Parallel target selection by trinucleotide threading

Pawel Zajac

Royal Institute of Technology
School of Biotechnology
Stockholm 2009
DNA is the code for all life. Via intermediary RNA the information encoded by the genome is relayed to proteins executing the various functions in a cell. Together, this repertoire of inherently linked biological macromolecules determines all characteristics and features of a cell. Technological advancements during the last decades have enabled the pursuit of novel types of studies and the investigation of the cell and its constituents at a progressively higher level of detail. This has shed light on numerous cellular processes and on the underpinnings of several diseases. For the majority of studies focusing on nucleic acids, an amplification step has to be implemented before an analysis, scoring or interrogation method translates the amplified material into relevant biological information. This information can, for instance, be the genotype of particular SNPs or STRs, or the abundance level of a set of interesting transcripts. As such, amplification plays a significant role in nucleic acid assays. Over the years, a number of techniques – most notably PCR – has been devised to meet this amplification need, specifically or randomly multiplying desired regions. However, many of the approaches do not scale up easily rendering comprehensive studies cumbersome, time-consuming and necessitating large quantities of material.

Trinucleotide threading (TnT) – forming the red thread throughout this thesis – is a multiplex amplification method, enabling simultaneous targeted amplification of several nucleic acid regions in a specific manner. TnT begins with a controlled linear DNA thread formation, each type of thread corresponding to a segment of interest, by a gap-fill reaction using a restricted trinucleotide set. The whole collection of created threads is subsequently subjected to an exponential PCR amplification employing a single primer pair. The generated material can thereafter be analyzed with a multitude of readout and detection platforms depending on the issue or characteristic under consideration.

TnT offers a high level of specificity by harnessing the inherent specificities of a polymerase and a ligase acting on a nucleotide set encompassing three out of the four nucleotide types. Accordingly, several erroneous events have to occur in order to produce artifacts. This necessitates override of a number of control points.

The studies constituting this thesis demonstrate integration of the TnT amplification strategy in assays for analysis of various aspects of DNA and RNA. TnT was adapted for expression profiling of intermediately-sized gene sets using both conventional DNA microarrays and massively parallel second generation 454 sequencing for readout. TnT, in conjunction with 454 sequencing, was also employed for allelotyping, defined as determination of allele frequencies in a cohort. In this study, 147 SNPs were simultaneously assayed in a pool comprising genomic DNA of 462 individuals. Finally, TnT was recruited for parallel amplification of STR loci with detection relying on capillary gel electrophoresis. In all investigations, the material generated with TnT was of sufficient quality and quantity to produce reliable and accurate biological information.

Taken together, TnT represents a viable multiplex amplification technique permitting parallel amplification of genomic segments, for instance harboring polymorphisms, or of expressed genes. In addition to these, this versatile amplification module can be implemented in assays targeting a range of other features of genomes and transcriptomes.

**Keywords**: trinucleotide threading, multiplex amplification, expression profiling, microarray, generic tag, short tandem repeat, microsatellite, electrophoresis, single nucleotide polymorphism, allelotyping, 454, Pyrosequencing

© Pawel Zajac
List of publications
This thesis is based on the following four papers, which are referred to in the text by their Roman numerals (I-IV). The papers are found in the appendix.


Contents

INTRODUCTION ................................................................................................................. 5

1. A journey from DNA to proteins .................................................................................. 5
   1.1. The central dogma ................................................................................................. 6
       1.1.1. Revisions of the central dogma ......................................................................... 6
   1.2. Large-scale DNA projects ...................................................................................... 6
       1.2.1. The Human Genome Project (HGP) ................................................................. 6
       1.2.2. Human variation ............................................................................................... 7
       1.2.3. Genome-wide association studies (GWAS) ...................................................... 7
       1.2.4. The HapMap project ....................................................................................... 8
       1.2.5. The ENCODE project ..................................................................................... 8
       1.2.6. A new definition of a gene? .............................................................................. 9
   1.3. RNA .......................................................................................................................... 9
       1.3.1. Noncoding RNA species .................................................................................. 10
       1.3.2. Messenger RNA (mRNA) ................................................................................ 10
   1.4. Proteins .................................................................................................................... 10
   1.5. A word about technology development ................................................................... 11

2. Amplification .................................................................................................................. 13
   2.1. Polymerase chain reaction (PCR) .......................................................................... 13
       2.1.1. Long-range PCR (LR-PCR) and real-time PCR (RT-PCR) .................................... 14
       2.1.2. Improved instrumentation .............................................................................. 14
   2.2. Global amplification ............................................................................................... 14
       2.2.1. Amplification of reproducible genomic fractions ............................................. 14
       2.2.2. Whole-genome amplification (WGA) ................................................................. 15
   2.3. Isothermal targeted amplification ........................................................................... 15
   2.4. Multiplexity issues with PCR ................................................................................. 16
       2.4.1. Two-stage amplification .................................................................................. 17
       2.4.2. Compartmentalization .................................................................................... 18
       2.4.3. Emulsion PCR ................................................................................................ 18
       2.4.4. Solid-phase PCR ............................................................................................. 18
       2.4.5. Targeted compartmentalization ....................................................................... 19
   2.5. Ligase-mediated amplification ................................................................................. 19
       2.5.1. Ligase chain reaction (LCR) ........................................................................... 19
       2.5.2. Gap-LCR ........................................................................................................ 20
       2.5.3. Padlock probes ................................................................................................ 20
       2.5.4. Multiplex renditions of padlocks .................................................................... 20
       2.5.5. Multiplex ligation-dependent probe amplification (MLPA) ................................ 21
   2.6. What happens post-amplification? ........................................................................... 21

3. Genotyping ..................................................................................................................... 23
   3.1. Historical overview ............................................................................................... 23
   3.2. Short tandem repeats (STRs) .................................................................................. 24
   3.3. Single nucleotide polymorphisms (SNPs) ............................................................... 24
   3.4. Low-scale genotyping ............................................................................................ 24
       3.4.1. Allele-specific oligonucleotide hybridization (ASOH or ASH) .......................... 24
       3.4.2. Counteract hybridization differences between probe sets .............................. 25
       3.4.3. Dynamic allele-specific hybridization (DASH) ............................................... 25
       3.4.4. Solution-phase hybridization ........................................................................ 26
       3.4.5. Allele-specific PCR (AS-PCR) ........................................................................ 26
       3.4.6. Invader ............................................................................................................ 26
       3.4.7. Oligonucleotide ligation assay (OLA) ............................................................... 27
       3.4.8. Specificity ....................................................................................................... 27
   3.5. Detection platforms ............................................................................................... 27
Parallel target selection by trinucleotide threading

3.6. Highly parallel SNP genotyping .......................................................... 29
3.6.1. Allele-specific extension (ASE) and single-base extension (SBE) .......... 29
3.6.2. Generic tag arrays ........................................................................... 30
3.6.3. GoldenGate ................................................................................... 31
3.6.4. Molecular Inversion Probes (MIP) ................................................... 31
3.6.5. Whole-genome sampling analysis (WGSA) ....................................... 32
3.6.6. Infinium ......................................................................................... 32
3.6.7. Conversion rates ............................................................................. 33
3.7. A possibility to cover all SNPs? ............................................................ 33
3.8. Summary table .................................................................................... 34

4. Sequencing ............................................................................................... 35
4.1. The first revolution ............................................................................... 35
4.1.1. The chemical degradation method of Maxam and Gilbert .................... 36
4.1.2. The enzymatic chain termination method of Sanger ............................ 36
4.2. Alternative approaches ......................................................................... 37
4.2.1. Pyrosequencing ............................................................................... 37
4.2.2. Sequencing-by-hybridization (SBH) and tiling arrays .......................... 37
4.3. Second generation platforms ............................................................... 38
4.3.1. The main players ............................................................................. 39
4.3.2. Protocol characteristics and interrogation principles .......................... 39
4.3.3. 454 .......................................................... 40
4.3.4. Illumina / Solexa ............................................................................ 41
4.3.5. Helicos ........................................................................................... 41
4.3.6. Sequencing by Oligonucleotide Ligation and Detection (SOLiD) .......... 42
4.3.7. What about sequencing-by-hybridization? ........................................ 43
4.3.8. Guidelines regarding second generation platform choice ..................... 43
4.3.9. Sanger versus second generation systems ........................................ 44
4.3.10. Combining several approaches ...................................................... 44
4.4. Third generation approaches ............................................................... 45
4.4.1. Complete Genomics ........................................................................ 45
4.4.2. Pacific Biosciences and VisiGen ...................................................... 45
4.4.3. Nanopores ...................................................................................... 46
4.5. Summary table .................................................................................... 47

5. Expression profiling .................................................................................. 49
5.1. Methods of differing magnitude ........................................................... 50
5.2. Historical overview of global expression analysis ................................. 50
5.2.1. Differential display ........................................................................... 50
5.2.2. EST sequencing ............................................................................... 51
5.2.3. Digital counting to analog relative quantification and back to digital counting ......................................................................................... 51
5.3. Microarrays .......................................................................................... 52
5.3.1. Spotted arrays featuring cDNAs, ESTs or pre-synthesized oligonucleotides .......................................................... 52
5.3.2. In situ synthesized Affymetrix arrays ............................................... 53
5.3.3. Other commercial arrays ................................................................. 53
5.3.4. Probe position considerations .......................................................... 53
5.3.5. Benefits and drawbacks? ................................................................. 54
5.3.6. Amplification strategies .................................................................... 54
5.4. Tag-based methods ............................................................................... 55
5.4.1. Serial analysis of gene expression (SAGE) and its longer tag cousins .... 55
5.4.2. 3′-SAGE and CAGE ....................................................................... 56
5.4.3. Paired-end ditag strategies (GIS and MS-PET) .................................. 56
5.4.4. Massive parallel signature sequencing (MPSS) ................................ 57
5.4.5. Polony multiplex analysis of gene expression (PMAGE) ................................................................. 58
5.5. Deep sequencing with second generation platforms ................................................................. 58
  5.5.1. Advantages and disadvantages of RNA-Seq ................................................................. 59
5.6. Validation methods ................................................................................................................... 59
  5.6.1. Northern blot .................................................................................................................. 60
  5.6.2. Reverse transcription quantitative RT-PCR ...................................................................... 60
5.7. Intermediary methods ............................................................................................................. 60
  5.7.1. BeadsArray for the Detection of Gene Expression (BADGE) ........................................ 61
  5.7.2. cDNA-mediated annealing, selection, extension and ligation (DASL) .......................... 61
5.8. Summary table ......................................................................................................................... 62

PRESENT INVESTIGATION ................................................................. 63

6. The papers ....................................................................................................................................... 63
  6.1. Trinucleotide threading ......................................................................................................... 64
  6.2. Initial publication .................................................................................................................. 64
  6.3. Expression profiling (Paper I) ............................................................................................. 65
  6.4. Short tandem repeat analysis (Paper II) ............................................................................. 67
  6.5. SNP allelotyping (Paper III) ............................................................................................... 67
  6.6. Targeted expression profiling with 454 sequencing (Paper IV) ................................ .... 69
  6.7. Future perspectives .............................................................................................................. 71

ABBREVIATIONS .......................................................................................................................... 73

ACKNOWLEDGEMENTS ............................................................................................................... 75

REFERENCES .................................................................................................................................... 79
INTRODUCTION

1. A journey from DNA to proteins

DNA. The blueprint. The drawing of all life. Everything, from the simplest bacteria to humans, via all facets of life, shares this essential macromolecule. This beautifully simple molecule, consisting of two entwined strands created from a repertoire of only four building blocks, was discovered by the Swiss doctor Friedrich Miescher in 1869 during his efforts to elucidate the makeup of leukocytes\(^1\). The fact that DNA represented the hereditary material was demonstrated in 1944 by Oswald Avery, Colin MacLeod and Maclyn McCarty\(^2\), but was regarded with a lot of skepticism as the common belief at that time was that proteins constituted the information unit passed from one generation to the next. Eight years later, a study conducted by Alfred Hershey and Martha Chase corroborated the role of DNA as the genetic material\(^3\).

DNA, or deoxyribonucleic acid, is composed of four basic units, the four nucleotides. Each nucleotide contains the sugar moiety deoxyribose linked both to a phosphate group and one of the four bases, the purines adenine (A) and guanine (G) and the pyrimidines cytosine (C) and thymine (T). Structurally, DNA is a double helix, with the two chains coiled around each other. The alternating sugars and phosphates of the nucleotides create the backbones of the strands, allowing the bases to face inwards thus enabling a base from one strand to form hydrogen bonds with a base on the opposing one. Not all base combinations are permitted – purines in one chain bond with pyrimidines in the other – an A pairing only with a T, with a C matching only with a G. This elegant structure was proposed by James Watson and Francis Crick in 1953\(^4\), based on X-ray diffraction patterns obtained by Rosalind Franklin\(^5\) and Maurice Wilkins\(^6\). In addition to the X-ray images, work by Erwin Chargaff on DNA base composition presented a couple of years earlier lent further support to the suggested double helical nature of this molecule. In particular, Chargaff unfolded that the amount of purines was equal to that of pyrimidines, and, furthermore, that A and T shared the same numbers, as did C and G\(^7-9\).
1.1. **The central dogma**

The elucidation of the structure of the hereditary material sparked an interest in the information flow from DNA to the other classes of macromolecules – RNA and proteins. In 1958, Crick postulated the central dogma, stating how biological sequence information is relayed between the macromolecules\(^1\). This theorem stipulates three widespread general transfers – 1) DNA → DNA (replication), 2) DNA → RNA (transcription), and 3) RNA → protein (translation). Thus, specific regions of DNA are transcribed into RNA, which subsequently acts as the template for production of proteins in ribosomes in a process called translation. Once the information has been encoded in the proteins, however, it cannot be conveyed back, either to protein or nucleic acid. New body of data in the following decade, for instance the deciphering of the genetic code where each triplet of bases in the DNA – a codon – corresponds to a specific amino acid in the protein sequence\(^11-14\), led to a reformulation, expansion and defense of this theorem in 1970\(^15\).

1.1.1. **Revisions of the central dogma**

As stated by the central dogma, proteins represent the final destination for the instructions embedded in the genome and, correspondingly, should execute all or nearly all functions acting as the working horses of the cell\(^16\). Undeniably, proteins are versatile macromolecules well adapted to perform the myriad functions necessary to keep a cell alive and well. However, a number of discoveries have painted a far more entwined picture. For example, the majority of the human genome is expressed at one time or another, yet only a miniscule proportion is encoding proteins\(^17\). This suggests novel roles for the unaccounted transcripts and there is accumulating evidence that these noncoding RNA species play essential regulatory functions, perhaps in a highly intricate regulatory network\(^18, 19\). Accordingly, the rather simplistic view of the central dogma is maybe only applicable to simple prokaryotes where both the functions and the regulation can be accommodated by proteins\(^16, 20\).

1.2. **Large-scale DNA projects**

1.2.1. **The Human Genome Project (HGP)**

Proteins and regulatory RNA moieties may be the molecules executing and controlling the crucial functions of the cell, but the fact remains that DNA encodes this multitude of vital components. To investigate the human genome in full depth, an ambitious sequencing effort was undertaken by the Human Genome Project launched in 1990. The initial time frame for this effort was estimated at 15 years but technological advances pertaining to automated DNA sequencing, as well as inter-laboratory collaboration, made possible the announcement of a draft genome already in 2000. Concurrently, the draft prepared by a private effort initiated in 1998 by the company Celera Genomics was proclaimed. Whereas the public endeavor employed a clone-by-clone sequencing strategy, a whole-genome shotgun approach was preferred by Celera\(^21\). Two separate drafts were thus presented in February 2001\(^22, 23\). The assembly by the Human Genome Sequencing Consortium reflected an assortment of numerous individual's DNA, whereas Celera Genomics released a reference sequence based on the genome from five individuals. A couple of years later, after filling in a large proportion of gaps still present in the draft sequence and resolving some of the ambiguities, the finished version saw the light of day\(^24\).

The human genome is composed of slightly over three billion bases. This staggering amount of information is perhaps brightly illuminated by a book analogy. If each base corresponds to one letter and each page in a 1000-page book contains 1000 letters, then
the genome would entail a collection amounting to 3000 books. Genomic DNA is spread out on 23 chromosomes. One such set is inherited from the mother and one from the father, making for 23 chromosom al pairs. 22 of these are autosomal pairs, while the remaining one determines the sex. The finished sequence did not entail the ‘complete’ genome as some repetitive chromosomal regions – telomeres and centromeres – as well as certain troublesome segments were not able to be elucidated. Instead, the finished version covered roughly 99% of the euchromatic portion of the genome, the lightly packed form of chromatin commonly related to gene-rich regions and active transcription. Furthermore, the genome was found to only encompass between 20,000 and 25,000 genes. This was a quite mesmerizing downshift from previous predictions. These protein-coding parts only account for 1.2% of the genome. 

1.2.2. Human variation

One of the goals of the genome sequencing initiatives was to create a catalog of human variation. It was illustrated that two individuals share more than 99.9% of their genomes, implying that the genetic component of what makes us unique is represented by a highly marginal fraction of our DNA. Of course, the environment and stochastic variations work in conjunction with the genetic architecture to shape all of our characteristics. Each specific DNA variant is dubbed an allele, and the particular alleles an individual possesses make up his genotype. Such variants can be classified in a number of ways, but this categorization is not always clear-cut. They can be pooled into common or rare variants, the difference relating to the minor allele frequency (MAF), that is the prevalence of the less common allele. If MAF is greater than 1% the variation is considered common – a polymorphism – and values less than this signify a rare variant. The classification can also be made according to the magnitude of the actual change in the DNA sequence. Here single nucleotide polymorphisms (SNPs) are distinguished from larger structural variants. SNPs – an alteration in only one base – are the most abundant genetic variations in the human genome with one SNP present, on average, in every 300 bp. 11 million such markers are estimated to be present. Structural variants, comprising a diverse group of alterations – insertion-deletions, block substitutions, inversions and copy number variations (CNVs) – have been analyzed to a lesser extent, but the scope of their contribution to the overall human variation is starting to become clear. It has been estimated that they represent at least 20% of all variants; however, they may constitute approximately 70% of the alternate bases, thus accounting for the majority of differing base pairs between two individuals. On average, the human genomes sequenced to date comprise about 4 million variants – approximately ¾ of which correspond to SNPs – which together make up roughly 12 Mb.

1.2.3. Genome-wide association studies (GWAS)

Several types of studies became the natural extension following acquisition of the human genome sequence and the catalogue of genetic variation. One of the goals in the post-genome era is to link genetic variants to phenotypical traits, for example physical characteristics or diseases. Such investigations are not a new phenomenon as similarly aimed family-based linkage studies or candidate-region association studies have been performed previously. One example is the identification of a marker linked to the gene behind Huntington’s disease, a single gene genetic disorder, in the beginning of the 80s. The scope of these studies has, however, widened to accommodate the multifaceted genetic etiology of the majority of common diseases.

The currently preferred approach is a genome-wide association study (GWAS), whereby markers, preferentially SNPs, across the whole-genome are assessed for association to a particular phenotype. By comparing marker frequencies in a case cohort possessing the
sought-after attribute to an unaffected control one, genomic regions linked to this attribute can be identified. In 2007 Science Magazine dubbed GWAS the ‘breakthrough of the year’ and the last couple of years have seen hundreds of loci associated with over 80 diseases and various other characteristics. At present, the online GWAS catalog contains 356 publications and 1640 SNPs. Clearly, such studies have provided a wealth of novel information about the underpinnings and potential pathways of diseases. However, these seemingly conceptually simple investigations are fraught with potential drawbacks, possibly diminishing their usefulness. The genomic region identified could prove to be quite substantial making the transition from the association to the causal functional variant quite complicated. Moreover, rare variants are not included in such studies but could well contribute to the trait. Additionally, the effect the common variants confer to a phenotype may be rather small, necessitating studies on tens of thousands of individuals in order to reach adequate statistical power. The two above-mentioned facts could account for why the associations encountered so far only explain a rather small fraction of the heritable risk. Lastly, the transfer of GWAS results from one population to another is not straightforward.

1.2.4. The HapMap project
The fact that human genomic DNA has a block-like architecture and the HapMap Project aiming at defining these blocks are important factors explaining the surge of popularity the GWAS have encountered in recent years. The individual alleles in each block are linked to each other creating haplotypes that are inherited as discrete units. New haplotypes are a result of recombination events breaking the linkage or transpire out of novel mutations. This arrangement implies that the allele status at one position can reveal the status at another location within the same haplotype block. As such, a carefully chosen marker set can lessen the genotyping burden while retaining most of the genotype information by allowing the assayed markers to act as proxies for the non-assayed ones. To create a human haplotype map describing the common patterns of DNA sequence variation the International HapMap Project was instigated in 2002. The first phase saw the genotyping of more than one million SNPs in 270 individuals from four distinct populations. This was complemented by an additional 2.1 million SNPs in the second phase, increasing the density from one marker per 5 kilobases of euchromatic DNA to around one every kilobase. The developed haplotype map represents a valuable resource for selection of tag SNPs, which can provide information about SNPs they are in linkage with. However, approximately 0.5 – 1% of all common variants cannot be substituted for others, predominantly by virtue of their location in recombination hotspots. Nevertheless, sound nonredundant tag SNP sets can dramatically reduce genotyping cost.

1.2.5. The ENCODE project
The human genome encompasses sequences with a myriad of functions. A vital, albeit tremendous, task in the post-genome era is to identify, organize and annotate these elements. The ultimate goal is a genomic map where the functionality and characteristics of every base are accounted for. Several initiatives have focused on some of these features providing significant new insights. The HapMap Project has, for instance, delved into variation and haplotype structure. Ensembl, an automated annotation tool probing mostly for genes, is another example.

The ENCODE Project (ENCyclopedia Of DNA Elements), launched in 2003, is the most extensive effort to provide a functional annotation. With an amalgamation of experimental techniques and computational algorithms the aim is to identify all functional elements embedded within the human genome. The project is divided into three phases – a pilot phase, a technology-development phase and a production phase. For the pilot
stage 1% of the human genome – approximately 30 Mb – distributed over 44 genomic segments was chosen for analysis. 14 of the regions were manually selected to represent regions with plentiful information already available, whereas the remaining ones were chosen at random but with certain constraints to cover a range of gene content and human-to-mouse conservation. The pilot phase was finalized in 2007 and noticeably revealed that the transcriptional landscape is far more complex than originally thought. The majority of the human genome is transcribed into primary transcripts with 93% of the bases experimentally shown to be transcribed. Of these, only a minor fraction corresponds to protein-coding segments. Additionally, considerable overlap was found between transcripts. Furthermore, the project generated novel insights about transcriptional start sites and the interplay between chromatin structure and the processes of replication and transcription. Perhaps the most striking finding was the large fraction of experimentally determined functional elements not under any evolutionary constraints. On the other hand, for 40% of the constrained regions the experiments did not reveal any function. The pilot study and the technology-development phase were to be run concurrently to assess the feasibility of the venture, as well as to devise and optimize the experimental and theoretical framework to accommodate a shift toward the high-throughput analysis of the rest of the genome.

1.2.6. A new definition of a gene?
The findings of the ENCODE pilot phase prompted a debate about the definition of a gene. Since 1909, when the word ‘gene’ was first coined stemming from the Greek genesis (‘birth’) or genos (‘origin’), its description has seen several reformulations and updates coinciding with the expanding knowledge in the field. The definition has predominantly been centered on protein-coding moieties and the current view of a gene is as a discrete genomic segment playing an integral part in a specific function or phenotype. However, the vast complexity of transcription illuminated by ENCODE has led to a proposal of a new gene designation: ‘The gene is a union of genomic sequences encoding a coherent set of potentially overlapping functional products’. The Consensus Coding Sequence (CCDS) project intending to integrate annotation information to arrive at a complete protein-coding set has currently identified 20,159 consensus coding regions stemming from 17,052 human genes. The total number of protein-coding genes is probably slightly higher with the latest known protein-coding gene estimate in Ensembl being 22,258. Additionally, about 20,500 protein-coding genes were reported by a study comprehensively analyzing the gene catalog to remove noncoding entries.

1.3. RNA
The transcriptional landscape shown by the pilot project of ENCODE and other studies is complex. The RNA species historically garnering the most attention is the messenger RNA (mRNA). This particular type of transcript transmits information from genomic regions encoding for proteins to ribosomes where these proteins are synthesized. Following transcription, the precursor mRNA is processed to a mature form. The most significant element of this process is splicing. The pre-mRNA contains protein-coding segments called exons that are interrupted by longer noncoding introns. As reported by the HGP, the mean exon and intron lengths are 145 bp and 3365 bp, respectively. During splicing the introns are excised leaving only a string of joined exons. Splicing can occur in alternative patterns giving rise to an extra layer of diversity as one genomic locus can give multiple transcripts with different combinations of exons. This is a pervasive phenomenon. The GENCODE subproject of ENCODE reported an average of 5.4 transcripts per genomic locus. Furthermore, 80-90% of protein-coding genes are subjected to alternative splicing.
1.3.1. Noncoding RNA species
Noncoding RNA can broadly be divided into infrastructural and regulatory species\(^{18, 19}\). The former category encompasses relatively well-known RNA types involved in several fundamental housekeeping processes within the cell. Both transfer RNAs (tRNAs), corresponding to the physical link between a coding triplet of bases and an amino acid, and ribosomal RNAs (rRNAs), essential components of the ribosomes, are crucial for translation. Small nuclear RNAs (snRNAs) play a part in the splicing process.

The regulatory RNA species fall into two classes based on their length. MicroRNAs (miRNAs) and small interfering RNAs (siRNAs) are approximately 20 nucleotides in length and affect gene expression by recognizing mRNA molecules via base pairing ultimately resulting in the protein-coding transcript being degraded or prevented from entering translation\(^{56, 57}\). The siRNAs have also been shown to be involved in histone modification. Albeit relatively similar in their mature form, miRNAs and siRNAs have disparate biogenesis. The former ones are created by a two-step cleavage of hairpin-shaped precursors, whereas the latter ones are a result of cleavage of longer double stranded transcripts. MicroRNA represents the most widely studied group of short regulatory transcripts and analysis of sequence conservation of miRNA target regions has estimated that more than half of the human protein-coding genes have the potential to be influenced by miRNAs\(^{58}\). In recent years, more and more attention has centered on long noncoding RNAs. These are somewhat arbitrarily defined as being longer than 200 nucleotides and play a part in gene regulation\(^{50, 59}\).

