Methods for Analysis of Disease Associated Genomic Sequence Variation

BY

LOVISA LOVMAR
Dissertation presented at Uppsala University to be publicly examined in Rudbecksalen, Rudbecklaboratoriet, Uppsala, Friday, October 8, 2004 at 09:15 for the degree of Doctor of Philosophy (Faculty of Medicine). The examination will be conducted in English.

Abstract

In Molecular Medicine a wide range of methods are applied to analyze the genome to find genetic predictors of human disease. Apart from predisposing disease, genetic variations may also serve as genetic markers in the search for factors underlying complex diseases. Additionally, they provide a means to distinguish between species, analyze evolutionary relationships and subdivide species into strains.

The development and improvement of laboratory techniques and computational methods was a spin-off effect of the Human Genome Project. The same techniques for analyzing genomic sequence variations may be used independent of organism or source of DNA or RNA. In this thesis, methods for high-throughput analysis of sequence variations were developed, evaluated and applied.

The performance of several genotyping assays were investigated prior to genotyping 4000 samples in a co-operative genetic epidemiological study. Sequence variations in the estrogen receptor alpha gene were found to be associated with an increased risk of breast and endometrial cancer in Swedish women.

Whole genome amplification (WGA) enables large scale genetic analysis of sparse amounts of biobanked DNA samples. The performance of two WGA methods was evaluated using four-color minisequencing on tag-arrays. Our in-house developed assay and “array of arrays” format allow up to 80 samples to be analyzed in parallel on a single microscope slide. Multiple displacement amplification by the Φ29 DNA polymerase gave essentially identical genotyping results as genomic DNA. To facilitate accurate method comparisons, a cluster quality assessment approach was established and applied to assess the performance of four commercially available DNA polymerases in the tag-array minisequencing assay.

A microarray method for genotyping human group A rotavirus (HRV) was developed and applied to an epidemiological survey of infectious HRV strains in Nicaragua. The method combines specific capture of amplified viral sequences on microarrays with genotype-specific DNA-polymerase mediated extension of capture oligonucleotides with fluorescent dNTPs.

Keywords: microarray, molecular medicine, single nucleotide polymorphism, whole genome amplification, breast cancer, endometrial cancer, human rotavirus

Lovisa Lovmar, Department of Medical Sciences, Akademiska sjukhuset, Uppsala University, SE-75185 Uppsala, Sweden

© Lovisa Lovmar 2004

ISSN 0282-7476
ISBN 91-554-6027-5
urn:nbn:se:uu:diva-4525 (http://urn.kb.se/resolve?urn=urn:nbn:se:uu:diva-4525)
TILL MIN FAMILJ
SUPERVISOR: Ann-Christine Syvänen, 
Professor of Molecular Medicine 
Department of Medical Sciences 
Uppsala University 
Uppsala, Sweden

FACULTY OPPONENT: Pui-Yan Kwok, 
Professor of Dermatology, 
Henry Bachrach Distinguished Professor and 
Investigator 
Cardiovascular Research Institute 
University of California 
San Francisco, CA, USA

REVIEW BOARD: Jan Dumanski 
Professor of Experimental Pathology 
Department of Genetics and Pathology 
Uppsala University 
Uppsala, Sweden

Martin Schalling, 
Professor of Medical Genetics 
Neurogenetics Unit 
Center for Molecular Medicine 
Karolinska Institutet 
Stockholm, Sweden

Agneta Siegbahn 
Professor of Clinical Coagulation Science 
Department of Medical Sciences 
Clinical Chemistry 
Uppsala University 
Uppsala, Sweden
LIST OF PUBLICATIONS

This thesis is based on the following papers, which will be referred to in the text by their roman numerals:

I  Homogeneous scoring of single nucleotide polymorphisms: The 5’-nuclease "TaqMan" assay versus Molecular beacon probes
   Täpp I, Malmberg L, Rennel E, Wik M and Syvänen A-C.

II a) Estrogen receptor alpha gene haplotype and postmenopausal breast cancer risk: A case control study

   b) Estrogen receptor alpha gene polymorphism and endometrial cancer risk
   (Submitted)

III  Quantitative evaluation by minisequencing and microarrays reveals accurate multiplexed SNP genotyping of whole genome amplified DNA
    Lovmar L, Fredriksson M, Liljedahl U, Sigurdsson S and Syvänen A-C.

IV  Silhouette scores for assessment of SNP genotype clusters: Comparison of DNA polymerases in minisequencing on Tag-arrays.
    Lovmar L, Ahlford A, Jonsson M and Syvänen A-C
    (Submitted)

V  Microarrays for genotyping human group A rotavirus by multiplex capture and type-specific primer extension
    Lovmar L, Fock C, Espinoza F, Bucardo F, Syvänen A-C and Bondeson K.

In addition some unpublished data is included.

Originals were reprinted with the kind permission of Eaton Publishing, BioMed Central, Oxford University Press and the American Society for Microbiology.
ABBREVIATIONS

AFLP  Amplified fragment length polymorphism
ASO   Allele-specific oligonucleotide
ATP   Adenosine triphosphate
bp    Base pairs
CAHRES Cancer hormone replacement epidemiology in Sweden
CCD   Charge-coupled device
cDNA  Complementary DNA
CGH   Comparative genomic hybridization
CI    Confidence interval
cTag  Complementary/capture Tag
CV    Coefficient of variation
DNA   Deoxyribonucleic acid
dNTP  Deoxynucleoside triphosphate
ddNTP Dideoxynucleoside triphosphate
ESR1  Estrogen receptor alpha gene
FP    Fluorescence polarization
HRV   Human rotavirus
LD    Linkage disequilibrium
MB    Molecular beacon
MDA   Multiple displacement amplification
NASBA Nucleic acid sequence-based amplification
NCBI  National Center for Biotechnology Information
OR    Odds ratio
PCR   Polymerase chain reaction
PEP   Primer extension preamplification
PMT   Photomultiplier tubes
PoGD  Power of genotype discrimination
RFLP  Restriction fragment length polymorphisms
RNA   Ribonucleic acid
RT-PCR Reverse transcriptase PCR
SDA   Strand displacement amplification
SNP   Single nucleotide polymorphism
TQ    TaqMan
WGA   Whole genome amplification
# CONTENTS

**THE GENOME** .......................................................................................................................... 9
  1  GENETIC VARIABILITY ............................................................................................................. 10
      1.1 Single nucleotide polymorphisms ..................................................................................... 11

**ANALYSIS OF SEQUENCE VARIATIONS** ............................................................................. 13
  2  ASSAYS ..................................................................................................................................... 14
      2.1 Hybridization with ASO probes ........................................................................................... 14
      2.2 Primer extension ................................................................................................................... 16
      2.3 Minisequencing .................................................................................................................... 17
      2.4 Ligation and enzymatic cleavage .......................................................................................... 18

**FORMATS** ............................................................................................................................... 19
  3  3.1 Microarrays .......................................................................................................................... 20

**LABELING AND DETECTION** .................................................................................................. 22
  4  4.1 Fluorescence detection .......................................................................................................... 23

**TEMPLATES FOR GENOTYPING** ........................................................................................... 26
  5  5.1 Multiple displacement amplification ...................................................................................... 27
      5.2 Primer extension preamplification ......................................................................................... 28

**POLYMERASE CHAIN REACTION** ........................................................................................... 30
  6  6.1 Reverse transcriptase PCR ..................................................................................................... 30
      6.2 Multiplex PCR ....................................................................................................................... 30

**ALTERNATIVE AMPLIFICATION STRATEGIES** ................................................................... 31
  7  7.1 Complexity reduction without PCR ....................................................................................... 31
      7.2 Signal amplification ................................................................................................................. 32

**GENETIC VARIABILITY AND HUMAN DISEASE** ................................................................. 34
  8  8.1 Linkage analysis ..................................................................................................................... 35
      8.2 Association and linkage disequilibrium ................................................................................ 35
      8.3 Haplotypes and the HapMap project ..................................................................................... 36

**MICROORGANISMS AS INFECTIOUS AGENTS** ................................................................. 38
  9  9.1 Genotyping microorganisms by microarrays ....................................................................... 38
      9.2 Vaccine studies and human rotavirus ................................................................................. 39

**PROLOGUE TO THE PRESENT STUDY** ............................................................................... 40
PRESENT STUDY .............................................................................................. 43
1 AIMS ........................................................................................................... 43
2 MATERIALS & METHODS ........................................................................ 43
  2.1 Study samples ...................................................................................... 44
  2.2 Assay design ....................................................................................... 45
  2.3 Whole genome amplification ............................................................... 46
  2.4 Microarray preparation ....................................................................... 47
  2.5 Genotyping ......................................................................................... 47
    2.5.1 Assay quality assurance ............................................................... 50
    2.5.2 Genotype scoring and signal analysis .......................................... 52
    2.5.3 Genotype discrimination and cluster quality ............................... 53
  2.6 Statistical analyses ............................................................................. 54
3 RESULTS AND DISCUSSION .................................................................... 55
  3.1 Assay performance ........................................................................... 55
    3.1.1 Summarized results on assay performance .................................. 55
    3.1.2 Homogeneous assays ................................................................. 55
      TaqMan assay versus Molecular beacon probes ............................... 55
    3.1.3 Semi-homogeneous assays .......................................................... 58
      Minisequencing using fluorescence polarization detection ............. 58
      WGA samples and minisequencing with FP detection ................. 59
    3.1.4 Microarrays ............................................................................... 60
      Silhouette scores for cluster quality assessment ............................. 61
      WGA samples and minisequencing on Tag-arrays ....................... 62
      Extension enzymes in minisequencing on Tag-arrays ..................... 64
      Genotype-specific primer extension on arrays for HRV genotyping .... 66
  3.2 Applications ....................................................................................... 67
    3.2.1 ESR1 polymorphisms in breast cancer ........................................ 67
    3.2.2 ESR1 polymorphisms in endometrial cancer ............................... 68
    3.2.3 Rotavirus genotyping .................................................................. 68
EPILOGUE ...................................................................................................... 70
EXPRESSIONS OF GRATITUDE .................................................................... 71
REFERENCES ............................................................................................. 73
The genome of an organism serves as central storage for the genetic information that is transmitted between generations. Even though species may be completely different on a phenotypic level, the fundamental genetic principles and other functions have remained remarkably conserved throughout evolution and form the basis for all life on earth (Cromie et al. 2001, Lander et al. 2001). The universality of genetic material is well illustrated by viruses, being obligate intracellular parasites which combine their own features with those of infected cells to translate, transcribe and replicate, their genomes thus multiplying and spreading themselves.

According to the classical “central dogma of molecular biology” the expression of genetic information is mediated through the transcription of deoxyribonucleic acid (DNA) into ribonucleic acid (RNA). The RNA is subsequently processed and translated into protein (Figure 1.).

![Figure 1. The central dogma of molecular biology. The universal steps of replication, transcription and translation are illustrated by solid lines whereas alterations such as reverse transcription and RNA replication are shown as dashed lines.](image)

Translation is performed by macromolecular complexes reading nucleotide triplets (codons) that direct the addition of a specific amino acid to the growing polypeptide. The central dogma is valid for most organisms; however, alterations in the steps of this principle are known and different classes of genomes have been found (Table 1). One example is the human mitochondrial genome. Each human cell typically contains thousands of copies of the double-stranded circular mitochondrial genome. The small number of proteins encoded by the mitochondrial genome has allowed the mitochondrial genetic code to drift from the nuclear code and for example two nuclear amino acid codons function as stop codons in the mitochondria.

While the principal functions of the other 98.5% remain to be elucidated, the proportion of the protein-coding sequence declines with developmental complexity and occupies only 1.5% of the genome in humans. RNA molecules have been suggested to play important roles in the function of higher organisms (Mattick 2004).
Table 1. Different classes of genomes

<table>
<thead>
<tr>
<th>DNA</th>
<th>RNA</th>
</tr>
</thead>
<tbody>
<tr>
<td>ds</td>
<td>ss</td>
</tr>
<tr>
<td>Single circular molecule</td>
<td>Bacteria, archa, mitochondria, chloroplasts, viruses</td>
</tr>
<tr>
<td>Single linear molecule</td>
<td>Bacteria, viruses</td>
</tr>
<tr>
<td>Multiple circular molecule</td>
<td>Bacteria, viruses</td>
</tr>
<tr>
<td>Multiple linear molecule</td>
<td>Eukaryotic nuclei, viruses</td>
</tr>
<tr>
<td>Mixed linear and circular</td>
<td>Bacteria</td>
</tr>
</tbody>
</table>

1 Table adapted from Strachan and Read, 2004. ds = double-stranded, ss = single-stranded

In 1953 James Watson and Francis Crick presented the double helix structure of the DNA molecule (Watson and Crick 1953) after which the unraveling of the genetic code started. The first fully sequenced DNA genome was that of the ΦX174 bacteriophage (Sanger et al. 1977a) with approximately 5,400 base pairs (bp). Since then the genomic sequences of an increasing number of organisms have been completed. The first free-living organism to be sequenced was Haemophilus influenzae (Fleischmann et al. 1995). Through the parallel efforts of the Human Genome Project (Lander et al. 2001) and the Celera company (Venter et al. 2001) the draft sequence of the human genome was published. Subsequently the quality of the information has improved and the reference sequence for 99% of the 3 billion base pairs in the human genome was made available in April 2003, with an accuracy of 99.9%.

