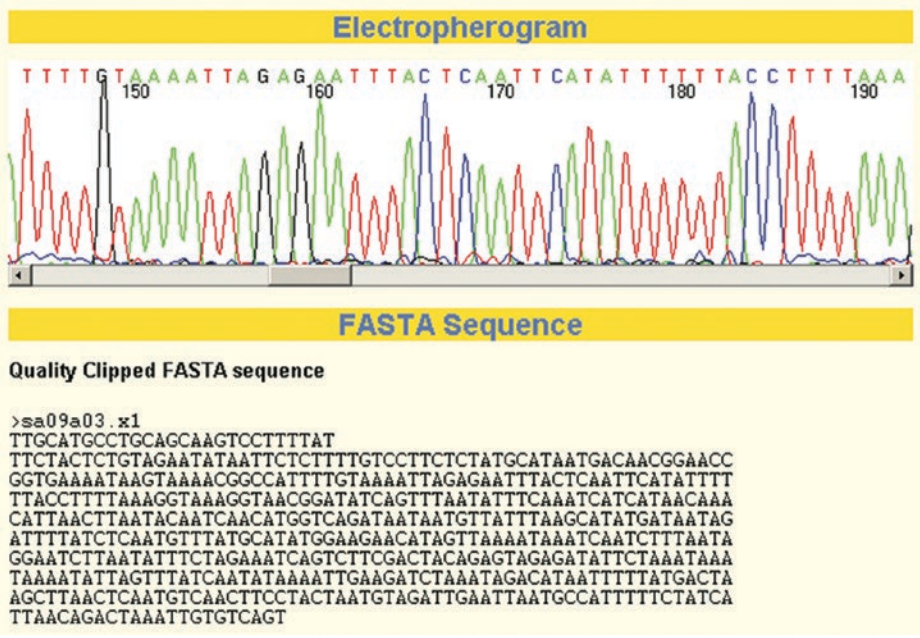

Fig. 4.1 Section of electropherogram from dideoxy DNA sequencing reaction with corresponding nucleotide sequence of expressed sequence tag (Clipping from Ensembl database, printed with kind permission from EBI, Hinxton, UK)

ESTs, being automatically generated, can be of poor sequence quality. Frequently, not just nucleotide changes occur, but also base insertions and deletions that lead to frameshift errors. Simply, it was feared that many public EST databases would be of poor quality.

Despite these criticisms, EST projects became widely accepted. In particular, the speed with which ESTs could be generated on a high-throughput scale (owing to the automation of DNA sequencing technology and plasmid DNA production) resulted in a real boom for EST projects. Important EST projects were initiated at Washington University (WU) [washington], for example. In collaboration with the American pharmaceutical company Merck & Co., Inc., in Kenilworth, New Jersey, USA, WU sequenced 580,000 human ESTs between 1995 and 1997. These ESTs were generated from cDNA libraries that had been made available by the Integrated Molecular Analysis of Genomes and their Expression (IMAGE) consortium, which is a merger of several academic research groups that produce high-quality cDNA libraries and make them available for other research, such as EST projects. The IMAGE consortium has the largest collection of publicly available cDNA libraries worldwide [image].

As a reaction to the huge increase in EST data, dbEST [dbest] was established at NCBI to collect all publicly accessible ESTs. In 1993, less than 50,000 sequences were stored in dbEST; today, however, more than 74 million ESTs from over 2,400 organisms are stored in this database (dbEST release 130101, January 2017). One drawback of dbEST is that it contains redundant ESTs, especially for strongly expressed genes like actin. For this reason, the UniGene database [unigene] was established in which all cDNAs and ESTs that originate from an identical gene are combined into a group or cluster. The result is a reduction in the number of entries down to the actual number of proteins produced in an organism. Because of its nonredundancy, UniGene is a useful basis for other databanks such as ProtEST and HomoloGene [homologene]. ProtEST is integrated into UniGene and provides information on whether cDNAs and ESTs that are assigned to a UniGene cluster are similar to known protein sequences upon translation. In contrast, the independent database HomoloGene provides information on whether human UniGene clusters have homologs in other species.

Another NCBI database, dbGSS [dbgss], stores Genome Survey Sequences (GSSs). Like ESTs, GSSs are partial nucleotide sequences with a length of up to 1,000 bases and generated by end-sequencing individual clones. The difference between GSSs and ESTs is the nucleic acid source material: GSSs are prepared from genomic libraries, whereas cDNA libraries are used for ESTs. Thus, GSSs differ from ESTs by potentially containing DNA fragments that lie outside of areas encoding genes. More than 35 million sequences from more than 1,000 organisms are stored in dbGSS (dbGSS release 130101, January 2017).

Although the importance of EST projects has diminished over the years, we will present a brief overview of how an EST project was done because the principal approach is similar to the modern high-throughput sequencing (► Sect. 4.6.3). The similarity of the two approaches is easy to see if we compare an EST project to whole transcriptome shotgun sequencing, also known as RNA-Seq (Wang et al. 2009). Both approaches start with the generation of a cDNA library. In addition, the following steps of high-throughput sequencing are easier to understand if we keep in mind how an EST project works.

4.3 EST Project Implementation

At the beginning of an EST project, the starting material for the construction of a cDNA library is selected. This can be cells, specific tissues, or even whole organisms (Fig. 4.2). From this material total RNA is isolated, which predominantly comprises ribosomal RNA (rRNA), transfer RNA (tRNA), and messenger RNA (mRNA). The most

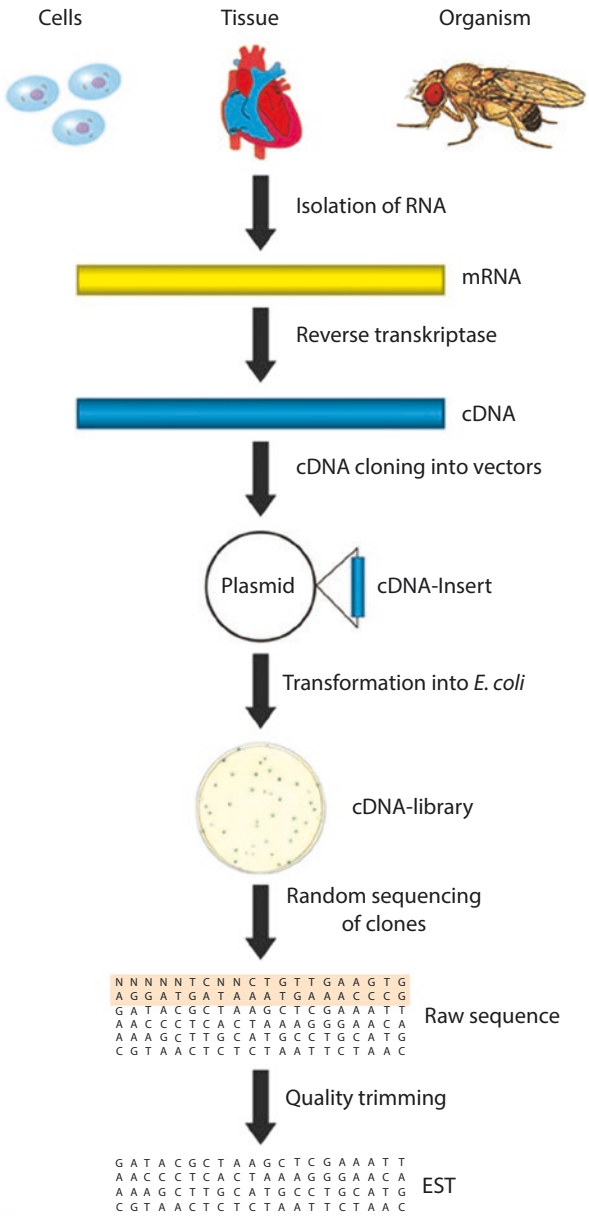

Fig. 4.2 Diagram for establishment of cDNA library and generation of EST sequences (*Drosophila melanogaster* from Patterson JT, Univ. Texas Publs 4313, 1943, printed with kind permission from the University of Texas; Heart from Schmidt, Thews Lang, Physiologie des Menschen, 28th edition 2000, printed with kind permission from Springer-Verlag, Heidelberg, Germany)