The preponderance of noncoding RNAs has been argued to reflect the fact that regulatory requirements of more complex organisms far surpass the regulatory capacities held by proteins. Simpler organisms such as prokaryotes, on the other hand, have much less noncoding DNA suggesting that, in these, proteins both act as effectors and regulators\(^{16, 20}\). A larger extent of noncoding portions of the genome seems to relate to increased organism complexity. Conversely, the number of protein-coding genes does not exhibit a clear-cut correlation with organism complexity\(^{60}\). For instance, the nematode *Caenorhabditis elegans*, one of the simplest multicellular eukaryotes composed of only about 1000 cells, and humans have a similar number of genes\(^{61}\). RNAs have several beneficial properties making them suitable for this regulatory task: the sequence specificity by which they can recognize their interacting partners, the compactness making these elements quite space-efficient in a genomic context and their mutability giving rise to flexibility and reconfigurability without exchanging the underlying protein components of the system\(^{20}\).

1.3.2. Messenger RNA (mRNA)
The noncoding regulatory transcripts have been the focus of attention in recent years. However, historically most of the efforts have centered on the protein-coding mRNAs and specifically their expression levels that might be indicators of the corresponding protein levels. The interest in this field rose sharply in the middle of the 1990s with the advent of microarrays, which were a cost-effective and scalable platform for expression analysis\(^{62, 63}\). Since then, evolution in array fabrication, experimental protocols and statistical evaluation has made possible rapid studies on the level of the transcriptome, the set of all mRNAs. Currently, microarrays are progressively exchanged in favor of novel sophisticated sequencing techniques able to sequence the whole repertoire of both coding and regulatory RNAs\(^{64}\).

1.4. Proteins
Proteins, the third important class of biological macromolecules, are challenging to study in a high-throughput fashion owing to the multitude of physical and chemical
characteristics they can possess. Nevertheless, ingenious endeavors have been initiated in the last couple of years to systematically study proteins, thus complementing the existent genomics and transcriptomics data. The most notable of these is the Swedish Human Proteome Resource (HPR) program with the goal of creating an atlas displaying the expression and localization of all human proteins\textsuperscript{65, 66}. High-throughput generation of mono-specific antibodies (msAbs)\textsuperscript{67} coupled with analysis on tissue microarrays (TMAs) has allowed a rapid systematic exploration of the protein space. The immunohistochemistry images showing the spatial distribution of the proteins are accompanied by validation of the generated antibodies with protein arrays and Western blots, as well as with immunofluorescent-based confocal microscopy images. Additionally, features of the proteins are displayed. 48 normal tissues, 20 cancer types and 47 human cell lines are assessed in the project. The first release of this protein catalog featured 718 antibodies covering 650 genes. The latest, fifth, release of the atlas is based on 8,832 antibodies used to produce 7,334,244 images\textsuperscript{68} representing in excess of 25\% of the protein-coding genes.

1.5. A word about technology development
All the exciting breakthroughs within biotechnology have been brought about by technological advances that have enabled researchers to pursue novel types of investigations or studies on a much larger scale. For instance, the introduction of the polymerase chain reaction in mid-1980s provided the scientific community with a convenient DNA amplification technique spurring experiments that beforehand were either cumbersome or impossible. The development of automated capillary sequencing instruments made possible the elucidation of the human genome and the list of additional examples is extensive. In other words, technological development and novel biological insights go hand in hand, why ingenious new platforms and approaches remain a vital part of this field.
2. Amplification

A single human cell contains only two copies of the genome, one inherited from the mother and the other passed on from the father. Although together comprising more than 6 billion bases, each segment is present in only two copies precluding its analysis with most methods. Correspondingly, either DNA from a plentitude of cells is required or the genomic region of interest has to be amplified. Since biological material is most frequently severely limited the former approach is simply not viable in the majority of cases.

2.1. Polymerase chain reaction (PCR)

The polymerase chain reaction (PCR) represents the paradigm shift method enabling scientists to reliably amplify a DNA portion of interest. Concomitantly, the complexity is greatly reduced enabling efficient downstream analysis. A pair of oligonucleotides is designed to encapsulate the desired DNA region and by a two-step or three-step thermal cycling the action of a polymerase is harnessed to exponentially amplify this region doubling the number of DNA molecules in each cycle, eventually yielding an enormous amount of identical copies of this DNA segment. During the first step the temperature is raised to allow for denaturation creating single strands to which the primers can anneal. The annealing is performed at a lower temperature. In the final step a polymerase extends the chain using the primer as an initiator. A standard thermal profile is composed of strand melting at 95°C, primer annealing at 50 – 72 °C and elongation at 72°C, the optimum temperature for the most frequently used *Thermus aquaticus* (Taq) polymerase.

This beautifully simple protocol was theoretically described in the beginning of the 1970s and according to the authors: ‘*experiments based on these lines of thought are in progress’*69. However, this group published no such results and, instead, researchers at the Cetus Corporation experimentally demonstrated the feasibility of this approach in 1985 amplifying a region of the beta-globin gene70. It was proposed that this technique could have numerous applications and far-reaching implications but it was at a conference a year later that PCR received the wide audience it certainly deserved71. One initial predicament was that the enzyme firstly used was thermolabile, thus degraded by the heat.
required in the denaturation step, necessitating replenishment before the next cycle could commence. However, this was resolved by employing the thermostable Taq polymerase72, 73.

2.1.1. Long-range PCR (LR-PCR) and real-time PCR (RT-PCR)
Since its inception, several modifications and refinements of the original PCR approach have been implemented. Long-range PCR (LR-PCR) uses optimized reaction conditions and a two-polymerase system with one enzyme possessing proofreading activity to amplify DNA regions up to 40 kb74, 75.

PCR is an end point method, meaning that only the end products are analyzed, which makes quantification problematical. This is because the plateau phase of amplification contains approximately equal amount of product irrespective of the initial number of template molecules. Real-time PCR allows for a continuous monitoring of the amplification reaction76, 77. In the first adaptation the reaction was supplemented with ethidium bromide (EtBr) that generated a fluorescent signal increasing in proportion to the amount of double-stranded DNA formed78. Currently, the favored dye is SYBR Green. Detection specificity is, however, a problem with these systems as the dyes bind to all double-stranded DNA, not discriminating between the desired product, undesired amplicons and primer-dimers77. This problem can be circumvented by using targeted probes with TaqMan probes79, 80 and Molecular Beacons81 being the most commonly employed ones. These are double labeled with a fluorophore and a quencher that effectively silences the probe when intact. In TaqMan the 5’ → 3’ exonuclease activity of Taq polymerase cleaves the specifically annealed probes thus separating the quencher from the fluorophore generating a signal. This physical separation is attained by a conformational change, from a stem-loop structure to an elongated linear form, upon probe binding in the case of Molecular Beacons.

2.1.2. Improved instrumentation
In addition to modifications of the PCR conditions, improved surrounding instrumentation can enhance performance. The conventional PCR instrument is a thermal cycler, in which a tube containing the reaction mixture is placed in a metal block that is repeatedly heated and cooled to achieve the required temperatures. With this setup the ramping times in-between the PCR temperature steps can be quite long. In continuous-flow PCR the reaction mixture is not stationary, but is instead flowed between spatially separated temperature zones in a microfluidic device, reducing the time required for a standard PCR to around 15 minutes82. These devices have, however, mostly been applied to amplification of short fragments and, naturally, necessitate investments in new infrastructure. Carryover contamination between successive runs in the same system is yet another problem, but this can be avoided by employing these devices in conjunction with emulsified water-in-oil droplets containing the reaction mixture83. An alternative high-speed PCR method relies on centrifuging the reaction while heating and cooling it84.

2.2. Global amplification

2.2.1. Amplification of reproducible genomic fractions
A modified PCR protocol, particularly pertaining to the pre-amplification sample preparation, can be utilized to amplify reproducible fractions of the genome. In this approach, genomic DNA is digested with a restriction enzyme, adaptors are ligated onto the digestion products and a PCR is performed using primers complementary to the adaptor sequences. As shorter fragments are preferentially amplified the outcome is a genomic representation of reduced complexity. The magnitude of this reduction depends
on the restriction endonuclease employed and on the conditions determining the size range of the fragments amplified. This method, one of its first uses being to isolate sequences bound to regulatory proteins, is currently most popular in the setting of parallel SNP genotyping in whole-genome sampling analysis (WGSA; Chapter 3). Originally, the initial step in this genotyping process entailed restriction with XbaI followed by amplification of fragments between 250 and 1000 bases, reducing the size from the 3,000 Mb of a complete genome to approximately 60 Mb, a 50-fold reduction. In a latter implementation, XbaI and HindIII treated genomes, coupled with a size selection in the interval of 250 to 2000 bases, represented 300 Mb of sequence complexity.

2.2.2. Whole-genome amplification (WGA)

In certain cases, particularly for global analysis involving biological material that is scarce, it becomes desirable to amplify the whole genome, not only a reproducible representation of it. Several PCR-based methods have been developed for whole genome amplification (WGA). Primer-extension preamplification (PEP) and degenerate oligonucleotide-primed PCR (DOP-PCR) represent two such approaches. In PEP, 15-mers are used to prime repeated extensions, whereas DOP-PCR relies on partly degenerate primers, containing a random hexamer in-between a specific 3’ 6-mer and 5’ 10-mer, and low annealing temperatures in the first PCR cycles. However, these approaches suffer from biases lowering the coverage, producing an uneven amplification across the genome, as well as a high proportion of unspecific artifacts.

An alternative WGA strategy – multiple displacement amplification (MDA) – relies on the DNA polymerase of phage ϕ29, a stable, proofreading and highly processive enzyme possessing strand displacement activity. It has been demonstrated that it allows for extension of over 70,000 bases during a single binding event. Additionally, it can retain its activity for periods longer than 12 hours. This polymerase has previously been employed in rolling circle amplification (RCA), faithfully amplifying circular templates yielding a long linear molecule composed of tandem repeats of the circular template. Prior to this, other enzymes, particularly E. coli DNA polymerase I, were demonstrated to be amenable for RCA. These studies were, however, limited to very small circular templates. In MDA, random hexamer primers together with the high processivity and strand displacement capability of ϕ29 DNA polymerase enable amplification all over the genome. Displaced amplified regions create new primer binding sites allowing the reaction to continue, eventually producing large branched products. As no denaturation is required to provide new primer annealing points, the amplification can performed isothermally. An assessment of an assortment of WGA methods using massive parallel 454 sequencing revealed that none of these is free from bias. However, MDA strategies introduce least bias and generate the highest amplification yields.

2.3. Isothermal targeted amplification

In addition to the above-mentioned RCA approach, several other methods for isothermal targeted amplification exist. Since these eliminate the need for a thermal cycler they could potentially aid in development of simple diagnostic devices to be used in the field or point-of-care. Additionally, these methods could prove valuable when handling sensitive samples prone to heat degradation. As a number of polymerases capable of efficient nucleotide incorporation at ambient and moderate temperatures have been identified, it is the heat-based strand-melting step that represents the major obstacle for an isothermal technique.
In helicase-dependent amplification (HDA) heat denaturation has been replaced by
unwinding of double-stranded DNA with a helicase enzyme. In its first version the
reaction was performed at 37°C and termed mesophilic HDA (mHDA)\textsuperscript{99}. Charac-
terization of a thermostable helicase enabled higher temperatures (60 – 65°C) leading
to increased sensitivity and specificity in thermophilic HDA (tHDA)\textsuperscript{100}. A
drawback of this method is that only relatively short fragments – shorter than 120 – 200
bases – can be amplified\textsuperscript{100, 101}. However, fusion of the helicase to the polymerase with a
coiled-coil creating a ‘helimerase’ to facilitate the synchronous action of the two enzymes
has made possible the amplification of approximately 2 kb long regions\textsuperscript{102}. Other
isothermal approaches are also available. Many of these suffer from complicated
protocols. For instance, strand displacement amplification (SDA) entails four different
primers and an alternating nicking and extension/displacement process\textsuperscript{103, 104}.

2.4. Multiplexity issues with PCR

Simplex PCR, amplification of a single fragment in a single reaction tube, is in the
majority of cases a rather straightforward task not necessitating extensive efforts of
optimization. However, most studies focus on a number of interesting regions and
selectively amplifying these one by one is time-consuming and expensive. Global analysis
of human exons illuminates the magnitude of such a one by one strategy. If the setup of
each PCR requires five minutes and if each of the approximately 20,000 protein-coding
genes encompasses 10 exons, 200,000 individual PCR amplifications have to be prepared
per individual to be analyzed, equaling to a time frame of one million minutes. This is
equivalent to about 700 days, or 2 years, given working round the clock. Moreover, the
time required for the thermal cycling and the high instrumentation demands are not
taken into account. Of course, the whole-genome amplification approaches outlined
above could be employed. But by amplifying on a genome-wide scale, thus not exclusively
the segments of interest, the ability to select for these regions only, as well as the reduction
of complexity in the reaction outcome, are lost. A targeted multiplex amplification can, on
the other hand, be achieved using multiplex PCR. It was first demonstrated in the end of
the 1980s, when 6 regions were amplified in parallel – a 6-plex – in a procedure to find
deletions in the human dystrophin gene\textsuperscript{105}.

The shift from simplex PCR to its multiplex variety, although conceptually simple – just
add all the necessary primer pairs and run – is, in fact, rather complicated and filled with
potential pitfalls. Every target to be amplified naturally necessitates the inclusion of an
appropriate primer pair, but each extra primer in the reaction increases the risk of
primer-dimer formation and of unspecific priming\textsuperscript{106–108}. Primer-dimers are viable and
amplifiable species created when two primers anneal in a way creating free extendable 3’-
ends. Several mechanisms are responsible for formation of these artifacts\textsuperscript{109}. The number
of potential primer-dimers is \((n^2 + n)/2\), where \(n\) corresponds to the total number of
primers\textsuperscript{110}. Amplification of 10 distinct regions (20 primers) results in 210 potential primer
interactions. This number rises to 20,100 when 100 segments are considered. Since these
artifacts are shorter than the desired products they can effectively take over the whole
amplification reaction. This is because multiplex PCR is essentially a competition for the
available resources between all the products, and the shorter ones are superior
contenders.

The problems of multiplex PCR are further exacerbated by unspecific priming. This
arises when one primer from a primer pair chosen to amplify a certain region generates a
spurious product together with a primer from another pair targeting a different locus.
This is referred to as cross-talk with the potential number of interactions being \(2n^2 + n\),
where \(n\) corresponds to the number of targets\textsuperscript{108}. For example, 10 targets imply a
possibility for 210 cross-reacted products. A 100-plex can produce 20,100 unspecific interactions. While primer-dimer artifacts can be overcome, at least to a certain extent, by careful primer design, unspecific priming is more difficult to tackle.

Even if the frequency of the above-mentioned artifacts is kept at a minimum, multiplex PCR can still generate inferior results by preferentially amplifying some regions while completely missing others\textsuperscript{106, 107}. This is related to the contest nature of this reaction, where all the amplicons compete for a limited supply of enzyme and nucleotides. This competition is influenced by a number of factors, such as product length, sequence and priming efficiency. For example, as mentioned above, shorter fragments are preferred to longer ones. Additionally, certain genomic regions may be more inaccessible then others, for instance because of secondary structures, leading to poor amplification. In general, the efficiently amplified segments outcompete the less efficiently amplified regions resulting in uneven product amounts.

The spurious reactions brought about by increasing the number of different primers imply that only a certain extent of multiplexing is possible. Commonly employed multiplex PCR reactions involve simultaneous amplification of about 10 to 20 loci. For instance, a 16-plex covering 15 STRs and the sex-determining marker amelogenin, is the mainstay in forensic DNA typing\textsuperscript{111}. Unsurprisingly, highly multiplexed reactions necessitate careful primer design and condition optimizations. Software packages aiding in the multiplex primer design process by assessing potential for primer-dimer and hairpin formation have been developed\textsuperscript{110, 112, 113}. One such computer program, minimizing primer complementarity at 3’-end regions and on a global scale, was used to evaluate primer interactions eventually permitting a >1000-plex. About 90% of the segments initially chosen were approved for the multiplex\textsuperscript{114}. This high degree of multiplexing is, however, somewhat of an exception. Also, it is important to bear in mind that meticulous empirical testing is still necessary, despite the availability of computer algorithms.

Several simple modifications can greatly improve the outcome of multiplex amplifications. Adjustment of primer amounts for each region and of the primer-to-template ratio can be beneficial\textsuperscript{106, 107}. Furthermore, a highly multiplex PCR can be subdivided into several smaller reactions. For example, in a SNP genotyping study, amplification of all 558 loci in a single amplification had a success rate of 50%. Subgrouping these into 24 bins of approximately 23 loci increased the rate of success to 92%. Intermediate partitioning into 12 sets of about 46 loci, and into 6 bins of about 92 loci, generated 90% and 85% success, respectively\textsuperscript{115}.

2.4.1. Two-stage amplification
To lessen the impact of primer-dimers a two-stage amplification can be performed where initial priming with locus-specific primers is replaced with universal primers at a later stage in the reaction. Such approaches require concatenated oligonucleotides, where the region-specific portion is positioned 3’ of the universal handle sequence. The conversion from the multiplex phase of the amplification to the ‘simplex’ one can be accomplished in a number of ways. The multiplex genotyping system (MGS) entails three PCR rounds, the two opening ones introducing one handle each, and the final round relying on a single primer pair complementary to the handles. An aliquot of the previous round serves as the template in the forthcoming round. 26 regions were successfully amplified with this system\textsuperscript{116}. The universal tag in Homo-Tag Assisted Non-Dimer System (HANDS) is designed to have a higher melting temperature than the specific primer portions, hence enabling a shift from multiplexed to universal priming simply by raising the annealing temperature. This method was applied in a 4-plex genotyping assay\textsuperscript{109}.
2.4.2. Compartmentalization

Compartmentalization – spatial separation of individual PCR reactions – is also a possibility for simultaneous amplifications. With this strategy, only a single primer pair is present in each compartment, thus mimicking regular simplex PCR and preventing uncontrollable primer interactions.

2.4.3. Emulsion PCR

Water-in-oil emulsions, easily created for instance by slowly dropping a water-based mixture to continuously stirred oil containing detergents117, encompass large numbers – about $10^{10}$ compartments in 1 ml of emulsion – of stable cells suitable for biochemical reactions118. Once the reaction is finalized the emulsion can be broken, pooling the aqueous compartments, returning to an oil-free system. The droplets drastically reduce the volume. A standard PCR is performed in 50 µl, whereas a 1 µm droplet has a volume of less than a femtolitre119. Initially, transcription and translation were performed inside such miniscule chambers in an assay for catalyst gene selection118. The first report on PCR amplification in emulsion emerged in 2001 in an approach for directed evolution of polymerases120.

PCR primers can be coupled to beads prior to the emulsification. This permits amplification on beads inside the droplet, eventually covering the whole bead surface with identical copies of the chosen DNA region. A primer present in the surrounding solution is used to increase amplification efficiency121. Such a bead-based emulsion PCR approach is currently employed by a number of massive parallel sequencing platforms, including 454122 and SOLiD123. However, emulsions have certain limitations. The bead number and template amount have to be thoroughly evaluated in order to maximize the number of reaction chambers containing exactly a single bead and single template, as multiple beads and templates can give rise to artifacts121. The polydisperse nature of the droplets is not optimally suited for Poisson statistics, necessitating experimental validation to find the most favorable ratios and leading to a many empty chambers. Another drawback is that it is relatively difficult to manipulate the contents of the compartments once the emulsion is created119. Nonetheless, novel microfluidic devices may be able to resolve these problems124.

2.4.4. Solid-phase PCR

Solid-phase PCR offers another means of spatial segregation of individual amplifications. In these approaches the reaction is performed with immobilized primers on a solid support, for instance on a glass slide or in a gel. In a former approach, the primers were in solution125. However, the reaction was carried out in a dense polyacrylamide matrix that hindered product diffusion generating demarcated gel regions containing multiple copies of the same template. These regions were referred to as PCR colonies or ‘polonies’. In addition, modification of one of the primers enabled attachment to the matrix. However, some primer-dimers were evident in this method.

Bridge amplification relies solely on immobilized primers on a glass slide126. Template annealing to a covalently attached primer and the following polymerase extension generates an immobilized complementary copy of the template. This molecule can subsequently anneal to the surface-bound second oligonucleotide of the primer pair forming a bridge or loop-like structure. Following another round of extension both strands are immobilized and this process is repeated in a cyclic fashion. One drawback of this approach is the low efficiency in comparison to traditional liquid-phase PCR. A type of solid-phase amplification is part of the sample preparation protocol for massive Illumina/Solexa sequencing127, 128.
2.4.5. **Targeted compartmentalization**

In emulsion PCR and the solid-phase methods mentioned above, the same primer is used in all droplets or all over the solid support. Multiplex microarray-enhanced PCR (MME-PCR) is a targeted approach where different regions are amplified in a compartmentalized fashion\[^{129}\]. In particular, it entails concatenated primers comprising both specific and universal regions. Primer pairs for different segments are deposited in discrete locations. The immobilized primers instigate the reaction, which is then driven by a solution-phase universal primer pair eventually resulting in numerous copies of each template attached to the solid support. 60 individual amplifications were simultaneously performed with this technique. A 75 SNP region set was amplified using an akin method termed MegaPlex PCR\[^{130}\]. However, with this approach, the products end up in the solution.

2.5. **Ligase-mediated amplification**

Ligases constitute a group of enzymes capable of joining adjacently annealed DNA molecules. In particular, these enzymes can link two oligonucleotides hybridized end-to-end on a template by creating a covalent bond between the 3\(^{-}\)OH group of the upstream oligonucleotide and a 5\(^{-}\)-phosphate on the downstream one. Perfect base pairing at the junction is a prerequisite for the reaction to occur\[^{131, 132}\]. This capacity of ligases was exploited already in the 1970s\[^{133}\]. In the late 1980s, the specific nature of this linking event was utilized for sequence variant detection\[^{134, 135}\]. Subsequently, a number of novel ligase-based amplification strategies emerged providing alternatives to conventional PCR.

2.5.1. **Ligase chain reaction (LCR)**

In 1989, the ligation amplification reaction (LAR) was described\[^{136}\]. In this approach, two oligonucleotides are designed to anneal in a side-by-side fashion to one template strand. Another pair of oligonucleotides is created to adjacently bind to exactly the same region, but on the complementary strand. Naturally, this region corresponds to the segment to be amplified. Perfect base matching at probe interfaces allows for ligation, and these joined molecules can act as templates in the next cycle by permitting the complementary oligonucleotide pair to anneal and be linked. Accordingly, a thermal cycling allowing denaturation, annealing and ligation enables the products of one cycle to be employed as templates in the next one resulting in an exponential amplification. Similarly, by utilizing only a single probe pair hybridizing to one of the template strands, linear amplification is obtained. The high discriminatory power of the ligase precludes ligation of probes not perfectly matched at the boundary. Interest in ligation-dependent amplification rose with the cloning of a thermostable ligase, enabling the reaction to be run at considerably higher temperatures, thus increasing the specificity\[^{137}\]. Another benefit was that the need to replenish the enzyme before commencing a new round of ligation was eliminated\[^{136}\]. This rather more practical method was named ligase chain reaction (LCR). The linear amplification strategy was labeled ligase detection reaction (LDR).

LCR has a lower efficiency compared to PCR\[^{138}\]. Another dilemma is the target-independent artifact formation by blunt-end ligation\[^{137}\]. In PCR, the unspecific primer-dimers are usually shorter than the desired products, and this length difference can be employed in assessing amplification success, for instance by gel electrophoresis. Conversely, the artifacts in LCR are of the same length as the specific products and hence indistinguishable\[^{139}\]. There are several strategies for reduction of non-specific ligation products\[^{136, 137, 139}\].
2.5.2. Gap-LCR
An alternative means to reduce unspecific amplification involves inclusion of a small gap between the two annealed probes. Addition of a polymerase to the reaction mixture leads to filling of this gap, after which the two, at this moment adjacent, fragments are linked by a ligase. The polymerase in this process has to lack 5' -> 3' exonuclease activities so as not to interfere with the annealed downstream oligonucleotide. The theoretical framework for such an approach was presented in 1991\textsuperscript{139, 140}. Subsequently, an asymmetric gap-LCR was developed to allow studies on RNA, in which the presence of all four nucleotides in the cDNA synthesis reaction could interfere with the precise gap filling required for the amplification. Here, the composition of the gaps is restricted to three or less of the four possible nucleotides leading to a much more controlled gap bridging\textsuperscript{141}.

2.5.3. Padlock probes
Padlock probes can be seen as an extension of the approaches relying on specific ligation of two adjacent oligonucleotides. Here, a single probe is used\textsuperscript{142}. Its 5'- and 3'-ends encompass locus specific sequences with a linker segment present in-between. Upon hybridization the specific portions end up in a side-by-side arrangement enabling ligation and, concomitantly, a probe closure creating a circular molecule. DNA structure characteristics lock the circularized probe onto the template. Accordingly, it can only move along the template axis. The first demonstration of the padlock principle did not entail any exponential amplification. However, cycling of the reaction permitted linear amplification, but only on linear templates. The inability of the circularized probe to slide off a circular template hampered amplification as the locked-in probes simply annealed to their templates before fresh probes had the possibility to do so. Since the critical hybridization/ligation is an intramolecular event the probe amount can be kept low. This helps in reducing interaction between probes of different loci, thereby conferring specificity\textsuperscript{142, 143}. Furthermore, enclosure of the template, as well as the circular nature of the ligated products, enable several discriminatory procedures differentiating circularized from linear probes, maximizing the input of correct product to the subsequent step. This can be accomplished by, for example, exonuclease treatment degrading non-circularized probes\textsuperscript{143}. A disadvantage concerning padlocks is the long probe length, which can compromise probe quality\textsuperscript{143}.