The effort to obtain more accurate and thorough information about the human genome sequence as well as of the sequences of other organisms is continuing. At the time of writing, the complete sequences of 168 bacteria, 19 archa, 20 eukaryotes and approximately 1,300 viruses were publicly available on the website of the National Center for Biotechnology Information (NCBI Genome, http://www.ncbi.nih.gov/Entrez/; August 8th 2004). In parallel with the sequencing efforts, an increasing interest has turned towards the elucidation of genetic variability to gain information on inter- and intra-species sequence variability.

1 GENETIC VARIABILITY

Intra-species sequence variability arises when a genomic sequence has been altered between generations. In a diploid organism, such as humans, this may occur either by transmission of a new mutation that has occurred for example by replication errors in germline cells (premeiotic mutations) or by recombination between the two parental genomes during oogenesis or...
spermatogenesis (meiotic mutations). Recombination may occur either between homologous regions; maintaining the general order of genes, or it may be site-specific; moving genetic elements or adding new information. Site-specific recombination takes place with transposable elements such as the retroviral-like retrotransposons. More than 50% of the human genome is composed of repeated elements that are mainly derived from transposable elements (Lander et al. 2001). Additionally, haploid organisms may utilize recombination. For example, homologous recombination in *Escherichia coli* is mediated through the enzymes encoded by the *rec* genes and is extensively used in recombinant DNA engineering. New viral types, *i.e.* genetic alterations, may arise through recombination between viruses but also between virus and host. If the virus carries a segmented genome, reassortment strains may arise from co-infected cells as when new influenza A virus infections arise.

The human genome has been found to be on average 99.9% similar between individuals. However, the variable part of the genome still covers three million bp and contributes to all genetic differences found between individuals (Li W.H. and Sadler 1991, Sachidanandam et al. 2001, Reich et al. 2002). Despite the large total number of sequence variations in humans, the majority of inter-individual genetic variability is inherited. This is due to the low mutation rate in mammalian genomes, estimated to be on average $2 \cdot 10^{-9}$ per bp and year (Kumar and Subramanian 2002). The mutation rates in mammals is in strong contrast to the viral RNA genomes that generally have a high mutation rate due to error prone RNA replication or reverse transcription. The high mutation rate of viral RNA genomes gives the evolutionary advantage of fast adaptation to changing environments *e.g.* the introduction of a vaccine, see section 9.2. Furthermore, the mitochondrial genome is highly mutational compared to the nuclear, probably a result of a more error prone replication and a significantly higher number of replications. However, on the contrary mitochondria do not have recombination.

Apart from directly causing or contributing to human disease, some genetic variations, or polymorphisms, are valuable as genetic markers in the search for genetic factors underlying complex diseases, see further section 8. The analyses of sequence variation provide a means to distinguish between species, to analyze evolutionary relationships, and to subdivide species in strains. In addition the comparison between species is also a powerful means to find important genomic regions as indicated by their conservation throughout evolution (Harris et al. 2003).

### 1.1 Single nucleotide polymorphisms

The first molecular polymorphism described in humans was the blood typing AB0 system, which awarded Karl Landsteiner with the Nobel Prize in Physiology or Medicine 1930. Since then a large number of genetic variation has been reported. The most abundant sequence variations are the single nucleotide polymorphisms (SNPs). SNPs occur on average once every 1,300 bp, but the density of SNPs varies extensively between different regions of the
genome (Sachidanandam et al. 2001, Marth et al. 2003). There has been a large effort to identify and make publicly available a complete list of the SNPs in the human genome during the last few years. Several databases have been set up to help researchers find SNPs in the gene of interest. Almost ten million SNPs, approximately half of them validated, are now present in the public SNP database of NCBI (dbSNP, http://www.ncbi.nlm.nih.gov/SNP/; build 121, June 1st 2004).

Depending on their genomic location, SNPs have different impact at the phenotypic level. SNPs in protein coding regions of a gene may be synonymous, resulting in an unchanged amino-acid, or non-synonymous, potentially altering the structure or function of the protein. SNPs in regulatory regions of a gene may alter the expression of the protein thus having an impact on the phenotypic level. Since only about 1.5% of the human genome constitutes of protein coding DNA (Lander et al. 2001, Levine and Tjian 2003), the largest abundance of SNPs are found in the non-protein coding regions and the functions of these SNPs are yet to be explored. However, even though most SNPs are not directly disease causing, they have the potential of serving as genetic markers in linkage and association studies, see section 8.

One of the spin-off effects of the Human Genome Project was the development and improvement of laboratory techniques and computational methods for rapid and accurate sequencing of large quantities of clones (Lander et al. 2001). Since the genome of an organism, be it virus or man, is constructed from common building blocks, the same techniques for analyzing sequence variations may be used independent of the original source of DNA or RNA.
ANALYSIS OF SEQUENCE VARIATIONS

In 1978 the first discovery of linkage between a disease causing mutation and a polymorphism in a non-coding DNA sequence was found between a mutation adjacent to the human β-globin gene and the sickle cell mutation (Kan and Dozy 1978). This discovery was preceded by the use of type II restriction enzymes (Kelly and Smith 1970, Smith and Wilcox 1970) in combination with size separation via gel electrophoresis (Danna and Nathans 1971) and hybridization to detect specific sequences (Southern 1975). The proof of linkage was an important breakthrough for the concept of using polymorphisms as genetic markers and the process also illustrates how SNP genotyping was performed in the early years. A range of important inventions and assay improvements have been made since then. Of perhaps largest important was the Nobel-prize awarded cell-free DNA cloning method of polymerase chain reaction (PCR), which enabled specific amplification of a selected region, providing the required sensitivity and specificity for a range of assays, see section 6. However, an emerging challenge is to directly investigate single molecules and a brief discussion of such techniques is given in section 8.3.

To interrogate the nature of a given nucleotide, different reaction principles, formats, detection and labeling strategies can be combined. The combination of these depends on the requirements of each specific target application, the available technology and the financial frame, for reviews see Kwok, 2001 and Syvänen, 2001. The range of these features as illustrated in Figure 2 will be discussed in the following sections.

Figure 2. Several reaction principles, formats, and detection & labeling strategies can be combined in SNP genotyping assays.
2 ASSAYS

During the 1990’s, following the invention of PCR, a large number of alternative reaction principles to genotype single nucleotide polymorphisms were developed. Six central allele-specific reaction principles exist: (i) hybridization of oligonucleotides i.e. ASO probes, (ii) primer extension, (iii) single nucleotide incorporation i.e. minisequencing, (iv) ligation, (v) invasive cleavage and (vi) restriction enzyme cleavage (Figure 3). The first three principles have been applied in the present study and are discussed in more detail.

![Diagram of reaction principles for SNP genotyping](image_url)

**Figure 3.** Reaction principles for SNP genotyping. An A-to-G transition is illustrated. All needed probes/primers are included but the assay outcome is only shown for one probe/primer. Figure adapted from Syvänen, 2001.

### 2.1 Hybridization with ASO probes

Genotyping with ASO probes relies on the ability to monitor the formation or denaturing of a probe-template duplex and the discrimination between genotypes relies on the power to detect differences inferred by the single alternate base (Wallace *et al.* 1979, Conner *et al.* 1983). No enzyme is involved in the allele discriminating step, making the mechanism the simplest of those presented.

An advantage of ASO mediated genotyping is the possibility to easily perform homogeneous assays allowing the genotypes to be deduced directly in the PCR tube without separation steps, see below section 3. The 5’nuclease assay, *i.e.* the TaqMan (TQ) assay (Livak *et al.* 1995), utilizes two probes, one for each allele, labeled with fluorescent dyes at both ends. The fluorophores are quenched through fluorescence resonance energy transfer (section 4.1).
when hybridized to their targets. Fluorescence is restored during the PCR extension step as the 5'-3' exonuclease activity of the DNA polymerase displaces and degrades the probe (Figure 4, panel A and C).

In the equally homogeneous Molecular beacon (MB) assay, the fluorescently labeled allele-specific probes are designed to have the central part complementary to the target sequence and to carry “self-complementary” arm-sequences on each end (Tyagi and Kramer 1996, Tyagi et al. 1998). In contrast to TQ probes, MB probes are designed to hybridize to their targets at a different temperature than in the annealing and extension steps of the PCR.

**Figure 4.** The principle for the TaqMan (TQ) (panel A) and Molecular beacon (MB) (panel B) assays. For each PCR cycle of the TQ assay more template molecules are formed and more matched and hybridized probes will be degraded causing the fluorescence to increase. The fluorescence is monitored on a real-time PCR machine (C). The fluorescence increase for MB probes with each PCR cycle can be monitored analogously to the TQ assay (D). The position of the variable nucleotide is indicated by a thicker line, F = fluorophore, Q = quencher. The results are shown for both allele-specific probes and parallel reaction with one sample of each genotype.
reaction; hence, the MB probes are not degraded by the DNA polymerase. When in solution and in absence of the target, the MB probes have a hairpin structure that brings the fluorophore and quencher (usually the non-fluorescent chromophor Dabcyl) very close together so that no fluorescence is emitted (Tyagi et al. 1998) (section 4.1). The hairpin formation also destabilizes mismatched probe-target hybrids which increases the power of the allelic discrimination. When the MB probe binds to its target the fluorophore becomes more distant from the quencher and fluorescence is emitted. Since the quencher is non-fluorescent the fluorescence is not quenched by FRET as with TQ probes. When an extra annealing step for the MB probes is added during the PCR, the amount of targets complementary to the probes can be followed for each cycle by monitoring the increased fluorescence (Figure 4, panel B and D).

Molecular beacons have been applied to study mRNA expression in vivo (Bratu et al. 2003, Nitin et al. 2004), and are also widely used for the detection of microorganisms, e.g. in combination with NASBA (Leone et al. 1998), see section 9.

A different ASO strategy is to monitor denaturation of hybrids over a temperature range to achieve a more robust genotyping, as is done in the dynamic allele-specific hybridization assay (Howell et al. 1999, Prince et al. 2001). Also, more sophisticated design algorithms, hybridization enhancing moieties (Afonina et al. 1997, Kutyavin et al. 2000), structural analogues (Kuimelis et al. 1997) and peptide or locked nucleic acids with a high affinity to form complementary DNA (Griffin et al. 1997, Ortiz et al. 1998, Orum et al. 1999, Svanvik et al. 2000) have been used to improve specificity.

Multiplexing with ASO probes is difficult, mainly because the thermal stability of the hybrid depends on the sequence flanking the SNP and the secondary structures of the target (Conner et al. 1983, Mir and Southern 1999). A low level of multiplexing is possible by using probes with different labels (Lee et al. 1999, Tyagi et al. 2000). The problem of multiplexing with ASO probes is overcome by a different strategy in which a large probe set for every interrogated SNP is immobilized on high-density microarrays (Chee et al. 1996).

2.2 Primer extension

Allele specific primers can be utilized to either directly amplify the target sequences by PCR (Gibbs et al. 1989, Newton et al. 1989, Wu et al. 1989, McClay et al. 2002) or to detect the variable nucleotide in an already amplified sample. Both strategies rely on the ability of the DNA polymerase to extend a primer only if the 3'-end has a perfect match. A benefit of using primer extension is the cost reduction achieved since labeled probes are not necessary; however, size separation, intercalating dyes or the incorporation of fluorescent deoxynucleotides (dNTPs) are required to visualize the genotyping result. The selectivity of the reaction may be increased by using allele-specific primers with an intentional mismatched nucleotide at the penultimate 3'-end nucleotide position (Li B. et al. 2004). Molecular beacon
probes have been altered to serve as primers in allele-specific PCR (Nazarenko I.A. et al. 1997, Whitcombe et al. 1999). The multiplexing problems are the same as for the ASO probes; thus the strategy is preferably used for large numbers of samples, but only a few interrogated variations. If performing the allele-specific primer extension on PCR amplified template, two "levels of specificity" are obtained. This strategy has been used with gel-pad arrays and microarrays (Dubiley et al. 1999, Erdogan et al. 2001). A different strategy is to use PCR primers with T7 RNA polymerase promoter sequences in a first amplification followed by a second linear amplification with T7 primers yielding RNA. The RNA is then hybridized to allele-specific immobilized primers that are extended by reverse transcriptase in the presence of labeled nucleotides which allow the genotypes to be deduced (Pastinen et al. 2000). The advantage of this strategy is that additional copies of single stranded template molecules that are more accessible for hybridization are generated. Another approach to increase the reaction sensitivity is to include the enzyme apyrase, which inactivates the nucleotides and competes with the allele-specific primer extension action of the DNA polymerase (Ahmadian et al. 2001).

Furthermore, allele-specific primer extension is applicable to sequence analysis of microorganism genomes. However, “sequence-specific” or “genotype-specific” is a more suited terminology because viruses and bacteria do not have diploid genomes. Primers for genotype-specific PCR can be designed to match different strain-(genotype) specific regions; thus, the presence of product indicates a specific strain. Multiplexed hemi-nested reverse transcriptase PCR has been applied to genotype rotavirus (Gouvea et al. 1990). In Study V of this thesis a new approach, applying genotype-specific primer extension post PCR, was used. Primers designed to be genotype-specific, especially at the 3'-nucleotide, were immobilized on a microarray to direct a strain-specific primer extension (Lovmar et al. 2003a).

2.3 Minisequencing
The incorporation of single allele-specific nucleotides, i.e. minisequencing or single base primer extension (Syvänen et al. 1990) is more specific than allele-specific primer extension and less dependent on reaction conditions than ASO probes. The reaction principle employs the ability of DNA polymerase to specifically extend a detection primer with nucleotides directed by the complementary strand. The detection primers are designed to end directly upstream of the SNP position, in contrast to allele-specific primer extension where the primers cover the SNP position.