interesting of these in the construction of a cDNA library is mRNA because it represents all active genes of a given cell or tissue. It is present only in very small amounts (approx. 3% of the total RNA). The very unstable mRNA is transcribed into the considerably more stable cDNA by the viral enzyme reverse transcriptase. The cDNA is then cloned into plasmids that serve as vectors. Usually cDNAs are cloned directionally, i.e., it is known at which end of the vector the 5′ and 3′ ends of the cDNA are located. Plasmids are amplified by transforming the bacterium *Escherichia coli,* resulting in the desired cDNA library, which can then provide the basis for generating EST sequences. The transformed bacteria are plated and grown on nutrient media, and plasmid DNA is isolated from randomly selected individual clones. The cloned cDNA can then be sequenced either from the 5′ or 3′ end or from both ends simultaneously. The identified nucleotide sequence is then exported to a computer, and the raw data are bioinformatically processed.

The quality of the data is first checked in a process called quality trimming. For example, quality trimming defines the minimum length that an EST must have and what number of ambiguous nucleotides (variable N) is allowed relative to the nonambiguous nucleotides (A/T/G/C). Modern sequencers permit the computation of quality scores that are a measure of the quality of the sequencing of each individual nucleotide. Using these values, sequence regions of poor quality, e.g., the ends of sequences, are removed. Finally, any contamination with sequences from vector and bacteria are also removed.

Curated ESTs are a collection of random cDNA sequences of different lengths, and many are derived from identical transcripts. Many ESTs will be found particularly for highly expressed genes. To eliminate redundancy, therefore, alignments of these ESTs are generated to form overlapping sequences that are as long as possible (◘ Fig. 4.3). These consensus sequences are compared again to other ESTs so that further identical ESTs are incorporated into the alignment. This iterative process is described as sequence assembly. Often sequence assembly programs such as CAP3 [cap] and Phrap [phrap] are used. Sequence assemblies are either contigs whose sequences correspond to the consensus sequences of the alignments or singletons that are not similar to other ESTs and, therefore, cannot be grouped into contigs.

For large EST data sets, it can be useful to subdivide ESTs into groups or clusters first. Those clusters displaying identical nucleotides for a given region are summarized into groups. Finally, within these groups, a more stringent sequence assembly is performed to generate consensus sequences. Thus, ESTs that descend from alternatively spliced forms are arranged into the same clusters, but different contigs, better depicting the EST relationships. One useful program for sequence clustering is stackPACK [stackpack].

4.4 Identification of Unknown Genes

Once ESTs are arranged into contigs, the corresponding consensus sequences can be used to identify unknown genes. For this purpose, annotation and sequence searches are carried out against various databases.

ESTs are usually first annotated, i.e., a potential function is assigned to both the level of the single ESTs and the assembled contigs by comparison with existing proteins of known function. Usually the BLASTx algorithm is applied whereby the EST nucleotide sequences are first translated into all six reading frames. This process is shown in

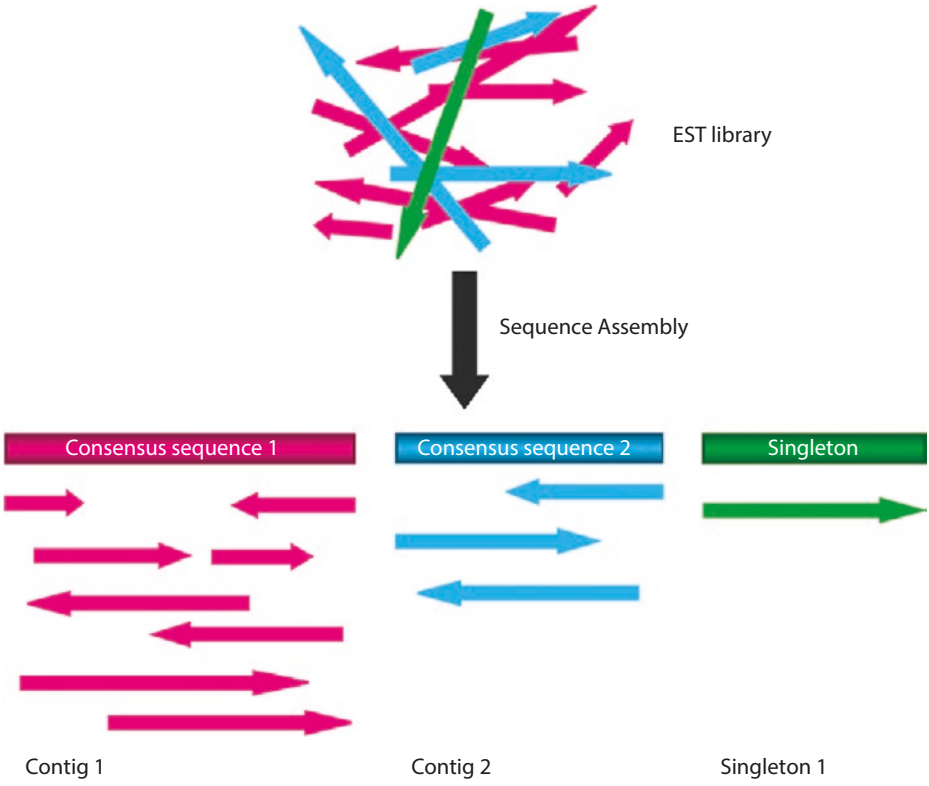

Fig. 4.3 Classification of ESTs into contigs and the formation of consensus sequences

Fig. 4.4 using an EST sequence obtained from bovine intestine. The EST was annotated by BLASTx against a nonredundant protein database and shows high similarity with part of murine caspase 6. Caspases are proteases that function during programmed cell death (apoptosis). Because of the similarity to caspase, it can be inferred that the gene transcript from which the EST is derived encodes either a true caspase or a protein containing a caspase domain. It is important to state that ESTs are usually partial gene sequences, and therefore, alignments may not contain the entire length of a deduced protein. Indeed, ESTs often encode only the untranslated region (UTR) of mRNA, and such ESTs are known as nonencoding ESTs (Fig. 4.5). These difficulties can be avoided, however, when ESTs are extended by sequence assembly, sometimes to the point where the entire protein can be identified.