The rolling circle amplification (RCA) strategy offers a means to amplification of ligated padlocks. The small circular products are perfect templates in RCA that generates long DNA stretches composed of tandem copies of the original probe\textsuperscript{96}. However, it was demonstrated that the locking of a circularized probe to its template might hinder efficient polymerization. The presence of a free end on the template in a vicinity of the probe annealing region was proposed to be a prerequisite for successful amplification. With this scenario, the polymerase can displace the hybridized probe, which is then free to slide off the template and participate in RCA\textsuperscript{95}. In solution, the RCA products form random coils, a practical property for detection purposes. Hybridization of fluorescent-labeled padlock-complementary oligonucleotides coupled with fluorescence microscopy can readily identify the coils\textsuperscript{144}. This principle was implemented in a microfluidic digital quantitation device\textsuperscript{145}. Furthermore, another detection strategy is based on alteration of magnetic properties of magnetic nanobeads as they hybridize inside the coils\textsuperscript{146}.

2.5.4. Multiplex renditions of padlocks
Two multiplex implementations of padlocks have been reported. One of these involves only a ligase for probe circularization\textsuperscript{143}. Non-ligated probes are removed by exonuclease digestion. All probes encompass universal amplification tags that are capitalized upon for a PCR reaction with common primers. However, as with RCA, PCR is hampered by
locked padlock probes, albeit to a less extent. Therefore, prior to PCR extension, the DNA has to be fragmented to provide free ends, which the circularized probes can slide off.

The second approach, utilizing molecular inversion probes (MIPs), complements the ligase with a polymerase. Specifically, probe annealing creates a one-base gap necessitating the concerted action of the two enzymes for circularization. Following exonuclease treatment eliminating linear probes, the circular products are released from their templates by cleavage at uracil residues, which were embedded in the probes during synthesis. As the probe opening occurs at a position distinct from the ligation site the probes are effectively inverted. As in the above-mentioned approach incorporation of universal tags in the linker sequences allows for simultaneous PCR of all probes. Since this method requires proper hybridization, gap polymerization and ligation, it offers an extra level of specificity. Additionally, for SNP genotyping purposes, it requires only a single probe per polymorphism, eradicating possible probe inequalities compared to a two-probe assay. However, it necessitates four separate reactions, one for each nucleotide. Conversely, the ligation-only approach is performed in a single tube at the expense of demanding two probes to score each SNP.

The ligation-only approach was initially performed as a 13-plex SNP analysis, whereas MIPs were employed to study over 1500 SNPs in parallel.

2.5.5. **Multiplex ligation-dependent probe amplification (MLPA)**

In multiplex ligation-dependent probe amplification (MLPA), two probes are designed for each locus to be analyzed. Adjoining annealing of this pair results in ligation. The probe ends not involved in ligation carry universal handles and linked probes of all regions can thus be PCR-amplified with a single primer pair. This resembles the two-stage PCR process described previously, but here, the initial step relies on ligation. Furthermore, each locus creates a product of a distinct length. This is achieved by inclusion of a stuffer sequence between the specific and universal portions of the downstream probe. This size distribution enables convenient analysis using electrophoresis. However, this strategy introduces a number of limitations. The stuffer sequences lengthen the probes, making their production troublesome. Additionally, PCR length bias acts on the differently sized products, and other inherent PCR predispositions, lead to nonuniform amplification limiting the magnitude of multiplexing.

An MLPA-study of 40 regions entailed products between 130 and 480 bp. The most widespread application has been copy number variation analysis, but it has also found uses in, amongst others, SNP genotyping and expression profiling.

2.6. **What happens post-amplification?**

Amplification plays a vital role in most assays. However, it represents only the first phase, the relevant information such as the genotype or level of gene expression acquired during a subsequent detection, interrogation or scoring step. Consequently, other methods take over translating the amplified material into data answering questions about various biological issues. There is a multitude of assays for SNP interrogation and analysis of gene expression and these will be described in the forthcoming chapters. Furthermore, the majority of current second generation massively parallel sequencing platforms incorporate amplification in the sample preparation pipeline. The different chemistries involved in producing the sequences will also be presented.
3. Genotyping

On a genomic scale, two persons are 99.9% alike, yet it is the remaining diminutive variable fraction that accounts for the differences between individuals. For example, these variations can offer an explanation to why we look different and why certain individuals are more prone to acquire particular diseases. Of course, the genetic makeup does not act alone. The environment acts against this genetic background shaping us to what we are. The process of elucidating an individual’s repertoire of variation – the genotype – is referred to as genotyping. The utility of this genotype information lies in the clues it can provide about a person’s phenotypical traits, such as physical characteristics or diseases.

Genetic variants come in a number of varieties. These, together with large-scale efforts to create catalogs of human variation, have been described in Chapter 1. The two types of common variants having been the subject of most studies are microsatellites and single nucleotide polymorphisms (SNPs). The difference between a common variant – a polymorphism – and a rare one – a mutation – pertains to the frequency of the minor allele, with 1% being the dividing frequency.

3.1. Historical overview

Historically, restriction fragment length polymorphism (RFLP) was the first genotyping method to be employed. It is based on enzymatic discrimination. In particular, the specific cleavage of restriction endonucleases only at well-defined recognition sequences, most often four or six bases in length, is used for interrogation. By choosing a polymorphism within such a recognition site, the enzyme is only able to cleave if one of the alleles is present. The other allele disrupts the recognition site, preventing action of the endonuclease. Accordingly, homozygotes and heterozygotes produce digestion products of varying lengths and these patterns can easily be visualized.

Minisatellites – tandem repeats with a 10-60 bp motif frequently displaying length polymorphisms – were introduced in mid-80’s. The introduction of the polymerase
Parallel target selection by trinucleotide threading

chain reaction shifted focus from the longer minisatellites to the shorter and hence more easily amplified microsatellites, in which the repeated unit is between 1 and 6 bp\textsuperscript{155-158}. Single nucleotide polymorphisms (SNPs), owing to their prevalence and potential for high throughput automated readout, represent the marker class most frequently used in the last decade. However, in the recent years structural variants, including copy number variations, have gained progressively more attention.

3.2. Short tandem repeats (STRs)
Microsatellites, or short tandem repeats (STRs), are plentiful markers composed of a 1-6 bp motif repeated in a tandem fashion, the variation relating to the precise number of repeated units at the particular STR locus\textsuperscript{155, 159, 160}. Since this number of repeats can vary within a wide range, microsatellites represent a highly informative marker group\textsuperscript{161, 162}. However, as the analysis of these markers is difficult to perform in a parallel fashion, they have progressively been replaced by SNPs, the interrogation of which is inherently easier to parallelize. Nevertheless, owing to their high variability, STRs are routinely employed for identity establishment, for instance within forensics or in paternity testing\textsuperscript{161}.

The dominating STR assay relies on multiplex PCR of a number of STR regions, followed by scoring using fragment analysis with capillary gel electrophoresis. The multiplex amplification step is carefully designed to minimize formation of spurious products, as well as to generate products of distinct lengths to avoid overlapping peaks in the detection step. Artifact formation and issues related to the need for resolved peak groups are both exacerbated when increasing the number of microsatellites in the multiplex PCR, and preclude highly scalable assays. For example, forensic DNA analysis is based on simultaneous amplification and fragment analysis of 15 STR loci and the sex-determining marker amelogenin\textsuperscript{111}.

3.3. Single nucleotide polymorphisms (SNPs)
SNPs – single-base variations where one nucleotide is exchanged for another – represent the most abundant class of markers. Over 10 million such polymorphisms are estimated to span the human genome\textsuperscript{28, 163}, implying one such marker every 300 bases. Most of the SNP loci have only two base options, for instance A/G, and are, consequently, biallelic. As such, SNPs are less informative than STRs, which can vary considerably in the number of repeated units. Conversely, as SNPs are a quite homogenous marker group, they are amenable for highly parallel interrogation.

Unsurprisingly, all of the SNP discrimination strategies rely on an event taking place when allele 1 is present, generating one type of signal. Likewise, a similar event, producing another type of signal, occurs only in the case of allele 2. There is a multitude of SNP scoring principles and these can be broadly classified either as hybridization-only or enzymatic depending on how the typing specificity is attained. In the majority of approaches the sample has to be amplified prior to the SNP-detection step.

3.4. Low-scale genotyping
Not all discrimination strategies are compatible with a high degree of multiplexing. Instead, some of these are better suited for investigating few diagnostic SNPs in a large number of samples. These are presented below, whereas the highly parallel assays, capable of interrogation of millions of SNPs, will be described later in this chapter.

3.4.1. Allele-specific oligonucleotide hybridization (ASOH or ASH)
Allele-specific oligonucleotide hybridization (ASOH or ASH) is a principle that is well adapted for interrogating small sets of SNPs, but as will be described further on, can also
be applied to large-scale genotyping. The discriminatory power of this strategy is bestowed by the higher thermal stability of perfectly complementary oligonucleotides compared to ones with a single mismatch\textsuperscript{164}. A probe perfectly matching one SNP allele will, obviously, carry a mismatch for the other one. Therefore, by designing two probes, complementary to the two alleles, and carefully choosing hybridization conditions, especially the temperature, only the fully matched probes will hybridize, thereby scoring the SNP.

In the beginning of the 1980s, ASH was employed to detect the sickle cell β-globin allele\textsuperscript{165}. In this study, hybridization was performed on digested and electrophoresis-separated genomic DNA. Even though this sample preparation reduces the complexity to some extent it is still challenging to interrogate a single SNP within this context. With the entrance of PCR\textsuperscript{70, 71} came the possibility to amplify the SNP-harboring region thus considerably reducing complexity in the hybridization step leading to more convenient assays. Two strategies were employed. In the normal procedure, PCR-amplified material was deposited onto a series of membranes, each of which were then assayed with labeled oligonucleotide probes\textsuperscript{166}. Conversely, the reverse procedure immobilized the probes letting labeled PCR-generated target hybridize\textsuperscript{167} in a manner reminiscent of more modern high-density arrays. The fundamental limitation of this technology stems from each probe set requiring different hybridization conditions depending on, among other factors, oligonucleotide length and composition\textsuperscript{138}. As proper hybridization is crucial for specificity, combining probes for several SNPs in multiplex assays may lead to spurious cross-hybridization introducing non-specific noise.

### 3.4.2. Counteract hybridization differences between probe sets

One way to counterbalance probe composition differences is to perform the hybridizations in tetramethylammonium chloride (TMAC), which brings the hybridization strength of A-T and C-G closer to one another\textsuperscript{168}. However, not all differences can be eliminated\textsuperscript{169}. Another possible solution is to use redundant probes for each SNP. For example, an Affymetrix approach from 1998 employed 10 oligonucleotides per SNP\textsuperscript{170}. In later approaches, 40\textsuperscript{87, 88} or 56\textsuperscript{86} 25-mer probes were employed. With the 40-probe strategy, five probe quartets target the sense strand and five the antisense strand. In one quartet for each strand, the central base corresponds to the SNP position, with two probes matching the two alleles, and the remaining ones covering the two mismatch situations. This is supplemented with four offset probe quartets, where the SNP position is no longer occupying the middle position. In these offset probes both the central positions and the SNP site are varied\textsuperscript{87}. The redundant probe strategy leads to a complicated translation of array intensities to genotype data. This is facilitated by computer algorithms\textsuperscript{171-173}. However, even with redundant probes, this hybridization-only approach has problems discriminating between the genotypes for a number of SNPs.

### 3.4.3. Dynamic allele-specific hybridization (DASH)

Dynamic allele-specific hybridization (DASH) is a method able to handle the diverse temperature requirements of different probe sets\textsuperscript{174}. In this approach, the amplified target is immobilized and a probe complementary to the SNP region is allowed to anneal at a low temperature. The fluorescent signal generated upon interaction of double-stranded DNA with a dye is monitored as the temperature is increased. The dissociation of the probe from the template is evident as a loss of this signal. This occurs at different temperatures depending on whether perfect complementarity (i.e. allele 1) or a single-base mismatch (i.e. allele 2) is formed. The number of peaks in a melt curve reveals the SNP genotype: one peak corresponds to a homozygote, whereas two indicate a heterozygote. The temperature differences range between 4°C and 15°C\textsuperscript{175}. By successively probing the
Parallel target selection by trinucleotide threading

template with oligonucleotides for both alleles more reliable data can be generated\(^\text{174}\). A limited throughput increase can be reached by sequential hybridizations with different probes on multiplexed PCR products. It was demonstrated that at least five such probings do not compromise assay quality. Another possibility is to deposit multiple simplex PCR products on a membrane and assessing the SNPs with probe cocktails\(^\text{175}\). However, these multiplexing strategies are cumbersome.

3.4.4. **Solution-phase hybridization**

The approaches mentioned above rely, in one form or another, on immobilization of either the probes or the targets. However, hybridization can also occur in homogenous solution-phase assays. With these, SNPs can simultaneously be amplified and scored in a real-time system. Hairpin loop structure Molecular Beacons, becoming linear upon hybridization, form the basis of one such approach\(^\text{176}\). They have been described in Chapter 2. By employing two beacons, each assigned to one of the alleles and labeled with a specific fluorophore, SNPs can be interrogated\(^\text{176}\). A beneficial property of the stem loop configuration is that this constriction confers additional specificity, as mismatched hairpin probes are considerably less stable than similar linear ones\(^\text{176}, \ 177\).

Another homogenous solution-phase technique uses TaqMan-probes, also discussed in Chapter 2\(^\text{180}\). As only perfectly matched probes will be degraded by *Taq* polymerase and generate a signal, a dual color two-probe assay, each probe corresponding to one of the alleles, can be used for SNP genotyping\(^\text{178}, \ 179\).

Both Molecular Beacons and the TaqMan probes offer simple, single tube assays that can be monitored in real-time. The dual-labeled probes necessary are, however, rather expensive.

3.4.5. **Allele-specific PCR (AS-PCR)**

A polymerase is required to discriminate SNPs in allele-specific amplification assays, such as allele-specific PCR (ASPCR)\(^\text{180}\) or amplification refractory mutation system (ARMS)\(^\text{181}\). Specifically, two primers whose 3’-ends coincide with the SNP position, one primer for each allele, are included in the reaction together with a common reverse primer. The 3’-mismatches prevent polymerase extension leading to generation only of products primed by perfectly complementary oligonucleotides. In this assay, the polymerase has to lack proofreading activity as this would correct the mismatched terminal nucleotides generating products of both alleles irrespective of target genotype. *Taq* polymerase is one enzyme that can be used\(^\text{180}, \ 181\). However, in some cases the polymerase initiates extension even from terminally mismatched oligonucleotides generating incorrect genotypes\(^\text{181}\). Moreover, this being a PCR-based approach, it suffers from the various biases and artifacts that were outlined in Chapter 2.

3.4.6. **Invader**

The Invader assay exploits the unique characteristics of a group of enzymes termed flap endonucleases (FENs)\(^\text{182}\). In particular, these possess the ability to specifically cleave invaded overhanging structures formed when an upstream probe invades and displaces a downstream probe annealed to the template. In the SNP scenario the 5’-terminal complementary base of the downstream oligonucleotide is at the polymorphism position. This probe additionally carries a 5’ non-matching segment creating an overhang. The upstream probe – the Invader – matches the target and ends at the SNP spot having a non-matching base at this 3’-terminal position. This probe design enables the Invader to invade the downstream probe, displacing the base at the polymorphic site, creating a conformation recognized and subsequently cleaved by the FEN. Similar to other
approaches, two downstream probes, one for each allele, are used. Only perfectly matched allele-specific probes create the structure necessary for FEN cleavage, achieving specificity. The cleavage products can be directly detected or participate in secondary reactions. Additionally, by performing the reaction near the melting temperature of the downstream oligonucleotides several of these are able to anneal and be cleaved contributing to signal amplification. The drawback of the Invader approach is its large template amount requirement.

3.4.7. Oligonucleotide ligation assay (OLA)
The specific linking of oligonucleotides by ligases only when these oligonucleotides adjacently anneal to a template creating perfect complementarity at the interface is the premise of oligonucleotide ligation assay (OLA). This approach, demonstrated in 1988, uses two upstream probes designed to anneal next to a downstream probe. Each of the upstream oligonucleotides corresponds to one of the variants and ends at the polymorphic base. Only probe duets fully matched at the junction will be ligated, whereas a single mismatch precludes this event. Originally, one cycle of OLA was used for detection following PCR amplification. Subsequently, sequential cycling of the ligation assay was enabled resulting in either linear or exponential amplification, as has been outlined in Chapter 2.

3.4.8. Specificity
Generally, SNP interrogation schemes involving enzymatic allele discrimination offer an additional layer of specificity as compared to technologies relying solely on hybridization. Furthermore, combinations of enzymes add another level of discriminatory power by channeling each enzyme’s particular specificity characteristics for allele scoring.

3.5. Detection platforms
Several detection platforms can be used in conjunction with the above-mentioned discrimination principles. For instance, in many assays, the various probes are labeled with fluorescent dyes, either single ones or in pairs that can be detected in fluorescence readers or with real-time PCR instruments. TaqMan and Molecular Beacon approaches rely on this kind of detection. Other platforms include mass spectrometric (MS) discrimination between products of different molecular masses, and size separation by various kinds of electrophoretic approaches. It should, however, be emphasized that the distinction between an SNP interrogation principle and detection system is, in some instances, not straightforward.

3.5.1. Pyrosequencing
Pyrosequencing is a detection system offering a possibility for rapid analysis of a small set of SNPs in many samples. Here, the SNP region is read by a sequencing-by-synthesis approach, which is described in Chapter 4. Briefly, incorporation of a nucleotide is converted to light emission by an enzymatic cascade, the light producing a peak. The sample to be genotyped is PCR-amplified, immobilized and made single-stranded. A primer is annealed and extended into the SNP region. Each allele gives an easily discernible discrete set of peaks. Sequencing over the SNP segment leads to flexibility in primer positioning, a benefit as the reaction, due to the thermolability of the enzymes, has to be run at modest temperatures. Furthermore, bases both preceding and following the SNP site can be read and used to calibrate the peak heights.

3.5.2. Microarray overview
Of the many detection platforms, DNA microarrays belong to the most commonly employed ones, predominantly because they enable miniaturized highly parallel assays.
Broadly, these can be classified into conventional planar chip arrays and bead-based arrays. The former category can be subdivided into spotted arrays and \textit{in situ} synthesized chips. The latter group encompasses random bead arrays and suspension bead arrays. Of these types, the suspension bead format is suitable for small to modest number of probes, whereas \textit{in situ} synthesized arrays and random bead systems represent the other end of the spectrum, allowing high level of parallelization. Spotted arrays fall somewhere in between\cite{188, 189}.

3.5.3. \textbf{Spotted arrays}

Spotted arrays are manufactured by printing robots, whose pins carry miniscule volumes of modified oligonucleotides from conventional 96-well or 384-well plates and deposit them onto defined regions of activated planar slides enabling a covalent coupling\cite{63}. The printing process can either be contact-based, the release of oligonucleotide being an effect of the pin touching the chip surface, or, performed in a non-contact form, where the probe is ejected onto the slide. Non-contact printing is capable of depositing smaller volumes and, additionally, missing spots can easily be filled in. One drawback with contact spotting is that meticulous optimization of conditions, such as temperature and humidity, is required for reproducible quality. Spotted arrays can be produced in house and are cheaper than the commercial alternatives. Additionally, since the probe synthesis and deposition are separated, classical and reliable oligonucleotide synthesis can be used, together with careful purification steps, to ensure maximum probe quality. However, printing large sets of probes is time-consuming, due to its ‘one probe per pin and event’-nature, leaving the probe plates exposed for prolonged time periods\cite{188}.

3.5.4. \textbf{In situ synthesized arrays}

With the \textit{in situ} systems the oligonucleotides are created from scratch directly on the slide surface. The first strategy employed photolithography and was initially used to synthesize peptides\cite{190}. Subsequently, oligonucleotide arrays were manufactured, first by liquid-based synthesis\cite{169} and then using light-activation\cite{191, 192}. Currently, a large number of different photolithographic arrays are commercially available from Affymetrix\cite{193}. By using blocked nucleotides able of light-activation, coupled with illumination of features where incorporation is desired while masking the other positions, oligonucleotides can be built up in a precise manner. This entails sequential addition of the blocked building blocks and a set of masks guiding the synthesis process. This manufacturing process allows for high-density arrays. For example, the custom GeneChip arrays can be fitted with 1.3 million unique probes\cite{194}. Additionally, the Genome-Wide Human SNP Array 6.0 is capable of analyzing approximately 1.8 million genetic markers, about half of which are SNPs and the rest copy number variations\cite{195}. However, the oligonucleotides are limited to 25 bases\cite{194}. Additionally, the requirement of masks severely limits the flexibility, as a new set of masks has to be created in order to add or replace probes. Naturally, this increases the costs associated with custom arrays\cite{188}.

One alternative is to use a maskless, digital micromirror array approach to control the light activation\cite{196}. This virtual mask system reduces cost and allows for flexibility. Furthermore, the synthesized probes can be attached both at their 5' and 3'-ends, whereas the Affymetrix procedure invariably attaches the 3'-end due to the use of 5'-photoprotected groups\cite{188, 191}. Nimblegen-arrays produced in this manner carry 2.1 million >60-mer oligonucleotides\cite{197}.

A third \textit{in situ} type of array is available from Agilent and involves liquid-based synthesis. However, this reduces the feature density as compared to the systems employing light\cite{188}. Arrays featuring 244,000 probes with a maximum length of 60 bases are marketed\cite{188}.
3.5.5. **Microsphere suspension arrays**

Luminex has developed a bead-based suspension array platform, where the hybridization occurs in solution instead of on a planar surface. The heart of this approach is the 100-microsphere set, each bead type internally labeled with varying amounts of two dyes and thus distinguishable. Each bead is coupled to a probe in separate coupling reactions, after which the beads are mixed and added to a sample. Following hybridization, an instrument equipped with two lasers scans the beads, one used to decipher the internal labeling, the other allowing for analyte quantification. Recently, a novel system entailing microspheres carrying three internal labels, enabling 500 parallel assays, was introduced. Introduction of magnetic coded beads has facilitated sample handling.

Other benefits of this platform are the fast hybridization kinetics and rapid analysis.

3.5.6. **Decoded random bead arrays**

Naturally, it is vital to know the precise location of each probe on an array, irrespective of the format. In the platforms outlined above these positions are known beforehand. With Illumina random bead arrays, the decoding takes place post-manufacture. As with the Luminex, each bead is linked to its probe in a separate reaction. All the beads are then mixed and dispensed onto micro wells on optical fiber bundles. The decoding process involves sequential probe hybridizations generating either a green or red fluorescence signal or no signal at all. These decoding probings are possible by inclusion of barcode sequences in the probes. About 1520 different bead types were distinguished by using eight rounds of hybridization, each round capable of producing one of three states (green, red, no signal). Increasing the number of states to four allows up to 65,536 beads to be identified. The decoding rounds additionally produce a unique code for each bead type that can be employed for error-checking purposes. Furthermore, each probe is represented on the array by about 30 beads allowing for more reliable signals and a sound statistical analysis.

Currently, custom arrays featuring plasma-etched wells on a silica substrate contain 200,000 bead types each in 12 discrete regions allowing 12 samples to be analyzed per chip. Additionally, about 1.2 million SNPs are interrogated in 2 samples on each Human1M-Duo v3 chip. The format encompassing optical fiber bundles carries up to 1536 bead types in a 96-sample design.

3.6. **Highly parallel SNP genotyping**

The high-density array formats are, as mentioned, tailor-made for highly parallel genotyping approaches and the four major commercial systems all employ one or another array variant for detection. Two of these systems – the Illumina GoldenGate assay and the Affymetrix molecular inversion probe (MIP) approach – involve generic tag arrays, offering flexibility and ease of customization. These enable up to 20,000 polymorphisms to be investigated in parallel. The remaining platforms - Affymetrix’s whole-genome sampling analysis (WGSA) and Illumina’s Infinium assay – are capable of genome-wide genotyping while sacrificing some of the flexibility of the former approaches. With these, about one million SNPs can be simultaneously assayed. As opposed to the SNP scoring principles already mentioned, which are problematical to scale up and therefore suitable to the ‘few SNPs in many samples’ scenario, the multiplexed approaches offer streamlined sample preparation, allele interrogation and detection steps to accommodate for the parallelization.

3.6.1. **Allele-specific extension (ASE) and single-base extension (SBE)**

Two discriminatory principles are well adapted to massive multiplexing. Allele-specific extension (ASE) is a variant of allele-specific PCR approaches described earlier. Here,
only the two allele-discriminating primers, with their 3’-ends located over the SNP position, are used to prime chain elongation. This commences only in the case of perfect complementarity. As the common reverse primer is omitted, no amplification can take place, making this purely a detection strategy.

One limitation of ASE is that a polymerase has the potential to extend even from terminally mismatched primers, implying that erroneous genotypes may be generated. This, however, only applies if the polymerase is given enough time, as elongation starting from a fully complementary oligonucleotide has much faster kinetics than that of a mismatched primer. Accordingly, one strategy to circumvent this ASE problem is to restrict the amount of time that an intact extension reaction persists by eliminating some of the components. Apyrase-mediated allele-specific extension (AMASE) involves addition of an apyrase, an enzyme capable of nucleotide degradation\textsuperscript{208, 209}. In this manner the nucleotides will be degraded before the spurious elongation can take place. Conversely, extension from matched primers is rapid and will transpire. Protease-mediated allele-specific extension (PrASE) represents a refinement of AMASE where the thermolabile apyrase is substituted for a thermostable protease, able to digest the polymerase\textsuperscript{210}. This alteration allows the reaction to be performed at elevated temperatures, both increasing the hybridization stringency and the selection of polymerases to employ.