Only one detection primer is needed per reaction in minisequencing compared to the two allele-specific oligonucleotides needed for hybridization or primer extension. This renders the minisequencing reaction independent of the nature of the variable nucleotide. Furthermore, this makes the reaction principle suitable for multiplexing when combined with the specificity of the DNA polymerase. A microarray based minisequencing assay reported
approximately one order of magnitude better genotype discrimination in a side-by-side comparison with ASO-probes (Pastinen et al. 1997).

In minisequencing, the DNA polymerase is the most important factor that governs the efficiency and specificity of the genotyping reaction. The reaction conditions in minisequencing reactions are often not optimal, e.g. when only two of the four nucleotides are present or modified nucleotides are used. If not all nucleotides are present, this causes the DNA polymerase to mis-incorporate nucleotides with a frequency much higher than the fidelity as measured under optimal conditions. Mis-incorporation properties may be improved by combing a 3’-5’ exonuclease proofreading DNA polymerase with protected minisequencing primers (Di Giusto and King 2003). When incorporating modified nucleotides, e.g. fluorescently labeled dideoxy-nucleotides (ddNTPs) (Pastinen et al. 1996) or acyclo-nucleotides (Trainor 1996), different enzymes, such as mutants of Vent and Taq DNA polymerases, have different incorporation efficiencies and preferences, probably due to different recognition of the sugar residues (Gardner and Jack 2002). Several DNA polymerases have been modified to incorporate ddNTPs and dNTPs with equal efficiency as required in Sanger sequencing (Sanger et al. 1977b). The ThermoSequenase DNA polymerase was originally engineered with a Phe667Tyr substitution to incorporate ddNTPs with a approximately two-fold preference over dNTPs (Tabor and Richardson 1995). Four enzymes optimized for Sanger sequencing were evaluated with regard to their performance in minisequencing reactions in Study IV of this thesis.

The Pyrosequencing approach is related to minisequencing. Several single base extensions are performed and detected sequentially, deducing the identity of several nucleotides following the primer (Ronaghi et al. 1996, Alderborn et al. 2000), see section 4.

The minisequencing principle has been shown to give excellent quantitative performance, especially with tritium-labeled nucleotides or when using mass detection (Syvänen et al. 1993, Buetow et al. 2001). These features are useful for determination of allele fractions in pooled DNA samples (Olsson et al. 2000, Matyas et al. 2002, Werner et al. 2002), to quantify chimerism (Fredriksson et al. 2004) or differentially expressed alleles (Karttunen et al. 1996, Pastinen et al. 2004).

2.4 Ligation and enzymatic cleavage

The oligonucleotide ligation assay relies on the power of DNA ligase to discriminate between a perfect match and a mismatched base at the junction between two primers (Landegren et al. 1988). The padlock probe is a development of this assay in which a single probe that is circularized by ligation upon hybridization of its two ends to the target molecule is used (Nilsson et al. 1994, Nilsson et al. 1997). Molecular beacons have been used to monitor ligation events in real-time PCR reactions (Tang et al. 2003). The ligation event is highly specific and has the power to genotype without prior target amplification, see section 7.
Invasive cleavage is applied in the invader assay (Figure 3). An allele-specific signaling probe with a 5´ non-target complementary flap is used together with an invader probe complementary to the sequence upstream of the variable nucleotide. A perfectly matched allele-specific probe, together with the invader probe, generates a three-dimensional structure recognized as a cleavage site by the 5´-FLAP endonuclease. The flap sequence is released and can be used to determine the genotype (Lyamichev et al. 1999) or as a template in a second signal amplification step (Hall et al. 2000). The invader assay has the potential to be multiplexed on solid-surfaces (Wilkins Stevens et al. 2001, Olivier et al. 2002) and can be used for genotyping without PCR amplification, see section 7. However, large amounts of DNA are needed and the assay has mostly been performed with PCR amplified DNA as the target (Mein et al. 2000).

An early and still widely used method to characterize sequence variations, especially if the complete sequence is unknown, is the use of restriction endonucleases for allele-specific cleavage of the target molecule at a recognition site. A target molecule with an altered recognition site introduced by the SNP or sequence variation remains uncleaved, or becomes cleavable. Traditionally, restriction enzyme cleavage has also been used for SNP discovery.

3 Formats

Most biochemical reactions benefit from being performed in solution allowing the components to interact freely. However, a solid support enables highly paralleled analysis of both samples and reactions; thus increasing throughput and saving time and money. Solid support also enables thorough washing and an easy change of reactants and reaction conditions. A major distinction of formats can be made between homogeneous and heterogeneous. The choice of format is determined by the available laboratory techniques, throughput required and economic limits.

Homogeneous formats lack separation steps and therefore allow the genotyping reactions to be performed and analyzed in the same test tube. This minimizes the laboratory handling and the risk of contamination between samples. In addition to the TQ and MB assays, discussed in section 2.1, formation of PCR products can also be analyzed directly in homogeneous format by melting curve analysis (Gundry et al. 2003).

There are a wide range of different heterogeneous assays most of which apply solid phase immobilization of target, template or a generic identification sequence. Microarrays have been a main focus during this thesis work and are extensively covered in the following section. Other solid-phases are microtiter plates; combined for example with minisequencing (Syvänen et al. 1993, Fredriksson et al. 2004) or ligation assays (Nickerson et al. 1990), and micro particles; combined with for example minisequencing (Syvänen et al. 1990, Cai et al. 2000) or the invader assay (Rao et al. 2003).
The semi-homogeneous is an “in-between format” where no separation is done and no solid-phase is involved but reagents are added at different reaction steps. The main reactions of pyrosequencing is performed in this format, with new nucleotides and enzymes being added after each sequencing step and the sequence flanking the primer being deduced base by base. However, the PCR-products in pyrosequencing are separated by solid-phase methods (Alderborn et al. 2000). Minisequencing (Chen et al. 1999), ligation assays (Faruqi et al. 2001), and the invader assay (Mein et al. 2000, Hsu et al. 2001b) can readily be performed in a semi-homogeneous format.

3.1 Microarrays
The use of microarrays has expanded rapidly in many fields of research since the idea was first presented in the early 1990’s (Southern et al. 1992, Drmanac et al. 1993). Microarrays is a miniaturized solid-support format carrying sets of ordered, immobilized molecules which facilitating parallel analysis of reactions. Glass surfaces are most commonly used as solid-support, but nylon membranes, magnetic beads, polyacrylamide gel pads etc. can also be used.

Today the RNA-expression arrays are dominating, but microarray technology is also extensively used for DNA analysis and is gaining importance for tissue, antibody, protein and cell applications. Comparative genomic hybridization (CGH) on arrays carrying genomic clones, complementary DNA (cDNA) clones, PCR products or oligonucleotides as the immobilized hybridization targets have allowed the detection of DNA copy number aberrations at a much higher resolution than previously possible using metaphase chromosomes, as reviewed by Mantripragada et al., 2004. Furthermore, in microorganism genotyping microarrays have contributed new perspectives allowing several strains or species to be analyzed for large regions or whole genomes in parallel, see section 9.1.

Microarrays are widely used in SNP genotyping and both in situ synthesis (Chee et al. 1996, Hacia et al. 1998) and immobilization of pre-synthesized oligonucleotides (Lindroos et al. 2001) are applied. If primer extension or minisequencing is to be performed on the immobilized molecules, a free 3’-end of the minisequencing primer is required and is usually accomplished by either using pre-synthesized oligonucleotides. It is also possible to perform a in situ 5’-3’ synthesis directly on the glass surface (Beier and Hoheisel 2002) or inverting 3’-5’ synthesized oligonucleotides (Kwiatkowski et al. 1999), the latter having the additional positive effect of eliminating truncated oligonucleotides.

Tag-arrays The concept of using “tagged” primers was first applied to identify new virulence genes in bacteria (Hensel et al. 1995) and to perform quantitative phenotypic analysis by directly labeling yeast deletion mutants (Shoemaker et al. 1996). Tag-arrays were initially applied to SNP genotyping using a ligase assay (Gerry et al. 1999). SNP genotyping based on minisequencing using either high-density oligonucleotides arrays with two-
color ddNTPs (Fan et al. 2000) or medium-density oligonucleotide arrays with three-color ddNTPs (Hirschhorn et al. 2000) to capture the extended and tagged minisequencing primers from cyclic minisequencing reactions performed in solution were then published. At the same time the principle was applied to fluorescent microparticles carrying Tag-sequences (Cai et al. 2000).

Tag-sequences of 20 bp length can theoretically give $4^{20}$ or $\sim 10^{12}$ unique probe combinations. The Tag-sequences should be designed to have limited homology with each other and with the assayed organism (Hirschhorn et al. 2000). The Tag-sequences should also be similar in thermodynamic properties (Fan et al. 2000). An advantage with the Tag-array approach compared to immobilized primers is its universal or generic feature, making it possible to use the same array design to interrogate any reaction if the product carries the Tag-sequences.

“Array of arrays” One of the major advantages of microarray technology is the parallel analysis. This is further emphasized if utilizing an “array of arrays” format, which enables several samples to be analyzed in parallel on individual subarrays that together form a microarray. The concept was first presented for allele-specific primer extension (Pastinen et al. 2000). Since then the format has been used in ligation- and apyrase- based assays (Baner et al. 2003, Gharizadeh et al. 2003), and devices for the “array of arrays” format has been commercialized. The “array of arrays” format is especially well suited for comparisons between samples, evaluating assay performance or doing quantitative analysis (Lindroos et al. 2002, Lovmar et al. 2003b). See Figure 5.

![Figure 5. Principle of minisequencing with tag-arrays in the “array of arrays” format. Either 80 (A) or 16 (B) subarrays can be arrayed on a standard microscope slide. Minisequencing is performed in solution with labeled ddNTPs and tagged minisequencing primers (C). The extended primers for all samples simultaneously are hybridized to their corresponding arrayed cTags and the genotypes can be deduced by measuring the fluorescence. For the sample (subarray) enlarged, SNP 1 is homozygous AA and SNP 2 is heterozygous CT (D). © Humana Press](image-url)
Labeling and detection

Detection was based primarily on radioactivity and/or gel electrophoresis in the early years of nucleic acid assays (Syvänen et al. 1993). Today, fluorophores are the dominating label and most common detection strategy, but detection of mass, luminescence and electrochemical changes are also employed. Fluorescence has been the main detection method in the present study.

Radioactivity and mass detection have the advantage of applying only slightly or totally unmodified nucleotides, minimizing the adverse effects inferred by bulky labels. This facilitates a good quantitative measurement since all molecules have essentially the same features which determine their behavior in signal detection (Olsson et al. 2000, Ross P. et al. 2000, Fredriksson et al. 2004). Mass spectrometry is perhaps the ideal detection since no modifications are introduced. However, it requires advanced instrumentation and either several assay steps or combinations of dNTPs need to be incorporated to achieve enough resolution to be able to detect single base alterations (Braun et al. 1997, Griffin et al. 1997, Ross P.L. et al. 1997, Sauer et al. 2000, Blondal et al. 2003). When combining mass spectrometry with the invader assay (see section 2.4), flaps of different length can be used to identify the presence of multiple interrogated alleles (Griffin et al. 1997).

In pyrosequencing luminescence is employed to monitor the reaction. The inorganic pyrophosphates formed upon a primer guided DNA polymerase mediated dNTP incorporation are converted to adenosine triphosphate (ATP) by firefly sulfurylase. The ATP subsequently fuels a luciferase reaction and luminescence can be detected (Nyren et al. 1993, Ronaghi et al. 1996). When sequencing several bases, apyrase is added to inactivate the remaining nucleotides and ATP before the addition of the next nucleotide (Alderborn et al. 2000). A drawback of the method is the need for multiple enzymes.

A still very young detection strategy, with several features that need to be refined, is to use electrochemical changes to monitor hybrid formation. Thus far the principle has been applied to solids-supports with oligonucleotides immobilized on microelectrodes (Wang J. et al. 1997) and to detect hybridization of nanoparticle linked probes to targets immobilized between microelectrodes (Park et al. 2002).

The format of the assay has consequences for the choice of label. Since no separation is performed in homogeneous formats, the incorporation of a label must infer a change when the allele-specific reaction has occurred. Fluorescence based detection is well suited because fluorophores can interact to reflect these molecular changes. This may be monitored for example through fluorescence resonance energy transfer or fluorescence polarization, see the following section. In heterogeneous formats the possibility to wash the solid support puts fewer demands on the applied label.
4.1 Fluorescence detection

Fluorophores are applied as labels in different applications and are easily detected using instrumentation such as fluorescence microscopes or fluorescence scanners with charge-coupled device (CCD) cameras or photomultiplier tubes (PMT). The most straightforward strategy is to simply label either probes or incorporated nucleotides and measure the fluorescence after a washing or separation step (Pastinen et al. 1996, Gerry et al. 1999, McClay et al. 2002). If using four differently labeled nucleotides, multiplex minisequencing reactions give the possibility to simultaneously analyze a diverse set of SNPs (Kurg et al. 2000, Lindroos et al. 2002). Instead of using labeled probes or nucleotides, indirect labeling can be used where the target sequence is biotin-labeled and a fluorescent streptavidin conjugate is bound prior to fluorescence detection (Chee et al. 1996).

Intercalating dyes, such as ethidium bromide, PicoGreen or SYBR Green, which fluoresce only when bound to double-stranded DNA, is the simplest fluorescence detection strategy and commonly used in gel electrophoresis. ASO assays are readily analyzed with intercalating dyes (Howell et al. 1999); however, there are some inherent problems: unspecific DNA hybridization may cause background, the need for single stranded templates which complicates the separation procedure and the use of a single dye which means that separate reactions must be performed for the different alleles. Intercalating dyes have also been used with allele-specific PCR (Moran et al. 1998, Germer and Higuchi 1999) and are commonly used to quantitate DNA or for measuring expression levels with real-time quantitative PCR (Higuchi et al. 1993, Wittwer et al. 1997, Whelan et al. 2003). In all PCR applications, spurious primer-dimer formations generate background signals.