By direct comparison of EST sequences between different organisms, similar or even new genes or proteins may also be identified. Generally, however, it is not advisable to attempt this at the nucleotide level (e.g., with BLASTn) because little similarity exists between species due to species-dependent codon usage (▶ Chaps. 1 and 7). However, sequences normally show greater conservation at the amino acid level. Therefore, sequences should be compared after translation of the nucleotide sequence into all six reading frames. For this, tBLASTx, which automatically carries out both the translation and the database comparison, is a good choice (▶ Chap. 3). However, when large

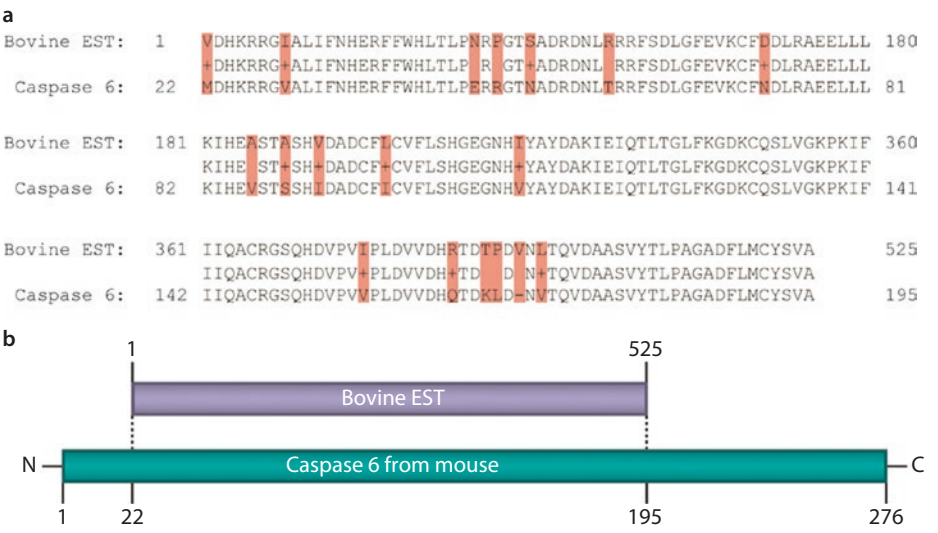

Fig. 4.4 Annotation of an EST sequence from bovine intestine. **a** The translated EST has an identity of 89% over a length of 175 amino acids (525 nucleotides) with murine caspase 6. Sequence differences are highlighted in red. The numbering of the EST sequence from 1 to 525 refers to nucleotides. In contrast, the numbering of the caspase 6 protein from 22 to 195 refers to amino acids. **b** Schematic of alignment of EST sequence with sequence of murine caspase 6

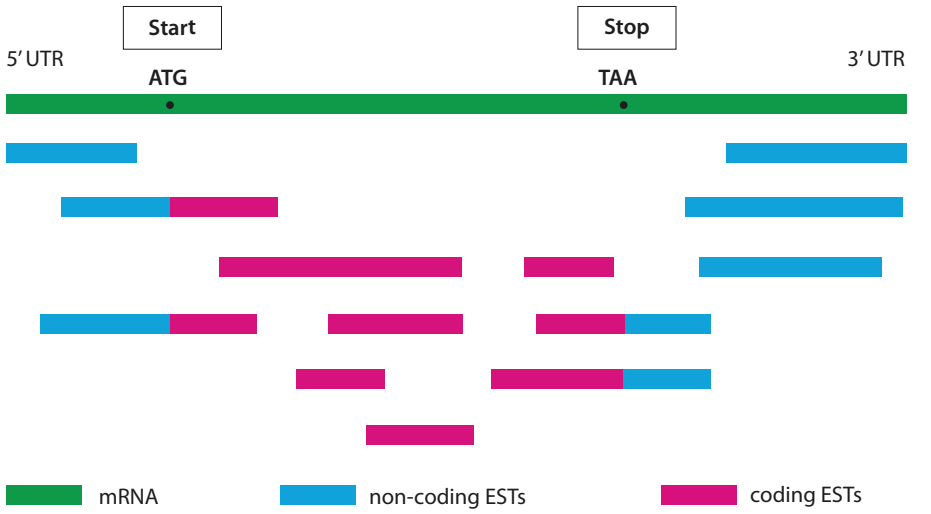

Fig. 4.5 EST sequences are derived from coding and noncoding segments of an mRNA

databases are being queried, some time may be needed. An interesting example of a large-scale comparison is the evaluation of EST sequences from different parasitic worms. At WU in St. Louis, a parasitic nematode sequencing project is under way in which more than 300,000 ESTs of different parasitic threadworms are being sequenced [nematode]. By comparison of these data sets it will be possible to find genes that are