Minisequencing, or single-base extension (SBE) as it is sometimes referred to, also involves extension. However, this strategy, devised in the boundary between the 1980s and 1990s, utilizes only a single primer ending at the base preceding the polymorphic site\textsuperscript{211, 212}. This arrangement implies that the first base incorporated by a polymerase will correspond to the SNP. The interrogation part is achieved by splitting the immobilized template and performing distinct extension reactions with only one labeled nucleotide present in each. Using separate dyes for the different nucleotides obviates the need of dividing the reaction\textsuperscript{213-214}. Consequently, the highly accurate manner in which a polymerase incorporates nucleotides bestows the method with discriminatory power.

Both ASE and SBE principles were employed in initial multiplex SNP genotyping implementations that eventually paved the way for the highly multiplex platforms of today. In these, ASE or SBE discrimination was used together with array-based readout\textsuperscript{207, 215, 216}.

The polymerase-relying distinction in ASE and minisequencing is in some approaches supplemented with the action of a ligase to provide another level of discrimination.

3.6.2. **Generic tag arrays**

A number of multiplexed genotyping assays incorporate hybridization to generic tag arrays to identify SNP loci\textsuperscript{147, 205-217}. The actual genotypes are, subsequently, obtained by scanning the array and reading the signals. The utilization of universal tag arrays is made possible by appending unique address sequences to the probes required for SNP discrimination. Naturally, each polymorphism is associated with one such sequence. By employing generic address handles, which are independent of the specific locus sequences, the same array can be utilized for all different assays. By simply exchanging one probe set for another, the same type of chip can be used to identify and interrogate diverse marker sets. This allows for increased flexibility and reduction of costs. Also, the address sequences can be meticulously chosen to possess similar characteristics and minimal cross-hybridization potential enabling reliable hybridizations at the same conditions. This is in stark contrast to ASO hybridizations, where every probe set requires distinct conditions, as outlined previously.
3.6.3. **GoldenGate**

The highly parallel GoldenGate system offered by Illumina incorporates many of the described principles and techniques\(^{205}\). It encompasses an allele-specific extension reaction followed by ligation to provide for discrimination. Two allele-specific oligonucleotides (ASOs) are allowed to anneal to immobilized genomic DNA, and the correctly matched primers will initiate allele-specific primer extension. These elongated oligonucleotides are subsequently ligated to a third annealed probe, referred to as a locus-specific oligonucleotide (LSO). The 5′-end of the two allele-distinguishing primers carries a different universal handle. A common tag is also present at the 3′-end of the LSOs, and in-between the specific portion of the probe and this tag is a unique 22- to 24-meric address sequence\(^{202}\). Following PCR with three universal primers – two distinctly labeled forward ones matching the two ASO handles and a reverse one complementary to the LSO-tag – the products are identified by hybridization of the unique address identifiers to generic tag random bead arrays. A single signal after scanning of the array indicates a homozygous sample, whereas dual coloring signifies a heterozygous SNP.

One interesting feature of this protocol is positioning of the amplification step after the extension-ligation SNP discrimination reaction. Most of the other approaches amplify relevant SNP-loci prior to genotyping. The post-discrimination amplification, however, enables reliable multiplex PCR with only three primers. As mentioned above, this method suffers to some extent from elongation from terminally mismatched primers. At the moment, the random bead arrays comprise 1536 bead types in a 96-sample format, making possible about 150,000 genotype data points in one run\(^{204}\).

3.6.4. **Molecular Inversion Probes (MIP)**

The rival system marketed by Affymetrix bears resemblance to the GoldenGate assay in several aspects\(^{147, 217}\). This approach involves molecular inversion probes (MIPs) that were outlined in Chapter 2. Each MIP contains two SNP-locus specific sequences at the ends. A linker sequence between these contains universal amplification handles and a unique address identifier tag. Upon correct annealing, the probe adopts a circle shape. However, a gap over the polymorphic base is present between the two annealed regions. Accordingly, to close the circle a single-base extension and a ligation must occur. This gap fill is performed in four separate reactions, each encompassing only one of the four nucleotides. Following circularization, non-reacted linear probes are digested with exonuclease. The probe is released from the template, de-circularized and inverted by cleavage at a uracil base, aligning the universal amplification handles for PCR. Single primer pair PCR is carried out, and the detection step entails hybridization of the address tag to corresponding anti-tag on an Affymetrix array.

As in the GoldenGate case, amplification occurs after the discriminatory step, both the action of a polymerase and ligase confer specificity and universal tag arrays enable flexibility. Here, however, only a single probe is sufficient for each SNP, eradicating the effects of different qualities of the two allele-specific probes. On the other hand, four separate reactions need to be performed. Initially, a single dye was used in all four reactions, which required hybridization to four chips. Via an intermediate two-dye-two-chip approach, nowadays four spectrally distinct dyes are coupled to detection on a single chip. It was demonstrated that this latter strategy was superior\(^{217}\). A benefit of not only analyzing the two possible SNP variants, but also the non-allelic bases, is that signals from the latter can be used as assay controls. In an initial study, 1517-plex genotyping was performed\(^{147}\). This level was subsequently increased to over 10,000\(^{217}\). Currently, Affymetrix offers custom assays targeting up to 20,000 SNPs\(^{218}\).
Padlock probes were used in a related genotyping strategy\textsuperscript{143}. Amplification using these circular probes has been described in Chapter 2. Two padlocks for the two SNP variants are annealed to the SNP region. Each of these has the SNP interrogation base in the 3'-terminal position. No gap is present between the two adjoining ends enabling ligation of perfectly matching probes. Following exonuclease treatment to remove unreacted padlocks and PCR amplification with common primers, the genotypes are discernible upon hybridization to a tag array. Compared to early MIP renditions, only one reaction is performed eliminating the variation that may occur when four separate typing reactions need to be carried out. Conversely, two probes are necessary for each SNP possibly introducing bias due to uneven probe qualities. In the padlock case, the ligation discriminatory step was also cycled.

### 3.6.5. Whole-genome sampling analysis (WGSA)

Whole-genome sampling analysis (WGSA) is a genome-wide genotyping system developed by Affymetrix\textsuperscript{86-89}. The unusual property of WGSA is the reliance solely on allele-specific hybridization for genotyping. To alleviate problems associated with cross-hybridization and unspecific signals, which are frequently encountered in simultaneous hybridization of multiple probe sets, a complexity reduction step is implemented reducing the number of unique bases to which the genotyping chip is exposed. Additionally, as outlined earlier, multiple probes, both completely complementary and carrying mismatches at various positions, are used to score each SNP. The complexity reduction has been described in Chapter 2. Briefly, the sample is digested with a restriction endonuclease, universal adaptors are ligated to the digestion products and a PCR with a single primer complementary to the adaptor is run. The conditions are carefully chosen to accommodate amplification only of products within a certain size interval. As the same restriction enzyme is used in all reactions, reproducible fractions of the genome are obtained. Following amplification, the sample is fragmented, end labeled and hybridized to an Affymetrix high-density photolithographic array.

Currently, the most parallel assay – the Genome-Wide Human SNP Array 6.0 – scores 906,600 SNPs and 946,000 CNVs. Two separate digestion reactions with \textit{NspI} and \textit{SstI} are carried out and the PCR step predominantly amplifies fragments between 200 and 1100 bases\textsuperscript{193}. However, in order to enable highly parallel assays, WGSA surrenders some of the ability to choose SNPs. Since a polymorphism has to reside in a digest of the correct size to complete the pipeline and be scored, this strategy puts restrictions to the SNPs that can be genotyped. Naturally, given the multitude of restriction enzymes presently available, a suitable one for a particular SNP can be found. Still, for large sets of polymorphisms, no enzyme covering all of these might exist. Pooling of final products from two different digestion reactions for each sample, as in the SNP Array 6.0, can offer a solution. Moreover, the whole genome has to be digested \textit{in silico} to assess which SNPs are assayable.

### 3.6.6. Infinium

Infinium, developed by Illumina, is a platform for whole-genome genotyping (WGG)\textsuperscript{206, 219, 220}. The distinguishing characteristic of this assay is the lack of complexity reduction as the interrogation reaction takes place in the presence of the whole genome. In other approaches, measures such as PCR are required to reduce the complexity in order to allow for robust genotyping, particularly pertaining to the hybridization-based detection step. The Infinium protocol begins with whole-genome amplification (WGA) followed by fragmentation. This ensures high concentration of appropriately sized products enabling specific hybridization to locus-specific probes attached to beads in a random bead array. The 25 most 5'-bases of the probes are used for the decoding hybridizations, while the 50
bases at the 3’-end target the locus ending just prior to the polymorphic site. Accordingly, SBE is recruited to interrogate the genotype and a signal-amplification step is implemented to increase sensitivity. A two-color strategy is employed with ddCTP and ddGTP sharing one dye, and ddATP and ddUTP another. In this manner four of the six SNP classes can be discriminated. This four represent over 83.5% of database SNPs. Nonetheless, the remaining categories – C/G and A/T – can be interrogated with SBE acting on ASE-probes. ASE was, in fact, the principle employed in Infinium I in a two-probe one-color assay, before switching to SBE in Infinium II. The rationale behind this was a 50% reduction in probe numbers / bead types and the increased robustness of SBE over ASE as it is an endpoint rather than kinetic measurement. It has been estimated that over 95% of the non-repetitive portions of the genome are assayable. At present, the most high-density Infinium assays genotypes 1.2 million SNPs in two samples on each Human1M-Duo v3 array.

3.6.7. Conversion rates
It is interesting to look at the conversion rates between the platforms. Conversion rate is defined as the percentage of SNPs that can be assayed and depends on both the quality of the selected SNPs and the platform. Both the GoldenGate and MIP assays reach high rates of over 80%, which by measures such as careful SNP selection and probe resynthesis can be pushed into the >90% territory. Hybridization-based methods have significantly lower rates.

3.7. A possibility to cover all SNPs?
The approximately one million SNPs that can be analyzed in parallel in a single experiment with the whole-genome genotyping platforms is an impressive number. Still, this only represents a minority of the more than 10 million SNPs scattered all over the human genome. Inclusion in the genotyping approaches of highly informative tagging SNPs increases the amount of information that can be obtained. However, there is only one answer if a ‘real’ genome-wide SNP survey is contemplated: sequencing of the whole genome. An added bonus is the interrogation of the allelic status of other marker classes, such as STRs and structural variants. About 20 years ago, sequencing the massive number of bases constituting the human genome was a daunting and extremely costly task. Rapid advancement in sequencing technology in recent years has seen the appearance on the market of novel instruments, making possible large-scale sequencing efforts considerably faster and cheaper. However, the costs associated with such efforts are currently still substantially higher than for the WGG assays. The sequencing field is, nevertheless, moving forward at breakneck speed and the array-era may come to an end when the prices for massive sequencing start to approach those of genotyping chips. The novel sequencing platforms are the subject of the next chapter.
3.8. Summary table

Table 1: The predominant SNP genotyping methods and their characteristics.

<table>
<thead>
<tr>
<th>Principle</th>
<th>Specificity</th>
<th>Platform flexibility</th>
<th>SNP throughput</th>
<th>Sample throughput</th>
<th>Cost per SNP</th>
<th>Cost per sample</th>
</tr>
</thead>
<tbody>
<tr>
<td>Allele-specific oligonucleotide hybridization (ASH)</td>
<td>ASH</td>
<td>Low</td>
<td>Dependent on array content</td>
<td>Low</td>
<td>Dependent on array content</td>
<td>High</td>
</tr>
<tr>
<td>Dynamic allele-specific hybridization (DASH)</td>
<td>ASH</td>
<td>Medium</td>
<td>High</td>
<td>Low</td>
<td>High</td>
<td>High</td>
</tr>
<tr>
<td>Solution-phase hybridization: TaqMan &amp; Molecular Beacons</td>
<td>ASH</td>
<td>High</td>
<td>High</td>
<td>Low</td>
<td>High</td>
<td>Low</td>
</tr>
<tr>
<td>Allele-specific PCR</td>
<td>ASE</td>
<td>Low</td>
<td>High</td>
<td>Low</td>
<td>High</td>
<td>Low</td>
</tr>
<tr>
<td>Invader</td>
<td>Invader / FEN cleavage of overhanging structures</td>
<td>High</td>
<td>Low</td>
<td>High</td>
<td>High</td>
<td>Low</td>
</tr>
<tr>
<td>Oligonucleotide ligation assay (OLA)</td>
<td>Allele-specific ligation</td>
<td>High</td>
<td>High</td>
<td>Low</td>
<td>High</td>
<td>Low</td>
</tr>
<tr>
<td>Pyrosequencing</td>
<td>Sequencing</td>
<td>High</td>
<td>High</td>
<td>Low</td>
<td>High</td>
<td>Low</td>
</tr>
<tr>
<td>Golden Gate</td>
<td>ASE and ligation</td>
<td>Low</td>
<td>High</td>
<td>High</td>
<td>Medium</td>
<td>Medium</td>
</tr>
<tr>
<td>Molecular Inversion Probes</td>
<td>SBE and ligation</td>
<td>High</td>
<td>High</td>
<td>Low</td>
<td>Medium</td>
<td>Medium</td>
</tr>
<tr>
<td>Whole-genome sampling analysis (WGSA)</td>
<td>ASH following complexity reduction</td>
<td>High*</td>
<td>Low</td>
<td>Genome-wide</td>
<td>Low</td>
<td>High</td>
</tr>
<tr>
<td>Infinium</td>
<td>SBE following WGA</td>
<td>High*</td>
<td>Low</td>
<td>Genome-wide</td>
<td>Low</td>
<td>High</td>
</tr>
</tbody>
</table>

* The specificity is high for the fixed set of SNPs on these platforms.

The ranking is performed according to a relative low-medium-high scale. In addition, since the assays can be performed with different settings and conditions, the ranking is only general and may not be applicable to all situations.
4. Sequencing

Sequencing – the elucidation of the order in which the four nucleotide building blocks appear in nucleic acid molecules – has played a significant role in biomedical research over the last three decades shedding light on a myriad biological processes and events. In 1972, the bacteriophage MS2 coat gene became the first gene to be fully sequenced. This was accomplished by arduous analysis of fragments generated by nuclease digestion of MS2 RNA. Four years later, with the sequencing of the replicase gene, the whole 3,569 bp genome of this bacteriophage had been unraveled earning this bacterial virus fame as the first entity with a fully sequenced genome.

However, the first revolution within this field came about in the end of the 1970s with the advent of two seminal sequencing techniques: the enzymatic chain termination approach of Sanger and the chemical degradation method of Maxam and Gilbert. The former technique became the favored one and advances in the following decade rendered it ready to take on the enormous task to sequence the three billion bases of the human genome in the Human Genome Project. The sequencing part of this effort required between 3 and 4 years at a cost of 300 million dollars. Of course, the years have seen the emergence of alternative approaches, but these have not been able to dethrone the conventional Sanger sequencing.

It was not until 2005 that the next paradigm shift was initiated. The launch of the 454 Life Sciences sequencing instrument, the first of the second generation massively parallel sequencers, marked the beginning of this new phase. Since then, several other systems have become available and put to use in a multitude of applications.

4.1. The first revolution

The two sequencing approaches presented by Sanger and Maxam and Gilbert in the end of the 1970s transformed this field, eventually firmly rooting sequencing as an indispensable tool.
4.1.1. The chemical degradation method of Maxam and Gilbert
The technique demonstrated by Maxam and Gilbert relies on chemical degradation of DNA\textsuperscript{225}. Specifically, the DNA region to be sequenced is radioactively end-labeled and divided into four separate reactions. Each of these entails a different set of chemicals that is capable of breaking the DNA only at specific bases or base types. As this degradation is partial, the outcome of each reaction is a collection of differently sized fragments being defined by the presence of the breakage-inducing base. The common denominator is hence that each fragment ends at a position coinciding with the base being degraded. The lengths of the degradation products of the four reactions are subsequently resolved in individual lanes on a slab polyacrylamide gel, the pattern of bands revealing the DNA sequence.

4.1.2. The enzymatic chain termination method of Sanger
The method developed by Sanger and colleagues bears resemblance to the Maxam-Gilbert technique\textsuperscript{223, 224}. However, the assortment of progressively longer fragments is attained by an enzymatic synthesis process rather than by chemical degradation. The DNA acts as a template in a primer-dependent extension reaction with chain-terminating nucleotides. In particular, four separate reactions are performed. Each encompasses the four normal deoxynucleotides, one of which is labeled, supplemented with one nucleotide type in its dideoxy form that precludes further chain elongation. The incorporation of this dideoxynucleotide occurs in a random fashion, thus generating fragments ending at every position that particular base is encountered. As in the Maxam-Gilbert case, slab polyacrylamide gels are used for readout, the banding pattern conveying the base order.

Owing to a better efficiency and the involvement of fewer hazardous chemicals, the Sanger method became the preferred choice over the Maxam-Gilbert approach. Accordingly, most efforts have revolved around improving the chain termination technique. A disadvantage with the early renditions of Sanger sequencing was the usage of potentially dangerous radioactive labeling for detection purposes. This was resolved in the middle of the 1980s with the introduction of fluorescent labeling strategies. Concurrently, a lot of interest focused on automating the protocol to enable a more expedient and rapid sequencing. Some approaches featured a single dye in conjunction with four separate reactions\textsuperscript{227, 228}. This was highly reminiscent of the original approach, but the radioactivity was exchanged for the much safer fluorescence. In other versions, the primers in each of the reactions were labeled with spectrally distinct dyes\textsuperscript{229, 230}. Here, the reaction outcomes could be combined necessitating only one gel lane. Finally, introduction of fluorescently labeled dideoxynucleotides enabled the fragment generation to be performed in a single tube as well\textsuperscript{231}. This dye-terminator chemistry eventually outcompeted the protocols necessitating multiple reactions.

Adaptation of capillary gel electrophoresis to length discrimination of Sanger fragments in the end of the 1980s represented another breakthrough as it eliminated the need for laborious slab gels\textsuperscript{232-234}. Additionally, the speed and efficiency of separation was increased. With arrays of capillaries, multiple sequencing reactions could be interrogated in parallel, one reaction being separated per capillary\textsuperscript{235}. Eventually, 96 and 384 parallel capillary instruments together with the fluorescent dideoxynucleotides formed the mainstay of sequencing and were crucial players in the elucidation of the human genome. Today, multicapillary machines are capable of read lengths up to approximately 1000 bases\textsuperscript{236}. At present, the appearance of second generation sequencing platforms has shifted focus away from the classic Sanger sequencing. However, advancements are still made. For example, miniaturized microfluidic devices integrating all steps necessary for sequencing have been developed\textsuperscript{237}. 

4.2. Alternative approaches

4.2.1. Pyrosequencing

Pyrosequencing is a sequencing-by-synthesis (SBS) approach meaning that base interrogation occurs by repeated cycles of nucleotide incorporation and detection. The detection principle in Pyrosequencing relies on an enzymatic cascade featuring ATP sulphurylase and luciferase that converts inorganic phosphate ($\text{PP}_i$), via intermediary ATP, to a light burst. This method was initially used to follow $\text{PP}_i$ levels. However, as nucleotide incorporation releases $\text{PP}_i$, it soon became evident that the same strategy could be employed to monitor this event. Specifically, an oligonucleotide-primed template is subjected to one type of nucleotide at a time. If the corresponding base is present in the template, the incorporation that occurs is detected by real-time light emission generated by the concerted action of the involved enzymes. After washing, another base is added and a new cycle initiated. However, with this strategy, the template has to be immobilized and the enzymes replenished for each sequencing round. To circumvent these limitations, apyrase was included in the mixture. This enzyme is able to degrade dNTPs and conditions can be set up to allow the polymerase to incorporate the correct nucleotide and sulphurylase to generate ATP faster than the apyrase-mediated degradation. In early adaptations, a reaction cocktail was flowed over a column with immobilized template and enzymes encountering the different components in the correct order. Later on, a real-time system was developed and refined. Here, dATP was replaced by deoxyadenosine α-thiotriphosphate (dATPαS) to minimize spurious signals. The latter analog functions well in the polymerase addition event but is a poor substrate for the detection machinery.

The vision was that this elegant approach, not requiring labeling or fragment separation, would replace conventional Sanger sequencing, but as the maximum read lengths were considerably shorter this switch never materialized. Only approximately 20 – 50 bases could be read with this approach owing to dephasing problems pertaining to the fact that the incorporation and degradation is not 100% complete. This implies that not all template molecules will be elongated when they should (denoted as ‘- frame shift’) and that some fragments will elongate faster than they are supposed to (+ frame shift) thus leading to progressively increased background signals, eventually precluding correct base identification.

Interestingly, the light signal generated is proportional to the number of nucleotides added by the polymerase allowing homopolymeric stretches to be identified. However, this is only possible to a certain extent as the correlation with emitted light becomes less precise with an increasing number of inserted bases. Homopolymeric regions longer than 4-5 bases are difficult to resolve.

4.2.2. Sequencing-by-hybridization (SBH) and tiling arrays

Hybridization is, given proper conditions, a specific process dictated by the precise base pairing between A and T, and C and G. Accordingly, hybridization between a known probe and a sample nucleic acid reveals the sequence of the latter. This forms the basis of simple, non-enzymatic sequencing-by-hybridization (SBH) approaches, which were proposed in the end of the 1980s.

Two main assay formats have been developed. In the first one, the interrogatory oligonucleotides of known sequence are attached to a solid support, most frequently a high-density array, and a labeled unknown target is allowed to hybridize. Following washing and scanning, the obtained signal intensities expose short bits of sequence of the
Parallel target selection by trinucleotide threading

sample. The bound probes can be either combinatorial libraries of n-mers allowing sequencing of practically any fragment\textsuperscript{169, 243}, or tailor-made to a specific already identified sequence in re-sequencing applications\textsuperscript{245-248}. In the latter case, the preferred assay design entails a tiled, overlapping pattern, where the probes are shifted in one-base intervals along a reference sequence corresponding to the region of interest. Errors in the reference, as well as unexpected sample sequence features such as insertions or rearrangements, can produce erroneous results\textsuperscript{246, 248}. Each base is probed with at least four oligonucleotides, each carrying a distinct base in the central position, which corresponds to the interrogated base. Only perfectly matched probes will participate in hybridization, revealing the base identity. With this probe setup only substitutions can be resolved. To investigate deletions or insertions, additional oligonucleotides have to be designed\textsuperscript{245-248}. A benefit is that this approach is amenable to parallelization, the limiting factor being the number of features that can be fitted on the array. However, as different probe sets have diverse hybridization properties the system is prone to cross-hybridization and unspecific signals\textsuperscript{247, 248}. This phenomenon, described in the setting of allele-specific hybridization in Chapter 3, represents the major limitation of this approach as it lowers the accuracy. Secondary structures present in the target can also compromise the hybridization\textsuperscript{247, 249}. Additionally, even with the highest-density arrays, large genomes cannot currently be probed using a single chip, making the protocol less convenient\textsuperscript{246, 248}.

To eliminate the detrimental spurious hybridization interactions, the assay format can be inverted\textsuperscript{244, 249-251}. By immobilizing the template and probing it with sequential hybridizations with individual probes from carefully chosen sets, optimal conditions can be used for each probe. As with the above-mentioned approach, this inverted strategy is also scalable, as many targets can be bound to a solid support. Moreover, as a generic probe set is used, a reference sequence is not strictly required. In practice, however, only resequencing is feasible, as it is highly problematic to produce an unambiguous sequence when the assembly is based on the very short sequence stretches obtained with this assay. Furthermore, the need to perform a large number of serial hybridizations renders this strategy relatively cumbersome. For example, 8,192 thoroughly selected 7-mers were used to sequence a 1.1 kb region of the p53 gene\textsuperscript{251}. This strategy affords redundancy as multiple probes query each base, albeit at different probe positions. The redundancy level is determined by the oligonucleotide length. For the 7-mers, each target base is involved in seven different hybridizations.

4.3. Second generation platforms

2005 marked the year when the sequencing field was radically transformed for the second time. In that year, 454 Life Sciences and Roche unveiled their massively parallel sequencing platform, capable of sequencing at a previously unimaginable rate\textsuperscript{122}. The inherent difficulty associated with parallelization and scale-up of the Sanger approach, precluding a truly high throughput output and thus a reduction in costs, provided the incentive that eventually materialized in the form of novel second generation approaches. Instead of the rather limited number of parallel reactions afforded by the chain termination-capillary gel electrophoresis pipeline, millions of reads can be generated in a parallel fashion. Of course, this implies an unprecedented sequencing rate. This ‘next-generation’ sequencing was celebrated as ‘Method of the Year 2007’ in the journal Nature Methods\textsuperscript{252}. Indeed, these sophisticated systems are partially responsible for a paradigm shift within biomedical research enabling large-scale efforts to be converted from mere dreams to reality. A multitude of areas has benefited from the second generation sequencers. Additionally, novel applications have sprung out\textsuperscript{253, 254}. The various applications do not only relate to sequencing in the traditional sense, where the obtained sequence represents the desired end information. Instead, the reads increasingly
serve intermediary roles acting as stand-ins for other types of data. For example, gene expression analysis using the massively parallel sequencers relies on counts of reads corresponding to a gene to infer its expression level\(^6\). The rapid increase in the amount of gigabases generated places a pressure on development of novel bioinformatics tools to manage the storage, handling and analysis of this vast body of data. Accordingly, the enhancements in chemistries and instruments have to coincide with faster software suites\(^255-258\).

It should, however, be emphasized that the emergence of the new instruments will not totally eliminate the conventional Sanger sequencing, which due to its cost-efficiency for small sequencing activities still finds widespread use in areas such as validation of PCR products and cloned inserts. For the foreseeable future, it is highly probable that the two will coexist\(^258\).

