Arrayed Identification of DNA Signatures

Max Käller

Royal Institute of Technology
Department of Biotechnology

Stockholm, 2005
ABSTRACT

In this thesis techniques are presented that aim to determine individual DNA signatures by controlled synthesis of nucleic acid multimers. Allele-specific extension reactions with an improved specificity were applied for several genomic purposes. Since DNA polymerases extend some mismatched 3’-end primers, an improved specificity is a concern. This has been possible by exploiting the faster extension of matched primers and applying the enzymes apyrase or Proteinase K. The findings were applied to methods for resequencing and viral and single nucleotide polymorphism (SNP) genotyping.

P53 mutation is the most frequent event in human cancers. Here, a model system for resequencing of 15 bps in p53 based on apyrase-mediated allele-specific extension (AMASE) is described, investigated and evaluated (Paper I). A microarray format with fluorescence detection was used. On each array, four oligonucleotides were printed for each base to resequence. Target PCR products were hybridized and an AMASE-reaction performed in situ to distinguish which of the printed oligonucleotides matched the target. The results showed that without the inclusion of apyrase, the resulting sequence was unreadable. The results open the possibilities for developing large-scale resequencing tools.

The presence of certain types of human papillomaviruses (HPV) transforms normal cells into cervical cancer cells. Thus, HPV type determination is clinically important. Also, multiple HPV infections are common but difficult to distinguish. Therefore, a genotyping platform based on competitive hybridization and AMASE is described, used on clinical sample material and evaluated by comparison to Sanger DNA sequencing (Papers II and III). A flexible tag-microarray was used for detection and the two levels of discrimination gave a high level of specificity. Easy identification of multiple infections was possible which provides new opportunities to investigate the importance of multiply infected samples.

To achieve highly multiplexed allele-specific extension reactions, large numbers of primers will be employed and lead to spurious hybridizations. Papers IV to VI focus on an alternative approach to control oligomerization by using protease mediated allele-specific extension (PrASE). In order to maintain stringency at higher temperatures, Proteinase K, was used instead of apyrase, leading to DNA polymerase degradation and preventing unspecific extensions. An automated assay with tag-array detection for SNP genotyping was established. First PrASE was introduced and characterized (Paper IV), then used for genotyping of 10 SNPs in 442 samples (Paper V). A 99.8 % concordance to pyrosequencing was found. PrASE is a flexible tool for association studies and the results indicate an improved assay conversion rate as compared to plain allele-specific extension.

The highly polymorphic melanocortin-1 receptor gene (MC1R) is involved in melanogenesis. Twenty-one MC1R variants were genotyped with PrASE since variants in the gene have been associated to an increased risk of developing melanoma. A pilot study was performed to establish the assay (Paper VI) and subsequently a larger study was executed to investigate allele frequencies in the Swedish population (Paper VII). The case and control groups consisted of 1001 and 721 samples respectively. A two to sevenfold increased risk of developing melanoma was observed for carriers of variants.

Keywords: apyrase, allele-specific extension, competitive hybridization, DNA sequencing, genotyping, human papillomavirus (HPV), MC1R, microarray, mutation, p53, protease, single nucleotide polymorphism (SNP), tag-array.
LIST OF PUBLICATIONS

This thesis is based on the papers listed below which will be referred to by their roman numerals.


# Table of Contents

## Introduction to DNA
- The Human Genome Project
- Genomic Variation
- The HapMap Project

## Identification of DNA Signatures
- Microarray Detection
- Suspension Arrays

### Sequencing Techniques
- Sanger DNA Sequencing
- Pyrosequencing
- Sequencing by Hybridization

### SNP Genotyping Techniques
- Hybridization-Based Techniques
- Oligonucleotide Ligation Assays
- Single Base Extension
- Allele-Specific Extension

### Whole Genome Analysis Techniques
- Multiplex Polony Sequencing
- The 454 Genome Sequencer 20 System
- Cloned Single-Molecule Arrays
- Molecular Inversion Probes
- Allele-Specific Oligonucleotide Hybridization
- GoldenGate and Infinium Assays

## Summary

## Investigations
- Apyrase-Mediated Allele-Specific Extension
- Resequencing of the p53 gene (Paper I)
  - The p53 Gene
  - Brief Assay Description
  - Results
Genotyping Human Papillomaviruses (Papers II and III) 35
Microbial Genotyping 35
Human Papillomavirus 36
Methods for HPV Type Determination 36
MUCH-LAMASE 38
Type-Specific Primer Design 39
Samples and Results of the Initial Proof-of-Concept Study (Paper II) 39
Samples and Results of the Subsequent Pilot Study (Paper III) 40

Genotyping of SNPs and Common Variations (Papers IV-VII) 41
Protease-Mediated Allele-Specific Extension (Paper IV) 42
Accuracy and Robustness of PrASE as Compared to Pyrosequencing (Paper V) 43

Screen for Variants that Increase Melanoma Risks 44
The Genetic Basis of Hereditary Malignant Melanoma 44
PrASE-Based Analysis of Melanoma Variants (Paper VI) 46
MC1R Variation in the Swedish Population (Paper VII) 46

FUTURE PERSPECTIVES 49

ABBREVIATIONS 51

ACKNOWLEDGEMENTS 53

REFERENCES 55

ORIGINAL PAPERS (Appendices I-VII)
INTRODUCTION TO DNA

The design for every living organism is contained in each of the respective organisms' cells. This information is stored in the form of deoxyribonucleic acid, DNA, and is sufficient to create the organism (Avery et al., 1944). The diploid human genome is organized into 23 pairs of chromosomes (46 single molecules), and one of each pair is inherited from each parent. At precise times in the cell cycle the chromosomes are coiled in a specific way and can be identified by their banding patterns (Figure 1). When the DNA is about to be used, the chromosome becomes untwined so that the enzymes in the cell machinery can access the appropriate parts and “read” the sequence. The double-stranded-helical structure of DNA was first described on the basis of x-ray diffraction patterns in 1953 (Franklin and Gosling, 1953; Watson and Crick, 1953). The revolutionary findings by Franklin, Watson and Crick demonstrated that the two strands of DNA are complementary, meaning that the base adenosine (abbreviated A) is always opposite to the base thymine (T) while cytosine (C) always sits opposite to guanine (G). By reading the sequence of these four bases, the cellular machinery translates the genetic information into proteins (Khorana, 1965; Nirenberg et al., 1965) that form the cells’ structures and control their functions.
Figure 1. Genomic organization. During mitosis the genetic material is condensed into 46 chromosomes as shown (top). Each chromosome contains on average 130 million base pairs packed in an orderly fashion (middle). At the lowest level of packing the two opposing strands of DNA adopt a double-helical structure (bottom). An individual's unique sequence can be read and variations between the chromosomes identified at single base-level (bottom right).

The Human Genome Project

The once-incredible task of sequencing the entire human genome of about 3.2 billion base pairs was first seriously considered 20 years ago and the Human Genome Project was initiated in 1990. The method used for sequencing DNA was invented in 1977 (Sanger et al., 1977), but with the sequencing rates attainable in 1990 the project would not have been finished today (Collins et al., 2003). However, the project plans envisaged that significant technological breakthroughs would be both needed and made. Accordingly, the major part of the project was spent inventing and refining new technologies and the actual production of data only took place during the final few years. This last phase was spurred by a private initiative from Celera Genomics, a company holding a number of innovative strategies (Venter et al., 1998). The Human Genome Organization and Celera presented the draft sequence of the human genome simultaneously in 2000 (Lander et al., 2001; Venter et al., 2001) and later the complete sequence two years ahead of schedule (The International Human Genome Sequencing...
Max Käller

Cons ortium, 2004). Since then the established sequencing pipelines have found new targets and currently over 300 complete genomes have been sequenced (www.genomesonline.org). The latest addition is the draft sequence of Clint the chimpanzee, showing that chimpanzees differ from humans in 35 million single base substitutions and 5 million insertions and deletions (The Chimpanzee Sequencing and Analysis Consortium, 2005).

Genomic Variation

Knowledge of the genetic sequence promises to facilitate the identification of disease-causing genes and thereby improve human health. One way of using the information is by looking for DNA signatures that differ between our genomes. The genomes of any two individuals have up to 99.9% identity, and the majority of the differences are known as single nucleotide polymorphisms (SNPs) which are defined as single base changes where both alleles (variants) are present in at least 1% of a given population. To date, there are more than 10 million SNPs in dbSNP and it has been estimated that there are 11 million SNPs in total (Kruglyak and Nickerson, 2001). This means that two randomly chosen human haploid genomes will differ at one in every 1331 bases, on average (Sachidanandam et al., 2001). These differences are thought to influence most simple phenotypic differences such as height (Hirschhorn et al., 2001), eye color (Rebbeck et al., 2002) and hair color (Valverde et al., 1995), but they may also be causally linked to disease, as in the case of sickle cell anemia (Pauling et al., 1949; Ingram, 1957), or disease susceptibility. Other applications of genotyping are in pharmacogenetics, since genetic variations can explain differences in rates of drug metabolism and adverse reactions (Lai et al., 2002; Rogers et al., 2002).

Most attention, to date, has been paid to SNPs located in coding sequences, which can cause variations in the structure and, thus, function of the corresponding proteins. However, SNPs located in promoters, enhancers and 3'- and 5'-untranslated sequences are also of great potential interest, since they may influence the expression of the proteins (Knight et al., 2003; Pastinen and Hudson, 2004; Pastinen et al., 2004; Lindstrom et al., 2005). Similarly, SNPs located at splice-sites may influence splicing and thus also alter the structure and expression of the affected proteins.
The HapMap Project

One of the most important applications of genomic variation analysis is for mapping genes that underlie disease. There are several strategies for gene mapping, but the main options are to perform linkage or association studies by a candidate gene or genome-wide approach (Ardlie et al., 2002; Hirschhorn and Daly, 2005). However, finding a disease allele through genetic markers requires them to be co-inherited. The concept of non-random association of two markers, linkage disequilibrium (LD), is central in linkage and association studies. Linkage is complete between newly formed mutations, but LD is lost over time due to mutations, recombinations and demographic processes, such as natural selection and migration (Ardlie et al., 2002).

Accordingly, in linkage studies family material is used and the objective is to find a marker that segregates with the disease or trait in the pedigrees. Linkage studies are therefore best for finding monogenic Mendelian inheritance patterns of rare alleles with relatively high penetrance (Hirschhorn and Daly, 2005). In fact, in a survey of 923 disease-associated genes that have already been discovered, 97% were found to be monogenic (Jimenez-Sanchez et al., 2001). Variants that contribute to disease susceptibility have also been found, for example susceptibility to inflammatory bowel disease (Hugot et al., 2001; Ogura et al., 2001; Rioux et al., 2001), schizophrenia (Stefansson et al., 2002) and type 1 diabetes (Nistico et al., 1996). However, the success of linkage studies has been limited and few true relationships between disease and specific genes have been found (Altshuler et al., 2001). For many diseases the genes found only partially explain the heritability, suggesting the involvement of other causal genes.

Instead, most common diseases are believed to be caused by complex interactions between multiple common genetic and environmental factors (Wang et al., 2005). To find genes involved in such interactions, whole-genome association studies have been proposed (Risch and Merikangas, 1996). Ambitious goals and very large sample groups can make the whole-genome approach prohibitively expensive. In many cases a candidate-gene approach can be much more economical. This approach is also usually necessary to fine-map regions found in linkage-studies or whole-genome association studies. In association studies case-control sample material is usually used to find alleles that are more common in one of the groups than the other. Association studies are generally better at finding connections to common alleles.
that give a modest increase in relative risk than linkage studies (Ardlie et al., 2002). Consequently, this approach has been successfully used to identify alleles that confer a 2-fold or less relative risk for several complex diseases (Altshuler et al., 2000; Gloyn et al., 2001; Ioannidis et al., 2001; Giallourakis et al., 2003; Lohmueller et al., 2003; Negoro et al., 2003; Stefansson et al., 2003; Ueda et al., 2003; Wang et al., 2005). Calculations of the size of the sample and control groups required can be made from the chosen detectable odds ratio, expected allele frequency and the established significance level (Zondervan and Cardon, 2004).