Fluorescence resonance energy transfer and contact quenching When using probes undergoing a conformation change or a degradation upon allele-specific interaction, fluorescence resonance energy transfer (FRET) is commonly used as detection principle. FRET is dependent on the distance between a donor and acceptor fluorophore which optimally are 20–100 Å. Instead of releasing a photon, the energy from an excited donor fluorophore is transferred to an acceptor fluorophore thus acting as a quencher. Both the TQ assay and the invader assay have applied this detection strategy. The MB assay instead utilizes contact quenching. Contact quenching occurs when the two moieties are brought closer than FRET distance and most of the absorbed energy is dissipated as heat and only a small fraction as light (Marras et al. 2002). A drawback of both of these detection strategies is the need for double labeled probes which introduces a high cost.

It is possible to detect an allele-specific primer extension by using labeled dNTPs for the elongation, eliminating the need for labeled probes, and detect the elongation by FRET (Takatsu et al. 2004). FRET has also been applied in minisequencing by combining labeled primers with dye-labeled terminators and FRET is observed as a ddNTP is incorporated bringing its fluorophore within FRET distance of the primer (Chen and Kwok 1997). In the ligation assay, the two probes may be fluorescently labeled, and FRET can be
observed when the ligation brings the fluorophores into close proximity of each other (Chen et al. 1998). If combining several emitter fluorophores at different distances along each ligation probe, it is possible to create fluorescence emission signatures allowing highly multiplexed assays (Tong et al. 2001). Allele-specific PCR has also been combined with FRET detection by applying a second detection primer set with a structure similar to MB probes (Myakishev et al. 2001).

**Fluorescence polarization.** Another fluorescence detection strategy that utilizes polarized light was first published in 1926, unfortunately in French (Perrin 1926). Fluorescence polarization (FP) has been applied to immunoassays, studies of protein-protein and protein-DNA interactions, and degradative assays (Checovich et al. 1995). Also the formation of DNA hybrids has been monitored by FP (Murakami et al. 1991) to detect as little as ten copies of the Mycobacterium tuberculosis (Walker et al. 1996). FP was introduced for the detection of genetic variation by Gibson et al. in 1997 and Chen et al. in 1999 (Gibson et al. 1997, Chen et al. 1999).

As detection strategy FP is based on the observation that when plane-polarized light is used to excite a fluorophore it also emits plane-polarized light if the molecule remains stationary between excitation and emission. The FP of a molecule is proportional to the viscosity of the solvent, temperature and volume of the molecule; thus, if temperature and viscosity are kept constant, only the molecular volume influences the FP. The molecular volume is proportional to the molecular weight. In summary, if the excited fluorophore is bound to a small molecule that will tumble and rotate quickly, the emission will be largely depolarized and FP lost. In contrast, a large molecule will tumble and rotate less and FP will be preserved (Chen et al. 1999).

When applied to minisequencing, the labeled ddNTPs in solution are smaller molecules than the extended labeled primer; thus, FP measures the proportion of bound label to free label. (Chen et al. 1999). The FP is generally given as an mP value obtained from the instrument and given by:

$$mP = 1000 \cdot \frac{I_{\text{parallel}} - G \cdot I_{\text{perpendicular}}}{I_{\text{parallel}} + G \cdot I_{\text{perpendicular}}}$$

where the emission intensities ($I$) are measured in two directions: parallel and perpendicular to the excitation light. $G$ is a factor close to 1.0 determined by the instrumentation used and the ratio is multiplied by 1,000 to obtain “easy-to-work-with” values. Figure 6 gives a schematic illustration of the FP detection principle for minisequencing.

Apart from the minisequencing assay described above, FP has been used for genotyping in combination with allele-specific PCR (Gibson et al. 1997), ASO probes in the TaqMan assay (Latif et al. 2001) and the invader assay (Hsu et al. 2001b).
Figure 6. Principle of minisequencing with FP detection. Polarized light is used to excite the ddNTP linked fluorophore. If the minisequencing primer has been extended the fluorophore is attached to a larger molecule (top) than if the fluorophore is only attached to the ddNTP (bottom). By detecting the emission intensities after filters in parallel and perpendicular directions compared to the excitation light it is possible to measure the FP. A larger proportion of the emission light will remain polarized in samples where the minisequencing primers have been extended. Figure adapted from Chen et al., 1999.
Both DNA and RNA can serve as templates in genotyping assays. Most of the currently used genotyping techniques rely on amplification of genomic DNA by PCR prior to genotyping, see section 6. PCR amplification provides specificity to the assay by enriching the region of interest. A high specificity is important to detect single base alterations in the complex human genome. PCR amplification also provides the sensitivity needed to detect a small amount of for example microbial genetic material present at low amounts in clinical samples. However, in the high-throughput reaction principles applied in SNP genotyping and processing many SNPs in parallel, PCR has become one of the major bottle-necks and causes a substantial part of the genotyping cost. During the last few years, several alternative assays, avoiding the need to perform PCR amplification of the target region, have been presented. Alternative amplification strategies and signal amplification are further discussed in section 7.

Despite PCR amplification, the amount of DNA obtainable from patient or population samples may become limiting for SNP mapping studies on the scale needed in genome wide association studies. Given the large efforts involved in the collection of DNA samples from well characterized patient or population cohorts, it is desirable that the samples could serve as a long-lasting resource for future genetic studies. The amount of DNA is often limiting in SNP genotyping studies on a more modest scale when the only available source of DNA is biobanked tumor or other tissue samples, buccal swabs or blood stains collected on filter paper. The whole genome amplification methods discussed in section 5 promises to obliterate these problems.

5 WHOLE GENOME AMPLIFICATION

The traditional approach to create an infinite DNA source is to immortalize the cell samples by transformation with Epstein-Barr virus. However, transformation is labor intensive and expensive to apply on a large scale. Moreover, it is not applicable to already existing biobanked DNA sample collections. A technically more feasible approach for increasing the amount of DNA is to apply a whole genome amplification (WGA). WGA is a challenge, since it requires faithful replication of three billion bases without the loss or distortion of any loci or allele (Lasken and Egholm 2003). Table 2 provides a summary of the performances of the main current used WGA procedures.

Table 2. Main features of commonly used whole genome amplification methods 1, 2

<table>
<thead>
<tr>
<th></th>
<th>MDA</th>
<th>PEP</th>
<th>DOP-PCR</th>
</tr>
</thead>
<tbody>
<tr>
<td>Amplification primers</td>
<td>Random hexamers</td>
<td>Random femtomers</td>
<td>Partially degenerate 22-mers</td>
</tr>
<tr>
<td>DNA polymerase (error rate)</td>
<td>Φ29 DNA polymerase (&lt;3·10⁻⁶) ³</td>
<td>Taq DNA polymerase (3·10⁻⁵), if combined with Pwo (9·10⁻⁶) ⁴</td>
<td>Taq DNA polymerase (3·10⁻⁵), if combined with Pwo (9·10⁻⁶) ⁴</td>
</tr>
<tr>
<td>Amplification conditions</td>
<td>Isothermal (30°C)</td>
<td>Cyclic amplification with ramping</td>
<td>Two step cyclic amplification</td>
</tr>
<tr>
<td>Product length (bp)</td>
<td>2 - &gt;10 kb ³</td>
<td>~100-1,000 ⁴</td>
<td>0.5 - &gt;10kb ⁵</td>
</tr>
<tr>
<td>Yield (per 100 µl)</td>
<td>30µg ³ - 80µg ⁶</td>
<td>0.4 - 0.8µg ³</td>
<td>6 - 8.5µg ³</td>
</tr>
<tr>
<td>Single cell amplification demonstrated</td>
<td>No</td>
<td>Yes ⁷, ⁸, ⁹</td>
<td>Yes ⁹</td>
</tr>
</tbody>
</table>

1 Abbreviations and original publications: MDA, multiple displacement amplification (Dean et al. 2002); PEP, primer extension preamplification (Zhang et al. 1992) including the improved PEP (Dietmaier et al. 1999); DOP-PCR, degenerate oligonucleotide primed polymerase chain reaction (Telenius et al. 1992) including LL-DOP-PCR (Kittler et al. 2002).

2 References indicated by roman numerals: ¹ (Nelson J.R. et al. 2002), ² (Roche 1999), ³ (Dean et al. 2002), ⁴ (Gillespie et al. 2000), ⁵ (Kittler et al. 2002), ⁶ (Lasken and Egholm 2003), ⁷ (Zhang et al. 1992), ⁸ (Paunio et al. 1996), ⁹ (Wells et al. 1999).

In the following subsections multiple displacement amplification (MDA) and primer extension preamplification (PEP), that were applied in Study III of the thesis, are described in more detail.

5.1 Multiple displacement amplification

In 1998 Lizardi et al. presented an isothermal procedure for a rolling circle amplification of DNA templates using DNA polymerase from the Φ29 bacteriophage (Lizardi et al. 1998). The method has later been adapted for the amplification of linear templates (Dean et al. 2002). The isothermal multiple displacement amplification procedure uses random hexamers as primers and relies on the processivity, fidelity and strand displacement ability of the Φ29 enzyme (Blanco et al. 1989, Esteban et al. 1993, Paez et al. 2004).

The random hexamers are phosphorothioate-modified to protect them from degradation by the proofreading 3'→5' exonuclease activity of the Φ29 DNA polymerase. This improvement has been shown to increase the yield 40-fold compared to standard random hexamers when applied to a double stranded circular M13 DNA template (Dean et al. 2001).
MDA is highly promising and has been the subject of many recent studies that has evaluated the performance for a range of sample sources using several genotyping assays, see Table 3 for a summary.

The MDA products, on average >10 kb, is well suited for further processing (Dean et al. 2002). A potential drawback of MDA amplification is that the yield may decrease when using template with a lower molecular weight, due to the dependence on hyperbranched amplification, as has been predicted by mathematical modeling (Lage et al. 2003). Another WGA strategy using balanced-PCR amplification was recently described and showed locus representation similar to MDA after amplification when “fresh” DNA samples were used. However, the balanced-PCR approach may be favorable when archived and degraded sample are used (Wang G. et al. 2004).

The isothermal feature of MDA has an advantage compared to other WGA procedures in avoiding sequence derived amplification differences between regions, due to for example differences in GC content. Using quantitative real-time PCR, a three-fold amplification bias between genomic loci was observed in MDA products (Dean et al. 2002). Additionally, using comparative genomic hybridization, a significant amplification bias between different genomic sequences was detected, especially at the ends of the chromosomes (Lage et al. 2003). However, it has been argued that this difference may have been due to the low amount of Φ29 DNA polymerase used (Hosono et al. 2003). Recently, the ability to detect allelic imbalance due to a loss of heterozygosity and copy number abnormalities in cancer tissue samples after MDA was evaluated with encouraging results (Wong et al. 2004). When using WGA to increase the amount of DNA for SNP genotyping, a more critical requirement is the balanced amplification of both alleles of each SNP at the same genomic loci, which was investigated in Study III.

5.2 Primer extension preamplification

Another WGA strategy is to use a thermostable DNA polymerase in combination with random or partly degenerate primers and cyclic amplification. In 1992 Zhang et al. described the primer extension preamplification that applied random femtomers (Zhang et al. 1992). The procedure has later been improved by using a high fidelity PCR system in combination with slightly modified thermocycling conditions (Dietmaier et al. 1999). The improved protocol has been shown to dramatically increase the genotype performance from buccal swab samples compared to using the extracted genomic DNA directly (Zheng et al. 2001).

Using quantitative real-time PCR, an amplification bias of $10^2$ to $10^4$ between genomic loci has been observed in PEP products, compared to the three-fold bias found for MDA products (Dean et al. 2002). PEP has shown good genotyping performance without the loss of heterozygous genotypes when using more than 30 ng of genomic DNA as starting material (Kuivaniemi et al. 2002). It has been estimated that the probability of
Table 3. Studies on multiple displacement amplification of linear genomes

<table>
<thead>
<tr>
<th>Reference</th>
<th>Main features and results</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dean et al. 2002</td>
<td>Original publication. Compares MDA, PEP and DOP-PCR. Amplification bias 4-6 orders of magnitude between loci, less for MDA. 1-10 copies sufficient for MDA amplification. 13 SNPs genotyped.</td>
</tr>
<tr>
<td>Gorrochotegui-Escalante and Black 2003</td>
<td>MDA of DNA directly from whole or parts of mosquitoes. Results concordant to genomic DNA as template.</td>
</tr>
<tr>
<td>Lage et al. 2003</td>
<td>Array-CGH evaluation of MDA. DNA from tumor cell lines. With 1,000 cells as MDA input, 3-fold gene dosage alterations can be detected.</td>
</tr>
<tr>
<td>Hosono et al. 2003</td>
<td>MDA on clinical samples. 47 analyzed loci were represented at 0.5 to 3-fold of the starting copy number. 5 SNPs and 10 microsatellites gave concordant genotypes.</td>
</tr>
<tr>
<td>Tranah et al. 2003</td>
<td>99.95% accuracy across 6 SNPs for 352 samples including buccal cells.</td>
</tr>
<tr>
<td>Lovmar et al. 2003b</td>
<td>Compares MDA and PEP. 45 SNPs dispersed over whole genome analyzed with 99.7% concordance for MDA DNA compared to genomic. Imbalance between SNP alleles depends on MDA input (&lt; 3ng).</td>
</tr>
<tr>
<td>Rook et al. 2004</td>
<td>MDA on laser capture microdissected cells. Pooling of reactions increases success and decreases allele bias.</td>
</tr>
<tr>
<td>Bannai et al. 2004</td>
<td>Compares MDA, PEP and DOP-PCR. 4 SNPs genotyped with 100% concordance</td>
</tr>
<tr>
<td>Barker et al. 2004</td>
<td>Compared MDA and OmniPlex² technology. 2,320 SNPs analyzed with an concordance of &gt;99.8%</td>
</tr>
<tr>
<td>Paez et al. 2004</td>
<td>Fidelity and genome representation of MDA tested by analyzing &gt;10,000 SNPs. 99.82% of the genome was covered and no evidence of amplification induced genotype errors was found.</td>
</tr>
<tr>
<td>Wong et al. 2004</td>
<td>Ability to detect allelic imbalance and copy number abnormalities analyzed on 11,560 SNP microarray using tumor samples. Average false-positive rate of 13.8%.</td>
</tr>
<tr>
<td>Yan et al. 2004</td>
<td>Sequencing analysis and SSCP-scanning for polymorphisms of MDA amplified samples. Successful PCR of fragments 10-12 kb long and with high GC content. 99.5% success rate in genotyping using sequencing.</td>
</tr>
<tr>
<td>Pask et al. 2004</td>
<td>345 SNPs genotyped by BeadArray Technology³ after MDA. Concordance rate 98.8% but with increased sample exclusion rate and inheritance errors.</td>
</tr>
</tbody>
</table>

Abbreviations: MDA, multiple displacement amplification; PEP, primer extension preamplification; DOP-PCR, degenerate oligonucleotide primed polymerase chain reaction; CGH, comparative genomic hybridization; SSCP, single-stranded conformation polymorphism.