4.3.1. The main players
Currently, four commercial systems have hit the market\(^226, 255\). About a year after the 454 Genome Sequencer, the Illumina Genome Analyzer became available. This system was originally developed by Solexa, which was acquired by Illumina. In the end of 2007 Applied Biosystems launched the SOLiD (Sequencing by Oligonucleotide Ligation and Detection) platform. In 2008 Helicos presented the HeliScope sequencer incorporating true Single Molecule Sequencing (tSMS). Furthermore, several non-commercial scalable sequencing platforms have seen the light of day. Amongst these the polony sequencing, published in 2005\(^259\), and the shotgun sequencing-by-hybridization (Shotgun-SBH), published in 2008\(^249\), are the most notable ones. The next years promise further advances within this field. For instance, the instrument currently under development by Pacific Biosciences is scheduled for launch in 2010\(^260\). In addition, several groups aim to realize the goal of nanopore-based sequencing\(^261\).

4.3.2. Protocol characteristics and interrogation principles
One of the distinctive features of the new platforms is the lack of cloning of the fragments to be sequenced. Instead, various adaptations of compartmentalized amplification (Chapter 2) are used to create multiple copies of the same single molecule. The huge numbers of such spatially separated clonal ensembles are subsequently sequenced by chemistries amenable to parallelization. The system offered by Helicos sequences single molecules, obviating the need for amplification altogether, eliminating the potential bias that may be introduced in the PCR. The dominating interrogation principles are sequencing-by-synthesis (SBS) used in the 454, Illumina and Helicos systems, and sequencing-by-ligation (SBL) employed by SOLiD. In the former, a polymerase is used to incorporate one base or a base type per cycle. This naturally only occurs if the complementary nucleotide is encountered in the target. The presence or absence of this addition is subsequently determined by various strategies, before a new cycle with another base is initiated. In SBL, the high discriminatory power of ligation is harnessed to identify target bases. Also in this case, the reaction is cycled, the subsequent ligations increasing the read lengths. Massive parallel signature sequencing (MPSS), described in Chapter 5, and the open-source polony sequencing were the first approaches exploiting the SBL principle. Furthermore, the third common interrogation method — sequencing-by-hybridization (SBH) — has also been employed for large-scale sequencing in the shotgun-SBH technique.

The massively parallel sequencers can operate on either fragment or paired-end libraries. With the former, the traditional scenario, a DNA molecule is sequenced from one end up to the point of disintegrating base quality. In paired-end sequencing, both terminal
regions of inserts of a defined length are read. The somewhat more cumbersome sample preparation steps are compensated for by a less complicated mapping and assembly. The various protocols generate mate-pairs originating from differently sized fragments and several studies use a combination of both conventional single-end sequencing and paired reads from differently sized inserts. For example, a recent Illumina-based human genome resequencing effort entailed both single-end, short insert paired-end (insert size of 130 or 390 bp) and long insert paired-end reads (insert size of 2700 bp).262

As the sequencing field moves rapidly forward, generating ever increasing read lengths and parallelization while cutting the costs, it is important to stress that the specifications given below for each method may quickly become outdated.

### 4.3.3. 454

The regular Pyrosequencing approach is cumbersome to scale up. However, spatial separation of immobilized single molecule-generated clusters accompanied by immobilized enzymes and flowing of the nucleotides has made possible a parallelization.122, 263 Namely, the 454 pipeline starts with random shearing of the target DNA with nebulization. Two adapters, one of which is biotinylated, are ligated onto the generated fragments. The biotin is used in a capture step to create single-stranded molecules. These are subsequently bound to beads, under carefully chosen conditions to maximize the number of spheres carrying a single template. An emulsification is performed, enclosing each bead in an aqueous microreactor surrounded by oil. In this droplet, a PCR reaction is performed eventually covering the bead surface with identical copies of the template fragment. The emulsion is thereafter broken, fragment-carrying beads enriched and deposited onto a picotiterplate. This plate contains a large number of picotiter-volume wells able to accommodate no more than one bead. The base interrogation is performed by Pyrosequencing and smaller beads carrying the enzymes required for the light generation are added. The surface is then flowed with one base at a time and the wells in which this base has been incorporated produce a light flash which is detected in real-time by a charged coupled device (CCD). The flowing of the bases obviates the need for an apyrase as the unused nucleotides rapidly drift away.263

The first generation instrument, the Genome Sequencer 20, was, when applied to the *Mycoplasma genitalium* genome, shown to produce approximately 25 Mb of data in a four-hour run.122 Over 200,000 high-quality reads with average read length of about 100 bases could be obtained. The following Genome Sequencer FLX was employed to sequence the genome of James D. Watson to 7.4x coverage at a cost of $1 million – a hundredth fold reduction compared to conventional Sanger sequencing.264 Additionally, this was completed in two months. Here, the reads were around 250 bases. The latest Genome Sequencer FLX Titanium sequences 400 – 600 Mb in one 10-hour run.265 More than 1 million reads of over 400 bases are generated. This performance increase has been achieved by using smaller beads together with higher density plates encompassing over 3 million wells, and an improved sequencing chemistry. In the original setup the beads were 28 µm in diameter and the picotiter plate featured 1.6 million wells.122 Finally, it has been speculated that the read lengths in next 454 version will approach 1000 bases, matching those of Sanger sequencing.266

The dominant type of sequencing error with the 454 strategy occurs as miscalcs of the lengths of longer homopolymeric stretches.122, 264 For such regions the amount of light emitted in the Pyrosequencing approach is difficult to precisely translate into the number of bases. This phenomenon has been described in the Pyrosequencing section.
4.3.4. **Illumina / Solexa**

The sequencing technology developed by Solexa and subsequently acquired and marketed by Illumina also relies on sequencing-by-synthesis but instead of flowing a single type of unmodified base per cycle, four distinctly labeled reversible terminators are employed extending all templates with one base in each round\(^{127, 128, 267, 268}\). The DNA to be interrogated is randomly sheared and two adaptors are appended. The fragments are attached to a planar surface and solid-phase bridge amplification is performed multiplying the single fragment to about 1000 identical copies. As the immobilization process is stochastic, a high-density of such clusters is attained. The four reversible terminators are introduced, alongside a polymerase and sequencing primer, and dictated by the template, a single base is incorporated into the growing chain. Naturally, all clusters are assayed simultaneously. Following image acquisition revealing the base, a chemical cleavage removes the blocking group and the dye, reinstating an extendable template permitting a new round of sequencing to commence.

As the individual fragment ensembles can be tightly packed a massive number of reads is obtained. The latest available version of the instrument – the Genome Analyzer IIx System – generates about 150 million 2 x 75 base long paired end reads totaling approximately 20–25 Gb of high-quality sequence in a 10-day run\(^{269}\). An older variant of the sequencer has been employed for human resequencing of four different genomes: a Nigerian HapMap sample\(^{267}\), an acute myeloid leukemia patient\(^{270}\), a Han Chinese\(^{271}\) and a Korean individual\(^{262}\). The three studies each produced about 3 billion reads, with an average read length of over 30 bases, amounting over 100 Gb of data. The coverage was roughly 30x. Interestingly, the Korean study used a new cleavage mix for removing the fluorophore and chain termination allowing reads of up to 106 bases\(^{262}\). The reagent cost of the HapMap sample sequencing performed by Illumina was estimated at $250,000 with the sequencing part taking eight weeks\(^{267}\).

The fact that all four nucleotides participate and ‘compete’ in each cycle of sequencing reduces the risk of spurious incorporation of the incorrect base compared to single-base type approaches\(^{127, 128, 267}\). However, as the polymerase elongation and/or the chemical cleavage are not 100% efficient some of the molecules making up the sequencing ensemble may not be extended or restored. Consequently, these fall out of phase and the gradual accumulation of individual templates in different phases ultimately leads to an inability to discriminate the base. These events limit the read length\(^{258}\). In contrast to 454, homopolymeric regions do not present any problems in the Illumina system\(^{127, 128}\). Instead, the dominating error type is that of substitutions\(^{258}\).

4.3.5. **Helicos**

True single molecule sequencing (tSMS), the strategy devised by Helicos, is quite similar to the Illumina one with respect to the reversible terminator sequencing-by-synthesis\(^{272-274}\). The distinctive feature of Helicos is sequencing of single molecules, thus evading the amplification step. As the various PCR approaches can introduce errors and bias, this system offers potentially more unaffected, ‘cleaner’ reads. This single molecule discrimination necessitates a very sensitive and high-resolution imaging system. As is the case for the other platforms, the protocol begins with fragmentation of the target DNA. An adapter, the future sequencing primer site, is added to the 5’-end and a terminal transferase polyadenylates the 3’-end. The surface of the flow cell encompasses attached oligo dT molecules, which hybridize to the poly-A tails of the fragments. Extension with an unmodified nucleotide set creates an immobilized copy of the original target having a primer annealing site at the 3’-end. The sequencing chemistry involves sequential additions of labeled ‘virtual terminator’ nucleotides. After rinsing, scanning and signal
detection, the label is cleaved and a new base type introduced. The sequencing proceeds with ‘quads’ – consecutive CTAG cycles. In contrast to the Illumina ‘four differentially labeled nucleotides-one cycle’ method, this approach implements a single dye in conjunction with four cycles. As the template is immobilized in the flow cell, two-pass sequencing can be performed by reading the sequence, melting and washing away the synthesized strand, and interrogating the same bound molecule once again. This results in increased accuracy.

The foundation for the Helicos system was laid in 2003 when it was demonstrated that single base incorporation could be detected\textsuperscript{275}. The proof-of-concept study entailed sequencing of the bacteriophage M13 genome. 280,000 individual molecules were read in parallel with an average length of 23 bases\textsuperscript{273}. Recently, the genome of Stephen Quake was investigated with the HeliScope instrument\textsuperscript{276}. Enhancements in parallelization and nucleotide chemistry produced 90\% of this genome at a coverage of 28x in four instrument runs costing \$48,000 in consumables\textsuperscript{277}. Billions of, on average, 32 bp reads were generated, the individual reads varying between 24 and 70 bases. According to Helicos information from December 2008, an 8-day routine run yields 600-800 million reads averaging 30-35 bases for a total of 21-28 Mb of output\textsuperscript{278}.

As this system is free from amplification, the sample preparation is simplified and no PCR predispositions interfere with sequencing integrity. Furthermore, the phase shift problems when interrogating ensembles of molecules are avoided. Here, if an incorporation event is missed in a cycle, the next round with the same base can still reveal the base identity. Instead, deletion errors have been encountered, owing predominantly to nucleotides with missing or inactive dyes\textsuperscript{272, 273}. In principle, homopolymeric regions should be easily determined. However, the terminating properties of the first version of nucleotides were not absolute, allowing a small number of bases to be added when homopolymeric stretches were encountered. This led to dye interactions and distorted signals precluding correct length discrimination\textsuperscript{273}. A novel set of nucleotides has addressed this issue\textsuperscript{272}.

\textbf{4.3.6. Sequencing by Oligonucleotide Ligation and Detection (SOLiD)}

Sequencing by Oligonucleotide Ligation and Detection (SOLiD) marketed by Applied Biosystems is different from the three other approaches in using ligase-based sequencing, instead of one depending on a polymerase\textsuperscript{123, 259, 279}. The sample preparation steps – fragmentation, adaptor ligation, emulsion PCR and bead enrichment – are similar to the other systems. However, the targets covering the enriched beads are 3’-modified to allow covalent attachment to the sequencing slide in a random fashion. Subsequently, the base identities are obtained by a stepwise ligation scheme. In particular, a sequencing primer is annealed and a 16-probe set of 8-mers, where the two first bases are known, the three following ones degenerate and the three final ones universal, is added. A four-color strategy is employed implying that four probes with different combinations of the two known bases share the same dye. This dye is attached downstream of the universal bases. The high discriminatory power of the ligase will only join the perfectly matching probe. Washing and signal acquisition is followed by capping of unextended sequencing primers to prevent dephasing problems, and cleavage at the interface between the degenerate and universal bases removing the label and preparing the sample for yet another round of ligation. Five such ligation-detection-cleavage cycles are performed, interrogating every fifth base, before the created fragment is melted away and a new sequencing primer shifted one base upstream is annealed. Several ligation rounds with progressively shifted primers are used to increase the read length. Two-base encoding, the fact that each base is interrogated twice, with two probes starting from offset sequencing primers, allows for efficient error correction\textsuperscript{280}. Substitutions represent the predominating error variety\textsuperscript{258}. 42
Applied Biosystems states that the SOLiD 3 system is able to sequence 20-30 Gb on 2 slides during a two-week run. This corresponds to 400-600 millions of 2x50 paired end reads. Lately, an African individual’s genome was obtained with the SOLiD technique. An 18x coverage was attained using up to 2x50 bp reads.

Polony sequencing is the predecessor to SOLiD and has been converted into the open-source Polonator instrument by Dover Systems. In the published approach, a paired-end originating from approximately 1000 bp inserts was created, emulsion PCR amplified and bead-hybridization enriched. The beads were deposited onto an acrylamide monolayer and ligation sequenced. Here, instead of offsetting the sequencing primers, the known base of the nonameric probes is exchanged between the cycles. A four-color readout is employed each known base corresponding to a single label. As the T4 DNA ligase used can detect six bases from the joining interface in the 5’→3’ direction, and seven in the other one, a total of 13 bases is sequenced from each end of the insert. Accordingly, 2x13 bases are obtained, in the original template molecule separated by about 1000 bp. Several aspects of polony sequencing were enhanced in a recent transcript level analysis study.

4.3.7. What about sequencing-by-hybridization?
Sequencing-by-hybridization (SBH) has been described in the setting of alternative approaches to Sanger. This principle is not incorporated into any commercial second generation platform. However, a technique dubbed shotgun-SBH has been applied to parallel resequencing of one viral and one bacterial genome. Scalability is achieved by immobilizing target fragments onto a slide. This occurs via circularization of moieties composed of 200 bases of target and a 50-meric linker sequence, annealing of the linker region to complementary oligonucleotides on the support and in situ rolling circle amplification. The final outcome is a collection of coils, each encompassing tandem copies of a single initial fragment. Parallel interrogation is done with successive probings with 582 pentamer oligonucleotides. The pattern of obtained signals for each coil is converted, analyzed and aligned to a reference using computer algorithms. To improve probe performance, locked nucleic acids (LNAs) were used and degenerated bases appended to the ends of the pentamers.

The fragment length in the initial circularization step represents the read length. In this study, 200 bp was chosen as a compromise between several different counteracting factors, the generation of adequate signals being the guiding principle. With this universal strategy the 48kb bacteriophage λ and 4.6Mbp E. coli K12 genomes were sequenced to, 96% and 83%, respectively. This method is simple and inexpensive and this proof-of-concept study – the maximum output generated being 1.6 Gbp of raw data in 5 days – leaves much room for improvements. A problem is, however, that secondary structures can prevent efficient hybridization. Moreover, it is unclear if this approach can tackle more complex genomes.

4.3.8. Guidelines regarding second generation platform choice
As outlined in the above sections, the performance of the various second generation platforms in the context of sequencing accuracy for different motifs varies. Naturally, these platforms have distinct advantages and limitations. However, in broad terms, the choice between them boils down to read length versus the number of reads. Approaches necessitating long reads, such as de novo sequencing, are well-matched to the 454 system, making full use of its >400 bp reads. Conversely, the number of reads is the critical factor in counting or identification studies, for instance transcript abundance analysis and ChIP-Seq (Chromatin Immunoprecipitation Sequencing). Accordingly, the higher output
methods of Illumina or SOLiD are the preferred choice. Although these reads are significantly shorter in comparison with the 454 system, they are of sufficient length for unambiguous identification of a transcript or transcription factor binding site.

Alongside the hardware and chemistry developments, the bioinformatics tools used for data analysis and sequence assembly are progressively improved. This, together with novel protocols enabling paired-end reads of defined size fragments and ever increasing read lengths, has allowed also the shorter read technologies to undertake complicated tasks, such as human genome resequencing. The shorter sequences imply that a greater coverage has to be achieved for proper alignment and assembly as compared to the longer 454-reads. However, this is offset by the much greater number of reads generated by the Illumina and SOLiD platforms.

Of course, costs associated with the techniques also represent an important parameter to consider. The various studies making use of the second generation platforms often present cost estimates. However, as different types of expenses such as consumables, personnel or instrument costs are included in these calculations, a comparison is not straightforward. Generally, however, the 454 system is more expensive per Mb than the Illumina and SOLiD offerings. The Helicos instrument is the most expensive, but per Mb costs are probably comparable to the other short-read strategies.

4.3.9. Sanger versus second generation systems
Comparing Sanger sequencing, the workhorse approach of the last decades, with the novel sequencers reveals drastic differences. Some of these have already been touched upon. The massive parallelization offers a much higher throughput, which coupled to the fact that traditional cloning is no longer required renders the new platforms cost-effective. As an example, the 454 sequencing of a human genome was accomplished at a hundredth of the cost of the same endeavor using the capillary Sanger technique. Additionally, the novel instruments have fewer infrastructure requirements. Conversely, the data filtering and analysis is rather more complicated, posing serious bioinformatics challenges.

On the negative side, the accuracy associated with the novel chemistries is lower than in the Sanger case. It has been estimated that Sanger base-calls are more than ten times more accurate. Furthermore, the read lengths are shorter, making sequence assembly more complicated and necessitating higher coverages. However, constant fine-tuning, enhancing both the lengths and the read numbers, paired-end strategies and new algorithms ameliorate this situation.

Also, although the per Mb cost of the new systems is low, they are run in an ‘all-or-none’ fashion, thus being prohibitively expensive for small scale studies or routine applications featuring precise smaller genomic regions.

4.3.10. Combining several approaches
As is clearly evident Sanger and the various second generation platforms have non-overlapping benefits and drawbacks. As such, combining techniques complementing each other can create more reliable sequences and better assembly characteristics. Here, the method’s advantages can be fully exploited, while its disadvantages are resolved using another platform.
4.4. Third generation approaches
The appearance of second generation sequencing systems does not represent the final destination. Research groups and companies all over the world aim at developing methods capable of sequencing longer, faster and more accurately. Additionally, the field is moving into single-molecule sequencing to avoid problems associated with ensemble-based approaches. In many instances, the exact underpinnings and details of the techniques are shrouded in a veil of mystery, the information being rather scarce. However, a broad overview of the principles can be attained.

In addition to the predominant trends of nanopores and monitoring the activity of single polymerase molecules, several other approaches are explored. For instance, scanning tunneling microscopy (STM) has recently been shown to identify guanine moieties in single molecules placed on a copper surface287.

4.4.1. Complete Genomics
Complete Genomics has taken a different tactic. Instead of marketing an instrument, the company provides a sequencing service at their core facility. In their approach, a circularized library encompassing target regions interspersed with adaptors is amplified in solution generating DNA nano-balls (DNBs) composed of tandem copies of the original circle288. This is reminiscent of the RCA of Shotgun-SBS described earlier. The amplification means that this is not a single-molecule approach. The DNBs are thereafter dispensed onto a slide with roughly one billion activated spots able to catch the balls. The company reports that in this process, over 90% of the spots carry a DNB. The base interrogation assay is dubbed combinatorial probe-anchor ligation (cPAL) and relies on ligation of perfectly matched probes. It resembles the scheme used in polony sequencing. However, 10 bases from each end are discriminated by the ligase, which together with multiple adaptors able to prime the reaction, result in 2x35 bp read lengths. One beneficial feature is that, unlike SOLiD, the primer-probe complex is melted and removed after each ligation cycle. In this manner the base determination becomes independent of potential incompleteness in the preceding cycles.

4.4.2. Pacific Biosciences and VisiGen
One technique that looms on the horizon is single-molecule, real-time (SMRT) DNA sequencing from Pacific Biosciences, a company founded in 2004289, 290. This platform is en route to a marketed instrument with a launch expected in 2010260. As with the Helicos method, sequencing is performed on single molecules, but instead of the repeated incorporation-imaging-cleavage cycles the activity of a single polymerase molecule is monitored in real-time. Zero-mode waveguides (ZMWs) are a critical ingredient in this approach291. These are miniscule holes, between 60 and 100 nm in diameter, created in a metal film deposited on a silica slide292. As the non-metal side of the silica slide is illuminated with light of a longer wavelength than the aperture, this light cannot penetrate the small cavities. However, some of the light traverses into the holes but is confined to a zeptoliter-scale \(10^{-21}\) liters) volume. By placing a single polymerase molecule within this space, its activity can be followed in real-time. A chemistry suitable for this immobilization process has been devised293. A template and four distinctly labeled nucleotides are added to the immobilized polymerase. In contrast to the normally used base-labeled nucleotides, here the dye is attached to the terminal phosphates294. During the elongation process the polymerase cleaves these phosphates, generating a native substrate for the next incorporation. In this manner, problems associated with steric hindrance are resolved. The diffusion of labeled nucleotides in the small detection volume is much faster than when a nucleotide is bound by a polymerase on its way of being incorporated into the growing chain. Accordingly, this latter nucleotide can be detected.
even in the presence of high concentrations of labeled nucleotides in the small hole\(^{291}\). By simply monitoring the polymerase and reading the signals, the incorporation events reveal the sequence. Parallelization is attained by creating several such ZMW structures on the same slide. The detection relies on an advanced optics system\(^{293}\).

Initial publications from Pacific Biosciences show promise, but at the same time illuminate several obstacles that have to be addressed. The employed \(\phi 29\) DNA polymerase in the ZMW wells can synthesize long stretches of DNA, up to 25 kb in length, in a time-efficient manner\(^{290, 293}\). This implies that long read lengths may be feasible, but these studies only entailed short and simple circular templates. Reads up to 3 kb have been reported by the company, the major limitation to further increase being the disruption of the polymerase molecules by the laser-based detection. Employing a strobe mode, turning the lasers on and off, DNA stretches are obtained with unresolved regions of defined lengths in-between, effectively increasing the amount of information\(^{296}\). This is reminiscent of the paired-end strategies of the other systems. On the other hand, only about a third of the ZMW have a single immobilized polymerase, the other wells being either empty or harboring several copies of the enzyme\(^{290, 293}\). This drastically hampers the parallelization. Additionally, the presented arrays contain only 3000 ZMWs\(^{290, 293}\). Furthermore, the strategy has problems with accuracy. When performing a test four-color sequencing, 131 of 158 bases were correctly interrogated\(^{290}\).

Another approach entailing real-time monitoring of polymerase action is pursued by VisiGen Biotechnologies\(^{297}\). This approach relies on fluorescence resonance energy transfer (FRET), an energy exchange phenomenon occurring between two overlapping fluorophores, a donor and an acceptor, when they are brought sufficiently close together. Accordingly, FRET modifies the signal that is emitted. In the VisiGen approach the slide-bound polymerase is tagged with a donor and each of the four bases carries a separate acceptor. Sequestering of the correct base by the enzyme just prior to incorporation brings these two entities in close proximity allowing for FRET, generating a specific signal. Also in this approach the conventional base-altered nucleotides have been exchanged for phosphate-labeled ones.

### 4.4.3. Nanopores

Nanopores offer a completely new avenue to single-molecule sequencing\(^{261}\). The basic principle is to thread a long DNA molecule through a nanometer-scale pore while detecting each of the four bases by virtue of changes in current or other electrical properties. Both naturally occurring pores, such as \(\alpha\)-hemolysin, or fabricated ones can be used. These techniques promise convenient, rapid and cost-effective sequencing able to generate long read lengths. However, several problems have to be addressed to bring this exciting new type of methods from the laboratory bench to a reliable instrument. The low resolution and the fact that the DNA molecules pass the pores too fast to enable dependable base identification are just two such issues.

There are several approaches to address the problems associated with the fast threading of DNA molecules through nanopores. For example, exonuclease treatment of the target DNA can be implemented. This cleavage generates a sequential flow of single nucleotides, the order dictated by the target, which can be identified while passing a protein nanopore\(^{298}\). Another manner to circumvent this limitation is to combine nanopores with hybridization. In the technique devised by LingVitae, each base of a 24 bp target is first converted into binary code represented by carefully chosen 10-meric oligonucleotides\(^{261, 299, 300}\). A combination of two such oligonucleotides is adequate to encode the four different bases. By assigning ‘0’ to one and ‘1’ to the other, the four bases can be
discriminated with the two-bit binary codes of 00 (A), 01 (C), 10 (G) and 11 (T). Consequently, each 24 bp target acts as template in generation of a 480 base long design polymer. Subsequent to the conversion, probes complementary to the 10-meric bits are allowed to anneal, and as they are separated from the template as it enters a nanopore they emit a signal that can be detected. Two consecutive signals reveal the identity of the target base.

4.5. Summary table

Table 2: The main sequencing platforms and their features.

| First generation | Sanger DNA sequencing | Enzymatic chain termination | Cloning / PCR | Long | Low | High | High | High | Indels in homopolymers | High | High |
| Pyrosequencing | SBS / Pyrosequencing | PCR | Short | Low | High |
| 454 | SBS / Pyrosequencing | Emulsion PCR | Medium | Medium | Medium | Indels in homopolymers | Low | Medium |
| Illumina / Solexa | SBS | Bridge amplification | Short | High | Medium | Substitutions | Low | Low |
| Helicos | SBS | No amplification | Short | High | Medium | Deletions | Low | Low |
| Polony sequencing / PoIonator | SBL | Emulsion PCR | Very short | High | Medium | Substitutions | Low | Low |
| Second generation | Applied Biosystems SOLiD | SBL | Emulsion PCR | Short | High | Medium | Substitutions | Low | Low |
| Third generation | Complete Genomics | SBL | Solution RCA | Short |
| Pacific Biosciences | Real-time monitoring of polymerisation | No amplification | Long |
| VisiGen | Real-time monitoring of polymerisation | No amplification | Long |
| Nanopores | Electrical property changes as DNA passes through nanopore | Long |

The ranking is performed according to a relative low-medium-high scale. Furthermore, since the sequencing can be performed with different settings and conditions, the ranking is only general and may not be applicable to all situations.
5. Expression profiling

The last two chapters have described a multitude of approaches to investigate different aspects and various features of genomic DNA. Methods for SNP scoring and the novel sequencing platforms broadening the scope of obtainable genomic information have been highly influential in shedding new light on the complexity of life. Certainly, the DNA codes for all this life, but there are two other classes of biological macromolecules with crucial functions. Proteins are the principal effectors of the functions encoded by the genomic DNA, and, as such, the protein makeup in a cell determines its role and characteristics. As such, analysis of the protein content of different cell types under diverse conditions can provide vital information. However, the 20 amino acid building blocks together with numerous post-translational modifications render proteins a multifaceted group of molecules. Accordingly, global studies of all proteins – the proteome – are problematical to perform in a parallel fashion, although several methods have been proposed\textsuperscript{301, 302}.