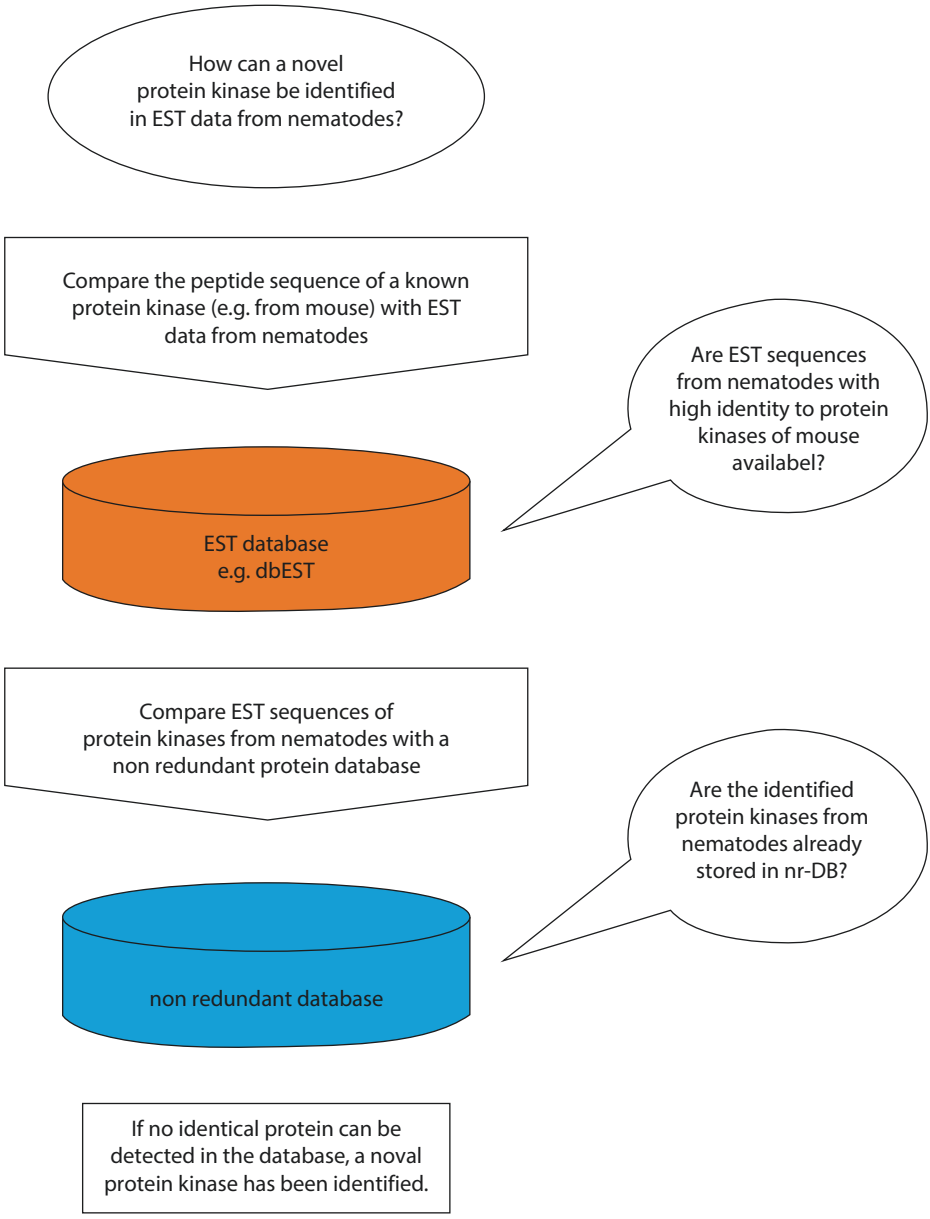

Fig. 4.6 Strategy for identification of new members of protein families

ubiquitous in all nematodes. Such an approach has been used to clarify evolutionary relationships within the phylum Nematoda (Blaxter 1998).

Using EST data, new members of a protein family can also be identified. The procedure to identify new protein kinases in the nematode EST data set is shown in Fig. 4.6. To start, one compares the peptide sequence of a known protein kinase (e.g., from mouse) with an EST database (e.g., dbEST). If a nematode EST sequence

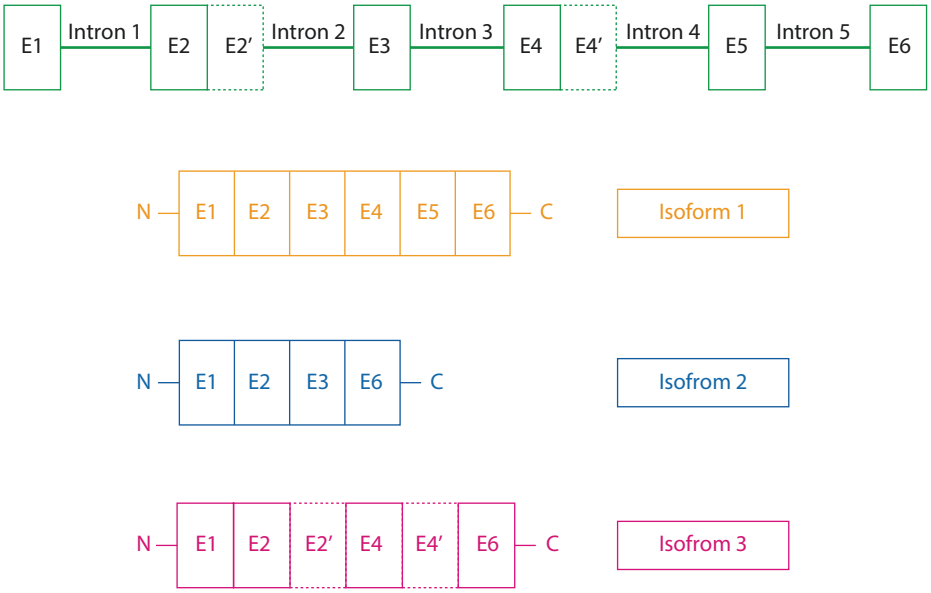

Fig. 4.7 Alternative splicing. The generation of several mRNA transcripts from a single gene by the combination of different exons (E) is called alternative splicing

of high identity to the mouse kinase is found, then it is likely that the EST encodes a protein kinase. To determine whether the identified protein kinase is novel, the ESTs must be compared with a nonredundant protein or nucleotide database. If no identical sequences are identified, then a new member of the protein kinase family has indeed been found.

4.5 The Discovery of Splice Variants

In addition to helping identify new genes, ESTs can also identify alternative gene splice variants. Alternative splice variants can arise upon gene transcription and during the processing of the RNA primary transcript. During splicing, noncoding introns are removed from the primary transcript and the remaining exons joined to form the mature mRNA (▶ Chap. 1). During alternative splicing, one exon can be replaced by another, thereby creating a new mRNA. In this way, different mRNAs, encoding different proteins, can arise from a single primary RNA transcript (Fig. 4.7). Alternative splicing, therefore, is an efficient strategy for producing several proteins from one gene. It is believed that alternative splice forms exist for one-third to two-thirds of all human genes (Yeo et al. 2004). For example, two mRNA transcripts are known for the F_c receptor that is important in immunology. During alternative splicing the cytoplasmic domain of the receptor is exchanged for a second form. Because the individual cytoplasmic domains are crucial for signal transduction, alternative splicing generates domains with very different cellular functions.

ESTs derived from fully processed mRNAs can give valuable hints as to the identification of unknown splice variants. ESTs are compared with nucleotide databases that contain information for mRNA transcripts (e.g., GenBank) or with protein databases (e.g., UniProt). In cases where otherwise identical sequences are found to differ in a few regions, e.g., by insertions or deletions, this can be evidence for alternatively spliced variants. Through such EST comparisons with known sequences in public databases, numerous alternatively spliced gene variants have already been discovered. At the University of California at Los Angeles, two databases called ASAP and ASAP2 of the Alternative Splicing Annotation Project have been established in which alternatively spliced genes, identified via EST sequences, are stored. Also, many gene prediction programs such as GrailEXP use EST sequences to correctly predict genes from sequenced genomes and derive information regarding splice sites [grailexp].

4.6 Genetic Causes for Individual Differences

A characteristic of eukaryotic genomes is the presence of mutations or genetic variations. These variations are responsible for the individual differences in a population. The most frequent variations are single-nucleotide polymorphisms (SNPs) caused by the exchange of a single nucleotide. Other polymorphisms are short deletions and insertions (deletion insertion polymorphisms) and variations due to repetitive sequences (short tandem repeats).

A consortium of commercial and noncommercial institutions has identified almost 1.8 million SNPs in the human genome (Thorisson and Stein 2003). Many of these SNPs lie outside genes and, therefore, do not alter cellular function. However, other SNPs lie within genes and are responsible for the occurrence of phenotypes. Example phenotypes are the color of eyes or hair, but also disease conditions. Functionally important SNPs are discovered by comparing the appearance of a phenotype with the frequency of a specific SNP. If a correlation is found, it is likely that this SNP is responsible for the phenotype. Because individuals are randomly selected for these correlation analyses, the strategy is simpler and faster than classical pedigree analyses, in which the appearance of phenotypes must be traced back in a family over several generations.

An example of a SNP-based disease is phenylketonuria in which the degradation of phenylalanine is disrupted. Point mutations in the enzyme phenylalanine hydroxylase lead to inactivation of the enzyme. Many different SNPs have been discovered in the human phenylalanine hydroxylase enzyme, and these are collated in the database Phenylalanine Hydroxylase Locus Knowledgebase [pahdb]. Because of the missing enzyme activity, phenylalanine accumulates in the brain of newborns and infants and ultimately leads to a mental defect. Newborns are therefore examined in many countries for high blood levels of phenylalanine. Disease symptoms are preventable by a phenylalanine-poor diet, allowing those affected to live a normal life.