Several large-scale SNP mapping projects have suggested that the genome is composed of a limited amount of haplotype blocks or “bins” (Daly et al., 2001; Johnson et al., 2001; Patil et al., 2001; Hinds et al., 2005). These suggestions led to the initiation of the HapMap project, an international effort to find the haplotype blocks in four populations (The International HapMap Consortium, 2003). The first phase of genotyping 600,000 SNPs in 269 samples has already been completed (The International HapMap Consortium, 2005) and a second gap-fill and resequencing phase is currently being undertaken. However, 1.6 million SNPs have already been genotyped by the private initiative of Perlegen (Hinds et al., 2005). Estimates from 71 samples indicate that block sizes are approximately 10, 20 and 25 kb in the African, European and Han Chinese populations, respectively. In addition, the minimal numbers of SNPs needed to distinguish common haplotypes (tag-SNPs) are estimated to be approximately 500,000, 300,000, and 250,000 in these populations, respectively (The International HapMap Consortium, 2005; Hinds et al., 2005). However, despite the impressive data produced by the HapMap project and Perlegen, there are skeptical voices. The block structure will be useful for choosing tag-SNPs required to maximize information content, but it may not be applicable beyond the chosen subpopulations (Lai et al., 2002; Tergiiler et al., 2002). Furthermore, since the map is based on common SNPs, rare and often causative SNPs might be missed entirely.
IDENTIFICATION OF DNA SIGNATURES

The introduction of this thesis reviews the main technological alternatives for unraveling genetic diversity. A wide range of approaches and systems can be used for assessing these variations, but the ideal choice for particular variations depends on what is known of the sequences surrounding them. To examine a sequence that is completely unknown before the study, \textit{de novo} sequencing strategies have to be used. However, to find a mutation in a known sequence, a re-sequencing technique can be used to save time and expense. SNPs and variants in pre-defined positions may be investigated with genotyping approaches. The borderline between these alternatives is, however, quite thin and some of them can be used for various purposes. In addition, the multitude of existing approaches indicates that no perfect technique that can be used in all situations at a reasonable cost has yet been invented.

The assays described in this chapter all require target sequences of the genomic DNA to be amplified in order to increase their abundance to detectable levels and to avoid complications arising from the complexity of genomic DNA (Syvanen, 2005). In addition, molecular reactions for labeling and separation of the two possible alleles under study and an appropriate detection system are required (Kwok, 2001; Syvanen, 2001; Ahmadian and Lundeborg, 2002). However, before describing sequencing, SNP genotyping and whole genome approaches, two detection systems will be presented that are generally applicable to many molecular reaction methods.
Microarray Detection

Several microarray platforms based on similar principles are available for analyzing DNA and RNA. DNA microarrays separate pools of different single-stranded DNA or RNA species by hybridization to a number of complementary and specific DNA sequences immobilized on a solid support (Southern, 1975). The sample sequences are usually fluorescently labeled and can be quantified in a scanner by exciting the labels with a laser and measuring the emitted light. The signals from individual DNA sequences can be identified according to the position of the signal.

There are three main types of DNA microarrays: arrays of microbeads, arrays made by in situ synthesis, and those printed robotically on an appropriate surface. The first DNA microarrays used the in situ delivery method for sequencing by hybridization (SBH, which is covered later) (Southern et al., 1992). Very high feature densities can be achieved using photolithographic methods developed for the integrated circuits industry, as pioneered by Affymetrix (Fodor et al., 1991; Pease et al., 1994). The surface is exposed through a photolithographic mask to selectively remove light-reactive protective groups. In this way a growing oligonucleotide chain is created in a stepwise fashion. The yield of each step is only 95%, which limits the useful oligonucleotide length to 25 bases. However the current feature size (5 µm) allows many probes to be used for each target to compensate for the low specificity. An alternative approach to photolithographic masking, which allows more flexible synthesis, is to use micro-sized mirrors (Singh-Gasson et al., 1999). Robotic printing of cDNA was pioneered at Stanford University for expression profiling (Schena et al., 1995; DeRisi et al., 1996). Today numerous robotic systems are available for direct-contact printing (Schena et al., 1995) and inkjet printing (Hughes et al., 2001) of single-stranded or double-stranded DNA onto glass slides at feature sizes down to ~150 µm. The third type of microarrays are the bead-arrays used in the platforms developed by Illumina (Leamon et al., 2003) and 454 Life Sciences (Margulies et al., 2005). These systems will be further described in a later chapter, however the beads involved, and thus the feature sizes, differ substantially (1.4 and 44 µm, respectively). Versatile tools that have been developed are microarrays of universal tag sequences or barcode sequences that can be used as capture probes for primers used in reactions performed in solution (Shoemaker et al., 1996; Hirschhorn et al., 2000).
Suspension Arrays

The cytometry-based suspension array platform from Luminex has been used for a multitude of purposes (Fulton et al., 1997). The system utilizes two fluorophores (one that emits red and one that emits infra-red light) coupled to microbeads and the amounts of fluorescent molecules on each bead are arranged in 10 discrete levels. The two fluorophores can be excited by a single laser and individually detected. Thus, 100 different bead types are available and can be analyzed with a flow cytometer. These beads can be used in the same way as a signature tag-microarray with 100 tag-sequences to separate a multiplexed reaction (Cai et al., 2000; Chen et al., 2000; Iannone et al., 2000). A second laser is used to analyze the result of a reaction on each identified bead.

Sequencing Techniques

Sanger DNA sequencing is still the gold standard for identifying the exact nature of DNA sequences, and the method that novel methods are compared to today. However, there are several methods for de novo sequencing, including Sanger's and Maxam-Gilbert's methods, sequencing by hybridization (SBH) and sequencing by synthesis (SBS) techniques such as pyrosequencing and its miniaturized version provided by 454 Life Sciences. 454 Life Sciences system, and other highly miniaturized parallel sequencing systems allowing whole-genome sequencing of bacterial genomes, will be described in a later chapter.

Sanger DNA Sequencing

DNA sequencing posed huge challenges to early researchers due to the enormous complexity of even relatively short sequences. However, two pioneering sequencing techniques, based on radically different solutions to break down large sequences into manageable portions, were published in the same year: the Sanger method of chain termination (Sanger et al., 1977) and Maxam and Gilbert's method of chain degradation (Maxam and Gilbert, 1977). Both techniques generate fragments of the queried sequence with all possible nucleotide polymer lengths present. These can then be separated in the right order to acquire the nucleotide sequence. The degradation step of the Maxam-Gilbert method involves hazardous chemicals and has therefore been much less commonly used than the Sanger alternative.
Arrayed Identification of DNA Signatures

Sanger DNA sequencing works by extending a DNA primer hybridized to the query single-stranded DNA template. A small proportion of the dNTPs used for this purpose are replaced by ddNTPs (dideoxyribonucleotide phosphates) and the polymerase extends the primer with dNTPs until it incorporates a ddNTP by chance, thus terminating the extension. In this way fragments of all possible lengths are generated. The first detection method involved the use of radiolabeled primers in four reactions, each with only one terminating ddNTP. The fragments were then electrophoretically separated according to size in four lanes of a polyacrylamide gel. The sequence could then be simply "read" from the sizes of the fragments in each lane.

Although the basic principle of chain termination has remained the same, the Sanger method has been significantly improved since it was first introduced. The introduction of thermostable DNA polymerases has enabled the fragments to be generated in a PCR-like fashion, but with linear amplification; a process termed cycle sequencing (Finnis et al., 1988; Carothers et al., 1989). In addition, four-color fluorescent labeling of the ddNTPs (dye-terminators) has enabled the automated detection of all four dNTPs in a single lane (Smith et al., 1986; Prober et al., 1987). Capillary electrophoresis and parallelization of 96 to 384 reactions have increased throughput significantly.

The long-lasting success of the Sanger DNA sequencing method, despite competition from many alternative approaches, may be due to two unique advantages it provides: relatively long read lengths of 500 to 1000 bases, and details of the exact nature of every alteration therein. Nevertheless, although the sequencing cost per base has been dramatically reduced since the early days of this method, from 10 USD in 1990 to 5 cents in 2003 per finished base (Collins et al., 2003), for visionary applications such as personal genome analysis, Sanger sequencing is far too expensive.

How much scope remains for further improvement is at best uncertain, but some groups are working on miniaturizing the detection method. The idea is to borrow technologies developed for microfabrication of electronic circuits to develop microelectrophoretic sequencing techniques (Koutny, 2000; Emrich et al., 2002; Paegel et al., 2003).
Pyrosequencing

A novel method for DNA analysis based on the principle of sequencing-by-synthesis was presented in 1998 (Ronaghi et al., 1998). The method, denoted pyrosequencing, was based on several earlier discoveries and involves the following steps. First, the four dNTPs are added one at a time to a reaction mixture containing a primed single-stranded template, either in a specific order determined by the investigated sequence (targeted dispensation) or in a cyclic manner. When a dNTP is complementary to the first base after the primer in the target template, the dNTP is incorporated and a cascade of enzymes generates light. The dNTPs that are not complementary will not be incorporated and do not generate light. The enzymatic chain starts when the polymerase (the Klenow fragment of DNA polymerase I from *E. coli*) incorporates a dNTP and cleaves off the diphosphate group, denoted PPi, from the dNTP. This PPi is then used together with adenosine phosphosulphate (APS) as substrate by a second enzyme, called ATP sulphurylase, generating ATP. Another enzyme called luciferase uses the ATP to emit light that can be detected with a CCD camera. The peak height of the light signal is directly proportional to the number of incorporated nucleotides. When an added dNTP is not complementary to the template, the last enzyme, called apyrase, makes it unusable for the polymerase by cleaving off the two phosphate groups one at a time. The apyrase also degrades ATP to stop the reaction, and is crucial to avoid cumbersome washing steps between nucleotide additions. This complex enzymatic cascade allows real-time monitoring of the light produced as the synthesis proceeds as follows with catalyzing enzymes in parentheses:

\[(\text{DNA})_n + \text{dNTP} \rightarrow (\text{DNA})_{n+1} + \text{PPi} \quad (\text{Polymerase})\]

\[\text{PPi} + \text{APS} \rightarrow \text{ATP} + \text{SO}_4^{2-} \quad (\text{ATP Sulphurylase})\]

\[\text{Luciferase} + \text{D-luciferin} + \text{ATP} \rightarrow \text{Luciferase-luciferin-AMP} + \text{PPi}\]

\[\text{Luciferase-luciferin-AMP} + \text{O}_2 \rightarrow \text{Luciferase} + \text{oxyluciferin} + \text{AMP} + \text{CO}_2 + \text{hv}\]

\[\text{ATP} \rightarrow \text{AMP} + 2\text{Pi} \quad (\text{Apyrase})\]

\[\text{dNTP} \rightarrow \text{dNMP} + 2\text{Pi} \quad (\text{Apyrase})\]
A major drawback with the pyrosequencing technique is that it only allows read-lengths of 20-40 bp with standard kits. However, some improvements have increased the possible read-length. For instance, the use of dATPαS instead of dATP reduces non-specific signals since dATP is a substrate for luciferase, but dATPαS is not (Ronaghi et al., 1996). Thus, by only using one of the two isoforms of dATPαS the read-lengths can be significantly increased, in some cases up to 100 bases (Gharizadeh et al., 2002). Several problems, such as positive or negative frameshifts and the generation of background sequences, may also occur in pyrosequencing. Background sequences are mostly due to mispriming of the sequencing primer, while frameshifts occur when some of the templates have more or fewer incorporated nucleotides than the majority and thus give background peaks. Positive frameshifts are due to insufficient activity of the apyrase caused by accumulated amounts of reaction products. This causes small residual proportions of dNTPs to be incorporated ahead of time. Negative frameshifts have the opposite effect and are due to insufficient activity of the Klenow polymerase. They can occur after a sequence of 4-5 or more identical nucleotides and cause incomplete incorporation of nucleotides. Finally, including single-strand binding protein (SSB) can increase read-lengths by 12 bases on average (Ehn et al., 2002), probably because it reduces the extent of unspecific binding at low temperatures (25 ºC). This measure has been incorporated into standard procedures in some laboratories, but the enzyme is not included in the commercial kits.