Messenger RNA (mRNA) represents the transitional molecule between the DNA code and the protein effectors. Instead of directly analyzing the abundance of various proteins, mRNA levels can be used as a surrogate with the assumption that the transcript levels correlate with those of the corresponding proteins. However, given the regulation occurring after the mRNA step, this correlation is not always straightforward. Nevertheless, the expression profile can give insight into a cell’s function, provide clues to the molecular underpinnings of diseases and processes, and aid in disease diagnosis or classification. Consequently, expression analysis represents an essential method group complementing techniques exploring DNA.

In a typical eukaryotic cell, the total RNA content is about 20 pg, divided into the various types of RNA species with mRNA constituting about 5%. This represents a couple of hundred thousand transcript molecules divided among the approximately 10,000 protein-coding genes expressed at a given moment. Commonly, a small number of mRNA molecules is highly abundant, whereas most messages are present only in low numbers\textsuperscript{303}.
As a whole, the human genome is estimated to harbor about 20,000 protein-coding genes\textsuperscript{46, 47}.

\section*{5.1. Methods of differing magnitude}

Generally, the various approaches to expression profiling can be divided based on the number of transcripts targeted. On one side, there are global techniques capable of analyzing all protein-coding mRNAs. The other side encompasses methods for studying a single, or at most, a small number of transcripts. These low-scale approaches, originating with Northern blots\textsuperscript{306, 307}, but nowadays primarily based on reverse transcription quantitative real-time PCR (RT-PCR)\textsuperscript{76}, are most frequently used for validation of significant results from larger studies. In addition, intermediary-scale approaches, residing somewhere in between the two major categories, have been devised. The cDNA-mediated annealing, selection, extension and ligation (DASL) assay\textsuperscript{308} and BeadsArray for the Detection of Gene Expression (BADGE)\textsuperscript{309} are two examples of such methods.

In addition to grouping the techniques based on the scale of the expression profiling studies, the underlying interrogation principle can be used as an alternative means of classification. In this regard, there are three main categories. Some methods, such as microarrays\textsuperscript{62, 63} and Northern blots\textsuperscript{306, 307}, rely on hybridization, while other utilize PCR, for example, differential display\textsuperscript{310} and quantitative real-time PCR\textsuperscript{76}. The third group encompasses sequencing-based approaches and includes, for instance, RNA-Seq\textsuperscript{64} and EST sequencing\textsuperscript{311}.

\section*{5.2. Historical overview of global expression analysis}

The beginning of the last decade of the 20\textsuperscript{th} century brought about the commencement of a gradual shift in the transcript analysis field. Previous methods such as Northern blots\textsuperscript{306, 307}, and subtractive hybridization – a complexity reduction method where mRNA species common to both the sample and the reference are depleted leaving behind only the ones unique to the former\textsuperscript{312} – were quite limited in scope. Now, instead of studying single transcripts, focus was centered on methods capable of a more comprehensive mRNA assessment.

\subsection*{5.2.1. Differential display}

One of the first approaches to such a systematic analysis was differential display (DD), introduced in early 1990s\textsuperscript{310, 313-315}. This method relies on PCR amplification with a primer pair consisting of oligo dT and an arbitrary oligonucleotide. A similar approach, called RNA arbitrarily primed PCR (RAP-PCR), entailed only random primers\textsuperscript{316}.

In differential display, the mRNA is first reverse transcribed starting from an oligo dT primer extending one base into the non-polyA tail region of the transcript. Accordingly, this primer aligns the synthesized cDNA at the polyA tail. To increase transcript coverage, three such primers, each with a different extending nucleotide, are usually used giving rise to three distinct cDNA populations. These are subsequently PCR-amplified using the oligo dT in conjunction with an approximately 10-15 bases long arbitrary primer. Naturally, for different mRNA species, this oligonucleotide anneals with varying distances from the polyA region generating products of different lengths. Again, a higher transcript coverage can be achieved by utilization of several arbitrary primers. Each oligo dT/arbitrary primer pair in this combinatorial scheme is selected to allow amplification of about 50 to 100 mRNA types. The products are separated and visualized with gel electrophoresis. Each sample produces a unique transcript fingerprint, both with respect to the number of bands and their intensities, clearly revealing any differences between, for example, two types of cells. Moreover, interesting bands can be recovered and analyzed...
Pawel Zajac

51

Further. One advantage of this approach is that no a priori knowledge about the transcriptional landscape is required.

5.2.2. EST sequencing
About the same time as the introduction of differential display, the concept of expressed sequence tags (ESTs) was coined\textsuperscript{311}. An EST was defined as a section of a complementary DNA (cDNA) clone matching a specific mRNA species. Therefore, conventional Sanger sequencing of such fragments was a suitable approach to identify novel genes and map them to the genome\textsuperscript{311, 317, 318}.

In the following years EST sequencing was adapted to larger-scale expression profile analysis\textsuperscript{319, 320}. Briefly, this entails cDNA library construction, corresponding to the transcript repertoire, followed by Sanger sequencing of a large number of randomly chosen entities from these libraries. This generates a count number for each mRNA, which is translated into its expression level. EST sequencing was the first rendition of a comprehensive sequencing-based technique. The reliance upon Sanger sequencing renders the method reliable, and, additionally, no prior knowledge about the transcripts is necessary. On the other hand, screening numerous library entries is rather costly, and given the varying abundance levels of different transcripts, a high sequencing depth is required to cover the most rare mRNAs.

5.2.3. Digital counting to analog relative quantification and back to digital counting
The idea that ESTs could be drastically shortened, while still accurately identifying a transcript led to the development of short tag sequencing methods. Of these, serial analysis of gene expression (SAGE) is the most seminal\textsuperscript{321}, and over the years several refinements of the original protocol have been implemented\textsuperscript{322, 323}. In SAGE and other approaches, the short fragments are concatenated, and thus sequenced as part of a larger molecule. In this manner, the capacity of Sanger sequencing is more fully exploited, with each read generating several data points instead of a single one as in the EST technique.

Even though SAGE enabled a significant cost reduction compared to EST technologies, the prices were still quite steep. In the mid 1990s the much more cost-effective method of DNA microarrays was introduced, eventually transforming and revolutionizing the gene expression field\textsuperscript{62, 63}. Hybridization of a labeled mRNA-derived sample to probes covalently attached to a solid support made possible a level of parallelization previously unimaginable, ultimately allowing the whole protein-coding transcriptome to be assayed simultaneously on high-density chips. The bound probes can be cDNAs/ESTs or oligonucleotides and, naturally, correspond to the transcripts of interest. The reduced costs present a substantial advantage, but these techniques are also associated with certain limitations. One in particular concerns the specificity, as cross-hybridization can generate spurious signals.

Despite the drawbacks of DNA microarrays, these methods have been used extensively during the last decade to study a multitude of diseases and processes\textsuperscript{324}. Currently, with the advent of second generation sequencing platforms, the gene expression field is experiencing another reform\textsuperscript{64}. The massive output generated by these sequencers can, in addition to establishing the transcript levels, reveal the structure of the mRNA molecules and any polymorphisms or mutations present in one and the same run. As such, the scientific community is offered a tremendously expanded wealth of information from a single experiment. With further advancements, progressively lowering the prices, the microarrays, the core platform of the last years, may become outdated.
An interesting pattern emerges when looking at the underlying principles of the different methods. The first attempts at global profiling relied on sequencing, the digital counts representing the expression levels. With the introduction of microarrays, the digital readout was exchanged for an analog one offering only a relative quantification. Nowadays, with massive sequencing, the field has come full circle reverting to the digital expression levels.

5.3. Microarrays

Microarrays as detection platforms have already been discussed in the setting of SNP genotyping (Chapter 3). The various array types, as well as their characteristics, benefits and disadvantages have been outlined. In brief, both solid-phase conventional planar systems and bead-based platforms are available. The former can be further subdivided into spotted arrays, where complete oligonucleotides or PCR products are deposited one by one by array printers, and in situ synthesized chips, involving simultaneous base-by-base production of the whole collection of probes.

The attachment of various probes to solid supports is not a new occurrence as it has been employed for diverse purposes since the end of the 1970s. However, in these initial approaches the scope was rather limited. During the first half of the 1990s, procedures and chemistries capable of array manufacturing at a much larger scale were introduced, enabling an increased number of assays to be performed in parallel. The field of expression profiling started to capitalize on these new inventions in the mid-1990s. Generally, two distinct approaches have been pursued. One direction has seen utilization of complete cDNAs or EST fragments as probes\textsuperscript{63, 325, 326}, whereas the other has relied on much shorter oligonucleotides produced on the chip surface\textsuperscript{62, 192}.

5.3.1. Spotted arrays featuring cDNAs, ESTs or pre-synthesized oligonucleotides

The cDNA/EST strategy was firstly demonstrated in 1995 with the expression of 45 Arabidopsis genes being investigated in parallel\textsuperscript{63}. The slides were prepared using a spotting robot transferring PCR-amplified and purified cDNAs from a 96-well plate to an array. Two distinctly labeled mRNA-derived cDNA specimens, a sample and a reference, were subsequently introduced to the array and hybridization allowed to take place. The fluorescent dyes were incorporated during a single round of reverse transcription. The slides were washed to remove unbound probes and then scanned. Following normalization to compensate for differing input mRNA amounts, the intensity ratio between the two dyes exposed the differently expressed genes.

One predicament with the cDNA/EST approach is the unspecific nature of the immobilization event, as multiple functional groups in the probes can attach to the surface\textsuperscript{325}. Accordingly, this process and the exact nature of the bound probes is only vaguely defined, possibly interfering with the hybridization. Deposition of pre-synthesized oligonucleotides harboring a single functional group offers a remedy to this problem. However, these are considerably shorter, possibly increasing problems associated with spurious cross-hybridization. The utilization of pre-produced oligonucleotides also obviates the need for labor-intensive cDNA/EST preparation, amplification and purification. Furthermore, this approach also relies on reliable traditional oligonucleotide synthesis making for high-quality probes\textsuperscript{181}.

Although the number of features on spotted arrays has increased, the general protocol, regardless of cDNA/EST or oligonucleotide variety, has not undergone any drastic
changes over the years. Still, arrays are produced by robotic printing, a labeled sample and reference is allowed to hybridize, the slide is scanned and an analysis is performed\textsuperscript{325}.

5.3.2. In situ synthesized Affymetrix arrays
Affymetrix has developed an alternative array strategy, where short 25-meric probes are prepared on the chip by photolithography using masks and light-activated nucleotides\textsuperscript{190, 192}. In the initial expression analysis publication, an array containing 65,000 probes 20 bases in length, with 300 perfectly matched / centrally mismatched probe sets targeting each gene, was employed to investigate expression levels of 118 murine genes\textsuperscript{62}. Such a high level of probe redundancy is, however, not necessary. For instance, the Affymetrix GeneChip Gene 1.0 Human ST Array System encompasses a median of 26 probes per gene\textsuperscript{327}. The first study hybridized a single labeled sample to each array. Current Affymetrix chips also adopt this single dye mode, thus differing from the two-color approach of spotted arrays. The sample preparation involves a cDNA synthesis and \textit{in vitro} transcription (IVT) mediated linear amplification and labeling, which is described below. As a last step prior to addition to the slide, the amplified and labeled antisense RNA (aRNA) is fragmented to minimize the extent of secondary structures potentially interfering with the hybridization.

The \textit{in situ} synthesized chips harbor shorter oligonucleotides potentially increasing the risk of unspecific hybridization compared to the longer probes of spotted arrays. On the other hand, the former strategy involves smaller features, thus offering a considerably higher density. Therefore, every gene can be covered with multiple probes, each corresponding to a different region. This redundancy can offset some of the specificity issues by allowing signal averaging over the probe set corresponding to a particular gene. Additionally, combinations of perfectly matched and mismatched probes can add another level of redundancy\textsuperscript{62, 192}. For instance, the Affymetrix GeneChip Gene 1.0 Human ST Array System has 764,885 probes targeting more or less the whole protein-coding transcriptome\textsuperscript{327}. However, as a new set of masks has to be created to produce altered content on the Affymetrix chips, this approach is significantly less flexible than spotted arrays where it is easier to include additional probes in a new printing run. Furthermore, the spotted arrays can be manufactured in house, whereas the \textit{in situ} synthesized arrays are only commercially available.

5.3.3. Other commercial arrays
The Affymetrix chips have been the most widely employed commercial arrays. However, other vendors offer alternative chips. For example, Illumina offers whole-genome expression assays based on direct hybridization of targets prepared in a manner similar to the Affymetrix protocol to 50-mer probes on decoded random bead arrays\textsuperscript{328}. The HumanWG-6 v3.0 beadchip assays 6 samples for approximately 48,000 transcripts each\textsuperscript{329}.

Agilent also offers both single and dual color platforms for gene expression relying on 60-mer oligonucleotide arrays\textsuperscript{330}. The Whole Human Genome Microarray Kit analyzes over 41,000 transcripts in four samples\textsuperscript{331}.

5.3.4. Probe position considerations
The probes are responsible for ‘capturing’ the correct transcripts, thus converting the amount of ‘capturing’ to an array intensity signal, which can subsequently be converted to a level of transcription. Accordingly, positioning of the probes along the transcript can affect the generated data. For example, utilization of oligo dT in the reverse transcription step exhibits a 3'-bias and positioning of probes in this 3’ region leads to most reliable
data. Alternatively, probes can be spread out over the entire length of the transcript, the expression level being an amalgamation of the signals of the individual probes targeting that gene. In addition, expanded knowledge about transcript structure can be used to generate exon arrays utilizing oligonucleotides corresponding to all exons.

A tiling design, covering large genomic regions with overlapping or non-overlapping probes, represents an extension of the exon-targeted probe positioning strategy. This approach does not require any information about the location of transcripts or exons, and can thus be applied to simultaneously study expression and map the detected transcripts. Naturally, the drawback is that these arrays are less space-efficient than their counterparts relying on already known transcript information.

5.3.5. Benefits and drawbacks?
As mentioned, the single greatest advantage of microarrays is the massive parallelization afforded by the highly miniaturized features. However, cross-hybridization is a problem with gene expression arrays, increasing the background. Together with signal saturation, these spurious interaction reduce the dynamic range, often leading to data compression when compared to, for example, real-time PCR that offers analysis in a far wider message range. The varying hybridization properties of molecules of different size and sequence has been the subject of discussion several times in this thesis. This implies that only relative message abundance can be investigated and prominent differences in expression usually have to be confirmed with other methods. Moreover, the normalization and statistical evaluation to translate the feature intensities to fold change data is far from straightforward. Another disadvantage is that the transcripts have to be known in order to be assayed with arrays. However, with the ever increasing number of genomes and transcriptomes elucidated, the repertoire of organisms that can be assessed steadily grows. Furthermore, the template input requirements are quite steep. These can, however, be reduced using amplification strategies.

5.3.6. Amplification strategies
One flaw, hampering the utilization of microarrays, is the high demand for total RNA input. In the seminal cDNA array study, between 1 and 2 µg of mRNA was applied to each slide. The total RNA requirement was estimated at between 50 and 200 µg. To circumvent and alleviate this limitation, a number of amplification strategies were devised. These strategies can either be implemented prior to hybridization, in the form of mRNA amplification, or post-hybridization, as ingenious signal intensification schemes.

One such scheme involves a two-step hybridization, the first one base-pairing the target to its correct probe, and the second introducing a multiply labeled structure increasing the obtained signal.

The mRNA amplification protocols result in either a linear or exponential increase in transcript amounts. As these procedures include enzymatic steps and modification of the transcripts, certain bias can be introduced, skewing the final data. Generally, the linear protocols have produced more reliable results as compared to exponential methods.

The classical linear strategy, frequently called the Eberwine method, relies on in vitro transcription (IVT). In particular, mRNA is converted to double-stranded cDNA using an oligo dT fused to a T7 polymerase promoter sequence. Addition of this polymerase initiates an IVT reaction generating several copies of antisense RNA (aRNA) from each template. The amplification can either be performed in a single round or in several rounds, which use random priming. In the original study, 80-fold amplification was observed. The initial Affymetrix expression study adopted the IVT strategy that
simultaneously amplified and labeled the target, reporting a reproducible 20- to 250-fold amplification without bias\textsuperscript{62}. A subsequent paper claims a 30- to 100-fold increase in product amount\textsuperscript{392}. Currently, the one-cycle cDNA synthesis/IVT-based amplification and labeling requires 1 to 15 µg of total RNA, while the two-step counterpart lessens the need to 10 to 100 ng\textsuperscript{337}.

A number of improved linear amplification procedures have been developed. The classic IVT method has a 3'-bias as oligo dT is used for priming purposes. To generate full length representation of the transcripts, a protocol utilizing the template switching property of reverse transcriptase was devised\textsuperscript{338}. This enzyme adds several dCTPs as it reaches the end of a template. Consequently, by supplementing the reaction with a primer encompassing a polyguanine stretch at its 3'-end, this template switch can be induced leading to amplification of intact cDNA. One round of amplification was reported to yield a $10^3$-fold amplification, which was increased to $10^5$ with a two-round procedure.

Exponential PCR-based amplification has also been adapted to the mRNA scenario. For example, one method uses a terminal transferase to incorporate a number of non-template dATP moieties at the 3'-end of an oligo dT synthesized cDNA\textsuperscript{339}. In this manner, a single polyT-containing primer can be employed in an exponential amplification, multiplying the material up to $10^{11}$-fold.

### 5.4. Tag-based methods

#### 5.4.1. Serial analysis of gene expression (SAGE) and its longer tag cousins

The elegant method of serial analysis of gene expression (SAGE) was presented in 1995\textsuperscript{321}. This digital expression analysis technique is based on three ideas. The first one is that a short signature sequence can unambiguously identify an mRNA species. Secondly, by joining several tags into longer molecules a single Sanger sequencing run can generate multiple data points. Both these notions offer a considerable cost benefit in comparison with EST analysis as the amount of sequencing necessary is markedly reduced. Lastly, the tag count is proportional to the level of expression. Naturally, a high count represents an abundant transcript, whereas a low count translates into a more rare mRNA.

These three guiding principles led to the development of a protocol to extract signature tags from an mRNA population\textsuperscript{321}. PolyA-tailed RNA species are first converted to double stranded cDNA with a biotinylated oligo dT. The biotin is used to immobilize the ds cDNA on beads. A restriction endonuclease, called the anchoring enzyme, is then used to cleave the transcripts, leaving the 3’ part attached to the beads. The reaction is then divided into two fractions. In one of these, one adaptor is ligated onto the digested end, while the other fraction entails another linker. The two adaptors have distinct sequences, but share a recognition site for another restriction nuclease – the tagging enzyme – capable of cleaving a small distance away from the recognition site. Consequently, digestion with this tagging enzyme produces a linker coupled to a short tag from the original transcript. The constructs are blunt ended, before the outcomes of the two reactions are pooled and ligated. This generates structures with different adaptors at each end and two transcript-based tags aligned end-to-end in-between. After amplification with primers matching the adaptors, the anchoring enzyme is used to remove these adaptors. The free di-tags are then concatenated, cloned and subjected to conventional Sanger sequencing. With 15-base tags, each Sanger read of about 700 bases covers more than 40 signatures.
The original approach entailed \textit{NlaIII} as the anchoring enzyme and \textit{BsmFI} as the tagging counterpart\textsuperscript{321}. This arrangement produces tags of about 13 bases in length originating from the 3’-region of transcripts. However, these short tags are not always sufficient to uniquely identify a transcript. Therefore, several adaptations to the protocol, pertaining particularly to the choice of the tagging enzyme, have been implemented. In LongSAGE, \textit{MmeI} is employed, increasing the tag size to about 20 bases\textsuperscript{340}. SuperSAGE uses \textit{EcoP15I} generating approximately 26 bp signatures\textsuperscript{341, 342}. The ends generated after cleavage with the tagging enzyme in LongSAGE are not blunt ended to provide for maximum tag length. This fact has been claimed to skew the data as it imposes certain restrictions on the ligation process producing the di-tags instead of it being fully random\textsuperscript{342}.

In broad terms the choice between shorter or longer tags is a tradeoff between the sequencing depth that can be achieved and the mappability of the tags to transcripts\textsuperscript{343}. The short tags are economical as a greater number can be sequenced, but some of the tags may map to several mRNAs precluding reliable identification. Conversely, the long tags necessitate more sequencing resources, but offer a superior recognition.

SAGE generally produces reliable digital quantitative data. To realize this, however, numerous tags have to be sequenced making the approach relatively expensive if conventional Sanger sequencing is used. For instance, over 100,000 signatures are usually collected\textsuperscript{343}. Also, there are some problems with unique matching of tags to transcripts, as discussed previously. In addition, while the protocol is elegant, it is rather complicated. Lastly, mRNAs without a recognition site for the anchoring enzyme cannot be assayed.

5.4.2. 5’-SAGE and CAGE

Regular SAGE derives its signature sequences from 3’-sections of the transcripts. Alterations in the procedure have led to methods where the tag instead originates from the 5’-end. As such, these approaches can be used to investigate transcriptional start sites (TSS) concurrently with establishing transcript frequencies.

Cap Analysis Gene Expression (CAGE) involves a cap trapper approach to isolate full-length cDNA\textsuperscript{344}. Following first-strand synthesis, a chemical biotinylation of the cap structure is performed, allowing anchoring of the double-stranded mRNA-cDNA on beads. After treatment with RNase I degrading single-stranded RNA species, the cDNA is released from the mRNA by RNA hydrolysis. A double-stranded linker, featuring an \textit{MmeI} recognition site is ligated to the 3’-end of the cDNA and the second cDNA strand is synthesized. A sequenceable CAGE library is then prepared by \textit{MmeI} restriction, appending of a second adaptor, PCR, isolation of the tags, concatemerization and cloning.

5’-end SAGE represents an alternative to CAGE\textsuperscript{345}. Here, instead of the chemical cap trapper method, an oligo capping approach is used to enrich intact mRNAs. This entails a collaborative effort of several enzymes ultimately leading to ligation of oligonucleotides to 5’-ends of transcripts. The rest of the protocol is highly reminiscent of the normal SAGE and CAGE schemes.

5.4.3. Paired-end ditag strategies (GIS and MS-PET)

The logical extension of the 3’-end regular SAGE and its 5’-end varieties were approaches simultaneously extracting signatures from both ends. Such approaches can find use in demarcating transcript regions as both the start and end sequences are obtained.
Gene identification signature (GIS) analysis of paired-end ditags (PETs), corresponding to the boundaries of mRNAs, is one method in this category. Not surprisingly, it bears resemblance to the various single-end tag extraction strategies. The major protocol difference is that following isolation of full-length cDNA flanked with appropriate adaptors, it is inserted into a vector, creating a full-length cDNA library, instead of directly being used for signature extraction. As with the other methods, Mmel sites are integrated into the adaptors and this restriction enzyme cleaves a small distance into the transcript regions. However, the produced 5’-end and 3’-end tags reside on the same plasmid, being physically linked by the plasmid sequence. Accordingly, these can be ligated creating PETs. The PETs are subsequently excised from the plasmids with another restriction endonuclease, concatenated, cloned and Sanger sequenced. Each PET structure encompasses 18 bases from the 5’-end and 18 bases from the 3’-end including an AA dinucleotide remnant originating from the transcript polyA tail. This AA dinucleotide allows for proper orientation assessment following sequencing.

The first PET publication saw 116,252 Sanger-generated PETs, corresponding to over 7,000 transcriptional units in mouse. The technology was able to faithfully map transcript start and end sites; however, some bias was evident for tag counting purposes. This bias is probably introduced in the cloning step necessary to establish the full-length cDNA library.

Minor changes in the PET protocol enabled this ditag strategy to harness the massive sequencing power of the 454 instrument, severely cutting sequencing costs compared to Sanger. The new strategy was dubbed multiplex sequencing of paired-end ditags (MS-PET) and generated diPET structures approximately 80 bases in length, thus matched to the read lengths of the first generation GS20 instrument. 462,626 tags were sequenced, representing 313,983 unique sequences. This was one of the first demonstrations of the impact on novel sequencers on the field of transcriptomics and a taste of what was to come.

Naturally, the benefit of these approaches is the illumination of the transcript boundaries. However, alternative splicing events, reshuffling the combination of exons present within these boundaries, are not detected. It has further been argued that information about both ends is superfluous in expression profiling applications, which are better served by the more cost-effective and simple SAGE and CAGE approaches. Moreover, the procedure is quite laborious and entails bias-prone cloning. On the other hand, the cDNA library represents a stable repository that can be probed with tag-sequence derived primers to isolate and define interesting novel transcripts.

5.4.4. Massive parallel signature sequencing (MPSS)
Massive parallel signature sequencing (MPSS) is a highly elaborate sequencing method capable to simultaneously read about 20 bases from millions of cDNA molecules bound to microbeads. The type of data is, hence, similar to that generated by SAGE, but MPSS offers an increased output.

Prior to the signature elucidation the mRNA-derived template is immobilized on beads. Briefly, double-stranded cDNA is synthesized and cleaved with a restriction enzyme. These fragments are then cloned into vectors harboring unique address labels and universal priming sites. PCR amplification produces linear molecules encompassing both the mRNA-originating region and the address handle. The address handle is rendered single-stranded and hybridized to a complementary sequence bound to a bead. Subsequently, a ligation occurs immobilizing the fragments onto the beads. 16.8 million handles are used and, consequently, 16.8 different anti-handles, one type per bead. This
Parallel target selection by trinucleotide threading

rather complicated protocol generates beads harboring about 100,000 copies of identical cDNA segments.