1 Randomly fragmented genomic DNA converted into a library (Langmore 2002).
2 Miniaturized format with oligonucleotides immobilized on microbeads assembled on fiberoptic bundles (Oliphant et al. 2002).
amplifying any sequence of the genome to a minimum of 30 copies when starting from a single haploid cell is \( \approx 0.78 \) (Zhang et al. 1992).

Paunio et al. amplified genomic DNA from a single cell by PEP, showing imbalanced allelic representation at all heterozygous nucleotide positions (Paunio et al. 1996). PEP has been further evaluated (Wells et al. 1999) and applied successfully in genetic preimplantation diagnosis (Kristjansson et al. 1994, Sermon et al. 1996), e.g. to analyze single blastomeres for a range of polymorphism causing \( \beta \)-thalassemia (Jiao et al. 2003). A whole genome amplification of single cells prior to genetic analysis increases the number of loci that can be analyzed and enables reanalysis in the case of ambiguous results.

6 POLYMERASE CHAIN REACTION

The polymerase chain reaction was first suggested in 1971 (Kleppe et al. 1971), but not put into practice until the mid 1980’s (Saiki et al. 1985, Mullis et al. 1986). PCR has revolutionized genetic research, especially after it was combined with the thermostable polymerase from the bacteria Thermus Aquaticus (Saiki et al. 1988). When applied prior to a genotyping assay, PCR amplification gives an abundance of the specific template and provides the specificity needed to detect a genetic variation of one base in the presence of the three billion bases that constitute the human genome.

6.1 Reverse transcriptase PCR

For analysis of RNA, reverse transcriptase PCR (RT-PCR) is used. RNA analysis is needed when analyzing the sequences of viruses carrying RNA genomes, as in Study V, and when quantifying gene expression. RT-PCR utilizes the RNA-dependent DNA polymerase (reverse transcriptase) encoded by some viruses to enable them to copy their RNA genomes to DNA in the infected cell. In RT-PCR an oligonucleotide primer hybridized to the RNA template is extended by reverse transcriptase to produce cDNA that is further amplified by PCR. The PCR product can then be used in standard genotyping procedures, as discussed in section 2.

In the SNP genotyping assay published by Pastinen et al. 2000, modeled from the Affymetrix gene chip ASO arrays (Chee et al. 1996), reverse transcription was used for the genotyping discrimination through allele-specific primer extension with RNA as template (Pastinen et al. 2000), see previously section 2.2.

6.2 Multiplex PCR

When several SNPs are investigated simultaneously, as in Studies III and IV, multiplex PCR is a beneficial solution. However, multiplex PCR requires a careful primer design and a range of optimizations are needed due to the risk of primer-dimer formation and preferential amplification of some products at
the cost of other. A range of strategies for primer design have therefore been suggested to aid in the multiplex PCR setup. Primer-dimer formation has been restricted by using primers with tagged and complementary ends that upon primer-dimer initiation forms panhandles that are unable to further amplify (Brownie et al. 1997). By using PCR primers with a common 5'-region one optimal set of reaction conditions can be applied (Shuber et al. 1995). Also the combination of restriction enzyme cleavage and adaptor fragment ligation (see section 7) with a single specific primer has been applied but requires additional steps compared to standard PCR (Broude et al. 2001).

Multiplex PCR can also be used as a genotyping or identification tool directly, especially to detect microorganisms as discussed by (Markoulatos et al. 2002). For example multiplex PCR has been applied to distinguish between six salmonella strains in clinical stool samples (Alvarez et al. 2004) and multiplexed hemi-nested RT-PCR has been applied to genotype rotavirus, as previously mentioned (Gouvea et al. 1990).

7 ALTERNATIVE AMPLIFICATION STRATEGIES

7.1 Complexity reduction without PCR
One main function for PCR when applied prior to genotyping is the obtained complexity reduction of the genotyping target. However, it is possible to perform this by other means than PCR. One option is to process the genomic DNA to the needed reduction, alternatively an amplification method other than PCR may be used.

The invader assay (see Figure 3) has been demonstrated directly on genomic DNA by immobilizing the probes on microspheres and detecting fluorescence by flow cytometry, but the assay requires large amounts of DNA (Rao et al. 2003).

Complexity reduction By performing complexity reduction of the genome by means similar to AFLP, the need for sequence-specific PCR amplification and thus multiplex PCR can be avoided. The genome is cleaved by restriction enzymes and adapter sequences are ligated using the overlapping ends of the cleavage product. The adapter sequences also provide universal primer sites for a subsequent PCR amplification, giving the complexity reduction and size selection (Vos et al. 1995). The PCR product can then be purified, fragmented and end-labeled prior to hybridization to microarrays for genotype deduction (Kennedy et al. 2003, Matsuzaki et al. 2004). A similar strategy for complexity reduction, but using double-sided adaptor sequences in combination with ligation to give circular molecules, has recently been described (Callow et al. 2004).

Target amplification methods In the ligase chain reaction, for each DNA strand, an allele-specific oligonucleotide is ligated to another oligonucleotide by a thermostable DNA ligase. Since both strands are used, the product from
the first cycle can serve as template in subsequent ligation cycles utilizing the thermostable properties of the DNA ligase and giving exponential amplification (Barany 1991). If using only one pair of oligonucleotides complementary to one DNA strand, a linear amplification is achieved. This detection strategy, named the ligase detection reaction, is often combined with PCR (Gerry et al. 1999), but has also been used directly on genomic DNA with the two oligonucleotides forming a MB probe when ligated (Wabuyele et al. 2003).

The isothermal amplification method called strand displacement amplification (SDA) was initially dependent on restriction enzyme recognition sites (Walker et al. 1992b), but has been modified to allow the amplification of any target. The modified SDA reaction utilizes the strand displacement activity of the exonuclease deficient Klenow fragment, in combination with HincII, that nicks the template at recognition sites introduced by the primers (Walker et al. 1992a). The SDA procedure has been combined with FP to detect Mycobacterium tuberculosis (Walker et al. 1996) and FRET to detect Chlamydia trachomatis and Neisseria gonorrhoeae (Little et al. 1999). The amplification strategy has also been combined with microelectronic chip technology (Westin et al. 2001).

Selfsustained sequence replication (Guatelli et al. 1990) also called nucleic acid sequence-based amplification (NASBA) (Compton 1991) depends on the isothermal amplification of target molecules using a combination of reverse transcriptase, RNase H, and DNA dependent RNA polymerase to specifically amplify single-stranded templates. The reaction is isothermal and performed at a temperature at which genomic DNA is not denatured. This feature eliminates the risk for contamination of example of host DNA when detecting viral sequences in clinical samples. However, by performing an initial denaturation step it is also possible to perform NASBA on genomic DNA (Guatelli et al. 1990). The reverse transcriptase and RNA polymerase (usually T7 RNA polymerase) may generate truncated or erroneous amplification products. An advantage is the possibility to detect RNA genomes or transcripts from an infectious agent in the presence of host or food DNA (Jean et al. 2001, Loeffler et al. 2001). NASBA has been performed in microchip format (Gulliksen et al. 2004) and mRNA quantitation can be performed when standards are included in the amplification reaction (Romano et al. 2001).

If circular template molecules are used, e.g. the padlock probes for genotyping, amplification can be achieved using circle-to-circle amplification. This methods utilizes rolling circle amplification (Lizardi et al. 1998), but the efficiency is increased through cycles of monomerization and ligation followed by a new replication step (Dahl et al. 2004).

7.2 Signal amplification
One option to avoid a PCR step, often multiplex PCR, prior to genotyping is to perform genotyping on genomic DNA directly and then to amplify the
signal from the genotyping reaction. This strategy requires a highly specific
detection step often combining two recognition events. 

Using a single common post-genotyping PCR as signal amplification is one
option. This has recently been applied to a range of ligation assays. In one
approach, locus-specific padlock probes are “gap-filled” with dNTPs, i.e.
minisequencing, in four separate reactions followed by fluorescent labeling.
The probes are then ligated, linear molecules degraded and the circularized
probes are released from their target by cleavage and amplified with primers
common to all probes. Finally the amplification products are captured on
microarrays and the fluorescence signals are measured (Hardenbol et al.
2003). This technology has been commercialized by ParAllele Biosciences.

In the SNPWave™ genotyping system, allele-specific padlock-probes are
ligated with a subsequent PCR amplification of the probes using primers with
two selective nucleotides, similar to primers used in amplified fragment length
polymorphism (AFLP) assays (Vos et al. 1995). The amplification products
are separated by size and fluorescently detected (van Eijk et al. 2004).

The SNPlex™ Genotyping System (Applied Biosystems) utilizes an allele-
specific ligation with two linear primers followed by a universal PCR
amplification of the ligation product with biotin labeled primers allowing
strand separation. Fluorescently labeled identification probes with different
electrophoretic mobility properties are hybridized to the target molecule and
subsequently washed, eluted and separated by capillary electrophoresis. Apart
from the above discussed ligation assays, Illumina has a genotyping system
where allele-specific primer extension is combined with a ligation step
followed by universal PCR amplification with fluorescently labeled primers
and detection by array hybridization using their BeadArray™ technology
(Oliphant et al. 2002). Also the invader assay has been used directly on
genomic DNA with a subsequent signal amplification (Lyamichev et al.
1999).
GENETIC VARIABILITY AND HUMAN DISEASE

Genetic variability exists as everything from deletions or duplications of whole chromosomes or large chromosomal regions, repeated segments of variable length and single base pair alterations or deletions. Large chromosomal aberrations can alter the phenotype of an individual, but even alterations as small as one base pair may have a dramatic impact on the phenotype by causing severe monogenetic disorders. A catalogue of human diseases following Mendelian inheritance patterns is found on OMIM – Online Mendelian Inheritance in Man (http://www.ncbi.nlm.nih.gov/Omim).

More than two decades ago it was recognized that variations in human DNA could be assayed directly and used as genetic markers in linkage studies (Botstein et al. 1980). The cystic fibrosis gene was the first gene to be positionally cloned through linkage analysis using restriction fragment length polymorphisms (RFLPs) (Kerem et al. 1989, Riordan et al. 1989). Since then many successful linkage studies have been applied to track markers that co-segregate with disease in affected families. Linkage describes co-segregation of characteristics in a pedigree due to their close physical location on a chromosome. The use of RFLPs was substituted by highly polymorphic repetitive sequences denoted microsatellites or short tandem repeats (Litt and Luty 1989, Weber and May 1989). Today the dense, but less polymorphic SNPs are gaining use as genetic markers.

Apart from alterations in the human genome affecting the spectra of disease, we are influenced by co-existing microorganisms. Sequence variability in both the microbial genomes and in the genome of human host-cells affects the outcome of disease. On some occasions the genomes have even evolved together as discussed by Woolhouse et al (Woolhouse et al. 2002). The most well known example of co-evolution of microbial and host genomes is perhaps the sickle-cell mutation giving the heterozygote carriers an advantage of resistance against malarial parasite infections (Miller et al. 1975, Gelpi and King 1976). Another interesting example is the suggestion that a strong selection for the deletion of 32 nucleotides in the gene of a human lymphoblastoid chemokine receptor (CCR5Δ32) was induced in European populations by the Yersinia pestis infections in the 14th century (Stephens et al. 1998). Today, the CCR5Δ32 allele gives decreased susceptibility to infections by human immunodeficiency virus.
8 SNPS AS GENETIC MARKERS

SNPs are evenly distributed over the genome, genetically stable and usually biallelic, making them suitable for high-throughput genotyping. Taken together, these features make SNPs highly interesting as genetic markers.

8.1 Linkage analysis

The initial success to identify the genes responsible for Mendelian disorders spurred optimism to find the genetic factors underlying complex diseases; however, the challenge has been proven to be much larger. Initially, there was optimism because several of the genes identified by linkage analysis in families were responsible for part of the common diseases, e.g. the BRCA-1 and -2 mutations were responsible for part of the susceptibility for familial breast cancer (Miki et al. 1994, Wooster et al. 1995); however, so far there has been limited success with identifying genes predisposing to complex diseases. The genes identified so far are either those with low frequencies and high displacement, i.e. Mendelian or “near-Mendelian” inheritance, or those present in isolated human populations with reduced genetic variation (Collins et al. 1997).