**Sequencing by Hybridization**

The inherent complementarity between DNA strands has been exploited in many different hybridization assays and for various purposes. Two examples are Southern blotting (its first application) for identifying target DNA sequences (Southern, 1975), and allele-specific oligonucleotide hybridization (ASOH) for discriminating between two alleles differing only in a single base position (Wallace et al., 1979). More pertinent for this thesis, hybridization also has potential utility for sequencing. De novo sequencing by hybridization (SBH) involves the use of all possible combinations of oligonucleotides of a given length; 65536 (4^8) if octamers are used, for example (Lysov Iu et al., 1988; Drmanac et al., 1989). The oligonucleotides are immobilized in an array format and the labeled target is hybridized. The sequence is then
reconstructed with the aid of a computer from the sequences of the overlapping oligonucleotides that give a positive signal.

However, despite its conceptual attractiveness, a number of problems are associated with SBH in practice. Most importantly, small differences in duplex stability between similar probes can give false positive results and repetitive sequences may be unreadable (Tibanyenda et al., 1984; Marziali and Akeson, 2001). To compensate for these drawbacks, up to 56 probes for each interrogated base are used (Kennedy et al., 2003). Due to the large numbers of probes involved the high-density microarray platforms supplied by Affymetrix are particularly suitable for such hybridizations (Chee et al., 1996; Maitra et al., 2004). Nevertheless, in an effort to resequence a whole chromosome, large regions were found to be inaccessible (Patil et al., 2001). However, for purposes other than de novo sequencing, SBH can be useful and it has been commercialized for applications such as resequencing of HIV, p53 and mitochondrial DNA (Kozal et al., 1996; Drmanac et al., 1998; Maitra et al., 2004).

SNP Genotyping Techniques

The methods described herein as sequencing technologies can, of course, also be used to establish the identity of single bases. However, for high-throughput SNP genotyping many other techniques have been developed that only examine the SNP position. Several hybridization-based methods have been developed for this purpose, such as ASOH, DASH, TaqMan and molecular beacons. There are also several methods that use the enzymatic recognition of DNA ligases or polymerases to increase specificity and to multiplex the detection of many SNPs in a single reaction. In addition, several methods based on gel mobility shifts have been used, such as single stranded length polymorphism analysis (Orita et al., 1989; Humphries et al., 1997) and heteroduplex analysis (White et al., 1992; Glavac and Dean, 1995), albeit with rather low throughputs to date. Their throughput can be increased using automated denaturing high-performance liquid chromatography detection systems (Choy et al., 1999; Gross et al., 1999), but these methods will not be further discussed here.
Hybridization-Based Techniques

Several other approaches rely on hybridization as the main mechanism for discriminating between alleles. In dynamic allele-specific hybridization (DASH) the stability of a template: oligonucleotide duplex is monitored as temperature is increased, eventually causing denaturation (Howell et al., 1999). The oligonucleotide used is designed to completely match one of the alleles and form a centrally located mismatch to the other, resulting in a difference in melting temperature depending on the target sequence. Fluorescently labeled oligonucleotides or an intercalating dye that fluoresces only when bound to duplex DNA can be used for detection. The derivative of the fluorescence is plotted and gives peaks at different melting temperatures for the different genotypes. To increase throughput, DASH has been used in various formats but its multiplexing capacity has been limited to two assays per well (Jobe et al., 2003).

TaqMan (or the 5'-nuclease assay) is a method for monitoring DNA amplification in real-time and can be used for DNA quantification (Holland et al., 1991; Lee et al., 1993; Livak et al., 1995). An allele-specific oligonucleotide is used that has a donor and an acceptor label at either end of the probe. When the labels are close to each other fluorescence resonance energy transfer (FRET) of the fluorescent light from the donor is quenched by the acceptor/quencher. This happens when the probe has hybridized to the amplified target after the denaturation step. In the subsequent extension phase of the PCR, the Taq DNA polymerase reaches the probe, its 5'-nuclease activity degrades the probe and the labels are separated, which increases the fluorescence.

Molecular beacons work similarly to TaqMan probes, but in this case the probe is designed to form a hairpin in solution (Tyagi and Kramer, 1996; Tyagi et al., 1998). This brings the labels close enough together to cause quenching of the signal. When the probe binds to the newly formed template, the hairpin breaks and the increase in distance between the labels causes the fluorescence to increase. A useful property of the TaqMan and molecular beacon assays is that the amplification can be monitored and extrapolated to calculate the amount of DNA present from the start of the reaction.
Oligonucleotide Ligation Assays

The use of DNA ligases for SNP genotyping was first called the oligonucleotide ligation assay (OLA) and was described in 1988 (Alves and Carr, 1988; Landegren et al., 1988). In the basic reaction, a PCR product is used and two oligonucleotides are hybridized to it directly adjacent to each other. One of these oligonucleotides is complementary to the sequence directly upstream of the SNP, while the 3'-base of the other oligonucleotide is complementary to the SNP. The latter is used in two variants, matching the two possible target alleles, with fluorescent 5'-end labels to distinguish them. The DNA ligase has the intrinsic ability to recognize matched 3'-ends and thus only joins the two adjacent oligonucleotides if they are matched. A biotin label on the other, general oligonucleotide is used to capture the probes on streptavidin-coated beads and thereby enable detection of any ligated oligonucleotides.

Thermostable ligases have also been used for amplification assays similar to PCR (Barany, 1991). In ligase chain reactions (LCR) the oligonucleotides are designed as above, for both strands of the DNA, and a DNA ligase is used instead of a polymerase. Only matched 3'-ends will be ligated and thus amplified. The reaction can be setup for amplification with oligonucleotides for all variants in the same tube or for genotyping with the oligonucleotides for each variant in separate tubes.

OLA has been further refined by using one probe containing the sequences of the two OLA probes as above, joined together with a linker, denoted a padlock probe (Nilsson et al., 1994). In this way the specificity of the assay is increased since the oligonucleotides have a lower risk of cross-hybridizing. The amplification can then take place after the genotyping reaction by including general PCR primer sequences in the linker sequence (Baner et al., 2003) or by rolling circle amplification (Baner et al., 1998; Lizardi et al., 1998). The padlock probes can also be used in situ for single molecule detection on histological slides (Larsson et al., 2004). Recently the multiplexing capacity of this kind of approach has been increased by combining the ligation step with a single base extension step, as discussed in the following chapter.

A new OLA-based assay called SNPlex has been introduced by ABI that has the ability to detect 48 SNPs in one reaction and with detection based on electrophoretic separation in capillaries (www.appliedbiosystems.com).
Single Base Extension

The current state of the art single base extension (SBE) or minisequencing assay uses incorporation of a ddNTP complementary to the SNP nucleotide in a PCR-amplified fragment (Syvanen et al., 1990; Lindroos et al., 2002). The four ddNTPs have different fluorophore labels to minimize the number of reactions needed. The extension reaction can be cycled, much like the Sanger sequencing reaction, to improve sensitivity. Several assay platforms have been devised, such as matrix-assisted laser desorption ionization-time of flight mass spectroscopy (MALDI-TOF MS) (Tang et al., 1999) and a microarray detection system in which the SBE-primes have 5'-located unique signature tags complementary to spotted oligonucleotides (Fan et al., 2000).

Allele-Specific Extension

In the late 1980's, several groups published papers describing similar techniques but with different names like ARMS, ASA, PASA and ASPCR: amplification refractory mutation system (Newton et al., 1989a; Newton et al., 1989b), allele-specific amplification (Okayama et al., 1989), PCR amplification of specific alleles (Sommer et al., 1989) and allele-specific PCR (Wu et al., 1989), respectively. The methods rely on the inability of the DNA polymerase to extend primers that have a 3'-nucleotide mismatch. This property can be used to analyze point mutations by PCR, by designing two variants of one of the PCR primers such that the 3'-nucleotides hybridize to alternative variants of the investigated SNP. The two versions of this primer are then used in separate reactions and form perfect matches to their respective complementary sequences. Thus, in the case of a heterozygous sample, both reactions will give a product that can be detected on an agarose gel, while only one of the reaction tubes will render a product with either homozygote.

However, discrimination by the DNA polymerase is not perfect in any of these methods, and spurious positive reactions were reported as early as 1989 (Newton et al., 1989a). The cited authors observed that certain primer-template combinations, such as G:T and A:C, were more prone to extension than others. Further authors have confirmed these findings (Kwok et al., 1990; Day et al., 1999a; Day et al., 1999b; Ayyadevara et al., 2000), but none have found any robust solution that did not involve meticulous optimization.
Whole Genome Analysis Techniques

Just as the HGP spurred development of the technology required for DNA sequencing, the HapMap project has spurred advances in SNP typing technology. Several highly multiplexed systems have been developed, based on both new and old concepts, mainly as private initiatives. The multiplexing capacities of these technology platforms are intended to facilitate whole-genome association studies, by allowing several hundreds of thousands of SNPs to be analyzed in a single reaction. A throughput of 12,000 SNPs in one reaction has already been reported, using the padlock probe-based molecular inversion probe (MIP) system provided by Parallelle Biosciences (see below) (Hardenbol et al., 2005), and 500,000 SNP assays have been set up by both Affymetrix (www.affymetrix.com) and Illumina (www.illumina.com). Also, very high throughput ASOH-based resequencing assays have been used (Hinds et al., 2005). In addition, rapid advances have recently led to the development of novel whole-genome sequencing systems with the capacity to sequence entire bacterial genomes in a single reaction.

Multiplex Polony Sequencing

At Harvard Medical School, a group headed by Church has worked on a promising new sequencing platform called fluorescent in situ sequencing (FISSEQ) (Mitra et al., 2003). FISSEQ has a short read length and is based on clonal amplification of a constructed library of DNA sequences performed in a polyacrylamide gel to prevent diffusion of the reaction products, resulting in polymerase colonies (polonies) (Mitra and Church, 1999). The sequencing reaction is based on the sequencing-by-synthesis concept of adding one fluorescently labeled dNTP at a time. After detecting the polonies that have incorporated the labeled dNTP, the fluorophore can be stripped from the nucleotide by breaking the disulfide linkage between them.