Genetic polymorphisms can also be an advantage. One example of this is the differential susceptibility of individuals to infection by the human immunodeficiency virus 1 (HIV-1). In addition to the surface protein CD4, the virus requires additional coreceptors, such as the chemokine receptor CCR5, to enter the cell. A mutant of this receptor with a deletion of 32 nucleotides was discovered in 1996. This mutation leads

to a shift in the reading frame and subsequently to the translation of a nonfunctional protein that is no longer present at the cell surface. Humans who are homozygous for this mutation (both alleles disrupted) are more resistant to HIV-1 infection. Those who are heterozygous for the mutation (one functional allele) will develop AIDS later and have a longer life expectancy than those who lack the frame shift mutation. In the Caucasian population of the USA, this polymorphism is homozygous at a frequency of 1%, with another 20% having a heterozygous allele. Unfortunately, among African and East Asian populations, this polymorphism is found only rarely (Berger et al. 1999).

SNPs are also excellent genomic markers because they are distributed over the entire genome and found at high density (on average every 300–500 nucleotides in the human genome). Moreover, SNPs have a low mutation frequency between generations and are detectable by high-throughput methods. SNPs, therefore, allow for the generation of precise genetic maps of high resolution. This resolution facilitates the discovery of disease genes, particularly if several genes are responsible for the emergence of complex illnesses like cancer or diabetes.

A number of methods exist for the detection of SNPs or genotyping. Microarray genotyping is based on the principle that the denaturation temperature of hybridized DNA strands will decrease if nonidentical nucleotides are present. The advantage of this high-throughput method is that it allows for the simultaneous and parallel analysis of many sequences. Other techniques for identifying SNPs are based on enzymatic reactions that show a very high specificity for their substrate and are, thus, more accurate than hybridization-based methods. A commonly used enzyme-based genotyping technique is pyrosequencing [pyrosequencing]. Short DNA segments are sequenced in real time without the necessity for time-consuming gel purification steps. The advantage of this method is that the entire vicinity of the SNP is sequenced and serves as an internal control for the sequencing reaction. An alternative enzyme-based technology is single-base primer extension, which provides precise quantitative results at a moderate cost. Short oligonucleotide sequences hybridize exclusively next to the SNP. These oligonucleotides then serve as primers for polymerases that incorporate a labeled nucleotide at the position of the SNP. The incorporated nucleotide is then detected using colorimetric methods or by mass spectroscopy. Furthermore, SNPs can also be determined in silico, i.e., by graphically comparing similar EST sequences from different individuals. Using such multiple alignments, nucleotide exchanges are very easy to recognize. However, caution is advised when describing new SNPs using ESTs because these can contain sequencing errors interpretable as SNPs.

The dbSNP database is the NCBI repository for polymorphisms [dbsnp]. Each entry contains details of the genetic variation, adjacent nucleotides, and frequency of the polymorphisms. It also includes data about the experimental method and conditions used to identify the SNP. The dbSNP contains approximately 780 million polymorphisms from 53 organisms, of which 545 million are human (September 2016). Moreover, a curated collection of human SNPs can be found in the GWAS Central, formerly known as the Human Genome Variation Database [gwas]. These SNP entries have been subjected to an additional quality check and are completely annotated.

4.6.1 Pharmacogenetics

Pharmacogenetics (or pharmacogenomics) deals with genetic variations that are responsible for how patients differ in their reactions to drugs. A study in the USA in 1994 reported that 2.2 million patients suffered from serious medication side effects and that over 100,000 patients died. Thus, there is a greater chance of dying from drug side effects than from most viral infections. Accordingly, the ability to predict how a patient might react to a drug prior to starting therapy would be a tremendous advance.

How a patient responds to drugs is a complex process involving many different proteins including the receptors and enzymes that bind to and metabolize drugs, respectively. Genetic variations in such proteins can result in decreased or absent drug binding or drug metabolism. Of particular importance are polymorphisms in proteins of the cytochrome P450 family. For example, the enzyme CYP2D6 is responsible for the metabolism of 20–25% of all prescription drugs. Mutations in CYP2D6 can influence the rate at which drugs are metabolized. Depending on the type of mutation, one can distinguish patients with ultrafast, extensive, medium, and slow drug metabolism. Clearly, therefore, genetic polymorphisms greatly influence the individual reactions of patients to drugs. Because SNPs represent by far the most frequent genetic variations, the search for SNPs that influence a drug's effect or metabolism is of central importance to pharmacogenetics.

As stated, a major aim of pharmacogenetics is to predict unwanted side effects of a drug in advance of therapy. An important prerequisite for this is the development of diagnostic tests to understand the genetic predisposition of patients and how they might react to a specific drug. In such diagnostic tests the genotype of every patient is established, i.e., whether relevant proteins such as drug-metabolizing enzymes show distinct polymorphisms. Patients can then be classified into corresponding groups and a suitable therapy selected based on their genotype (◘ Figs. 4.8 and 4.9). This is also referred to as stratified medicine because therapy is optimized and tailored to every patient belonging to a distinct responder group. An example already practiced in many countries is the chemotherapeutic treatment of patients with acute lymphatic leukemia (ALL). Mercaptopurine and thioguanine are frequently used as drugs that are incorporated into the DNA of proliferating cells (especially cancer cells), leading to their eventual death. One enzyme responsible for the metabolism of these compounds is thiopurine-S-methyltransferase. Clinical studies have shown that genetic polymorphisms greatly influence the activity of thiopurine-S-methyltransferase and, therefore, the toxicity and efficacy of mercaptopurine and thioguanine. Patients deficient in thiopurine-S-methyltransferase accumulate these drugs in blood cells at high concentrations, which eventually causes death. By contrast, in patients with high thiopurine-S-methyltransferase activity, mercaptopurine and thioguanine must be used at higher doses. Therefore, each patient is examined for polymorphisms in the gene encoding thiopurine-S-methyltransferase and the most effective dose determined before treatment with mercaptopurine and thioguanine.

In addition to patients in the clinic, pharmacological research has also benefited from pharmacogenetics. Prior to approval for use in patients, every new drug candidate must be tested in extensive clinical studies using the strictest safety and efficacy criteria. Pharmacogenetics offers the possibility of excluding those patients unlikely to react to therapy or who might experience undesired side effects before the start of each