The signature sequencing is also quite intricate\textsuperscript{349, 350}. The cDNA-carrying beads are deposited onto a flow cell and the bases interrogated, in a parallel fashion, by a ligation-based strategy. After restriction endonuclease cleavage and a fill-in of a single nucleotide, an adapter containing a $B_{	ext{be}}I$ site is ligated onto each fragment. This adapter is necessary for sequencing initiation. $B_{	ext{be}}I$ restriction cuts into the template producing a four-base overhang, which is then probed with four partially double-stranded 256-probe sets, each set targeting one position of the overhang. Only perfectly matching probes are annealed. As four sets are used, four different probes will be ligated onto fragments of each bead, one per interrogated position. Besides the querying single-stranded portion of the probes each carries one of 16 encoder protrusions at the other end. Accordingly, following ligation, a second labeled decoding probe set is used to delineate the identity of the interrogation probe and thus the base. The interrogation probes carry carefully positioned $B_{	ext{be}}I$ sites, which enable digestion exposing the next four bases. The sequencing is thus performed in cycles of ligation of interrogation probes, detection with decoder probes and cleavage.

MPSS is able to produce a large number of tag sequences. For example, over 1.6 million signatures were obtained from human THP-1 cell cDNA\textsuperscript{349}. However, as is clearly evident, this approach suffers from high complexity.

5.4.5. Polony multiplex analysis of gene expression (PMAGE)

Polony multiplex analysis of gene expression (PMAGE) is yet another tag approach for transcript abundance studies\textsuperscript{283}. It relies on the polony sequencing-by-ligation strategy (Chapter 4). Briefly, immobilized double-stranded cDNA is cleaved with $N_{	ext{lai}}$III and a linker, encompassing a recognition site for $AcuI$, is appended. Restriction with the latter enzyme generates a 14 bp tag consisting of 10 transcript-specific bases and the four-base $N_{	ext{lai}}$III recognition sequence. Another adaptor is ligated onto the $AcuI$-digested end and the adaptor-enclosed tags are purified, loaded onto beads and emulsion PCR amplified. The beads, each carrying multiple copies of a single tag, are thereafter subjected to ligation-based sequencing.

The PMAGE strategy sequenced over 4 million tags in parallel, and the obtained expression data correlated well with that of SAGE and reverse transcription RT-PCR\textsuperscript{283}. In addition, it was estimated that most transcripts could be analyzed using an output of about 2 million tags. A broad dynamic range was demonstrated with individual mRNA species occupying a count interval of between 0.3 and approximately 29,000 tags per cell.

5.5. Deep sequencing with second generation platforms

The novel second generation massively parallel sequencing instruments, described in Chapter 4, can be used for a multitude of applications. One type of investigations uses these sequencers to investigate various aspects of RNA, constituting a field collectively referred to as RNA-Seq. In several aspects, RNA-Seq combines advantages of the different global approaches: the highly parallel analysis of microarrays with the reliable quantification of EST-sequencing and tag-based approaches. However, while digitally measuring levels of gene expression, it simultaneously offers access to completely new avenues of data. For example, RNA-Seq can illuminate individual exons, shedding light on transcript structure and alternative splicing, and identify mutations and polymorphisms in transcribed regions. As such, one single experiment can provide a plethora of information. This resembles the paired-end tag strategies, which can analyze
expression levels while providing transcript boundaries, and tiling arrays, capable of supplementing expression with mRNA structure, but the new sequencers take this to an unprecedented level.

RNA-Seq studies have been performed on all four commercial platforms – 454, Illumina, SOLiD, and Helicos – targeting RNA from a diverse set of organisms. These studies are conceptually similar, although, of course, they each entail certain distinguishing and differing features. In broad terms, double-stranded cDNA is synthesized from an RNA population and fragmented to a desirable size. Sequencing adaptors are ligated onto these fragments, which are then enrolled into the sequencing pipeline. This pipeline differs between the platforms as has been discussed in Chapter 4. The single-end or paired-end reads are then either bioinformatically aligned to a reference or assembled de novo, generating a comprehensive view of the transcriptional landscape that can be used to infer the types of information outlined above. Most commonly, polyA-enriched RNA is used as input material, but other selection strategies have also been employed. Additionally, the RNA molecules can be fragmented prior, as opposed to after, the cDNA synthesis step.

5.5.1. Advantages and disadvantages of RNA-Seq
As mentioned, the benefits of RNA-Seq are far reaching. In addition to the wealth of data generated in a single experiment and the deep coverage these approaches offer a superior dynamic range. One study reports a dynamic range of at least 8000-fold, to be balanced against the about 60-fold for arrays. Other investigations have shown accurate data over a range of four or five orders of magnitude. The expression data compares well with the ‘gold standard’ method of quantitative PCR. Additionally, the approaches have low sample amount requirements. In one study utilizing carefully selected ‘not-so-random’ (NSR) hexamers, 1 µg of total RNA was sufficient. Moreover, an exponential amplification strategy akin to the one described for microarrays, was used to study the transcriptome of a single cell.

There are some limitations associated with RNA-Seq. Even though the cost related to the prosperity of data is very low, each run is still quite expensive, precluding a more widespread use. Additionally, the various sample preparation strategies can introduce bias, favoring certain transcript regions over others. For example, in a polyA-based RNA-Seq study, a 3'-bias was observed, probably because of enhanced selection and priming at 3'-ends. Furthermore, handling, storage and bioinformatics analysis of the large quantities of data are quite demanding.

5.6. Validation methods
The approaches discussed above are aimed at providing a global view of a cell’s protein-coding transcriptional landscape. When comparing two kinds of cells, or a single cell type under diverse conditions, the majority of genes exhibits rather small fold change differences, while a small number of transcripts shows signs of over- or underexpression. These differentially expressed genes are promoted to the limelight as they can provide clues to biological function or processes. However, it is customary to validate and confirm the global-scale results with an alternative platform. This pertains especially to DNA arrays, which only provide a relative quantification and are negatively influenced by high cross-hybridization induced backgrounds.

Two methods have been considered ‘gold standards’, generally agreed to generate reliable results. The first of these is Northern blotting devised in late 1970s. Quantitative real-time PCR (RT-PCR) represents the other, currently favored, approach. The
Parallel target selection by trinucleotide threading

drawback of these techniques is that only one, or at most a small set of transcripts, can be assessed simultaneously.

5.6.1. **Northern blot**
The DNA analysis protocol of Southern blotting is not amenable to study RNAs and, consequently, Northern blotting represents an adaptation of the Southern technique to address this obstacle. In particular, RNA is size-separated with gel electrophoresis and then transferred and attached to specially modified paper strips or membranes. Following the immobilization, the strips or membranes are surveyed with labeled probes complementary to the transcript of interest. Lastly, after washing an imaging step takes place, revealing the bands to which the probes have hybridized, thus providing a measure of the expression level. Additionally, as a size marker is employed in the electrophoretic separation, the lengths of the transcripts can be estimated, offering clues as to the mRNA structure, for example alternative splicing events. The Northern blotting technique is specific and the covalently bound samples can be sequentially probed with different oligonucleotides. However, the method necessitates rather high RNA amounts. Moreover, the early approaches were cumbersome and included hazardous chemicals. Some of these issues have been resolved with later adaptations.

5.6.2. **Reverse transcription quantitative RT-PCR**
Presently, reverse transcription quantitative real-time PCR (RT-PCR) is considered to be one of the most reliable quantitative expression analysis techniques. Furthermore, it has excellent sensitivity as RNA from single cells can be analyzed. RT-PCR has been described in the Chapter 2. Briefly, in contrast to conventional PCR, the amplification is monitored concurrently with its occurrence. To detect product formation, unspecific dyes or specific probes are utilized. For gene expression, the RNA is converted to complementary DNA (cDNA) and a PCR targeting a region specific to the desired transcript is thereafter followed in real-time. By allowing one of the PCR primers to double for cDNA synthesis, the reverse transcription and amplification can be merged in a one-step strategy. A two-step approach, separating the cDNA synthesis from the amplification phase, allowing tailor-made conditions for each, tends to improve the sensitivity.

A threshold level of signal is decided upon, and the cycle number when PCR has generated enough material to surmount this level is registered. Naturally, low threshold cycle values are indicative of large amounts of the corresponding transcript, whereas high threshold cycle numbers signify rare mRNA species. Normally, the data is normalized to a reference gene, primarily a housekeeping one, and correlated to a reference sample via the $2^{-\Delta\Delta C_t}$ method.

The major limitation of RT-PCR is the limited transcript processivity, as a maximum of only a couple of genes can be assayed simultaneously. However, real-time instruments operating in 384-well formats offer a great sample throughput. Moreover, a recent single cell RT-PCR study utilizes an oligo dT-based cDNA library attached on beads. By simply collecting and washing the beads, multiple real-time amplification assays can be performed in a row.

5.7. **Intermediary methods**
On a global scale, the gene expression differences among specimens, for instance between a diseased and control sample or two different cell types, are rather subtle. Only a relatively small set of genes exhibits considerable differential expression, whereas the great
majority of transcripts is present at similar levels. Consequently, following a large-scale survey of gene expression and corroboration of the significant upregulated and downregulated results with a ‘gold standard’ validation method, the obtained differential transcript profile can serve classification, discrimination or identification purposes. As such, the scale of experiments can be dramatically reduced, while preserving the most vital information. For example, to develop an expression fingerprint to distinguish acute myeloid leukemia (AML) from the related acute lymphoblastic leukemia (ALL), Affymetrix chips interrogating over 6000 genes were used for an initial scan. These produced a set of 1100 differentially expressed genes as possible candidates for the classification task. Of these, the 50 most significant ones were chosen and were shown to dependably differentiate between AML and ALL. Furthermore, sets comprising between 10 and 200 genes were argued to have similar distinction potential.

However, to unleash the full promise of the transcript profiles, especially for diagnosis and prognosis in a clinical setting, intermediary expression platforms are needed. These should be capable of faithfully capturing the expression levels of moderate gene sets, approximately between 10 and 100, at a high sample throughput and ease of operation.

5.7.1. **BeadsArray for the Detection of Gene Expression (BADGE)**

BADGE, or BeadsArray for the Detection of Gene Expression, represents one intermediary platform. It combines a hybridization-based scheme with the microsphere suspension array offered by Luminex (Chapter 3). Oligonucleotides, 25 bases in length, matching the transcripts of interest are coupled to the microspheres in individual reactions with one probe being coupled to one type of bead. Following the bead preparation, the different types of microspheres are combined to allow for parallel analysis. Total RNA is converted to double-stranded cDNA, which is in vitro transcribed into biotin-labeled antisense RNA (aRNA). The aRNA is then mixed with the microsphere set and a solution-phase hybridization allowed to occur. The microspheres, now carrying labeled aRNA in an amount proportional to the expression level, are then passed, one by one, through the Luminex instrument that simultaneously decodes the bead identity and detects the target amount.

This system was used to monitor expression of 20 *Arabidopsis* genes, and was demonstrated to produce results agreeing with ones generated with Affymetrix chips. In addition, it was stated that 10 µg of input total RNA was sufficient to analyze moderately abundant transcripts, defined as being present at between 10 and 30 copies per cell.

The current state of the art Luminex system encompasses 500 distinctly internally labeled beads, implying that 500 genes can be assayed simultaneously. In addition, the method is rapid, simple and flexible.

5.7.2. **cDNA-mediated annealing, selection, extension and ligation (DASL)**

cDNA-mediated annealing, selection, extension and ligation (DASL), and its forerunner, RNA-mediated annealing, selection and ligation (RASL), are techniques amenable for analyzing moderate gene sets in a high-throughput manner. The unconventional feature is that they utilize enzymes to add specificity, in a manner resembling SNP-interrogation assays. In DASL, cDNA, created using total RNA and a mixture of biotinylated oligo dT and random hexamers, is immobilized on a solid support. For each transcript, two oligonucleotides are allowed to anneal with a small gap present between them. The concerted action of a polymerase and ligase closes the gap and joins the two probes generating a longer entity. The ends of the primers carry
universal amplification handles, allowing all the longer entities to be simultaneously PCR amplified. In addition, the downstream primer carries an address tag, matching an anti-tag coupled to a bead on an Illumina random bead array. Consequently, following PCR amplification, one of the employed common primers being fluorescently labeled, the products are hybridized by virtue of their address tags and a signal, corresponding to the expression level, is obtained. The RASL-assay is analogous, but RNA is used directly instead of being converted into cDNA. In addition, in RASL the two oligonucleotides anneal end-to-end, relying only on a ligase. The polymerase-ligase approach improves the specificity.

The bead arrays in a 96-format carry 1536 different address tags, making this the upper limit of multiplicity. However, usually three sequences are queried per transcript allowing over 400 genes to be assayed\(^3\). The dynamic range was estimated to cover 2.5 orders of magnitude and adequate results are produced from 25 ng of total RNA. As with BADGE, DASL is flexible and can be performed in a high-throughput manner in a 96-format.

Currently, both a whole-genome DASL assay and targeted counterparts are marketed by Illumina. The former covers over 24,000 transcripts by DASL chemistry, but hybridizes the specific transcript regions directly instead of relying on address sequences\(^3\). 10 to 200 ng of total RNA is required. The targeted approaches are amenable to investigate up to 1536 mRNA types in 16 or 96 sample formats\(^3\). However, this encompasses a single site per transcript. Utilizing three such sites, 516 species can be analyzed in parallel. The total RNA requirement is about 100 ng.

5.8. **Summary table**

**Table 3:** The principal expression analysis approaches and their features.

<table>
<thead>
<tr>
<th>Validation approaches</th>
<th>Interrogation principle</th>
<th>Background noise</th>
<th>Dynamic range</th>
<th>A priori knowledge necessary?</th>
<th>Number of assayed genes</th>
<th>Number of assayed samples</th>
<th>Cost per gene</th>
<th>Cost per sample</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intermediary approaches</td>
<td>BADGE</td>
<td>Hybridization</td>
<td>Medium / High</td>
<td>Yes</td>
<td>Medium</td>
<td>High</td>
<td>Medium</td>
<td>Low</td>
</tr>
<tr>
<td></td>
<td>DASL</td>
<td>Extension and ligation</td>
<td>Low / Medium</td>
<td>Yes</td>
<td>Medium</td>
<td>High</td>
<td>Medium</td>
<td>Medium</td>
</tr>
<tr>
<td>Global approaches</td>
<td>EST sequencing</td>
<td>Sequencing</td>
<td>Low</td>
<td>High</td>
<td>Low</td>
<td>High</td>
<td>Low</td>
<td>High</td>
</tr>
<tr>
<td></td>
<td>Signature tag approaches: SAGE and various renditions</td>
<td>Signature tag sequencing</td>
<td>Low</td>
<td>No</td>
<td>High</td>
<td>Low</td>
<td>Medium</td>
<td>High</td>
</tr>
<tr>
<td></td>
<td>RNA-Seq</td>
<td>Deep sequencing</td>
<td>Low</td>
<td>No</td>
<td>High</td>
<td>Medium*</td>
<td>Low</td>
<td>High</td>
</tr>
<tr>
<td></td>
<td>Microarrays</td>
<td>Hybridization</td>
<td>High</td>
<td>Narrow</td>
<td>Yes</td>
<td>High</td>
<td>Low</td>
<td>Medium</td>
</tr>
</tbody>
</table>

* Given the massive output, expression analysis of many samples in a single sequencing run is feasible. The samples could be barcoded and pooled prior to sequencing. Alternatively, each sample could occupy a single lane on a multi-lane sequencing plate.

The ranking is performed according to a relative low-medium-high scale. Moreover, as the expression analysis assays can be performed with different settings and conditions, the ranking is only general and may not be applicable to all situations.
PRESENT INVESTIGATION

6. The papers

The Introduction part of this thesis has delved into amplification of nucleic acids and the predominant groups of methods employed to infer biologically relevant data from the amplified segments. For the majority of techniques, amplification represents a prerequisite step before the genotype, sequence or expression level can be elucidated. Accordingly, it is a crucial component in the portfolio of molecular biotechnology approaches.

The papers constituting this thesis all incorporate the trinucleotide threading (TnT) multiplex amplification strategy. TnT is capable of simultaneously amplifying several DNA segments in a highly specific fashion, and, consequently, complements the repertoire of parallel approaches outlined in Chapter 2.

TnT is strictly an amplification module, allowing integration of a number of upstream sample preparation approaches, as well as a multitude of detection platforms. As such, a variety of assays targeting interesting DNA or RNA features can be implemented. In Paper I, the original TnT protocol for single nucleotide polymorphism (SNP) genotyping was adapted to the expression profiling setting. The readout relied on generic tag arrays. In Paper II, TnT was recruited for amplification of short tandem repeat (STR) regions together with capillary gel electrophoresis scoring. Paper III saw TnT in conjunction with the massively parallel Pyrosequencing platform developed by 454/Roche be applied to determine the allele frequencies of SNPs in a cohort. Finally, in Paper IV, a TnT-454 sequencing approach was employed for targeted digital transcript profiling. Additionally, in this paper a simplified hybridization strategy was demonstrated.

In summary, this thesis displays the application of TnT amplification for gene expression analysis, as well as for STR and SNP genotyping. Furthermore, different detection platforms – conventional planar arrays, capillary gel electrophoresis and second generation sequencers – were combined with the TnT module.
6.1. **Trinucleotide threading**

Nucleic acid amplification comprises a diverse set of methodologies capable of addressing various needs (Chapter 2). Conventional PCR is, for instance, perfectly suited for targeting a single or a small number of regions. The opposing corner embraces whole-genome amplification (WGA) approaches, randomly multiplying the entire genomic complement. The space between the two categories is filled with a variety of approaches aiming at targeted multiplex amplification, thereby to some extent combining the distinguishing features of the PCR and WGA worlds.

Trinucleotide threading (TnT) adds to the assortment of targeted multiplex amplification methods. It harnesses the specificity of a polymerase, acting on a restricted trinucleotide set, and a ligase to faithfully amplify several regions in a parallel manner. The high specificity afforded by the collaborative effort of the two enzymes minimizes formation of artifacts and spurious products. TnT proceeds in two phases: a linear followed by an exponential. Two probes are designed for each segment of interest. Proper annealing of the upstream extension primer and downstream thread-joining primer to the same strand of the template molecule leaves a small gap between the probes. The placement of the TnT primers is carefully selected to enable a gap composed of only three out of the four possible nucleotide types. The reaction mixture encompasses a polymerase and the corresponding trinucleotide set allowing elongation over the gap region until the thread-joining primer is encountered. Following gap bridging, a ligase, also present in the reaction, ligates the two moieties forming a complete DNA thread. This is enabled by virtue of a phosphate group appended to the 5'-ends of all thread-joining primers. After one round of thread formation, the mixture is heat denatured and brought back to the threading temperature enabling a fresh pair of TnT probes to anneal and participate in the extension-ligation event. This cycling, made possible by the very low tendency of spurious product formation, permits a controlled linear amplification and, consequently, an increased sensitivity. One type of thread is generated for each targeted region. All threads share terminal universal amplification handles and after a clean-up step, they are subjected to PCR with a single primer pair. This represents the exponential phase of TnT. Care is taken during the assay design to keep the length of the extension region (i.e. the gap) restricted within a narrow interval. This enables a thread population of homogenous lengths, and hence not susceptible to inherent PCR length bias. Subsequent to the exponential amplification, the reaction products are analyzed, scored or interrogated, depending on the nucleic acid characteristic under study.

Trinucleotide threading integrates several measures to ensure a high degree of specificity. In particular, formation of a complete thread necessitates primer annealing, polymerase extension and ligase joining. Accordingly, primer misannealing is not sufficient to give an artifact as the vast majority of such events will create gaps composed of all four nucleotides and thus unbridgeable with the restricted nucleotide set. The generated partial thread species lack the necessary terminal handles and can, therefore, not participate in the ensuing exponential PCR amplification. Taken together, three distinct checkpoints have to be overridden to give rise to a spurious amplifiable thread.

6.2. **Initial publication**

Initially, the TnT amplification module was applied in a protocol for SNP genotyping.\(^{374}\) In the SNP scenario, two threading reactions featuring two distinct trinucleotide sets have to be set up to enable interrogation over all SNP classes. A set comprising 75 cancer-related polymorphisms was selected. To maximize the number of SNPs working in the assay consideration was taken with regard to nearby SNPs potentially interfering with priming, as well as to troublesome genomic segments such as repetitive regions. 39 of the
SNPs partook in a reaction comprising a CTG set, while AGT was used for the remaining 36. The TnT primers and the selection of the optimal trinucleotide sets were performed using a developed software. Following TnT, the SNPs were scored using protease-mediated allele specific extension (PrASE)\textsuperscript{210}, further described in Chapter 3, together with hybridization onto generic address arrays. This interrogation bestowed the TnT-genotyping assay with additional specificity.

Three polymorphisms produced only weak signals even in simplex reactions. Of the remaining 72 SNPs, 68 gave well-defined genotype calls. This corresponded to a conversion rate of about 94%. The call rate was approximately 98%. An input of 1 ng of genomic DNA, equivalent to roughly 150 cells, produced reliable results. Consequently, the TnT faithfully amplified the chosen genomic segments and the amplified material was ample and of adequate quality to permit SNP interrogation.

6.3. **Expression profiling (Paper I)**

In many cases it is superfluous to analyze the expression levels of all the genes. Instead, limiting the analysis to carefully selected genes displaying significant differential expression between two samples is still highly informative while dramatically decreasing the scale of the involved experiments. Usually, intermediary gene sets, roughly comprising between 10 and 100 entries, are employed offering a cost benefit and a potential to profile samples in a high throughput manner. Such moderate collections of appropriate genes can find uses in disease diagnosis or subgrouping, prognosis prediction or to distinguish two cells from each other. What is necessary to reap the promises of intermediate gene sets is a method capable of parallel analysis of the chosen genes in a convenient and reliable manner. Such a technique would complement the global and validation expression approaches outlined in Chapter 5. A number of methods have been developed to address this need. The cDNA-mediated annealing, selection, extension and ligation (DASL)\textsuperscript{308}, commercialized by Illumina, and BeadsArray for Detection of Gene Expression (BADGE)\textsuperscript{309} are two examples further discussed in Chapter 5.

In Paper I, a method for parallel analysis of intermediary gene sets incorporating the multiplex TnT amplification strategy was presented. The underlying principle was to convert the mRNA species of interest to DNA threads maintaining the original transcript proportions. The conversion relied on TnT regions, each of which was selected to uniquely match to a particular desired transcript. In addition to unambiguously identifying the genes, the TnT segments encompassed central trinucleotide sections acting as the bridgeable gaps in the threading reaction. Accordingly, assessing the amount of each type of DNA thread would provide a measure of the expression level of the corresponding gene. Naturally, high signals would indicate an abundant message, whereas low amounts would signify rare mRNA molecules.

To evaluate the TnT expression profiling approach an 18-gene set was selected. 15 entries corresponded to randomly chosen sample genes, while the three remaining ones served housekeeping functions and were included for normalization. The unique TnT regions were not chosen from full-length transcripts, but from protein epitope signature tag (PrEST) sequences. PrESTs are the basic units of the Swedish Human Proteome Resource (HPR) program aiming at producing a global human protein atlas\textsuperscript{66, 375} (Chapter 1). Each such unit corresponds to a unique approximately 100 to 150 amino acid stretch of a protein, and hence provides a clear-cut protein identifier. To facilitate the TnT probe design, a script implemented in Java/BioJava was developed. Given a set of user-defined parameters, this software tool screens the input sequences for suitable TnT
Parallel target selection by trinucleotide threading

regions and assigns a probe set for each identified one. In this investigation, an extension region (i.e. the gap) of between 10 and 12 bases was preferred.

Total RNA from eight different cancer cell lines was used as template together with a universal human reference RNA. Two cDNA synthesis priming strategies were applied, each encompassing 2 µg of total RNA. Recruiting oligo dT generates a representation of all polyA+ RNA species, whereas priming with gene-specific oligonucleotides only targets the desired genes. An amount corresponding to about 200-250 ng of total RNA entered the TnT reactions. Following the thread formation reaction, clean-up and exponential PCR amplification with a common primer pair, the levels of each type of thread were assessed with a detection primer extension reaction with a labeled nucleotide mix. Here, the extension regions of the threads formed the mainstay for detection primer annealing. Additionally, as each detection primer carried a 5'-address sequence, the scoring step employed hybridization to a universal tag array, scanning and retrieval of the individual array signal intensities. These intensities were subsequently normalized and converted to fold changes.

To validate the TnT-generated results, the same gene set was profiled with the ‘gold standard’ method of real-time PCR. To normalize the RT-PCR data with respect to all three housekeeping genes simultaneously and to derive fold change data, a computation scheme applying the 2ΔΔCt method was formulated. Moreover, the TnT and RT-PCR fold changes were complemented with expression data from a comprehensive cDNA microarray in a cross-platform comparison.

Firstly, the two priming approaches employed in the cDNA synthesis were evaluated. A high degree of conformity was observed with the average Pearson correlation coefficient over the eight cell lines being 0.84.

The benchmarking against RT-PCR revealed a good concordance. The average Pearson correlation coefficient across the eight cell lines was 0.86 with the genes generally displaying analogous expression patterns. This comparison was performed on 14 genes as one exhibited considerable deviation in a preliminary assessment and was therefore removed from subsequent analysis. However, this incongruent gene also showed fold change discrepancies in a comparison of whole-genome cDNA array and RT-PCR. Overall, TnT compared most favorably to RT-PCR in the cross-platform comparison, whereas the array platform exhibited inferior correlations. Another interesting observation pertained to the degree of data compression. RT-PCR generated the highest fold change magnitudes, while the extent of underestimation was largest for arrays. The TnT method showed intermediary compression.

The TnT technique bestows the expression profiling approach with a high specificity. An additional layer of specificity is offered by the detection primer extension step preceding the array readout. In particular, there is a miniscule risk that TnT probes of different genes anneal to undesired regions in a manner creating a trinucleotide gap. This will generate a spurious thread that will complete the TnT pipeline. However, as this thread is not matched to any detection primer, it will not give rise to any labeled detection products and, consequently, will not interfere with the scoring. Moreover, cycling of the thread formation reaction produces a controlled linear amplification, which lends the approach additional sensitivity.