Recently, however, the group has abandoned the FISSEQ strategy in favor of another method (Shendure et al., 2005). This novel DNA sequencing method has been used for resequencing an entire E. coli genome. Again, a library is constructed from genomic DNA by using restriction enzyme digestion and ligation of general oligonucleotides. This results in two 17-18 bp sequences that were separated by ~1 kb in the genome and are linked together and surrounded by three universal sequences used for general PCR and sequencing priming sites (Figure 2A). Each library molecule is then clonally amplified by PCR on 1 µm paramagnetic
beads in separate water bubbles in an oil emulsion (Dressman et al., 2003). Using the universal linkers, beads with amplified DNA can be enriched by hybridization to another bead type. The beads are then immobilized in an acrylamide gel and mounted in a flow-cell. A cyclic sequencing method of degenerate ligation is used for this, as follows (Figure 2B). First, a sequencing primer is hybridized to one of the four sites, then a pool of four degenerate nonamers where only the base complementary to the query base is known, are added. The four nonamers have different fluorophores attached to identify the query base. A ligation is used to discriminate correctly hybridized nonamers and then individual polonies are identified by imaging with an epifluorescence microscope and computerized analysis of the images. The ligated primers are then stripped off before the next sequencing round. By changing the position of the known base in the degenerate nonamer, sequencing is possible up to 6-7 bases from the sequencing primer. Thus, 26 bp from each genomic fragment can be read, assembled and compared to a reference sequence. The 30 Mb raw data generated from the resequenced E. coli genome was sufficient for 91% 1X coverage with 98% accuracy, or 75% 3X coverage with complete accuracy (Shendure et al., 2005).

![Diagram of sequencing process]

(A) Query sequences

Sequencing primer sites

In cycle 1 added degenerate nonamers:
- NNNNNNNA
- NNNNNNGN
- NNNNNNNC
- NNNNNNTT

Sample-sequencing primer complex:
CTAATGTC...GATTACA...GATTACAGA...

Hybridization and ligation:
NNNNNNNNCTAATGTC...GATTACA...GATTACAGA...

In cycle 2 added degenerate nonamers:
- NNNNNNAN
- NNNNNNGN
- NNNNNNAC
- NNNNNNTT

Readout: C
Figure 2. Organization of sequences (A) and degenerate nonamer design (B). Two fragments that were separated by approximately 1 kb in the sequenced genome are flanked and linked together by general sequences where sequencing primers can hybridize (A). In each sequencing cycle one base is interrogated by ligation of a degenerate nonamer to a sequencing primer (B). The four degenerate nonamers have unknown sequences in all but the query base.

The 454 Genome Sequencer 20 System

The 454 Life Sciences platform is another alternative for whole-genome sequencing. Using this platform, researchers have been able to sequence the entire genomes of *Mycoplasma genitalium* and *Streptococcus pneumoniae* of 0.6 Mb and 2.1 Mb, respectively, in single sequencing reactions (Margulies et al., 2005). The sample preparation used is quite straightforward; genomic DNA is randomly fragmented and linkers for PCR amplification and sequencing are ligated. The fragments are then immobilized on 28 µm sepharose microbeads, ideally with 1 fragment per bead. Subsequently, an emulsion-based PCR strategy is used to separate each bead into a microreactor for clonal amplification (Dressman et al., 2003). The beads are then loaded in a picotiterplate that consists of 75 µl, 44 µm diameter wells (Figure 3) (Leamon et al., 2003). A 6X6 cm picotiterplate contains 1.6 million wells and is manufactured by etching out the cores from a bundle of optic fibers. A 0.8 µm layer of microparticles and 2.8 µm paramagnetic beads with immobilized luciferase and ATP sulfurylase are also added to the wells. The plate is then mounted in a flow chamber in the sequencing instrument. The actual sequencing reaction is based on pyrosequencing chemistry, the reagents that are not immobilized are supplied to the flow cell by a liquid handling system and incorporations are detected with a CCD sensor. The average read-lengths are greater than 100 bases (with promises of 460-1000 bases in the future), which is a considerable improvement compared to conventional pyrosequencing, attributable to the washes that remove the byproducts. The generated sequences are then assembled from the 20 Mb of raw sequence generated from one instrument run. Coverage of 96% and 92%, with accuracy of 99.85% to 99.96%, respectively, was achieved for the genomes of the two investigated microorganisms. The method relies on the power of over-sampling to generate the required coverage and, thus, several of the steps have significant failure rates. For instance, the number of clonal fragments per bead may be more or less than one and many of the picotiterplate wells do not generate any signals.
However, as the number of sequence reads are extremely high, sufficient sequence data can be obtained to cover a complete bacterial genome in a single reaction.

![Flow of A, T, C, G](image)

Figure 3. The picotiterplate used in the 454 Genome Sequencer 20 system.

**Cloned Single-Molecule Arrays**

Solexa is currently developing a sequencing-by-synthesis method to sequence clonally amplified single DNA molecules (Bennett et al., 2005) (www.solexa.com). However, no scientific publication or product has yet been released. Short read-sized DNA fragments with ligated linkers are captured on a surface by hybridization of the linkers to immobilized oligonucleotides. Amplification is then performed in situ, resulting in discrete, micrometer-scale colonies, each containing about 1000 clonal fragments. The colonies are then sequenced with single incorporations of terminating four-color dNTPs. Fluorescence is read from each colony in every cycle by a sensitive CCD detector. The fluorophore is then removed before incorporation of the next dNTP. A read length of 25 bases is achieved in as many as hundreds of millions of colonies. The reads are then aligned to a reference genome for scoring variants. This approach generates Gb-scale raw data which is the required scale for whole genome resequencing of human genomes.
Molecular Inversion Probes
The specificity of the padlock probe-based SNP genotyping method has been further increased by using a combination of ligation and single base extension (Hardenbol et al., 2003). In the molecular inversion probe (MIP) assay developed by ParAllele Biosciences, the padlock probe lacks the single base complementary to the query base. This base is incorporated in the SBE step and then the ligation step is used to increase specificity. Furthermore, an exonuclease step is used to degrade probes that have not been circularized. The two alleles are assayed in separate reactions and PCR with universal primer sequences is used to amplify the circularized probes before tag-array detection (Figure 4). Using this approach, a 12,000-plex SNP genotyping reaction has been performed and described in a peer-reviewed journal (Hardenbol et al., 2005) and Affymetrix claims to have performed a 100,000-plex MIP SNP genotyping reaction (www.affymetrix.com).

![Diagram of Molecular Inversion Probes](image)

Figure 4. Outline of the molecular inversion probe assay.

Allele-Specific Oligonucleotide Hybridization
Very high throughputs in ASOH-based genotyping and resequencing assays have been attained, despite the limitations discussed above, using two different amplification strategies: long-range PCR and general PCR on fragmented DNA with a ligated universal PCR primer
(Figure 5). With the first strategy, 9 kb amplicons (on average) are pooled before hybridization to a large array of oligonucleotides synthesized *in situ* and ~50,000 SNP genotypes per array can be genotyped (Hinds et al., 2005). This assay has been used by Perlegen Biosciences for both SNP genotyping and discovery (the latter may also be called resequencing) (Hinds et al., 2005). The second PCR strategy introduced by Affymetrix is less labor intensive but gives a less flexible choice in SNP selection since the fragments are produced by restriction enzymes and the PCR reaction favors amplicons of certain sizes (Kennedy et al., 2003; Matsuaki et al., 2004).

![Diagram of the fragment-based amplification strategy for ASOH-genotyping.](image)

**GoldenGate and Infinion Assays**

The GoldenGate method developed by Illumina uses a combination of two enzymatic events to achieve specificity directly on genomic DNA (Fan et al., 2003). In this assay, the first enzymatic reaction is an allele-specific extension step, in which the incorporated nucleotides reach a second downstream oligonucleotide (Figure 6). In the second enzymatic event a ligation takes place between the incorporated nucleotides and the second oligonucleotide. The downstream oligonucleotide has a 3'-located universal sequence for PCR amplification and a tag-sequence for detection. The two ASE-oligonucleotides, directed towards the two possible
alleles, have different 3'-located PCR-primer sequences to introduce different fluorescent labels in PCR. The same PCR-primer sequences are used for all SNPs, allowing general PCR amplification of all extended and ligated oligonucleotides in one reaction. The genotypes are detected by hybridization of the unique tag sequences to a bead array that is manufactured with similar optic fibers to the 454 picotiterplate (Gunderson et al., 2004). The bead array consists of ~50,000 beads locked in individual wells, each ~1.4 µm in diameter. There are 1520 bead types, which have complementary sequences to the tag-sequences immobilized to them. These are randomly ordered and are decoded with a cyclic hybridization assay prior to actual use. By hybridizing to the array the PCR products can be separated, and the genotypes can be assigned to 1500 SNPs following fluorescence analysis.

![Diagram of the GoldenGate assay](image)

Figure 6. The GoldenGate assay.

Illumina’s very large-scale assay system, called Infinium, utilizes an isothermal whole-genome amplification as a first step before random enzymatic fragmentation (Gunderson et al., 2005). The DNA is then hybridized directly to the bead array where the immobilized oligonucleotides have a decoding sequence and a 50-base long 3'- termini allele-specific sequence. The two possible alleles for each SNP are in this case detected on different beads. A single ASE reaction is performed with biotin-labeled dNTPs that can be detected with a streptavidin-conjugated fluorophore. This assay has been performed with bead arrays to type
100,000 SNPs in multiplex and 250,000 and 500,000 SNP assay systems are close to being launched.

**Summary**

The sequencing and genotyping techniques presented herein have different advantages and drawbacks. Generally, factors such as accuracy, flexibility, throughput and cost have to be considered in selection of an assay platform. Obviously, the recent technological advances in the massively paralled reactions have realized throughputs that were not imaginable only a decade ago which opens fantastic possibilities. Still, the cost per sample much limits their usefulness however it may be improved. Instead, methods with moderate throughput but with significantly lower costs per sample are more affordable and may be more attractive for most applications of sequencing and genotyping, especially in clinical settings where sample throughput and cost are important factors.

In Table 1, some of the leading commercial genotyping techniques, that obviously have their roots in assays developed by academic efforts, have been compared. As shown, the characteristics of the techniques are very different in terms of sample throughput, genotype throughput, platform and flexibility. Whole genome association (WGA) studies require typing of tens to hundreds of thousands of SNPs in a single assay, for this application the Infinion and GeneChip platforms have to be considered as the assays of choice. However, these assays have a fixed set of SNPs and for candidate gene and fine mapping applications, these would not be useful. Instead, the assays which have a flexible choice of SNPs such as the GoldenGate assay and the MIP assays, still type a large number of SNPs in each reaction. Nevertheless, for an association study to be well-powered, it has to be performed in hundreds to thousands of samples, which will be immensely expensive with these technologies. The more wide-spread and well-used assays MassARRAY, SNPstream and several academic in-house assays not mentioned in Table 1 have much higher sample throughput and are so far better suited for these applications. The new SNPlex might also become a good choice under such circumstances.
Table 1. SNP genotyping techniques comparison.