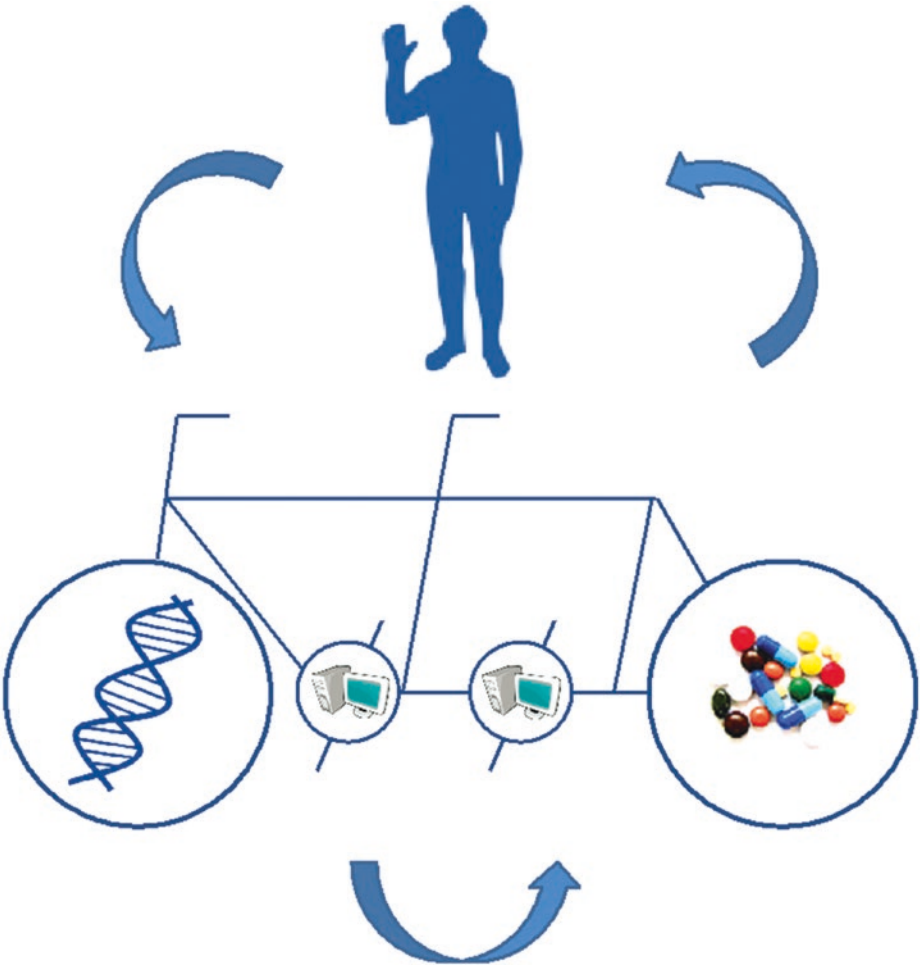

Fig. 4.8 Pharmacogenetics: Diagnosis and therapy are applied in tandem. A patient's genetic predisposition influences the effect of a given drug. Analysis of the patient's genetic predisposition can help in choosing a reasonable drug

study. This process increases the chances that a drug will reach the market through the appropriate selection of patients who will benefit from the drug without unpleasant or even dangerous side effects. A list of drugs that cannot be used unless their suitability for individual patients is tested can be found at the Web site of the *Verband Forschender Pharmaunternehmen* [vfa-personalisiert]. Moreover, pharmacogenetics will allow the development of new drugs for patient groups that do not respond to existing therapies or will allow for the stratification of the therapy. Patients with an ultrafast metabolism who metabolize a given drug extremely fast could be given an alternative drug or a higher dose of the same drug. Accordingly, patients with a slow metabolism, where dangerous plasma levels of the drug could be reached, can be medicated using an alternative drug or a lower dose. Such drug incompatibilities can be observed relatively

Fig. 4.9 Genotyping of patients by detecting SNPs

often for prodrugs, i.e., drugs that are converted to active drugs by being metabolized (e.g., Tamoxifen), as well as for drugs that are not prodrugs (e.g., antidepressants like Mirtazapine). Some companies, like Humatrix AG [humatrix, stratipharm], specialize in diagnostic tests to evaluate the suitability of drugs for individual patients.

It should also be considered that individual reactions to drugs can only be partly explained by genetic variations and that other factors may also influence drug efficacy and safety (Everett 2016). These include the patient's age and nutritional status, the consumption of alcohol, the status of the patient's microbiome, whether additional disease conditions coexist, and whether other drugs are being taken. Moreover, the existence of a genetic variation can, but need not, lead to a metabolic variation.

If we want to increase the success rate of individual medicines, we therefore need to keep in mind not only the patient's genetic predisposition but also her individual metabolic profile. Metabolic profiling using physicochemical methods has been going on for some 10 years now. In recent years, the terms *metabonomics* and *metabolomics* have been introduced. However, the initial studies showed that the metabolic profiles of two patients after medication with the same drug are critically dependent on the patients' metabolic profiles before the medication. The microbiome of a patient's gut influences the metabolism of the drug and, therefore, its effect. This knowledge led to a new discipline, pharmacometabonomics. This discipline predicts the effect of a drug based on a patient's metabolic profile before medication is administered using mathematical models. The application of pharmacogenomics and pharmacometabonomics should lead to a higher quality of individual medicine (Everett 2016).

4.6.2 Personalized Medicine and Biomarkers

The adjustment of a therapy to a patient's genetic predisposition and individual metabolic profile is often also referred to as personalized medicine. This term can be found quite often in the scientific literature starting around 2000, after which time it has gained

importance, although a clear definition of the term remains elusive, which leaves room for interpretation. Schleidgen and coworkers (Schleidgen et al. 2013) derived a common understanding by comparing 653 scientific publications defining the term using lexical approaches. Based on their investigation, personalized medicine strives to optimize stratification, i.e., the evaluation of current risk, and treatment on the basis of a knowledge of biological information and knowledge of biomarkers on the level of molecular metabolic pathways, genetics, proteomics, and metabolomics. Indeed, this definition is somehow cumbersome. In the end, it just means that personal biological properties are considered for the therapy of each individual patient. Special attention is thereby paid to biomarkers, which, among other methods, are determined on a genetic basis.

Biomarkers are simply parameters that can be used for diagnosis, prognosis, and therapy. Commonly known examples are parameters in hemograms, which are used by physicians for diagnosis and to monitor therapeutic success or to modify therapy. Given the great complexity of diseases like cancer, such global biomarkers are no longer sufficient. Biomarkers are needed that allow for a finer-grained picture. Here we find a connection to pharmacogenetics, where knowledge of a patient's genetic predisposition, e.g., a polymorphism in thiopurine-S-methyltransferase, is used as a biomarker to optimize therapy. In a rather similar way, biomarker research looks for regions in genomic DNA, mRNA, or proteins that are correlated with diseases or that can be correlated to the reaction to some therapy. Once such biomarkers are established, they can be used for diagnosis, prognosis, and therapy. Knowledge of a patient's genome is paramount for genetic biomarkers. Until recently, it was impossible to sequence individual genomes. Only with the emergence of next-generation sequencing (NGS) did the sequencing of genomes become a suitable diagnostic method that could be carried out in a few days at a cost of only a few thousand U.S. dollars, instead of taking 10 years and at a cost of approx. USD 3 billion.

How does one identify biomarkers? One possibility is a genome-wide association study (GWAS) [nhgri-gwas]. The goal of a GWAS is to identify alleles that are correlated with given diseases, i.e., alleles that are found if some disease is present and not found in the absence of the disease. If such a correlation is found, at first only an association is established. Whether it is a causative association must be determined by further biochemical and molecular biological studies. For a GWAS, two study groups are formed, one with individuals possessing some special property, e.g., a disease, and one made up of individuals not possessing that property. DNA samples of both groups are then analyzed for genetic variations. Either the whole genome or just special marker areas, i.e., defined SNPs, are analyzed. Recent technical progress in DNA sequencing (► Sect. 4.6.3) and falling costs now allow for sequencing more patient genomes. This makes it possible to use GWASs for diagnostic purposes on the one hand, e.g., in pharmacogenetics (► Sect. 4.6.1), and on the other hand for predictive purposes, e.g., to search for known allele-property associations in a given patient's genome, even if the disease has not yet emerged. The identification of such an association however, does not mean that the disease must eventually emerge; it just means there is a certain probability that it will. For instance, consider hemochromatosis, which is connected to a homozygous mutation in the HFE gene. The probability that the disease will indeed emerge is only 30–50%, i.e., of 100 patents showing the mutation in the HFE gene, only 30–50 patients show clinical signs of the disease. That said, it becomes clear that the ever-growing number of known

human genomes not only has advantages for patients and society but also raises social and ethical questions. How does a patient cope with the knowledge of a risk factor, for instance, or what should insurance companies do with this information? This textbook is not the proper forum for answering such a question; however, it is important to keep in mind that despite all the euphoria over the technical possibilities, public dialog is necessary.