A common approach for linkage analysis is to first do a scan in family material using microsatellites markers spaced about 5-10 centimorgan and then to fine-map regions indicating linkage using a more dense set of markers. However, the resolution is not only limited by the markers, but also by the number of meioses i.e. sample size, and the final region might still cover several genes. Linkage disequilibrium mapping (section 8.2) using SNPs as markers can be used to narrow down the candidate region further and haplotypes (section 8.3) may further increase the power of fine-mapping (Botstein and Risch 2003). A recent example of a study applying this approach is the association of genetic variations in the DLG5 gene with inflammatory bowel disease (Stoll et al. 2004).

8.2 Association and linkage disequilibrium

The simplest form of association analysis is to genotype a group of cases and controls for a given genetic variant and then assess whether there is a difference in the genotype or allele frequencies between the two groups, as was done in Study II. There are several reasons for why an association may be found in a population: the mutation can be directly causative, natural selection for a beneficial or against a deleterious allele, population stratification, imprecision due to small sample size and/or multiple testing which gives rise to “chance findings” and linkage disequilibrium (LD).

If alleles at distinct loci occur together in a population more often than by chance, the loci are in LD with each other and are associated. This phenomenon is also denoted population allelic association (Risch 2000). Several statistics describes LD. The two most frequently used are \(D'\) and \(r^2\). A \(D'\) value of 1.0 between two marker alleles concludes that there has been no
new mutation, gene conversion or recombination between the alleles and they are in “complete linkage”; hence, the allelic association is as strong as possible, given the allele frequencies. For the two polymorphisms to be perfectly correlated, the minor allele frequencies need to be the same, which would give an $r^2$-value of 1.0 and show that the alleles are in “perfect LD” (Devlin and Risch 1995, Carlson et al. 2004).

If the genotyped variation is suggested to be the causative variation, then association studies using coding or promoter variants are called the direct approach (Risch and Merikangas 1996). In the indirect approach, the genotyped variant is regarded as a marker only, selected to be evenly spaced and expected to track disease loci through linkage disequilibrium (Collins et al. 1997). The statistical power to detect an un-assayed, disease-associated polymorphism depends on the extent of LD between the un-assayed site and the genetic marker.

The estimated number of SNPs needed to perform genome-wide scans for association has been widely debated, and range of $5 \times 10^4$ to $>5 \times 10^5$ SNPs (Kruglyak 1999, Reich et al. 2001). Since LD is a property of a population, the demographic and social history strongly affect the extent of LD and the number of SNPs needed. There have been suggestions that the pattern of LD within populations could be characterized, and a subset of the markers could be used to “tag” a region. This would dramatically lower the number of SNPs needed to be genotyped in a genome-wide association study, which is one of the major goals of the on-going Haplotype mapping (HapMap) project.

8.3 Haplotypes and the HapMap project

Haplotypes are defined as a series of alleles at linked loci on a single chromosome, which means that these loci are inherited together in a non-random pattern. Haplotypes are formed when a new mutation occur or when the paternal and maternal chromosomes undergo recombination, as reviewed by Pääbo, 2003. LD has a block-like structure in the genome and it has been suggested that the human genome can be divided in haplotype blocks with high LD and limited haplotype diversity; thus, only a small number of SNPs would be required to capture all genetic variation in an individual or population in these regions (Daly et al. 2001, Patil et al. 2001, Dawson et al. 2002, Gabriel et al. 2002, Phillips et al. 2003). SNPs used for this purpose have been denoted as haplotype tagging SNPs (Johnson et al. 2001). The usefulness of haplotype blocks for identification of disease genes is intensely debated and still remains to be demonstrated (Cardon and Abecasis 2003, Schulze et al. 2004).

The HapMap project is a joint effort by leading academic and commercial teams of scientists in Canada, China, Japan, Nigeria, the United Kingdom and the United States (US) that have formed the International HapMap Consortium. Samples from Han Chinese, Japanese, Nigerian Yoruba and an US Utah population of Northern and Western European origin are analyzed. The goal of the HapMap project is to determine the common patterns of DNA sequence variation in the human genome, and to make this information
freely available. Hopefully the project will provide a tool allowing association studies of any gene, region or ultimately the whole genome to scan for disease risk factors (HapMapConsortium 2003).

A criticism raised against the HapMap project is that because a limited number of chromosomes (a total of 540) are analyzed, only more common variants in a population will be detected. The HapMap project will be useful only if common genetic variants are responsible for the majority of the complex diseases that are common in the population (Reich and Lander 2001, Lohmueller et al. 2003); however, for diseases caused by rare alleles it will be of limited usefulness (Cardon and Abecasis 2003). A spin-off benefit of the HapMap project is that a large number of SNPs are detected and validated as is already seen in the public SNP database; see Figure 7 in the Prologue.

Molecular haplotyping Today, haplotypes can be determined by either analyzing family members to acquire phase-information or by prediction using computer algorithms based on observed experimentally genotype frequencies in a population. However, the true haplotype is optimally determined by direct molecular haplotyping either by physically separating the diploid genome prior to genotyping or by analyzing individual DNA molecules directly. Molecular haplotyping is perhaps the largest challenge for the development of future genotyping technology. The current approaches and future challenges have been reviewed by Kwok and Xiao, 2004.

A straight-forward cloning approach has been used for haplotyping in a high-resolution scan of chromosome 21 (Patil et al. 2001). Patil et al. created somatic rodent-human hybrid cell lines with complete human chromosomes. Long range PCR products were hybridized to high-density oligonucleotides to discover SNPs and define the haplotype structure of chromosome 21.

A different strategy is to amplify multiple individual molecules that have been physically separated from each other using thin acrylamide gels (Mitra et al. 2003). If low concentrations of DNA as well as PCR reagents are included when polymerizing the gel the chromosomal fragments will be physically separated in the gel. An in-gel PCR amplification results in a PCR colony (polony) and because diffusion is limited by the gel, the polony will stay at the location of the chromosome fragment. If one PCR primer is modified to enable covalent attachment to the gel, the other strand can be removed leaving a single-stranded template available for genotyping.

Instead of amplifying pre-separated molecules, recent technological advances offer the promise to analyze single molecules; however, the molecules may require amplification to acquire statistical confidence and a sensitive detection is also needed. The DNA strand has to be investigated for both SNP allele identity and the order of alleles. If these two measurements are managed, a system with only three labels would be sufficient for haplotyping. This would be possible if analysis of linear molecules is achieved by stretching them either in a flow system detecting the molecules as they pass by a detector or in a stationary system with the molecules immobilized on a solid support (Kwok and Xiao 2004).
9 MICROORGANISMS AS INFECTOUS AGENTS

In 1840 the German pathologist Friedrich Henle suggested criteria to prove that microorganisms are responsible for causing human disease. His theories were later confirmed by Koch and Pasteur, who proved that tuberculosis is caused by microorganisms. In the 20th century, the possibility to cure these diseases emerged, with the famous Nobel-prize rewarded discovery of penicillin by Alexander Fleming in 1928 as an important landmark. In 1946 viruses were cultivated in cells by John Enders, which facilitated large-scale vaccine production. Viruses, bacteria, fungi and parasites all have the capability to infect humans and cause disease.

Viruses are intracellular parasites requiring host cells for replication. The type of cell infected partly dictates the clinical manifestations. For example, human rotavirus (HRV) is activated by the environment in the gastrointestinal tract and enters the host enterocytes on the tips of the small intestinal villi. The virus then replicates and assembles inside the cytoplasm and in the rough endoplasmic reticulum of the infected cell and is released by vesicular transportation to the apical part of the enterocytes. Diarrhea is a result of the damage caused by the infection of the epithelial cell lining and probably by the triggering of the enteric neural system by a specific viral gene product (Lundgren et al. 2000).

Microorganisms, i.e. viruses, bacteria, fungi and parasites, may cause disease by either an exogenous infection, e.g. influenza virus infection of the respiratory tract, or by an endogenous infection, e.g. Escherichia coli causing a urinary tract infection.

9.1 Genotyping microorganisms by microarrays

Microarrays are a promising platform for the analysis of microorganisms in infectious diseases as reviewed by Bryant et al., 2004. They are applicable both to study microorganism sequences directly for diagnosis or epidemiology and to study interactions between host and pathogen in expression studies.

Microarrays possess the possibility to simultaneously study several pathogens, which has been done with 70-mer oligonucleotide hybridization arrays with the potential to detect hundreds of viruses following RT-PCR using partially random primers (Wang D. et al. 2002a). Also, a microarray with the ability to detect 18 pathogens from prokaryotic, eukaryotic and virus origin by hybridization of pathogen-specific multiplex PCR products has been constructed (Wilson et al. 2002).

Genotype-specific hybridization has been applied to several different microorganisms. In an early study by Gingeras et al., 705 bp of a specific nucleotide sequence of Mycobacterium tuberculosis were determined using in situ synthesized high-density oligonucleotide arrays (Gingeras et al. 1998). These arrays allowed accurate detection of resistance-mediating mutations in all strains known to harbor them by dideoxynucleotide sequencing. The genetic diversity of Helicobacter pylori and Campylobacter jejuni has been examined using microarrays with immobilized PCR products (Salama et al.
Human papillomavirus has been genotyped in microtiter format, in principle similar to microarrays (Roda et al. 2002) and rotavirus has been genotyped using custom-made microarrays with immobilized genotype specific oligonucleotides (Chizhikov et al. 2002). Furthermore, hybridization microarrays have also been applied to identify new vaccine-derived poliovirus strains and to determine their evolutionary divergence (Cherkasova et al. 2003).

Performing an enzymatic reaction increases the specificity of the genotyping assay. Genotype-specific primer extension has been applied to human papilloma virus with subsequent capture on tag-arrays (Gharizadeh et al. 2003). Genotype-specific primer extension performed on immobilized oligonucleotides was applied to genotype human rotavirus using in Study V of this thesis (Lovmar et al. 2003a).

### 9.2 Vaccine studies and human rotavirus

Rotavirus accounts for more than 125 million cases of infantile gastroenteritis and approximately 500,000 deaths per year in children younger than five years, mainly in developing countries. The incidence of rotavirus infections is similar world-wide; thus, improvements in hygiene and water supply are not sufficient to limit infections. Instead, vaccine intervention would be beneficial to protect children from the severe diarrhea induced the rotavirus infections (Parashar et al. 2003).

The first recombinant DNA vaccine to be approved in the US was a hepatitis B vaccine produced in Saccharomyces cerevisiae (McAleer et al. 1984, Hilleman 1987). A rotavirus vaccine was approved in the US in 1998 and was subsequently introduced to the market. However, it was withdrawn after approximately one million administrated doses due to an adverse effect of intussusception in some infants (Murphy et al. 2001). The development of new vaccines is currently underway, but they are complicated by rotavirus strain diversity being greater than first estimated and varying worldwide. Rotavirus diversity seems more complex in developing countries where the clinical outcome of infection is also more severe. This indicates that vaccine trials and epidemiological studies of strain diversity prior to, and after vaccine interventions should be performed at locations where the need for vaccines is most urgent (Cunliffe et al. 2002). This was the long-term aim for application of the method developed in Study V in Nicaragua.
New methods for accessing sequence variations were developed and applied in this thesis (Table 4 in the Materials & methods section). During the course of the work, substantial progress in the field of genotyping and analysis of sequence variations has occurred. This progress is evident by the abundance of references from the late 1990’s and the beginning of the 2000’s. When my laboratory work was initiated, the spectra of techniques and their throughput were dramatically smaller than today, merely six years later.

Apart from the technical development, an important milestone during my thesis work was the completion of the human genome sequence and the large increase in publicly available sequence and SNP data. During the past three years, the quality of the information has improved substantially (Figure 7).

Figure 7. Increase in amount and quality of available information in dbSNP from build 106 (2002-08-09) to build 121 (2004-06-01). Number of entries are given for the total submitted SNPs (ss, black bars), non-redundant refSNPs (rs, gray bars), and validated SNPs (white bars) (dbSNP, http://www.ncbi.nlm.nih.gov/SNP).

A practical example of the improvement of the public databases is the SNPs genotyped in Study II that were selected from the literature when the study was initiated and that were well characterized at the time. However, it was
not until 2004, through the efforts of the HapMap project, that all four SNPs were included in the public SNP database at NCBI.

The large data output from high-throughput genetic studies sets extraordinary demand on data handling and processing. More than 30,000 genotypes were produced in my thesis work. Although this amount of data is not comparable to the large scale whole genome studies that are now conducted elsewhere, data handling is still a challenging task and was developed in parallel with the genotyping methods in my study.

In my thesis work the genotyping assays were applied in an epidemiological context ranging from breast cancer and endometrial cancer to rotavirus infections. Breast cancer is the most common cancer among women in developed nations. Approximately 6,500 Swedish women are diagnosed every year and the incidence is increasing by approximately 1.6% per year. Ductal forms of breast cancer have a higher prevalence than lobular. Endometrial cancer is the most common gynecological cancer and it occurs chiefly among postmenopausal women. Approximately 1,200 cases of endometrial cancer are diagnosed per year in Sweden (Roche 1999) (Figure 8).

Among the known risk factors for both breast cancer and endometrial cancer are early menarche, late menopause, and nulliparity, which indicates that estrogens play a central role in the cancer etiology. In contrast to breast cancer, the effect of estrogens on endometrial cells is counteracted by progestin (Purdie and Green 2001, Cuzick 2003).