<table>
<thead>
<tr>
<th>Technique and provider</th>
<th>SNPs / run</th>
<th>Samples / run</th>
<th>Flexibility</th>
<th>Amplification method</th>
<th>Molecular reaction</th>
<th>Detection platform</th>
</tr>
</thead>
<tbody>
<tr>
<td>SNPstream Beckman-Coulter</td>
<td>12</td>
<td>384</td>
<td>flexible</td>
<td>multiplex PCR</td>
<td>SBE</td>
<td>capillary</td>
</tr>
<tr>
<td>MassARRAY Sequenom</td>
<td>12 - 29</td>
<td>384</td>
<td>flexible</td>
<td>multiplex PCR</td>
<td>SBE</td>
<td>MS</td>
</tr>
<tr>
<td>SNPollex ABI</td>
<td>48</td>
<td>384</td>
<td>flexible</td>
<td>universal PCR</td>
<td>OLA</td>
<td>capillaries</td>
</tr>
<tr>
<td>MIP Parallel</td>
<td>1.5 - 10 k</td>
<td>1</td>
<td>flexible</td>
<td>universal PCR</td>
<td>SBE + ligation</td>
<td>capillary</td>
</tr>
<tr>
<td>GoldentGate Illumina</td>
<td>1.5 k</td>
<td>96</td>
<td>flexible</td>
<td>universal PCR</td>
<td>ASH + ligation</td>
<td>bead-array</td>
</tr>
<tr>
<td>Infinium Illumina</td>
<td>100 k - 500 k</td>
<td>1</td>
<td>fixed</td>
<td>WG amplification</td>
<td>ASH</td>
<td>bead-array</td>
</tr>
<tr>
<td>GeneChip Affymetrix</td>
<td>100 k</td>
<td>1</td>
<td>fixed</td>
<td>WG PCR</td>
<td>ASOH</td>
<td>array</td>
</tr>
<tr>
<td>Perlegen Sciences</td>
<td>1 M</td>
<td>0.02</td>
<td>fixed</td>
<td>LR-PCR</td>
<td>ASOH</td>
<td>array</td>
</tr>
</tbody>
</table>

Table 2 presents the most widely used and the most promising techniques for DNA sequencing. Even these tools have completely different application areas and consequently attract different fields of research. The techniques provided by Affymetrix and 454 Life Sciences and the multiplex polony sequencing are very convenient and cost-effective alternatives for whole genome bacterial sequencing while Solexa’s approach (if promises are filled) may be suited for resequencing of more complex genomes such as the human. In addition, if the envisions of 454 Life Sciences for increasing read length from 100 to 1000 bases are realized, the platform can be adopted for whole genome resequencing of complex genomes. However, in very near future, the Sanger DNA sequencing and Pyrosequencing will be dominating techniques in research and routine diagnostic laboratories. This fact can be attributed to the high accuracy and low cost per analyzed sample.

Table 2. Sequencing techniques comparison.

<table>
<thead>
<tr>
<th>Technique</th>
<th>Provider</th>
<th>Read length</th>
<th>Reads / run</th>
<th>bps / run</th>
<th>Cost / sample</th>
<th>Cost / genome</th>
</tr>
</thead>
<tbody>
<tr>
<td>SangerDNA sequencing</td>
<td>ABI</td>
<td>500-1000</td>
<td>96-384</td>
<td>384 kb</td>
<td>low</td>
<td>high</td>
</tr>
<tr>
<td>Pyrosequencing</td>
<td>Biotage</td>
<td>25-40</td>
<td>96-384</td>
<td>15 kb</td>
<td>low</td>
<td>very high</td>
</tr>
<tr>
<td>Sequencing by hybridization</td>
<td>Affymetrix</td>
<td>1</td>
<td>30 k</td>
<td>30 kb</td>
<td>intermediate</td>
<td>intermediate</td>
</tr>
<tr>
<td>Multiplex polony sequencing</td>
<td>(Shendure et al., 2005)</td>
<td>13 + 13</td>
<td>~1 M</td>
<td>30 Mb</td>
<td>high</td>
<td>low</td>
</tr>
<tr>
<td>454 genome sequencer 20</td>
<td>454 Life Sciences</td>
<td>100</td>
<td>~250 k</td>
<td>25 Mb</td>
<td>high</td>
<td>low</td>
</tr>
<tr>
<td>Cloned single-molecule array</td>
<td>Solexa</td>
<td>25</td>
<td>~100 M</td>
<td>2.5 Gb</td>
<td>high</td>
<td>low</td>
</tr>
</tbody>
</table>
INVESTIGATIONS

The work presented in this thesis was all based on techniques that aim to define individual DNA signatures by controlled synthesis of nucleic acid multimers. The oligomerization reaction was harnessed using two different strategies that will be described in detail. The papers can be grouped according to both the techniques used and the biological applications. Paper I concerns resequencing of the p53 tumor suppressor gene, which is often mutated in cancers, using in situ AMASE. Papers II and III describe the development of a method for viral and microbial genotyping, the utility of which was demonstrated by typing Human Papillomavirus (HPV) samples, based on a combination of multiplex competitive hybridization (MUCH) and AMASE on microbeads. Papers IV to VII deal with a second approach to control oligomerization, denoted PrASE. The papers present a series of studies in which SNP genotyping techniques were explored, validated and improved, notably by raising their throughput capacities. In Paper IV multiplex PrASE is introduced, while Paper V describes studies in which the method was used to type SNPs that had been shown to influence the progression of thrombosis in a large sample of patients. The method was also compared (especially in terms of its accuracy) to pyrosequencing. In Paper VI the PrASE approach was compared with Sanger DNA sequencing in terms of its utility for screening for variants of the MC1R gene associated with increased risks of developing malignant melanoma. Paper VII describes a full clinical study of MC1R screening in several groups of malignant melanoma patients and controls.
Apyrase-Mediated Allele-Specific Extension

In 2001 Ahmadian et al described a way to avoid the spurious extensions that cause false positive signals in allele-specific extension assays (Ahmadian et al., 2001). A milestone experiment was performed in which real-time measurement of light from a luciferase-based reporter system was used to detect the incorporation of nucleotides (Figure 7). The results showed that the incorporation of nucleotides was much faster with a matched 3'-terminus primer than with a mismatched primer. This kinetic finding was exploited to distinguish the two events using an enzymatic approach. The enzyme apyrase breaks bonds between phosphate groups and was used to degrade dNTPs to dNMPs. As the DNA polymerase is not able to incorporate dNMPs, the two enzymes have to compete for the dNTPs. Thus, the amounts of the reagents and enzymes had to be tuned to allow the DNA polymerase to incorporate dNTPs after a fast 3’-match reaction, but not after a slow 3’-mismatch reaction.

Figure 7. The classic Ahmadian experiment. Raw bioluminometric allele-specific extension data for the matched G-allele (left curves) and mismatched T-allele (right curves). Without the use of apyrase the curves look similar, albeit with different slopes, thus the T-allele primer is mistakenly extended (top curves). With the use of apyrase only the G-allele gives a peak.
The principle of apyrase-mediated allele-specific extension (AMASE) was also demonstrated with another detection platform; in situ extension on microarrays with fluorescence detection (O'Meara et al., 2002). Both studies demonstrated that AMASE can correctly genotype SNPs, while conventional allele-specific extension methods could generate erroneous results with certain mismatches.

**Resequencing of the p53 gene (Paper I)**

**The p53 Gene**

The tumor suppressor gene *p53* spans 20 kb of chromosome 17p13.1 (Benchimol et al., 1985) and contains 11 exons. The gene product is a 393 amino acid phosphoprotein with three main domains: an N-terminal transcription activation domain (Fields and Jang, 1990), a central specific DNA-binding domain (Pavletich et al., 1993; Wang et al., 1993) and a C-terminal oligomerization, nuclear localization and non-specific DNA-binding domain (Addison et al., 1990; Foord et al., 1991). The major function of p53 is to stop cells with damaged DNA from proliferating by inducing DNA repair, cell-cycle arrest or apoptosis (Melnikova and Ananthaswamy, 2005). It is also activated under various forms of cellular stress.

*P53* mutation is the most frequent event in human cancers (Hollstein et al., 1991; Hollstein et al., 1996; Olivier et al., 2002), being mutated in more than 50% of cancers of the lung, colon, stomach and esophagus (Soussi, 2000). Mutation rates in cancers of the skin vary; it is mutated in 90% of squamous cell carcinoma cases (Brash et al., 1991; Ziegler et al., 1993), ~50% of basal cell cancers (Brash et al., 1996; Soehnge et al., 1997), but usually not in melanomas (Yang et al., 2001; Chin, 2003). In addition, in cervical carcinoma, which is also epithelial, p53 is not usually mutated. Instead, it is targeted for degradation by the human papillomavirus protein E6 (Scheffner et al., 1990; Werness et al., 1990).

**Brief Assay Description**

A novel AMASE-based strategy for resequencing was investigated. To test the robustness of the assay, a difficult target sequence was chosen, consisting of 15 bps in exon 5 in the *p53*
gene, containing a 6 bp homoplymeric C-stretch with 67% GC content. For each position, four oligonucleotides were designed, differing only in their 3'-terminal bases. The oligonucleotides were spotted in duplicate in 16 identical subarrays on glass slides and covalently attached by 5'-located amino groups. A silicone mask separated the subarrays during hybridization to allow 16 assays to be performed on each slide (Pastinen et al., 2000). Strand-specific elution of the non-coding strand of the PCR products was performed automatically by a Magnatrix 1200 pipetting robot. This robot is equipped with a magnet enabling it to handle super paramagnetic streptavidin-coated beads to allow immobilization of biotinylated targets, such as PCR products amplified with a biotinylated primer. After alkaline treatment, to remove the non-biotinylated strand, the immobilized single-stranded DNA was eluted by incubating the beads in water at 80 °C for 1 minute. This treatment breaks the biotin-streptavidin bond (Holmberg et al., 2005a). The released DNA was then briefly hybridized to the spotted probes before washing to remove any unhybridized material. The AMASE-reaction was then performed in situ on the primed templates, starting by adding a solution containing Klenow fragment DNA polymerase with or without the apyrase. All four dNTPs were then added (including a 50:50 mixture of Cy5-labeled and unlabeled dCTPs) to initiate the polymerization simultaneously with the degradation of the dNTPs. Three allelic fractions (AFs) for each position were then estimated by dividing the fluorescence intensity from the wild type probe by the intensity from the wild type + possible mutation 1, 2 or 3 probes, respectively. Any ratio deviating significantly from 1 indicates a mutation.

Results

Initially the effect of the apyrase was investigated with a synthetic wild type template, giving AF values of at least 0.89 (mean, 0.96). It was found that the mismatch primers gave much higher signals in comparison to the match (wild type) when apyrase was not present (mean AF 0.81, lowest 0.43). In fact, without the inclusion of apyrase, the resulting sequence was unreadable. The linearity of the relative quantification was then investigated using a second oligonucleotide with an inserted mutation. In addition, the relative detection limit of mutated template was found to be low compared to that of Sanger DNA sequencing. Normally, the abundance of a mutated form of the template, relative to that of the most abundant form, needs to be at least 25% for detection by Sanger sequencing (Demeter et al., 1999; Schuurman et al., 1999), while the corresponding proportion for AMASE is just 10%.
To investigate the effect of a mutation on sequence reading, three oligonucleotides with inserted mutations were analyzed. The mutations were readily detected, but disturbed the sequence reading of the following bases. As a result of a template with 100% mutation, all four oligonucleotides at the subsequent position will have a 3'-1 mismatch, which is strongly unfavorable to extension. However, the sequencing will not be affected if the mismatch is located at 3'-2. Insertions and deletions may also cause problems, since they will be detected as substitutions, but could be correctly classified if read from both directions.

To conclude the study, clinical samples from microdissected skin lesions (squamous cell carcinoma and p53 patch (Ren et al., 1996; Rebel et al., 2001)) were amplified by PCR and mutations were studied with the AMASE assay. The detected mutations were all consistent with the results of previous studies in which the samples had been sequenced by Sanger DNA sequencing (Ahmadian et al., 1998). These findings suggest that the approach offers promising opportunities for developing and exploiting large-scale resequencing chips with up to ~2.5 million probes per array (Matsuzaki et al., 2004). In this way more than 500 genes (with mean coding sequence of 1200 bases) could be resequenced in a single AMASE-based analysis.