4.6.3 Next-Generation Sequencing (NGS)

As mentioned earlier, NGS allows for the rapid sequencing of whole genomes. Moreover, it is also possible to sequence RNA (RNA-Seq, see ▶ Sect. 4.2.2), identify splice variants and splice sites, quantify mRNA very precisely. It also allows one to study the microbial diversity in humans or the environment. NGS, therefore, has become a very important tool in everyday research. Several methods are available, which follow a similar basic principle. In a first step, a DNA library is created. This library consists of short DNA fragments (fragmentation), which are elongated by short DNA stretches of known sequence on both 5′ and 3′ ends (adaptation). These adapters are used in the next step to fix the DNA fragments to solid reaction media and to amplify them (amplification). In this step, several methods are used that in the end form clusters of identical DNA fragments. The actual sequencing then takes place in the individual clusters. The last step is the data presentation, and all methods present the data in the form of a DNA chip.

The principal difference in the various systems lies in the technical details of the sequencing. In principle, four systems can be defined:

- *Pyrosequencing*: During the sequencing reaction, a pyrophosphate is set free, which via a sequence of chemical reactions leads to the emission of light. This light is detected by a camera. The bases are added subsequently, and the camera detects whther light is emitted. Before adding the next base, a wash step is carried out.
- *Sequencing by synthesis*: This method involves the use of nucleotides that are bound to a terminator and a fluorescent dye. After the nucleotide has been added, the fluorescent dye is excited and the emitted fluorescent wavelength is recorded. Subsequently, the terminator is removed and the next nucleotide can be added.
- *Sequencing by ligation*: Instead of a DNA polymerase, 16 different oligonucleotide probes are used. Each of the nucleotide probes carries one of four different fluorescent dyes at its 5′ end. Each octamer consists of two specific and six general bases. For the sequencing a specific primer is bound to the adapter of the DNA sequence and a fitting probe is ligated using a DNA ligase. After a washing step, the fluorescent signal is recorded and the last three DNA bases and the fluorescent dyes are removed. This is done seven times, followed by a denaturing step. The process is started anew with a primer that binds, shifted by one nucleotide. Five different primers are used in total.
- *Ion semiconductor sequencing*: This method is similar to pyrosequencing. Instead of recording the release of a pyrophosphate, the release of a proton is detected. The DNA clusters are bound to a semiconductor chip capable of measuring the surrounding pH value. Once a nucleotide is added, a single proton is released, leading to a change in the pH value that can be detected by the semiconductor chip.

Each of the methods has advantages and disadvantages, e.g., different read lengths, reagent costs, error rates, acquisition time, and coverage. Coverage means the number of reads in a sequence assembly necessary to reproduce a reference sequence. For a complete genome, the minimum coverage is 30. Today, only two methods can reach this level of coverage for a human genome – sequencing by synthesis and sequencing by ligation. Pyrosequencing is suitable for bacterial genomes and for simple eukaryotes, e.g., *Arabidopsis thaliana* [ngs-movie, ngs-knowledge-base].

The amount of data generated by NGS constitutes a formidable challenge. Even a compressed FASTQ file – a specialized file format that per sequence contains the sequence identifier, the sequence itself as in a FASTA file, and two additional lines, one for comments and one for quality scores – easily reaches a size of 200 GB for a human genome with a coverage of 60. A project with 10–20 genomes therefore uses approximately 4 TB of disk space. Therefore, not only is storage not trivial, but transferring this amount of data between different research groups represents a challenge. A cloud solution seems appropriate. The National Institutes of Health (NIH) has two cloud systems called Biowolf and Helix [nih-biowolf]. In Europe, the EMBL is working on a cloud system based on the Helix Nebula cloud [helix-nebula].

Another challenge is the realignment or mapping of short reads to the reference genome. Because of the short length of the reads, they fit several positions of the reference genome. Moreover, the reference genome is large, and it can thus be difficult to find the correct position. Because of sequencing errors and the nature of SNPs, a certain variability of the mapping process is necessary. Errors can be distinguished from real variants later on. Several algorithms are available for read mapping, e.g., BWA, Bowtie, SNP-o-matic, NextGenMap, and BLAT, a far from comprehensive list; an exhaustive list of algorithms can be found at the HTS-Mapper Web site [hts-mapper]. The output of many mapper programs is in SAM/BAM format, where the BAM format is a compressed binary version of the human-readable SAM format. BAM files can be indexed, allowing for rapid access to any region of a sequence. Special tools, e.g., SAM tools [sam-tools], allow for the analysis, modification, and visualization of sequences.

Once the mapping is complete, the genome information can be analyzed, e.g., single-nucleotide variants like SNPs can be identified. Also for this step several tools are available, such as SAM tools MAQ, VariationHunter, and destruct, for instance. An overview of tools and methods can be found in the Wikibook *Next Generation Sequencing (NGS)*, which is constantly being updated [wikibooks-ngs].

4.6.4 Proteogenomics

With the emergence of NGS it quickly became clear that the number of splice variants and nucleotide polymorphisms must lead to a much greater number of variations in the proteome than are stored in standard databases. The goal of proteogenomics is to study the actual connection between the genome and the proteome. For this, proteins are "sequenced" by generating a protein fingerprint using mass spectrometry (MS), which is compared to a database of theoretically derived protein fingerprints. If an experimentally derived fingerprint is identical to a theoretically derived one, the proteins are identical and the sequence of the unknown protein is revealed. The databases of theoretical

protein fingerprints are built based on NGS data connecting the genome to the proteome. The method is much older than its name, which was coined only in 2004. Already in the 1990s and 2000s shotgun proteomics was being used, where MS data were used to search protein databases. In 2004, Jaffe and coworkers (Jaffe et al. 2004) used a six-frame translation of the *Mycoplasma* genome as the protein database and coined the term *proteogenomics*. The concept has quickly been used for more complex organisms and is nowadays, in combination with NGS, of crucial importance for identifying and studying human protein variants in biological and medical research (Sheynkman et al. 2016).

Different kinds of nucleotide data are suitable for this method. First EST data were used, which are either translated to three or six reading frames, depending on the knowledge of the actual orientation. If genomic data are used, they are translated to six potential reading frames. In addition, RNA-Seq data are used, as are data of ribosomal sequencing, where mRNA molecules bound to ribosomes are sequenced. Last but not least, special databases focused on specific variations, e.g., splice variants or SNPs, are used (Sheynkman et al. 2016; Nesvizhskii 2014).

Although a number of protein variants have been discovered so far, both methods are based on fragments, i.e., in the case of NGS, DNA or RNA fragments, and in the case of proteogenomics, enzymatically digested proteins. The complete and unimpaired sequence, therefore, cannot be revealed with 100% confidence. Therefore, it is at least possible that more variants remain undiscovered. However, both methods will improve over time, allowing for the discovery of intact sequences.