By the age of 5 nearly every child has had a rotavirus infection (Parashar et al. 2003). The rotavirus particle is non-enveloped, triple-layered and has a genome comprising 11 double-stranded RNAs that encode 6 structural viral proteins (VP) and the enzymes needed to transcribe, replicate and package...
the genome. The inner core is composed of the proteins VP1, VP2 and VP3, the middle capsid of VP6 and the outer capsid of VP7, with spikes from VP4 protruding from the surface. The structure and function of rotavirus are reviewed by Jayaram et al., 2004. The viral structure is illustrated in Figure 9.

Figure 9. Electron cryomicroscopy reconstruction of the triple-layered rotavirus particle. The six structural viral proteins (VP1, 2, 3, 4, 6 and 7) are indicated by arrows as is the double-stranded RNA (dsRNA) genome. Reprinted with the kind permission from B.V.V. Prasad.

Most human infections are mediated by rotavirus from serogroup A which is defined by the VP6 protein. In addition, the sequences of the RNA encoding VP7 (G-type) and VP4 (P-type) are used to subdivide rotavirus strains into genotypes (Fischer and Gentsch 2004).
PRESENT STUDY

1 AIMS

In my thesis work I developed and evaluated methods for high-throughput analysis of sequence variations to enable large scale genetic epidemiological studies or clinical diagnosis. Specifically, the aims for the different studies were:

- to compare the performance of two homogenous genotyping assays (Study I)
- to genotype the human estrogen receptor alpha gene (ESR1) in a large sample material and to detect its association to breast cancer (Study IIa) and endometrial cancer. (Study IIb)
- to evaluate whole genome amplified products as templates for multiplex PCR with subsequent tag-array minisequencing (Study III).
- to apply a cluster validation algorithm on genotype data and to use the algorithm to compare four DNA-polymerases in the tag-array minisequencing assay (Study IV).
- to develop a microarray method for genotyping human group A rotavirus and to apply it to survey infectious strains in Nicaragua (Study V).

Additionally, the performance of MDA products, amplified from DNA extracted from buccal cell, in minisequencing with FP detection was assessed (MDA-FP study).

2 MATERIALS & METHODS

The genotyping assays that were the targets of studies included in the thesis are listed in Table 4.
Table 4. *Genotyping assays used in the different studies I-V*  

<table>
<thead>
<tr>
<th>Study</th>
<th>I</th>
<th>II</th>
<th>III</th>
<th>IV</th>
<th>V</th>
<th>Application</th>
</tr>
</thead>
<tbody>
<tr>
<td>TaqMan assay</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Assay evaluation</td>
</tr>
<tr>
<td>Molecular beacon assay</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td>Assay evaluation. Genotyping of 2 exon ESR1 SNPs</td>
</tr>
<tr>
<td>Solid-phase minisequencing</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
<td>Reference method and SNP allele frequencies</td>
</tr>
<tr>
<td>Minisequencing with FP detection</td>
<td></td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td>Genotyping of 2 intron ESR1 SNPs</td>
</tr>
<tr>
<td>Minisequencing with Tag-arrays</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td>WGA evaluation. DNA polymerase comparison</td>
</tr>
<tr>
<td>Primer extension on microarrays</td>
<td></td>
<td></td>
<td></td>
<td>X</td>
<td></td>
<td>Assay development. HRV genotyping</td>
</tr>
</tbody>
</table>

1 Abbreviations: ESR1, estrogens receptor alpha gene; WGA, whole genome amplification; HRV, human rotavirus.

2.1 Study samples

In Studies I and IV, DNA was extracted from whole blood from Swedish volunteer blood donors using the Wizard genomic DNA purification kit (Promega, Madison, WI). To evaluate the power to resolve minority alleles of the two methods in Study I, a two-fold dilution series from 100% to 0% of one homozygous sample into a sample homozygous for the other allele was prepared.

For Study III, DNA samples from nine last generation children and their predecessors in two generations from a Centre d’Etude du Polymorphisme Humain (CEPH) family were obtained from the Coriell Cell repositories (http://arginine.umdnj.edu).

To assess the performance of MDA products in minisequencing with FP detection samples in the MDA-FP study samples from a biobanked collection were used. DNA had been extracted from buccal cells sampled from approximately 250 individuals. Buccal swabs provide only sparse quantities of DNA with a large variation in concentration. A selection of 56 samples with DNA concentrations ranging from approximately 2–200 ng/µl were analyzed both from the original DNA source and after MDA amplification.

**CAHRES study** Study II encompassed a subset of the participants in a nationwide population-based case-control study “Cancer Hormone Replacement Epidemiology in Sweden” (CAHRES). All primary invasive female breast cancer and endometrial cancer patients aged 50 to 74 that were residents of Sweden between October 1993 and March 1995 (breast) or January 1994 and December 1995 (endometrial) were asked to participate in the CAHRES study by their physicians. If they consented, a questionnaire asking for a range of detailed information was sent to them (Magnusson et al. 1999, Weiderpass
et al. 1999). A subset of the participants in the CAHRES study was invited to participate in the genetic investigations by donating DNA samples. Approximately 1,550 breast cancer patients, 700 endometrial cancer patients and 1,550 controls contributed with either blood samples or pathological tissues samples. The controls were randomly selected in 5-year age strata from the Swedish Total Population Register. DNA was extracted from whole blood using the Wizard genomic DNA purification kit (Promega, Madison, WI) and from non-malignant cells in paraffin embedded tissue using a standard phenol/chloroform/isoamylalcohol protocol.

**Rotavirus sampling** Fecal specimens from children with severe diarrhea were collected in cooperation with physicians from the community hospital and health centers in León, Nicaragua and from Uppsala University Hospital, Uppsala, Sweden. Forty samples determined to be rotavirus antigen positive by enzyme-linked immunosorbent assay were used in Study V. Rotavirus RNA was extracted from a fecal suspension in phosphate-buffered saline using the QIAamp RNA Mini kit (Qiagen, Hilden, Germany). RNA was amplified by RT-PCR prior to genotyping.

### 2.2 Assay design

For Studies I and II, the ESR1 polymorphisms were selected based on literature searches (McGuire et al. 1992, Yaich et al. 1992, Roodi et al. 1995, Sano et al. 1995). For Study III SNPs were selected to be dispersed throughout the human genome. We primarily chose SNPs already in use in our laboratory. SNPs located in repetitive elements were excluded using the RepeatMasker program (http://www.repeatmasker.org/). The Swedish allele frequencies of the SNPs in Study III were determined to vary between 0.05 and 0.5 using solid-phase minisequencing of pooled DNA samples (Fredriksson et al. 2004). In Study IV, a subset of 26 SNPs from Study III was used. For the MDA-FP study, we used four SNPs with assays already optimized by the genotyping service facility of our research group.

PCR primers were designed either using the Oligo Primer Analysis Software (Molecular Biology Insights Inc, Cascade, CO) or the Primer3 Software (http://frodo.wi.mit.edu/cgi-bin/primer3/prim3_www.cgi). To amplify the HRV samples in Study V using RT-PCR previously published Beg9 and VP7 primers were used (Gouvea et al. 1990, Ushijima et al. 1992). The PCR products spanning the SNPs were analyzed by BLAST analysis (http://www.ncbi.nlm.nih.gov/BLAST/) to ensure that the amplified product would be unique (Altschul et al. 1997).

To genotype the ESR1 exon SNPs using the MB assay in Studies I and II, two fluorescently labeled allele specific hybridization probes for each SNP, carrying the variable nucleotide in the middle of the loop region were designed. We used the web-based mfold program (Zuker 2003) (http://www.bioinfo.rpi.edu/applications/mfold/) to estimate the stability of the stem and loop structure of the MB probes. For the TQ assay in Study I, we designed two fluorescently labeled probes with the variable position close
to the center of the probe. The probe was designed in collaboration with experts on the two techniques: Dr. Sanjay Tyagi, Public Health Research Institute, USA, the inventor and major developer of the MB assay and Dr. Lars Melin, senior technical expert at PE Biosystems, Sweden, at the time of the study.

For the minisequencing assays in Study I, II, III and IV we designed minisequencing primers annealing immediately adjacent to each SNP. For the intron ESR1 SNPs in Study II and for all SNPs in Study III and IV minisequencing primers for both DNA polarities were used. In Studies III and IV, applying the Tag-array format, the 5'-end of the minisequencing primers contained a 20 bp Tag-sequence from the Affymetrix GeneChip® Tag collection (Affymetrix, Santa Clara, CA). Absence of hairpin loop formation in the tagged minisequencing primers was assessed using the NetPrimer software (http://www.premierbiosoft.com/netprimer/index.html). The capture oligonucleotides (cTags) complementary to the Tag-sequences contained 15 T-residues as a spacer and a terminal NH₂-group at their 3'-ends to mediate covalent coupling to the amino-reactive slides.

In Study V the HRV capture oligonucleotides were designed based on published sequences for HRV and the MS2 phage, respectively, using the MACAW program (Schuler et al. 1991). Sequence alignments of the VP7- and VP4-genes of reference strains used in previous studies were utilized for a preliminary primer selection (Gentsch et al. 1992, Chizhikov et al. 2002). The primers were then tested for their 3'-specificity by using them as query sequences in BLAST searches against GeneBank (Altschul et al. 1997), confirming that high scoring sequences corresponded to the intended genotype. The capture oligonucleotides were designed to have 15 T-residues as a spacer and a terminal NH₂-group at their 5'-end to allow covalent attachment to the slides.

2.3 Whole genome amplification

**MDA** Multiple displacement amplification in Study III and of the DNA from buccal cells, was performed using the GenomiPhi DNA Amplification Kit (Amersham Biosciences, Uppsala, Sweden). The amplification kit includes the Φ29 DNA polymerase, nucleotides, thiophosphate-modified random hexamers (5'-NpNpNpNpNpN-3') and buffers. To investigate the relationship between the original amount of genomic DNA used as template for MDA and the genotyping results, genomic DNA in a range from 3 ng to 3 pg of genomic DNA was subjected to MDA. The genomic DNA was from an individual chosen to be heterozygous for four assayed SNPs.

**PEP** Primer extension preamplification in Study III was performed according to the improved protocol published by Dietmaier et al. and using the Expand High Fidelity (EHF) PCR System (Roche Diagnostics, Basel, Switzerland) and random femtomers (Dietmaier et al. 1999). In addition to Taq DNA polymerase, the EHF Polymerase mix also contains Pwo DNA polymerase that has 3'-5'-exonuclease proofreading activity to enhance fidelity. The
thermal cycling was performed in a Thermal Cycler PTC-225 (MJ Research, Watertown, MA).

2.4 Microarray preparation

Microarrays were used in Studies III, IV and V. The microarrays were manufactured in-house, and all immobilized oligonucleotides were covalently coupled to the glass surface through NH2-groups. The oligonucleotides were printed in an “array of arrays” format of up to 80 subarrays per microarray slide, allowing several samples to be analyzed on each slide. Initially in Study V, teflon-lined slides (Erie Scientific Company, Portsmouth, NH) with 10 reaction wells of 7 mm diameter were used: Later in the study “standard” microscope slides were also used. The HRV microarray slides in Study V were activated essentially using the procedure described by Guo et al., except that 3-aminopropyltriethoxysilane was used for silanization instead of its methoxy-derivate (Guo et al. 1994, Lindroos et al. 2001). Based on the results by Lindroos et al., we used CodeLink™ Activated Slides (Amersham Biosciences, Uppsala, Sweden) in Studies III and IV according to the manufacturer’s instructions.

The first slides in Study V were arrayed with a custom built arrayer with two TeleChem CMP-3 (TeleChem International Inc., Santa Clara, CA) pins and a spot center to center distance of 250 µm. In Study V and also in Studies III and IV, a ProSys 5510A instrument (Cartesian Technologies Inc., Irvine, CA) with four Stealth Micro Spotting Pins (TeleChem International Inc., Santa Clara, CA) and spot to spot distance down to 185 µm was used. All spots were printed in duplicate. In Study V the oligonucleotides were dissolved in 400 mM sodium carbonate buffer to a final concentration of 25 µM prior to printing. In Studies III and IV the cTags were dissolved in 150 mM sodium phosphate buffer to a final concentration of 25 µM. After printing, the remaining amino-reactive groups of the slides were blocked, the microarrays were rinsed and dried by centrifugation.

In Study V the genotype-specific primer extension reactions were performed directly on the microarray surface with the samples separated by the teflon lining. In Studies III and IV a silicon rubber grid was used to form the separate hybridization wells of the “array of arrays”. The reusable silicon rubber grid is made by pouring polydimethylsiloxane (Elastosil® 601 or 625, Wacker-Chemie GmbH, Munich, Germany) into an inverted v-bottom microtiter plate used as a mold. The arrayed microarray slide and silicon rubber are assembled in a heat conductive aluminum rack with a plexiglas cover to form individual hybridization chambers for the tag-array minisequencing hybridization step (Figure 10). The principle of the “array of arrays” format is illustrated in Figure 5.

2.5 Genotyping

The principles of the genotyping methods used in this thesis were illustrated in Figures 3-6 and 11.
Homogeneous hybridization assays

Both the MB and the TQ assays in Studies I and II were performed on an ABI Prism 7700 Sequence Detection System (Applied Biosystems, Foster City, CA) in 96-well microtiter plates. The assays were optimized for temperature, MgCl₂ and glycerol concentration. The allele-specific probes were 5’-labeled with Fam or Tet and 3’-labeled with Dabcyl or Tamra for the MB and TQ assays, respectively. The fluorescence signals were measured at each cycle and corrected for the passive Rox-labeled reference oligonucleotide supplied in the buffer. The signals of the last cycle were used for genotype assignments.