**Genotyping Human Papillomaviruses (Papers II and III)**

**Microbial Genotyping**

Detecting various microbes and viruses has become very important for diverse environmental and health-related reasons. However, traditional methods of identification involve cultivation of the potential pathogen(s), and delays that are no longer acceptable. Many types, especially of viruses, cannot even be distinguished by cultivation and morphological characterization. Instead, they need molecular analysis, which is generally performed at the nucleic acid level. Many pathogens require exact identification for correct diagnosis and therapy. Determining the 16 S rRNA sequences of bacteria is a frequently used approach, for which Sanger DNA sequencing is a logical choice, but several other approaches are available. For instance, serological analyses with antibodies that have high specificity for species-specific antigens are commonly applied in routine clinical work, but are not available for all applications.
**Human Papillomavirus**

Direct cytological or histological identification of subspecies of the Human Papillomavirus is impossible, at least with current technology, and the virus cannot be cultivated (van Doorin et al., 2001). The definition of a type is based on the sequence similarity of the L1 gene and can be at most 10% similar between types, 2-10% similar between subtypes and less than 2% similar between variants (De Villiers et al., 2004). Over 100 HPV types have been defined and classified and, interestingly, different groups of types infect different exposed tissues of the body, where they can cause either warts or epithelial cancer, depending on their type. In fact, the presence of human papillomaviruses are known to be required for the transformation of normal epithelial cells into cervical cancer cells (Schiffman and Castle, 2003). HPV types that infect anogenital sites are divided into high-, intermediate- and low-risk types. Interestingly, different high-risk types are associated with different lesions, for example high-risk HPV types 16 and 18 predominate in squamous cell- and adeno-carcinomas of the cervix, respectively (McKaig et al., 1998). However, much remains to be elucidated about the differing etiology and persistence of infection of specific types, mainly due to methodological limitations. Furthermore, infections with multiple HPV types are common, but their role in the development of cancer is largely unknown (Hubbard, 2003).

**Methods for HPV Type Determination**

Several methods have been developed and are used clinically for the molecular diagnosis of papilloma infection. Among these is the Hybrid Capture II system (HCII), which is based on hybridization of a cocktail of RNA probes to the viral DNA. An antibody that recognizes DNA-RNA heteroduplexes is then used for detection and signal amplification. The method does not require amplification of the target, but at least 5000 copies are needed for detection, and thus it is less sensitive than PCR (Smits et al., 1995; Cope et al., 1997). HCII cannot distinguish individual genotypes, but can only separate high-risk from low-risk type HPV infections (Smits et al., 1995; Cope et al., 1997). *In situ* hybridization (ISH) (fluorescent and non-fluorescent) can detect infections directly in histological sections and is now used to test for integration of the viral genome into the host cells, which determines the persistency and gravity of the infection (Jeon and Lambert, 1995; Evans and Cooper, 2004). However, the
method has low maximal throughput and requires large amounts of target DNA for type
determinations (Hubbard, 2003). Target amplification with PCR is currently the best
alternative to improve sensitivity and throughput. In most cases it can also improve the
specificity of HPV testing. PCR amplification can be done using either of two approaches:
type specific amplification or consensus amplification. The former specifically amplifies a
single HPV type by using type-specific primers. However, use of this approach has been
hampered by the fact that each sample requires a large number of PCR reactions to cover all
of the clinically important HPVs. On the other hand, by combining this PCR setup with real-
time detection the viral load can be estimated, which is an important clinical outcome factor
(Moberg et al., 2003). The consensus amplification approach involves the use of consensus or
general PCR primers. Several different consensus PCR primer sets are available for
amplification of HPV genotypes, and a number of different techniques can be adopted to
determine the genotypes of these PCR-amplified HPV-DNAs. These techniques include
restriction fragment length polymorphism (RFLP) analyses, DNA sequencing methods,
hybridization assays (Bernard et al., 1994) and a combination of hybridization and enzymatic
assays (Delrio-Lafreniere et al., 2004). RFLP assays are the most low-throughput, time-
consuming and labor intensive of the PCR approaches. In recent years, relatively rapid DNA
sequencing methods have been developed for routine examination of clinical samples.
Nevertheless, it should be noted that direct DNA sequencing techniques are not suitable for
resolving multiple infections in a specimen (Goncalves et al., 1999; Kleter et al., 1999;
Gharizadeh et al., 2001). In addition, sequences that represent a minority of the PCR products
(10-25%) may remain undetected. In samples with a suspected multiple HPV infection, the
cumbersome approach of cloning the PCR products and sequencing a large number of clones
is necessary. Another drawback associated with direct HPV sequencing is the generation of
non-specific PCR products that can give rise to background signals. However, hybridization
techniques in various formats are being developed that allow increasingly high throughputs,
such as the easily handled Line Blot assays with a membrane strip substrate (van Doorn et al.,
2001) and spotted glass substrate microarrays (Park et al., 2004). Other microorganisms have
recently been subjected to tests using microarray platforms designed to detect multiple
different organisms, either with short 20-nt probes synthesized in situ (Wilson et al., 2002) or
with spotted 70-mers (Wang et al., 2002). Both short and long hybridization probes have
advantages and disadvantages, as reviewed (Bodrossy and Sessitsch, 2004). To circumvent the
problems associated with hybridization approaches, increased specificity has been achieved by including enzymatic steps (Lovmar et al., 2003; Delrio-Lafreniere et al., 2004).

**MUCH-AMASE**

A genotyping platform was established for identifying HPV types using hybridization of short type-specific probes. To account for the inherent difficulties in achieving perfect hybridization, specific extension was used as a second discriminatory step. The hybridization step was designed to be as specific as possible by immobilizing the viral biotinylated PCR product on streptavidin-coated beads and employing multiplex competitive hybridization (MUCH). The non-biotinylated strand was removed by strand-specific elution with sodium hydroxide to avoid any unwanted cross-reactivity. Type-specific oligonucleotides designed for all HPV types detectable in the assay were then added in large abundance. These oligonucleotides were allowed to hybridize by increasing the temperature to 70°C and then cooling slowly to room temperature. In this competitive hybridization the oligonucleotides that best match the sample templates will hybridize first and occupy the spot before other oligonucleotides can. In the reverse scenario, using immobilized oligonucleotides and annealing template, some of the templates may unspecifically hybridize to the spotted oligonucleotides due to the lack of competitive hybridization. However, some unwanted hybridization may still occur in cases where the difference in annealing temperature is too low to be effective. In such scenarios, coupling a specific enzymatic reaction, such as AMASE, to the competitive hybridization can ensure correct typing. In the specific extension with AMASE only oligonucleotides with matched 3'-ends will be extended by incorporation of fluorescent dNTPs and be detected, while 3'-end and 3'-1 mismatch extensions will be prevented.

Following the MUCH-AMASE reactions, the extension products were alkali-eluted, neutralized and hybridized to a tag-microarray via 5'-localized tag sequences (Hirschhorn et al., 2000). Using the heterogeneous format with bead-immobilized samples, unhybridized oligonucleotides could be washed away and prevented from influencing the later hybridization to the tag-array.
Type-Specific Primer Design

A custom Perl script was used to design the type-specific oligonucleotides used for the specific hybridization and extension. The L1 region sequences of all 39 types that can be amplified with the consensus PCR primers were aligned (about 150bp). Three 19-20 bp regions with high heterogeneity were selected for primer design, and checked for cross-hybridization and mismatch extension. These oligonucleotides were all compared against each other’s annealing sites in the amplified HPV sequences. The number of mismatches each oligonucleotide made to all possible HPV templates was recorded. In order to introduce high specificity, any number exceeding four was considered as being discriminatory in the hybridization step. Clearly, the position of the mismatch would influence this step, but less than four mismatches was a cautious assessment. The 3’-ends of the oligonucleotides were subsequently checked to examine if they would be extended in the AMASE-step after hybridization. In some rare cases neither of these two steps would be discriminatory, resulting theoretically in a false positive signal. However, the system is designed so that no more than one out of the three oligonucleotides for the three regions would show up as a false positive and would thus be identified as such. In this way all 39 oligonucleotides for the three regions were compared in silico to the 39 target template sequences.

Samples and Results of the Initial Proof-of-Concept Study (Paper II)

In the study described in Paper II, in which three L1 regions were used for each HPV type, the type-specific oligonucleotides for 10 genotypes were included. PCR amplifications of nine different whole genome plasmids, and DNA extracted from 15 HPV positive clinical samples, were analyzed. All analyzed samples gave the expected results. The expected false positive rates from the in silico evaluation proved to be overcautious and only one out of five gave a false positive result. The true positive signals were compared to the local background signals and were found to be 14 to 1648-fold stronger (335 times on average). Using this assay is very beneficial when investigating multiply infected samples. Plasmid samples were mixed to simulate multiple infections and two multiply infected clinical samples were successfully analyzed. All genotypes were confirmed with pyrosequencing and Sanger DNA sequencing.
Samples and Results of the Subsequent Pilot Study (Paper III)

The positive results from the first study indicated that using three type-specific primers would not be more informative than using only two primers. Therefore, the primers for the first (and least informative) L1 region were omitted so that more HPV types could be fitted with the limited number of tag sequences. In this way 23 genotypes could be detected. Ninety-two samples with HPV infections (confirmed by gel-electrophoresis) were genotyped by MUCH-AMASE and confirmed with Sanger DNA sequencing and, in some instances, cloning. The new strategy was capable of typing all samples, irrespective of whether they contained single infections (71 samples) or multiple infections (21 samples: 14 doubles, four triples and one each with quadruple and quintuple infections). Only one gel-electrophoresis-positive sample could not be genotyped, and it was found to represent a HPV 62 infection, which is very uncommon and not included in the 23 detectable types. All but two of the samples were also sequenced, and as expected only 63 of the 90 samples gave identical results. The genotypes of the samples that were easily identified as multiple infections with MUCH-AMASE were considerably more difficult to determine by sequencing.

A custom Excel script was developed to automatically score the number of infections and quantify them. The inputs to this script are the fluorescence intensities from the tag-array calculated by Genepix software. The script uses several criteria to score a sample. Firstly, the signals have to be stronger than a threshold value defined as two times the sum of the median background and spot intensities. Secondly, the intensity of the signals from both type-specific primers for one HPV type has to be higher than the threshold intensity.

Thirdly, in the case of multiply infected samples, the least abundant type has to be present at a proportion of at least 5% of the total, as calculated from the relative intensities. The reason for not including lower amounts is that signals are then quite weak. In fact, several of the investigated samples had infections with relative abundances lower than 5%, and to verify them four multiply infected samples were cloned and sequenced. The cloning results verified 14 of the 15 HPV types in these samples with relative abundances as low as 2%. The relative abundance of the only HPV type that was not found with this procedure was scored at 2%, in a sample with five infections. However, only 30 clones could be sequenced, and thus the chance of finding this last type was relatively low.
The comparison of results obtained with the MUCH-AMASE and Sanger DNA sequencing, and their verification with the cloning procedure, shows the utility of the developed method. MUCH-AMASE can easily detect multiply infected samples that are very difficult to analyze by ordinary sequencing. In fact, the presence of only two HPV types in a sequence makes the distinction very difficult when an ordinary sequencing approach is employed. The sensitivity of MUCH-AMASE is also much higher and can probably be improved further. This new tool provides novel opportunities to investigate the clinical importance of multiply infected samples. For instance, the possibilities that infection by several HPV types increases the risk of developing cervical carcinoma, and that certain combinations increase the risk of transformation, could be investigated.

The accuracy of relative quantifications by MUCH-AMASE was investigated by analyzing mixtures of two PCR products in known proportions. This analysis showed that by comparing the type-specific oligonucleotides for one of the regions, exact quantifications could be made across the proportional range from 50:50 to 95:5. However, the other type-specific oligonucleotide did not result in an exact quantification. These results indicate that the quantification of the types by MUCH-AMASE should be considered approximate.

**Genotyping of SNPs and Common Variations (Papers IV-VII)**

These papers describe a continuous effort to improve the ASE platform for SNP genotyping. In highly multiplexed ASE reactions large numbers of allele-specific primers have to be hybridized simultaneously, and this will lead to non-specific hybridizations. More specific hybridization can be achieved at higher temperatures, but this would cause problems for the AMASE assay since apyrase is a thermo-labile enzyme. In order to maintain stringency at higher temperatures, the relatively thermostable enzyme Proteinase K, which has the ability to degrade proteins at temperatures up to 72 °C, was used. Thus, instead of degrading the dNTPs, as in the AMASE approach, the DNA polymerase is degraded, which also prevents unspecific extensions of 3'-mismatched primers.