In addition to the high level of specificity, the modular nature of the assay renders it flexible. For example, the conventional planar array detection can be replaced with
another readout platform. The assay is also user-friendly with many of the steps being automated. The utilization of the 96-plate format together with biotin-streptavidin interaction for purification purposes offers further avenues to automation.

In summary, this proof-of-concept investigation demonstrated the suitability of the gene expression analysis incorporating the TnT multiplex amplification strategy for parallel expression profiling of intermediary gene sets. The method offers a high specificity and the obtained results compare favorably with those of the established RT-PCR technology. Additionally, the individual elements adding up to the complete assay are clearly separated allowing for substitutions with alternative steps, making possible tailor-made combinations for specific tasks.

6.4. Short tandem repeat analysis (Paper II)

Short tandem repeats (STRs), or microsatellites, constitute an abundant marker class that due to the high variability has found widespread use for identity establishment. Both forensic analysis and paternity testing have benefited from these informative variants. However, as STRs are more complex than SNPs it is more difficult to analyze them in a parallel fashion. Currently, the predominant approach entails multiplex PCR amplification followed by detection by length separation with capillary electrophoresis. As was outlined in Chapter 2, PCR is grossly affected when the multiplicity increases. As such, only a small degree of parallelization can be attained before the reaction is dominated by artifacts.

Therefore, Paper II describes the adaptation of trinucleotide threading to multiplex amplification of STR regions. In this proof-of-concept study, three markers from the extensively used FBI CODIS set were simultaneously assayed. To enable a comparison, equivalent simplex TnT amplifications were performed. The three STRs – TPOX, CSF1PO and D18S51 - have in common a tetrarepeat consisting of the AGT nucleotide set. Consequently, they were perfectly suited for TnT whereby the gap bridged during thread formation coincided with the repeated region. In addition, the prevalent repeat numbers of these three markers were nonoverlapping, making for length-separated thread species and a straightforward capillary gel electrophoresis readout.

Electrophoretic separation of the multiplex amplifications clearly displayed three length-separated peak groups, one for each analyzed STR. The D18S51 marker producing the longest thread species generated weaker signals than the other two STRs. This is not surprising given that PCR preferentially amplifies shorter products. In case of a single peak for a locus, the sample was homozygous, whereas dual peaks represented heterozygotes. The peaks from the multiplex amplifications matched the ones from the simplex reaction validating the parallel strategy. Finally, the inferred fragment lengths were in agreement with the prevalent published repeat numbers. Accordingly, the TnT technique represents a viable approach for STR amplification.

As already mentioned, the advantages are the high specificity and the flexibility afforded by the modular nature. For instance, the capillary electrophoresis platform vulnerable to overlapping fragment lengths could be replaced by 454 sequencing, dramatically increasing the output.

6.5. SNP allelotyping (Paper III)

The emergence of second generation platforms has revolutionized biomedical research permitting researchers to pursue novel studies and revealing the underpinnings of biological phenomena. The current commercial platforms as well as their impact are
described in Chapter 4. The novel sequencers offer a massive output. However, the costs associated with second generation sequencing are still rather steep. In order to maximize the information content of each sequencing run, upstream preparation technologies have to be provided to ensure that the sequencing resources are allocated for the proper targets. In this manner, the expenses can be kept at a minimum.

In Paper III, the SNP genotyping approach of the original TnT publication was recruited for allelotyping relying on digital readout with the 454 massively parallel sequencer. Allelotyping is defined as allele frequency estimation in a cohort. These studies simultaneously genotype a large number of individuals, making them perfectly matched to the 454 data output. Array studies along similar lines of thought – employing pooled genomic DNA – have suffered from relatively poor accuracy due to the inherent relative quantification provided by this platform.

The same cancer-associated 75 SNP set as used in the initial study was supplemented with 75 SNPs implicated in obesity. However, three polymorphisms in the obesity set exhibited unspecific signals in a preliminary array-based analysis and were consequently removed from further investigation. Additionally, some of the SNPs did not generate any detectable array signals or did not yield compelling clusters. However, to fully assess the performance of the 454 strategy, they were still included in the final list. Accordingly, this list entailed 147 entries.

In the original publication, the cancer set gave conversion and call rates of 94% and 98%, respectively. With the array readout in this study, the obesity genes generated analogous percentages of 88% and 99%. The lower conversion rate for the obesity set was not surprising given a higher presence of low-complexity regions in that set. Furthermore, one of the SNPs was present in both sets. As the TnT oligonucleotides for the one of the polymorphism ensembles were PAGE-purified, whereas those of the other were not, primer quality effects on the reaction outcome could be assessed. The PAGE purification was shown to generate improved array results. For instance, the obesity set generated 2.6-fold higher signal intensities on arrays compared to the cancer one.

Equal amounts of genomic DNA from 462 individuals were combined. 45 ng of this pool (corresponding to about 13,500 alleles) entered the TnT-based SNP genotyping approach. In addition to the array protocol, 454 adaptors were introduced and the population of threads ready to enter the sequencing protocol was quality assessed. The 454 protocol involved clean-up, emulsion PCR and parallel Pyrosequencing in the Genome Sequencer 20 (GS20), the first version of the 454 instrument. Following sequencing, custom scripts were used to assign the reads to hypothetical threads and to translate this data into allele frequencies.

The GS20 run generated over 204k reads. With the average read length being 90 bases this totaled to approximately 18.5 Mb of sequence. 61.6% of the reads could be matched to hypothetical threads without a single mismatch. Allowing for slightly more relaxed assignment criteria, the percentage of reads mapping to desired threads was increased to 81%. Of the remainder of the 454 sequences, 5.6% were found to correspond to an illegitimate thread formed by a one-base extension between non-matching TnT primers. This species did not influence array-based readout owing to the additional PrASE scoring step. The rest of the reads comprised short sequences and reads failing to meet the defined assignment parameters. In total, 86.6% of the obtained sequences were identified as threads.
Of the 147 targeted SNPs, 126 gave compelling allele frequency data based on the read counts for each individual allele. This corresponded to a conversion rate of 86.3%. The allele frequencies correlated well to those obtained via arrays. The failure of the remaining polymorphisms could primarily be attributed to low or no sequence counts. There was concordance between 454 and array readout with respect to nonfunctioning and poorly performing SNPs. However, some markers were found to produce successful results in the sequencing approach but not on arrays. Vice versa, a number of SNPs gave compelling clusters on arrays, but produced low count numbers with the 454 detection. One potential explanation for the latter observation was the presence of homopolymeric stretches around these markers as these segments present the major obstacle for Pyrosequencing. Furthermore, the number of sequences assigned to the obesity SNPs (82% of the true thread reads) was far greater than for the cancer polymorphisms. This was, however, expected based on the prior array investigation. There was also count variation between individual SNPs. It should, nevertheless, be emphasized that many of the observed negative effects were uncovered with initial analysis and could have been avoided in the 454 experiment. However, as the intention was to comprehensively assess the 454 platform for allelotyping purposes, even the poorly functioning primers in the preliminary array investigation were included.

Taken together, combining targeting desired polymorphisms, pooling of genomic DNA and counting 454-generated reads corresponding to each allele produces accurate frequency estimations, which can facilitate identification of marker SNPs. Naturally, the pooling precludes gathering of individual genotypes. However, this is counterbalanced by the drastically lessened workload and reduced time frame.

### 6.6. Targeted expression profiling with 454 sequencing (Paper IV)

Second generation sequencing has had a far-reaching impact on several biomedical fields. Examples of how these novel platforms are transforming the way research is conducted and how the massive output leads to an increased and more thorough understanding of biology are scattered throughout this thesis. For instance, as described in Chapter 5, the field of RNA-Seq, where highly parallel sequencing is used to comprehensively study RNA, has provided new insights on various aspects of the transcriptome. In Paper IV, the expression profiling variant of TnT was evaluated as an upstream transcript selection procedure for 454 sequencing, and thus for a targeted RNA-Seq approach. As such, the investigation constituted a combination of the TnT expression analysis pipeline of Paper I and the 454 readout of Paper III. Furthermore, transcript profiling with second generation sequencing offers a superior dynamic range that could be capitalized on to allow for reduced input material requirements. This topic was also examined in this paper.

To assess the viability of the TnT-454 targeted transcript profiling approach, a gene set featuring 32 entries related to basal cell carcinoma (BCC) was selected. Total RNA from two different cell lines – EFO-21 and SK-MEL-30 – entered the same TnT protocol as in Paper I. Total RNA dilution series were used to determine sample input requirements. Amounts corresponding to 350 ng, 35 ng (10x-dilution), 3.5 ng (100x-dilution) and 0.35 ng (1000x-dilution) of total RNA were employed in the threading reactions. As each cell encompasses approximately 20 pg of total RNA this equals to about 20, 200, 2000 and 20,000 cells, respectively. Parallel to the 454 readout, a direct hybridization scheme was implemented. This enabled a cross-platform comparison and benchmarking.
In contrast to the original expression profiling publication where detection involved thread-templated primer extension followed by hybridization of the extension products onto generic tag arrays, here, the threads were directly hybridized onto complementary probes. Omission of the primer extension step resulted in a simplified detection protocol. However, a high level of specificity was still retained as the array oligonucleotides matched the 10-14 bp extension regions of true threads. Spurious thread species harboring different extension regions lack docking sites among the array probes and will not be detected.

The direct hybridization strategy showed a clear tendency for increased signal intensities compared to the scheme involving an additional primer elongation step. Moreover, a total RNA input of 3.5 ng generated dependable transcript abundance data. The Pearson correlation coefficient compared to the data obtained from the highest amount sample was 0.91 for EFO-21 and 0.79 for SK-MEL-30. However, the 1000x-dilutions produced too weak array signals to yield reliable data.

The 454 sequencing generated roughly 15,500 reads for each sample. In the case of the highest total RNA input, over 70% of the reads were mapped to correct thread species. However, with decreasing total RNA amounts, the percentages of reads matching true threads were reduced. For the samples featuring 100 times less starting material, 6.6% (EFO-21) and 2.8% (SK-MEL-30) of the reads corresponded to genuine threads. With an additional 10x dilution, this dropped to about 0.3 – 0.4% yielding too few thread counts to enable analysis. Even with the diminishing count numbers, the transcript abundance data was still highly correlated. Comparing the highest input amount with the 100x dilutions produced Pearson correlation coefficients of 0.94 and 0.96 for EFO-21 and SK-MEL-30, respectively. Moreover, the data sets obtained with arrays and 454 exhibited good concordance. The same genes were found to be highly and lowly expressed with both platforms, although the magnitudes were somewhat different, not exhibiting any clear-cut pattern. The relative nature of the array data together with comparison of adjusted raw signals might explain part of the differences. It should be emphasized that one of the genes exhibited disproportional signals and was removed in this analysis. Overall, the Pearson correlation coefficients between the two readout platforms were over 0.8 with many values reaching above 0.9.

With decreased starting total RNA amounts progressively more undesired reads were identified. These were defined as mapping to regions of several true threads. However, as outlined above, they did not influence the obtained abundance levels. Nevertheless, these threads employed a large proportion of sequencing resources. Analysis of the most prevalent undesired thread species revealed that these were formed both during the threading reaction and the subsequent PCR amplification. However, the same primer concentration was utilized irrespective of total RNA amount. As the fractions of undesired reads were small when higher total RNA amounts were used, overabundance of primers with lower total RNA inputs might explain the high percentage of these unwanted species. Consequently, lowering the primer amounts, thus lessening the potential for spurious interactions, might resolve this issue.

In conclusion, TnT is a viable strategy to select transcripts of interest for downstream 454 sequencing. About 200 cell’s worth of total RNA is sufficient for analysis. At present, sequencing is rather costly why the direct array readout strategy could provide an alternative for smaller studies. For larger investigations, individual TnT reactions could be tagged and pooled prior to sequencing, allowing parallelization both with respect to sample and transcript number. Furthermore, second generation platforms with greater read numbers could improve the expression profiling. The superior dynamic range
afforded by the massive number of reads could further reduce the total RNA amount required; something that is most likely rather difficult to achieve with an array readout. The dilemma of undesired reads employing a large fraction of sequencing resources will be addressed with lower primer amounts.

6.7. Future perspectives

The papers constituting this thesis describe integration of the trinucleotide threading (TnT) multiplex amplification technique in assays investigating various facets of DNA and RNA. Initially, this amplification strategy was applied for parallel SNP genotyping. The articles presented herein supplement the inventory of TnT-incorporating approaches with expression profiling and STR analysis. TnT therefore offers a versatile means of amplification and could be implemented in protocols targeting other aspects of genomes and transcriptomes. For example, by placing the TnT probes in exon junction regions, analysis of alternative splicing can be envisioned. Moreover, it may be feasible to study several nucleic acid attributes simultaneously. For instance, given adequate readout, both expression analysis and interrogation of SNPs within transcripts can be assessed in a targeted manner in a single reaction. Such multi-feature studies could be beneficial when sample material is scarce.

The level of multiplexing represents a significant parameter for parallel amplification methods. Currently, TnT has been employed for a 147-plex allelotyping study with 126 polymorphisms producing adequate allele frequencies. The other investigations, being of a proof-of-concept nature, have entailed modest levels of multiplexing. Consequently, it would be interesting to apply TnT to a comprehensive high-plex study.

A more thorough assessment of sample input material would be advantageous. For SNP genotyping it was previously shown that 1 ng of genomic DNA, corresponding to the content of 150 cells, produced reliable data. Similarly, for expression profiling 3.5 ng of total RNA, equivalent to about 200 cells, was shown to be sufficient. However, low amounts of total RNA generated an unfavourable ratio between real threads and undesired ones. This could be attributed to an overwhelming primer amount that due to the scarcity of the template facilitates erroneous primer interactions. Fine-tuning the parameters might thus channel the maximum of the TnT approach. Accordingly, the starting material figures could most likely be further reduced enabling analysis of only a handful, or perhaps even single cells.

As TnT is strictly an amplification approach, it can be combined with numerous readout methods. In this thesis, various renditions of conventional planar arrays, capillary gel electrophoresis and massively parallel 454 sequencing have been used. These could, however, be exchanged for alternative methods. For instance, multiplex STR analysis would benefit from 454 sequencing, which generates read lengths well matched to those of microsatellite regions. In the case of gene expression analysis, the immense number of reads provided by the short read second generation platforms would be advantageous. At present, however, the novel sequencers are quite costly to operate. One strategy to use these in a cost-effective manner is to sequence barcoded and/or pooled samples, as a single run in this scenario generates information about a number of individuals.

Lastly, the scripts and algorithms associated with probe design and conversion of raw data into meaningful information could be improved and merged together to create a software suite. This would greatly facilitate experimental setup and allow for an efficient generation of results. Moreover, a sound statistical evaluation could be integrated into this suite.
No one knows what the future might bring. However, it is not too far-fetched to assume that the second generation sequencing platforms will continue their crusade into other forays of biomedical research. Moreover, the third generation platforms are knocking on the door with the first ones scheduled for launch in 2010. It is also not too implausible to imagine that ambitious efforts such as the 1000 Genomes project will generate novel insights and produce DNA or RNA signature profiles associated with certain diseases or cellular states. These will prove useful in a clinical setting and require smaller-scale methods. The approaches forming this thesis are based on interchangeable blocks where TnT represents the multiplex amplification block. This modular structure, where a sample preparation or detection module can be replaced with a superior one when it becomes available, makes TnT well suited for future applications.
ABBREVIATIONS

A  adenine
AMASE  apyrase-mediated allele-specific extension
ALL  acute lymphoblastic leukemia
AML  acute myeloid leukemia
aRNA  antisense RNA
ASE  allele-specific extension
ASH/ASOH  allele-specific oligonucleotide hybridization
ARMS  amplification refractory mutation system
ASPCR  allele-specific PCR
ATP  adenosine triphosphate
BADGE  beadsarray for the detection of gene expression
BCC  basal cell carcinoma
bp  base pairs
C  cytosine
CAGE  cap analysis gene expression
CCD  charge coupled device
CCDS  Consensus Coding Sequence
cDNA  complementary DNA
ChIP  chromatin immunoprecipitation
CNV  copy number variation
CODIS  Combined DNA Index System
cPAL  combinatorial probe-anchor ligation
DASH  dynamic allele-specific hybridization
DASL  cDNA-mediated annealing, selection, extension and ligation
dATP  deoxyadenosine triphosphate
dATPαS  deoxyadenosine α-thiotriphosphate
dCTP  deoxyctidine triphosphate
DD  differential display
ddATP  dideoxyadenosine triphosphate
ddCTP  dideoxyctidine triphosphate
ddGTP  dideoxyguanosine triphosphate
ddUTP  dideoxyuridine triphosphate
DOP-PCR  degenerate oligonucleotide-primed PCR
DNA  deoxyribonucleic acid
DNB  DNA nano-ball
dNTP  deoxynucleotide triphosphate
ds  double-stranded
ENCODE  ENCyclopedia Of DNA Elements
EST  expressed sequence tag
EtBr  ethidium bromide
FEN  flap endonuclease
FRET  fluorescence resonance energy transfer
G  guanine
Gh  giga bases
Gbp  giga base pairs
GIS  gene identification signature
GWAS  genome-wide association study
HANDS  homo-tag assisted non-dimer system
HDA  helicase-dependent amplification
HGP  Human Genome Project
HPR  Human Proteome Resource
IVT  in vitro transcription
kb  kilo bases
kbp  kilo base pairs
LAR  ligation amplification reaction
LCR  ligase chain reaction
LDR  ligase detection reaction
Parallel target selection by trinucleotide threading

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>LR-PCR</td>
<td>long-range PCR</td>
</tr>
<tr>
<td>LSO</td>
<td>locus-specific oligonucleotide</td>
</tr>
<tr>
<td>MAF</td>
<td>minor allele frequency</td>
</tr>
<tr>
<td>Mb</td>
<td>mega bases</td>
</tr>
<tr>
<td>Mbp</td>
<td>mega base pairs</td>
</tr>
<tr>
<td>MDA</td>
<td>multiple displacement amplification</td>
</tr>
<tr>
<td>MGS</td>
<td>multiplex genotyping system</td>
</tr>
<tr>
<td>mHDA</td>
<td>mesophilic helicase-dependent amplification</td>
</tr>
<tr>
<td>MIP</td>
<td>molecular inversion probe</td>
</tr>
<tr>
<td>miRNA</td>
<td>micro RNA</td>
</tr>
<tr>
<td>MLPA</td>
<td>multiplex ligation-dependent probe amplification</td>
</tr>
<tr>
<td>MME-PCR</td>
<td>multiplex microarray-enhanced PCR</td>
</tr>
<tr>
<td>MPSS</td>
<td>massive parallel signature sequencing</td>
</tr>
<tr>
<td>mRNA</td>
<td>messenger RNA</td>
</tr>
<tr>
<td>MS</td>
<td>mass spectrometry</td>
</tr>
<tr>
<td>MS-PCR</td>
<td>multiplex sequencing of paired-end ditags</td>
</tr>
<tr>
<td>msAb</td>
<td>mono-specific antibody</td>
</tr>
<tr>
<td>NSR</td>
<td>not-so-random</td>
</tr>
<tr>
<td>OLA</td>
<td>oligonucleotide ligation assay</td>
</tr>
<tr>
<td>PAGE</td>
<td>polyacrylamide gel electrophoresis</td>
</tr>
<tr>
<td>PCR</td>
<td>polymerase chain reaction</td>
</tr>
<tr>
<td>PEP</td>
<td>primer-extension preamplification</td>
</tr>
<tr>
<td>PET</td>
<td>paired-end ditag</td>
</tr>
<tr>
<td>PMAGE</td>
<td>polony multiplex analysis of gene expression</td>
</tr>
<tr>
<td>PP</td>
<td>inorganic phosphate</td>
</tr>
<tr>
<td>Pr:ASE</td>
<td>protease-mediated allele-specific extension</td>
</tr>
<tr>
<td>Pr:EST</td>
<td>protein epitope signature tag</td>
</tr>
<tr>
<td>RAP-PCR</td>
<td>RNA arbitrarily primed PCR</td>
</tr>
<tr>
<td>RASL</td>
<td>RNA-mediated annealing, ligation and selection</td>
</tr>
<tr>
<td>RCA</td>
<td>rolling circle amplification</td>
</tr>
<tr>
<td>RFLP</td>
<td>restriction fragment length polymorphism</td>
</tr>
<tr>
<td>RNA</td>
<td>ribonucleic acid</td>
</tr>
<tr>
<td>RNA-Seq</td>
<td>RNA sequencing</td>
</tr>
<tr>
<td>tRNA</td>
<td>ribosomal RNA</td>
</tr>
<tr>
<td>RT-PCR</td>
<td>real-time PCR</td>
</tr>
<tr>
<td>SAGE</td>
<td>serial analysis of gene expression</td>
</tr>
<tr>
<td>SBE</td>
<td>single-base extension</td>
</tr>
<tr>
<td>SBH</td>
<td>sequencing-by-hybridization</td>
</tr>
<tr>
<td>SBL</td>
<td>sequencing-by-ligation</td>
</tr>
<tr>
<td>SBS</td>
<td>sequencing-by-synthesis</td>
</tr>
<tr>
<td>SDA</td>
<td>strand displacement amplification</td>
</tr>
<tr>
<td>siRNA</td>
<td>small interfering RNA</td>
</tr>
<tr>
<td>SMRT</td>
<td>single-molecule real-time</td>
</tr>
<tr>
<td>SNP</td>
<td>single nucleotide polymorphism</td>
</tr>
<tr>
<td>snRNA</td>
<td>small nuclear RNA</td>
</tr>
<tr>
<td>SOLiD</td>
<td>sequencing by oligonucleotide ligation and detection</td>
</tr>
<tr>
<td>STM</td>
<td>scanning tunneling microscopy</td>
</tr>
<tr>
<td>STR</td>
<td>short tandem repeat</td>
</tr>
<tr>
<td>T</td>
<td>thymine</td>
</tr>
<tr>
<td>Tag</td>
<td>Thermus aquaticus</td>
</tr>
<tr>
<td>tHDA</td>
<td>thermophilic helicase-dependent amplification</td>
</tr>
<tr>
<td>TMA</td>
<td>tissue microarray</td>
</tr>
<tr>
<td>TMAC</td>
<td>tetramethylammonium chloride</td>
</tr>
<tr>
<td>TnT</td>
<td>trinucleotide threading</td>
</tr>
<tr>
<td>tRNA</td>
<td>transfer RNA</td>
</tr>
<tr>
<td>tSMS</td>
<td>true single molecule sequencing</td>
</tr>
<tr>
<td>TSS</td>
<td>transcriptional start site</td>
</tr>
<tr>
<td>WGA</td>
<td>whole-genome amplification</td>
</tr>
<tr>
<td>WGG</td>
<td>whole-genome genotyping</td>
</tr>
<tr>
<td>WGSA</td>
<td>whole-genome sampling analysis</td>
</tr>
<tr>
<td>ZMW</td>
<td>zero-mode waveguide</td>
</tr>
</tbody>
</table>
ACKNOWLEDGEMENTS


Mr. Bipster, Mårten, vi hade stora visioner med looparna, men istället ställde de till en massa bekymmer under exjobbet. Det var ett öde att handleda dig och kul att du fortsatte och började doktorera. Nu har looparna ersatts med en drös andra projekt, bland annat med TnT, och jag är övertygad om att den hamnat i goda händer. Och du, vi måste ta upp fredagscykelturena igen.

Christine, tack för STR projektet, det blev ett pek till slut. Simone, thanks for the colorectal cancer project. It was a lot of fun, and a lot of MBS runs, and I am glad we have kept in touch after the project was completed. Och du, vi måste ta upp fredagscykelturena igen.


Tack för skönt fika- och lunchsällskap de gånger jag behövtt rensa hjärnan lite innan det har varit dags att ta itu med nya försök. Mattias, eller Mos som jag kom att kalla dig, skönt att kunna diskutera bland mycket annat, Bajens uppgång och fall, musik, film, löpning och alla konstiga idéer vi har haft. Det blev dessutom en och annat ‘epic journey’ som man har kunnat prata om dagar, veckor, månader, ja, till och med år efteråt. Den

Tack till KTH för finansiering.

Det känns som att alla personer på Plan 3 på ett eller annat sätt bidragit till att jag trivts så bra. Tack!


Na koncu, chciałbym podziękować rodzicom za wszystko, za pomoc ze wszystkim i za to ze zawsze we mnie wierzyli. Dziękuję! Ett stort tack även till världens bästa bror. Kul att även du är på väg att bli en doktor. Efter det ser vi till att erövra världen! Tack för allt!

Tack även till Dig som läst denna avhandling. Om du känner dig bortglömd bland denna tack-kavalkad, så ber jag om ursäkt. En stressad hjärna som med alla mått försöker få fram en färdig avhandling till en deadline som kryper närmare med oroväckande fart är inte optimalt lämpad för att minnas alla som förtjänar att tackas.
Parallel target selection by trinucleotide threading
REFERENCES

Parallel target selection by trinucleotide threading


139. Barany, F. The ligase chain reaction in a PCR world. PCR methods and applications 1, 5-16 (1991).
Parallel target selection by trinucleotide threading

157. Miesfeld, R., Krystal, M. & Arnheim, N. A member of a new repeated sequence family which is conserved throughout eucaryotic evolution is found between the human delta and beta globin genes. Nucleic acids research 9, 5931-5947 (1981).
194. Affymetrix GeneChip CustomExpress Array Program: Data Sheet.
195. Affymetrix Genome -Wide Human SNP Array 6.0: Data Sheet.

85
Parallel target selection by trinucleotide threading

218. Affymetrix Custom and Catalog Genotyping with the Affymetrix GeneChip Scanner 3000 Targeted Genotyping System.


Parallel target selection by trinucleotide threading


Parallel target selection by trinucleotide threading