4.7 Exercises

Exercise 4.1
How many ESTs does the database dbEST at NCBI contain (► http://www.ncbi.nlm.nih.gov/dbEST/index.html)? What two organisms have the most entries, and what is their percentage of the total number of entries?

Exercise 4.2
Determine by querying dbEST how many *Mangifera indica* ESTs there are. Note: Enter the name `Mangifera indica` on the home page of dbEST, then repeat the search, this time entering `Mangifera indica [ORGANISM]`. Explain the differences between the two results.

Exercise 4.3
Save the result of the second search in FASTA format.

Exercise 4.4
Using the saved sequences perform a sequence assembly. Use the CAP3 sequence assembly program of the PRABI-Doua Institute (► http://doua.prabi.fr/software/cap3). Note: The server accepts a maximum of 50,000 bases. How many contigs are built? How many ESTs do the contigs contain? Also, are there ESTs that are not grouped into contigs (singletons)?

Exercise 4.5
Annotate the ESTs by comparing the contigs with a nonredundant protein database using the BLASTx algorithm. To do this, go to the NCBI BLAST home page. Can reliable hits for all contigs in the protein database be found?

Exercise 4.6
Search for an EST with the accession number (AN) AI590371 using the database query system Entrez at NCBI. Save the sequence in FASTA format.

Exercise 4.7
Compare the saved EST sequence with the nonredundant nucleotide database of the NCBI. To do this, use the NCBI BLAST home page. How many reliable nucleotide sequence hits can be found in this database?

Exercise 4.8
How many EST sequences are stored in the UniGene database for the first hit (sequence ID NM_080870.3)? In which disease is this protein involved and in which human population is this disease prevalent?

Exercise 4.9
What can be learned about the expression of the protein from the EST of ► Exercise 4.8?

Exercise 4.10
Using the database query system Entrez at NCBI, locate the protein sequence of the mouse proto-oncogene c-myc with the AN P01108. Save the sequence in FASTA format.

Exercise 4.11
Compare the saved sequence of the protein c-myc with an EST database from the mouse. Use the NCBI BLAST home page to do this. Are mouse ESTs found in the database? What is noticeable about the distribution of the ESTs, and how can this be explained?

Exercise 4.12
In addition to very good hits (alignment score >200, red bars), many hits with an alignment score of 80–200 (magenta bars) are found. Do these ESTs also encode the protein c-myc? Give reasons for the result. Note: Compare the nucleotide sequences of this EST with the protein database UniProtKB.

Exercise 4.13
At the NCBI book collection find *Genes and Disease*. In that book you will find information about phenylketonuria. On which human chromosome is the gene for phenylalanine hydroxylase found? Click on the hyperlink to the database Entrez Gene. What information does this database provide?

Exercise 4.14
In the dbSNP database at NCBI (► http://www.ncbi.nlm.nih.gov/SNP/) search for the reference cluster with the ID `rs334`. In which organism is this SNP found? Examine

the category GeneView. Compared to the reference sequence (contig reference), which nucleotide exchange is found? Does it result in an amino acid exchange? If so, which one? What gene is affected by this SNP? Follow the link of the gene name to the database Entrez Gene. What disease is caused by the mutation?

References

Adams MD, Kelley JM, Gocayne JD, Dubnick M, Polymeropoulos MH, Xiao H et al (1991) Complementary DNA sequencing: expressed sequence tags and human genome project. Science 252:1651–1656

Berger EA, Murphy PM, Farber JM (1999) Chemokine receptors as HIV-1 coreceptors: roles in viral entry, tropism, and disease. Annu Rev Immunol 17:657–700

Blaxter M (1998) Caenorhabditis elegans is a nematode. Science 282:2041–2046

Everett JR (2016) From metabonomics to pharmacometabonomics: the role of metabolic profiling in personalized medicine. Front Pharmacol 7:297 und darin enthaltene Referenzen

Ezkurdia I, Juan D, Rodriguez JM, Frankish A, Diekhans M, Harrow J, Vazquez J, Valencia A, Tress ML (2014) Multiple evidence strands suggest that there may be as few as 19,000 human protein-coding genes. Hum Mol Genet 23:5866–5878

Jaffe JD, Berg HC, Church GM (2004) Proteogenomic mapping as a complementary method to perform genome annotation. Proteomics 4:59–77

Mouse Genome Sequencing Consortium (2002) Initial sequencing and comparative analysis of the mouse genome. Nature 420:520–562

Nesvizhskii AI (2014) Proteogenomics: concepts, applications, and computational strategies. Nat Methods 11:1114–1125

Schleidgen S, Klingler C, Betram T, Rogowski WH, Marckman G (2013) What is personalized medicine: sharpening a vague term based on a systematic literature review. BMC Med Ethics 14:55

Sheynkman GM, Shortreed MR, Cesnik AJ, Smith LM (2016) Proteogenomics: integrating next-generation sequencing and mass spectrometry to characterize Human Proteomic variation. Annu Rev Anal Chem (Palo Alto, Calif) 9:521–545

Thorisson GA, Stein LD (2003) The SNP Consortium website: past, present, future. Nucleic Acids Res 31:124–127

Wang Z, Gerstein M, Snyder M (2009) RNA-Seq: a revolutionary tool for transcriptomics. Nat Rev Genet 10:57–63

Yeo G, Holste D, Kreiman G, Burge CB (2004) Variation in alternative splicing across human tissues. Genome Biol 5(10):R74

Further Reading

cap. http://doua.prabi.fr/software/cap3
dbest. https://www.ncbi.nlm.nih.gov/dbEST/
dbgss. https://www.ncbi.nlm.nih.gov/dbGSS/
dbsnp. https://www.ncbi.nlm.nih.gov/SNP/
dbsts. https://www.ncbi.nlm.nih.gov/dbSTS/
ebi-gwas. http://www.ebi.ac.uk/gwas/
grailexp. http://compbio.ornl.gov/grailexp/
gwas. http://www.gwascentral.org/
helix-nebula. http://www.helix-nebula.eu/usecases/embl-use-case
homologene. http://www.ncbi.nlm.nih.gov/homologene/
hts-mapper. http://www.ebi.ac.uk/~nf/hts_mappers/
humatrix. https://www.humatrix.de/
image. http://imageconsortium.org/
nematode. http://www.nematode.net/
ngs-knowledge-base. https://goo.gl/HIaY1W
ngs-movie. https://www.youtube.com/watch?v=jFCD8Q6qSTM

nhgri-gwas. https://www.genome.gov/20019523/
nih-biowolf. https://hpc.nih.gov/
pahdb. http://www.pahdb.mcgill.ca/
phrap. http://www.phrap.org/
pyrosequencing. http://www.pyrosequencing.com/
sam-tools. https://en.wikipedia.org/wiki/SAMtools
stackpack. http://genoma.unsam.edu.ar/stackpack.old/index.html
stratipharm. http://www.stratipharm.de/
unigene. http://www.ncbi.nlm.nih.gov/UniGene/
vfa-personalisiert. http://www.vfa.de/personalisiert/
wikibook-ngs. https://en.wikibooks.org/wiki/Next_Generation_Sequencing_%28NGS%29