The SNP genotyping in these studies was performed by employing the 48-well tag-array system. Furthermore, the genotyping assay was automated with a pipetting robot from
Arrayed Identification of DNA Signatures

Magnetic Biosolutions with an upgraded program allowing the preparation of 48 samples. However, the actual genotyping reaction took place on the single-stranded target DNA immobilized on super paramagnetic beads.

**Protease-Mediated Allele-Specific Extension (Paper IV)**

In the study described in paper IV, 13 SNPs were amplified by PCR in a nested multiplex reaction and then genotyped by protease-mediated allele-specific extension (PrASE)-based analyses. The nested approach was employed to ensure specificity with minimal optimization of the PCR and multiplexing in order to reduce the number of PCR reactions. Furthermore, since the SNPs were initially chosen for forensic purposes, it would be valuable to be able to amplify all the loci of interest when a limited amount of sample material is available.

First, the number of nucleotides incorporated in the PrASE reaction was investigated by designing four synthetic templates with only one G-residue at different lengths (5 to 20 bases) from the extended primer. The extension reactions were performed with Cy5-labeled dCTP, native dATP, dTTP, and dGTP plus varied amounts of Proteinase K (0 to 80 µg). The results indicate that the signal intensities decrease with increasing amounts of Proteinase K, as expected. Furthermore, all levels of Proteinase K and all four templates produced sufficiently strong signals for genotyping, but for robust clustering in standard applications the use of 20 µg of Proteinase K was recommended. The use of a solid phase to immobilize the PCR products facilitated washing of the bound complex. Unhybridized allele-specific primers could be removed by this procedure and the effect of washing on sensitivity could be investigated. As these primers would otherwise hybridize to the tag-array it was expected that the sensitivity of the assay would be increased, and there would be less need to optimize the amounts of allele-specific primers as compared to the generated amounts of PCR products. This approach was tested with six samples for all 13 SNPs, and at least a 15-fold increase in average fluorescence intensity was observed with removal of excess primers.

The effects of carrying out the genotyping reaction in multiplex were investigated by typing a sample in 12 reactions starting with only one pair of allele-specific primers and a stepwise increase of complexity up to 12 in the last reaction. All genotypings were correct and the
signal intensities from each SNP were constant in all reactions. These findings indicate that the multiplexing capacity limit is much higher than 12 SNPs, without any loss of sensitivity.

Next, the improvement in accuracy obtained by including a protease was explored. Thirty-six samples were analyzed with and without protease and the genotyping clusters and calls were compared. Eight samples were also genotyped by pyrosequencing to resolve inconsistencies. All pyrosequencing results matched the PrASE results. Several difficulties arose when using ASE, particularly for SNP 10 (which did not give any separation of clusters) and generally considerably less separation was obtained for all SNPs. In conclusion, higher assay conversion rates and more robust clustering are possible with PrASE.

**Accuracy and Robustness of PrASE as Compared to Pyrosequencing (Paper V)**

To further characterize the PrASE assay, 10 SNPs associated with cardiovascular disease were chosen (Lane and Grant, 2000; Endler and Mannhalter, 2003; Humphries and Morgan, 2004). These SNPs had been previously investigated in a candidate gene association study undertaken to find differences in allelic frequencies between two patient groups diagnosed with acute coronary syndrome (Holmberg et al., 2005b). The difference between the groups was in the outcome of the thromboses: the first group’s resolved quickly (so-called unstable angina) while the second group’s progressed to myocardial infarction.

An assay similar to the approach described in Paper IV was applied. As reported in Paper IV, the accuracy of PrASE was higher than that of ASE since two SNPs could not be genotyped with conventional ASE. In both cases, one of the homozygous genotypes clustered with the heterozygous samples due to false positive extensions.

PrASE was used to blindly type the 10 selected SNP positions in 442 samples. The results were compared to pyrosequencing results and 99.8% concordance was found. Eight out of the 4420 genotypes obtained with the compared techniques did not match. The discordant genotypes were individually distributed across assay dates, plates and SNPs. The robustness of the results was illustrated by including all genotyping data in one cluster diagram for individual SNPs and all SNPs together. The results displayed distinct, reproducible patterns for each
SNP. Differences can be attributed to the sequence context affecting the hybridization of signature tags as well as the extension efficiency. However, despite these differences, clustering of all the results for all 10 SNPs in a common diagram gave three distinct and correct clusters, further indicating the robustness of the approach. In fact, one of the SNPs is a variant of either 5 or 6 thymidine residues that was expected to be difficult to genotype. However, no additional effort was expended on optimizing the genotyping of this locus, even though the allele-specific primers ended with 5 T:s followed by a C or 6 T:s, respectively. These primers were expected to have very different 3'-end stabilities.

Twenty-four samples were each analyzed twice on one occasion and the same tag-array slide. In addition, 12 samples were assayed twice on different dates with different batches of reagents and tag-array slides. Both experiments showed small standard deviations for the results from individual samples, indicating good reproducibility.

In conclusion, the results presented in paper V indicate that PrASE gives higher assay conversion rates than the ASE assay, and equal accuracy and robustness than pyrosequencing.

Screen for Variants that Increase Melanoma Risks

The Genetic Basis of Hereditary Malignant Melanoma

Cutaneous malignant melanoma is the type of cancer that shows the most rapid increase in incidence in Caucasians and leads to the second highest mortality rates after lung cancer (Slominski et al., 2001). An estimated 10% of melanoma cases are familiar, i.e. the patient has a first or second-degree relative with melanoma (Hayward, 2003). Two high-penetration genes have been found that, when mutated, can cause melanoma: CDKN2A, encoding cyclin dependent kinase inhibitor 2A (Kamb et al., 1994; Hayward, 2000) and CDK4, encoding cyclin dependent kinase 4 (Zuo et al., 1996; Soufir et al., 1998). CDK4 is a protein that negatively regulates pRb, thereby causing it to release E2F transcription factors, initiating the G1 to S transition. The CDKN2A locus has a unique and fascinating organization; the four exons give rise to two unrelated protein products, p16\textsuperscript{INK4A} and p14\textsuperscript{ARF}, through alternative splicing and translation of the products in different reading frames (Stone et al., 1995). p16\textsuperscript{INK4A} is a tumor suppressor and normally inhibits CDK4. Accordingly, mutations in these two genes have
Max Käller

similar clinical effects (Goldstein et al., 2000). The other protein, p14ARF, is also a tumour suppressor but acts through a different pathway. It binds Mdm2 and restrains it from inactivating p53 (Zhang and Xiong, 1999). Thus, CDKN2A mutations can affect either or both of the gene products and pathways. However, CDKN2A accounts for only 25-40% of the melanoma heritability (Hayward, 2003) and only three families worldwide have been reported to carry CDK4 mutations (Zuo et al., 1996; Soufir et al., 1998). Clearly, additional melanoma predisposition genes are likely to exist (Hayward, 2003). Human pigmentation genes have been studied in attempts to find low-penetrance variants that may be associated with melanoma susceptibility (Sturm et al., 2001), and genome-wide association studies are currently being carried out by the International Melanoma Genetics Consortium (Kefford and Mann, 2003).

The best established low-penetrance variants to date are mutations in the highly polymorphic melanocortin-1 receptor gene (MC1R). A defined phenotype called the red hair color (RHC) phenotype has been associated to certain variants of MC1R (Valverde et al., 1996; Palmer et al., 2000; Kennedy et al., 2001) and to an increased risk of cutaneous malignant melanoma (Valverde et al., 1995). Traits associated with this phenotype include red hair, fair skin and an increased tendency to freckle. The receptor is located on melanocytes, which supply melanin to the skin’s keratinocytes and thus help to protect the skin from ultraviolet radiation (Sturm et al., 2001). MC1R is highly polymorphic in the Caucasian population and more than 40 variants have been found (www.ncbi.nlm.nih.gov/SNP/). Three amino acid changes (Arg151Cys, Arg160Trp, and Asp294His) have been reported to correlate well with the RHC phenotype and a 2-fold odds ratio of developing melanoma, (Valverde et al., 1996; Palmer et al., 2000). These are located in the turns on the cytosolic side between transmembrane α-helices in the receptor. The presence of such a variant results in the incapacity to change red/yellow pheomelanin to black/brown eumelanin production, leading to a weakened tanning response and less protection against ultraviolet radiation. Carriers of both MC1R variants and CDKN2A mutations have been shown to have a lower age of melanoma onset than carriers of either mutation alone, indicating that these variants act synergistically, further increasing the risk of melanoma developing (Box et al., 2001; van der Velden et al., 2001).
PrASE-Based Analysis of Melanoma Variants (Paper VI)

An assay based on PrASE was established for typing 21 MC1R and 2 CDKN2A variants that have been suggested to affect risks of melanoma. The choice of MC1R variants was based on initial sequencing of 100 samples and previously published studies (Valverde et al., 1995; Borg et al., 1996; Moro et al., 1999; Flanagan et al., 2000; Bastiaens et al., 2001; Box et al., 2001; Hashemi et al., 2001; Jimenez-Cervantes et al., 2001; Fargnoli et al., 2003; Pastorino et al., 2004). PrASE was used to type these variants in samples from 42 healthy controls and 50 malignant melanoma patients. The extension reactions were performed using Taq DNA polymerase at an assay temperature of 62 °C. MC1R contains a single exon that was amplified by PCR, together with its promoter, in a 1326 bp long fragment. Since many of the MC1R variants were located in close proximity to each other, they could cause the extension primers to overlap, which could influence hybridization and extension. For this reason the variants were typed in two separate sets and pooled before hybridization to the tag-array. The CDKN2A variants were typed together with one of the sets.

The PrASE and Sanger DNA sequencing results were in complete concordance and no additional variants were found with DNA sequencing. In total, heterozygous samples were found for 16 of the variants typed and double mutants were found for eight variants. The calculated allele frequencies were close to those reported in previous studies, however the results should not be directly compared since the sample material here was chosen to contain as many variants as possible. The study shows the feasibility of using PrASE with Taq DNA polymerase and long PCR fragments.

MC1R Variation in the Swedish Population (Paper VII)

The assay described in Paper VI was used to genotype germline variants in Swedish cutaneous melanoma cases (1001 samples) and controls (721 samples). The melanoma samples were stratified into several groups: 158 multiple primary tumor samples, 298 hereditary melanomas from 178 different families, 485 sporadic melanomas, and 60 early onset (adolescence/childhood) melanomas. The healthy controls were also from several different groups: 202 newborns from Northern Sweden, 203 from a random selection from Southern Sweden and 115 healthy blood donors from the Stockholm region and 201 blood donors from whom phenotypic data were also collected.
Allele frequencies of the Swedish population were similar to what has been found in other Caucasian populations and confirms an association to increased melanoma risk for carriers of variants. Allele frequencies of MC1R in the case groups (and in all cases together as one group) were used to test for significant associations. A two to sevenfold increased risk was observed and odds ratios (ORs) were generally higher for the hereditary and multiple melanomas groups. Carriers of any two coding variants had a higher risk than carriers of one and carriers of RHC alleles had a further increased risk.

Associations of individual alleles were found between, mainly to the RHC alleles R151C and R160W, and the groups with sporadic, multiple and all melanomas, with odds ratios (ORs) of ~2.0 (ranging from 1.8-2.2, p<0.0001; 95% CI). There was also a significant association and a non-significant tendency towards association between the hereditary group and the alleles R151C and R160W, respectively. When analyzing all patients together, the non-RHC alleles V60L and R142H were weakly associated to CMM. The difference in allele frequencies between patients and controls of the V60L allele was not so pronounced but because this variant was relatively common in the study population it reached statistical significance. On the contrary, variants like D84E and D294H, that had a larger difference in allele frequency between the patient group and the control group, did not reach statistical significance maybe due to the low frequencies of these variants in the population.

The results are consistent with previous reports for Caucasian populations and these alleles are well established disease alleles (Sturm et al., 2001; Hayward, 2003). In the control group 66% had at least one variant and 21% had at least two variants. These numbers were significantly higher in all the case groups, except the early onset group, with ORs ranging from 2.1 to 4.0 (p<0.0001; 95% CI). Significances were calculated using Pearson χ² test, Fisher's exact test and Mantel-Haenszel χ² test in SAS 9.1.

Haplotype calculations were done with the expectation maximization algorithm in Arlequin. Most variants were only found in haplotypes with wt-alleles except the promoter variant -226 which was found with R160W and R163Q, and the silent mutation T314T which was found with V92M and I155T.
FUTURE PERSPECTIVES

In this thesis techniques for identification of genetic variations in several different contexts have been presented. The promising results in the area of resequencing with \textit{in situ} AMASE can be used on a larger scale. Paper I presents a proof-of-concept study for the resequencing of part of the \textit{p53} gene which is commonly mutated in human cancers. The assay can readily be expanded to sequence the entire \textit{p53} as a printed microarray with 40,000 spots allows 10 kb to be resequenced. With the \textit{in situ} synthesized arrays manufactured by Affymetrix and Nimblegene a much higher density is possible and thus resequencing of 600 kb is possible in one reaction. For instance, this could be used to resequence the coding parts of approximately 1000 genes for various applications.

A method for detecting and distinguishing human papillomavirus types has been developed (Papers II and III). This can be used to further investigate the transformation efficiency as well as monitoring the persistence of infection of different HPV types. The method has been very successful in detecting infections of several HPV types in single samples. The clinical consequences of multiple infections have not previously been fully investigated. The technique could be further improved to measure viral load by using an internal control of an amplified human DNA fragment. Other improvements are subtype classification of HPV 16, the most common high-risk type, and the use of general detection probes towards HPV types that are PCR amplified but currently not detectable.

Several accurate SNP genotyping assays based on PrASE have been set-up (Papers IV to VI). In Paper VI and VII, PrASE was used to type variants in the \textit{MC1R} gene. Certain variants of this gene are associated with an increased risk of attracting melanoma. 21 such variants were genotyped in the Swedish population and the distribution of alleles was found to be similar to that of other Caucasian populations. The PrASE assay is flexible and could easily be used for association studies of other cancer or complex disease genes. Also, to make the PrASE assay
more cost effective, amounts of reagents can be further optimized and reduced, which is a task not yet fully performed.

The results in Papers IV and V indicate that a higher assay conversion rate is possible than with plain ASE. These results are generally applicable and could be used on different ASE-based platforms such as both Illumina's assays or in microfluidic platforms such as the system developed by Gyros. Recently, the PrASE platform has been used with an alternative DNA amplification method denoted tri-nucleotide threading which has facilitated the amplification and consequent genotyping of 70 SNPs in one PrASE reaction. These results indicate that PrASE can be used to type many more SNPs in a single reaction on whole-genome amplified DNA.
ABBREVIATIONS

AF  allelic fraction
AMASE  apyrase-mediated allele-specific extension
APS  adenosine phosphosulphate
ARF  alternative reading frame
ARMS  amplification refractory mutation system
ASA  allele-specific amplification
ASE  allele-specific extension
ASH  Allele-Specific Hybridization
ASOH  allele-specific oligonucleotide hybridization
ASPCR  allele-specific PCR
ATP  adenosine triphosphate
bp  base pairs
CCD  charge coupled device
CDK4  cyclin dependent kinase 4
CDKN2A  cyclin dependent kinase inhibitor 2A
CI  confidence interval
Cy5  cyanine 5
DASH  dynamic allele-specific hybridization
dATP  2'-deoxyadenosine triphosphate
dATPαS  α-thio dATP
dCTP  2'-deoxyctydine triphosphate
dGTP  2'-deoxyriboguanosine triphosphate
DNA  deoxyribonucleic acid
dNMP  2'-deoxyribonucleotide monophosphate
dNTP  2'-deoxyribonucleotide triphosphate
dTTP  2'-deoxyribothymidine triphosphate
Gb  giga base pairs
HPV  human papillomavirus
kb  kilo base pairs
Arrayed Identification of DNA Signatures

- LCR: ligase chain reaction
- LD: linkage disequilibrium
- MALDI: matrix-assisted laser desorption ionization
- Mb: mega base pairs
- MC1R: melanocortin-1 receptor gene
- Mdm2: mouse double minute 2 homolog
- MS: mass spectroscopy
- MUCH: multiplex competitive hybridization
- OLA: oligonucleotide ligation assay
- OR: odds ratios
- PASA: PCR amplification of specific alleles
- PCR: polymerase chain reaction
- PPI: pyrophosphate
- PrASE: protease-mediated allele-specific extension
- RFLP: restriction fragment length polymorphism
- RNA: ribonucleic acid
- SBE: single base extension
- SBH: sequencing by hybridization
- SNP: single nucleotide polymorphism
- SSB: single-strand binding protein
- TOF: time of flight
- WG: whole genome
- WGA: whole genome association
ACKNOWLEDGEMENTS

Vad ska man säga, man harar runt i godan ro och så vips är det dags att skriva avhandling. Hur gick det här till? Ja, det hade alltid gått så här ling utan mina två fenomenala handledare Afshin och Joakim.

Afshin, mitt alla hjärtafläkte tack för allting, ditt oändliga tämligter, intresse och kunande om allt vad DNA kan ha för sig. Jag borde tro på allt du säger. Tack för allt gemensamt arbete med projekten, pekältande, avhandlingskrivande mm.


Jag vill också särskilt tacka följande som på något sätt har bidragit till denna avhandling:

Mathias, det är ju din anda som svävar över hela bygget som är DNAcorners, gör det unikt och sätter det på kartan. Det är trevligt att vara en del av det hela.

De goa professorerna (och assoc. prof) Per-Åke, Stefan och Sophia. Ni fem (med Mathias och Joakim) är som en superhjältegrupp: Speciella och med unika superkrafter. Per-Åke med sina varmkorvar och tänkarpöserna, Stefan som mina kompisar inte tror på att han är professor när de ser honom och cellospelande Sophia.

Sådana mina favvisskillar på labbet, Henrik, Olof, Erik och Björn, utan er hade jag inte stått ut.

Björn, det är så härligt att ha dig som piggar upp på labbet och förser med matsäck och ruffi. Vi har ju varit med om ett antal klassiker du och jag till exempel hotellincidenten i Boston och hamburgersutblandet på Island.

Erik (Skäggis), indiebammen som lever för Dire Straits och flätade, röd-svarta högtalarkablar. Härligt att ha dig att sura tillsammans med.

dr Heinrich, hur många lurrar har det blivit? Alla behöver ha en bra lunchkompis! Glömmer aldrig Satisfaction på megafon!

Ollis, energiknippet som har smörgås på möret och är ett verkligt föredöme för oss alla. Vi får gå och simma någon dag.

Emilie för gott samarbete med AMASE, MUCH-AMASE och PrASE. Också bra att ha någon att dela indignation med ibland. Erik P, trevligt att ha någon mer att prata Prase och annat med.

Veronica och Rainer för allt slät med MC1R, det har ni gjort bra. Det är så kul att någon annan använder det man har själv har jobbat fram.
Goofy-gängen: Maria Ehnis, Sophia, Afshin, Olof, Joakim och Pyroginget. Med många små steg framåt sällan och trevliga möten med goda lunchmackor kan man komma framåt!

Annelie för all hjälp med chipsen. Angela också förstås.

Kicki och Annika för hjälp på labbet med robotpippetterande mm. Jacob för prover. Mina samarbeten med Balack, Ason, Johan H, Suzanne, Keng-Ling.

Elitkvällar. Förstärkande som säger pretentioner är något dåligt.

Konferensresor med Maria (tack för bröllopsförorten), Hela Figgies Uppsalagång och cancergruppen som bara var glada att vara själva vid poolen under hudenkonferensen. Gissi och Esther med den oövarblödiga fotomaten i Malaga, Afshin och Nippe som uppskattade Bostons alla t-shirtaffärer.

Labassarna som hade del i att få mig att börja doktorera, Nader N och Tove Andersson.

Hela bandygänget. Henrik och Valtteri tjaade tills jag följde med och gjorde mina kompisar smutt chockade. Från riktiga oldtimers som Klubbknäckar-Peru, Martin, Magnusson, Henrik, Olof, Erik och Björn, till nyare förmågor som Tove Alm, Marcus, Emile, Rockan, Johan, Goran aka Jojie, Hanna, Sofia, Mårten, Sara, Jochen…

My och Mats, det har varit bra med två föregångare att prata avhandling med.

Tack alla trevliga prickar på Avd. som livat upp alla events på labbet, DNA hörnet, dispfester, julfester, musikkvällar.

Mina musketörer Johan, Peter och Andreas, tack för att ni är så bra. Trevliga damer har ni också, Frida, Anna och Karin.

Katarina, tack för allt! och för bilden.

Fler essentiella vänner, Gustav och Anna, Emma, Moberg, Viktor (kemi hade ju inte gitt utan dig).

Karl, le temps à Paris était vraiment quelque chose, beaucoup grâce à toi. J’espère que toi et Inge viens pour la fête!

Alla fester genom åren på Djurgårdslätten i vist och med balalajka tillsammans med Åsa Snås, Jonas, Johanna (Palle), Johanna (Ludde), Monika, Åsa K m.fl.

John, min bundsförvant så länge jag minns. Bästa tänkbara brorsan!

Mamma och pappa helt enkelt för att ni är så bra och alltid finns där.
REFERENCES

Arrayed Identification of DNA Signatures


Arrayed Identification of DNA Signatures


probe genotyping over 10,000 targeted SNPs genotyped in a single tube assay.

Max Käller


Hirschhorn, J.N. and Daly, M.J.: Genome-wide association studies for common diseases and

Hirschhorn, J.N., Lindgren, C.M., Daly, M.J., Kirby, A., Schaffner, S.F., Burtt, N.P., Alshuler,
in multiple populations reveals several regions with evidence of linkage to adult

Hirschhorn, J.N., Sklar, P., Lindblad-Toh, K., Lim, Y.M., Ruiz-Gutierrez, M., Bolk, S.,
Langhoorst, B., Schaffner, S., Winchester, E. and Lander, E.S.: SBE-TAGS: an array-
based method for efficient single-nucleotide polymorphism genotyping. Proc Natl

polymerase chain reaction product by utilizing the 5'----3' exonuclease activity of

Hollstein, M., Shomer, B., Greenblatt, M., Soussi, T., Hoving, E., Montesano, R. and Harris,
C.C.: Somatic point mutations in the p53 gene of human tumors and cell lines:


biotin-streptavidin interaction can be reversibly broken using water at elevated

Holmberg, K., Persson, M.L., Uhlen, M. and Odeberg, J.: Pyrosequencing analysis of


940-5.

Hughes, T.R., Mao, M., Jones, A.R., Burchard, J., Marion, M.J., Shannon, K.W., Leftkowitz,
microarrays fabricated by an ink-jet oligonucleotide synthesizer. Nat Biotechnol 19

Hugot, J.P., Chamaillard, M., Zouali, H., Le Sage, S., Cezeurd, J.P., Bekchie, J., Almer, S., Tysk,

polymorphism analysis with high throughput modifications, and its use in mutation
detection in familial hypercholesterolemia. International Federation of Clinical


Arrayed Identification of DNA Signatures


Arrayed Identification of DNA Signatures