Minisequencing assays

Prior to minisequencing, the region spanning the SNPs were amplified by PCR. In Studies III and IV multiplex PCR amplification was used. The PCRs were optimized with regard to primer concentration, MgCl₂ concentration and annealing temperature. In some cases DMSO was included in the reaction mixture.

In the solid phase assay, used as reference method in Studies I and II and to genotype part of the samples in Study II, the minisequencing reactions were performed separately for each SNP and nucleotide. A biotin-labeled PCR primer was used, and the PCR product was allowed to hybridize to streptavidin-coated microtiter wells. The PCR product was denatured and the minisequencing reaction was performed with the immobilized strand as template. The DNA polymerase incorporates a tritium-labeled dNTP at the end of the minisequencing primer. The elongated primers were released by alkaline treatment and the radioactivity was measured.

The semi-homogeneous minisequencing assay with fluorescence polarization detection also named template-directed dye-terminator incorporation (FP-TDI) was used in Study II. The two ESR1 intron SNPs of Study II were analyzed in both DNA polarities and were amplified in one
PCR fragment. A PCR aliquot for each SNP and polarity were transferred to black 384-well microtiter plate wells (ABgene, Rochester, NY) and treated with exonuclease I and shrimp alkaline phosphatase to remove remaining primers and dNTPs. The minisequencing reactions were performed by adding a reaction mixture with the minisequencing primer at 5 µM concentration together with fluorescently labeled and unlabeled ddNTPs. The two ddNTPs relevant for the interrogated SNP were included at a 1:5 ratio of labeled to unlabeled ddNTPs, to a total concentration of 0.125 µM. The fluorescent dye R110 was used to label ddGTP and ddCTP, and Tamra was used to label ddATP and ddUTP (PerkinElmer Life and Analytical Sciences, Boston, MA, USA). ThermoSequenase DNA polymerase (Amersham Biosciences, Uppsala, Sweden) was included at 0.08 units/µl. Cyclic extension reactions were performed for 40 cycles of 92°C for 10 seconds and 55°C for 30 seconds. Fluorescence signals were measured directly from the 384 well microtiter plates with an Analyst AD™ instrument (Molecular Devices Corporation, Sunnyvale, CA, USA). An essentially identical procedure was performed for two of the SNPs in the MDA-FP study, whereas the AcycloPrime™-FP SNP detection kit (PerkinElmer) was used to genotype the other two SNPs.

For the microarray format used in Studies III and IV, the multiplex PCR products were pooled and in Study V also concentrated and enzymatic PCR clean-up was performed as above. The multiplex cyclic minisequencing reactions were performed in solution with minisequencing primers for both DNA polarities of each SNP. Up to 90 tagged minisequencing primers were included at 10 nM concentration each together with fluorescently labeled ddNTPs at concentrations from 0.1 to 0.2 µM (ddATP-Texas Red, ddCTP-Tamra, ddGTP-R110 and ddUTP-Cy5). The thermo stable DNA polymerase was added at 0.064 units/µl and the reactions were performed for 33 or 55 cycles of 95°C and 55°C for 20 or 30 s each. In Study IV TERMIPol (Solis BioDyne), Therminator (New England BioLabs, Inc.), and KlenThermase (Gene Craft) DNA polymerases were used at equal reaction conditions and performance were compared to ThermoSequenase (Amersham Biosciences). The extended minisequencing primers were allowed to hybridize to arrayed slides carrying cTag oligonucleotides and the slides were washed. The fluorescence signals were measured using a ScanArray® 5000 instrument (PerkinElmer Life and Analytical Sciences, Boston, MA, USA) with the laser power kept constant at 80% and the photo-multiplier tube gain adjusted to obtain equal signal levels from reaction control spots in all four laser channels (see section 2.5.2). The signal intensities were determined with the QuantArray® analysis 3.1 software (PerkinElmer Life and Analytical Sciences, Boston, MA, USA) to enable transfer to a spreadsheet computer program for further processing.

**Genotype-specific primer-extension microarray assay** For Study V, genotype-specific extension reactions were performed on microarrays (Figure 11). Regions of VP4 and VP7 as well as the control MS2 gene were amplified by RT-PCR and the products were combined and ethanol precipitated. The PCR product pool was denatured and applied to two wells of a preheated slide.
The annealing reaction was allowed to proceed in a humid chamber for 15 minutes at 37°C. The arrays were quickly rinsed and heated. A primer extension reaction mix containing 1.0 µM each of dATP, dCTP, dGTP, 0.6 µM of Cy5-dUTP (Amersham Biosciences, Uppsala, Sweden), and 0.063 units/µl of Taq DNA polymerase (Amersham Biosciences, Uppsala, Sweden) was added to each reaction well. The reactions were allowed to proceed for 10 minutes at 60°C. The microarrays were rinsed, and treated with alkali prior to washing. The fluorescence signals were measured using instrumentation as described above.

![Diagram](image.png)

**Figure 11.** Assay principle for rotavirus genotyping in Study V. Capture oligonucleotides are immobilized on a teflon-lined microscope slide with 10 subarrays (A). RT-PCR products are hybridized to genotype-specific capture oligonucleotides and a DNA polymerase extends perfectly matched oligonucleotides with fluorescently labeled nucleotides (B). The array is scanned to obtain fluorescent signals. A positive G1[P8] HRV sample and a negative reaction with only the internal control MS2 phage complementary capture oligonucleotides giving signals are illustrated (C).

### 2.5.1 Assay quality assurance

**Quality control** For the TQ and MB assays in Study I and II, the fluorescence intensity curves from both allele specific probes were inspected visually. This ensures that the PCR reactions were successful and that the signals used for genotyping were well above the background level, as defined by the signals from samples lacking the analyzed allele and negative controls. See Figure 4 for an example.

Several oligonucleotides were used to control for the different steps in minisequencing with Tag-arrays in Studies III and IV (Table 5).
Table 5. Control oligonucleotides used in minisequencing with Tag-arrays

<table>
<thead>
<tr>
<th>Oligonucleotide</th>
<th>Design and function</th>
</tr>
</thead>
<tbody>
<tr>
<td>Spot control</td>
<td>A fluorescently labeled oligonucleotide carrying a NH₂-group at the 3’-end and that is printed in each subarray on the microarray.</td>
</tr>
<tr>
<td>Print control</td>
<td>A fluorescently labeled oligonucleotide designed to hybridize to all cTags that is hybridized to subarrays from each batch (5’-AAA AAA AAA ANN NNN NNN NN–Fluorophore-3’).</td>
</tr>
<tr>
<td>Reaction control templates</td>
<td>Four synthesized single-stranded oligonucleotides differing at the one position included in the minisequencing reaction to mimic a “four-allelic” SNP.</td>
</tr>
<tr>
<td>Reaction control minisequencing primer</td>
<td>A tagged minisequencing primer, complementary to the common sequence of the four reaction control templates. Included in the minisequencing reaction. A corresponding cTag is included in the array.</td>
</tr>
<tr>
<td>Hybridization control</td>
<td>Two differently fluorescently labeled hybridization control oligonucleotides, with corresponding cTags, added in an alternating pattern over the microarray during hybridization to ensure that no leaking had occurred between subarrays.</td>
</tr>
</tbody>
</table>

In Study V, generic capture oligonucleotides recognizing the VP4- and VP7-genes, independent of genotype, served as positive controls, indicating the presence of HRV template. The same generic oligonucleotides, but with a mismatch at the 3’-end, served as negative controls to ensure that the stringency of the reaction conditions allowed discrimination of a 3’-mismatch. MS2 phage specific capture oligonucleotides served as controls for the spotting procedure, the efficiency of the hybridization and the primer extension reaction for each sample, since PCR products derived from the phage MS2 were added to the HRV derived RT-PCR product. The MS2 control was also used for normalization of the signals between arrays and experiments.

Validation All samples in Study I were genotyped using solid-phase minisequencing (Syvänen et al. 1993) in addition to the TQ and MB assays. Solid-phase minisequencing also served as a reference method to validate 2-3% of the SNP genotypes in Study II. Approximately 3% of the MB assays and approximately 30% of the minisequencing assays with FP detection in Study II were repeated. Additionally, the two intronic SNPs genotyped by minisequencing with FP detection were analyzed in both polarities.

The SNPs in the microarray based Studies III and IV were also analyzed in both polarities. The genotypes of the CEPH family samples in Study III were evaluated for Mendelian inheritance errors.

All fecal samples of Study V had been shown to be rotavirus antigen positive, and a subset of the samples was also analyzed by sequencing and genotype-specific multiplex RT-PCR (Gouvea et al. 1990).
2.5.2 Genotype scoring and signal analysis

A fluorescence signal fraction was calculated to assign the genotypes both in the TQ assay of Study I, the MB assay of Studies I and II and in minisequencing on tag-arrays used in Studies III and IV. The fluorescence signal from one of the two possible SNP alleles was divided by the sum of the signals from both alleles. The fluorescence signal curves for a MB assay is shown in Figure 12, panel A & B. For genotyping with minisequencing using tag-arrays in Studies III and IV, the fraction was plotted on the x-axis and the logarithm of the sums on the y-axis. (Figure 12, panel C). The SNPSnapper software (http://www.bioinfo.helsinki.fi/SNPSnapper/) was used to make these graphs and to assign the genotypes. In the publications, this fluorescence signal fraction has been denoted the signal ratio and thus I will use this nomenclature throughout the thesis.

As output from minisequencing with FP detection used in Study II, mP values are obtained as described previously in section 4.1. Using the software AlleleCaller, supplied with the fluorescence detection instrument, the mP values corresponding to each allele were plotted against each other and genotypes were assigned to the formed clusters (Figure 12, panel D).

\[
\text{log(\text{SignalAllele}_1 + \text{SignalAllele}_2) / \text{SignalAllele}_2}
\]

\[
\text{Hom 11, negatives and failed}
\]

Figure 12. Principles for genotype assignment. The fluorescent signal curves from the MB assay of 96 samples are shown for allele 1 (A) and allele 2 (B). The three genotypes are distinguished directly from the curves. For minisequencing on tag-arrays the signal ratio is plotted on the x-axis and the logarithm of the sum of the signals on the y-axis. The genotypes are then assigned based on the cluster formation (C). In minisequencing with FP detection, the mP values from the two labels are plotted against each other and genotypes are assigned based on the cluster formation (D). The genotypes are denoted Hom 11 (Homozygous allele 1), Het 12 (heterozygous), and Hom 22 (homozygous allele 2).
In Studies III and IV the signals were further analyzed to evaluate the performance of different templates or. To obtain signal to noise ratios for each SNP, all signals corresponding to the allele given by the genotype of the sample, were considered as signal and all other fluorescent signals from the same sample as noise. In Study III the signals were also analyzed to assess how a lower amount of input genomic DNA in the MDA reaction affected the balance between alleles in the genotyping assays. Four parallel MDA reactions were performed on genomic DNA from an individual selected to be heterozygous for four SNPs that were interrogated in both polarities and genotyped triplicate reactions.

The genotypes of the rotavirus samples of Study V were deduced by visual examination of the fluorescent spot pattern on the arrays. However, genotyping was also possible with the aid of the numeric fluorescent signal intensities determined with the QuantArray analysis 3.1 software (Perkin-Elmer Life and Analytical Sciences, Boston, MA, USA). The normalized ratios for each sample were obtained by dividing the mean signal for the duplicate capture oligonucleotide spots by the mean signals from the quadruplicate MS2 spots. Background signals were calculated for each capture oligonucleotide from a corresponding spot in a negative control reaction well. In Studies III and IV the background was calculated for each subarray from printed cTags without corresponding tagged minisequencing primers.

2.5.3 Genotype discrimination and cluster quality

In Study I the term “power of genotype discrimination” (PoGD) was introduced and defined as the difference between the signal ratios from the homozygous and heterozygous samples. The PoGD was also used in Study III and is displayed graphically for all SNPs and all three templates in Figure 5 of the published study. Since clusters were used to define the genotypes the PoGD was defined as the average distance between the homozygous genotype clusters and heterozygous genotype clusters as measured on the signal ratio-scale along the x-axis.

The quality assessment of the fluorescent signal clusters used to assign genotypes in Study III and in the evaluation of MDA products from buccal cells in minisequencing with FP, was refined further by using a silhouette algorithm (Rousseeuw 1987) as described in Study IV. For each given data point (i) in a scatter plot, the silhouette \( s(i) \) is defined as:

\[
s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}}
\]

where \( a(i) \) is the average distance to all data points in the same genotype cluster and \( b(i) \) is the average distance to all data points in the closest cluster (Figure 13).
**Figure 13** Principle for silhouette s(i) calculations for a given point (i) in a scatter plot used for genotype assignment. The average distance to all points in the same cluster a(i) and to all points in the other two clusters b1(i) and b2(i) is illustrated. The scatter plot is obtained by plotting the fluorescent signals from an array experiment as previously described.

The silhouette width is calculated by averaging the s(i) for each genotype cluster and a silhouette score (also termed the overall average silhouette width) from each SNP assay is obtained as the average of the silhouette widths for all clusters in the scatter plot. The silhouette score can range from -1.0 to 1.0 with 1.0 being the optimal value.

### 2.6 Statistical analyses

The statistical analysis in Study II was performed at the Karolinska Institutet. We tested the SNP genotypes for Hardy-Weinberg equilibrium among cases and controls separately. We also calculated all pairwise linkage disequilibrium values |D'| and r^2. In Study IIa, since lobular and ductal breast cancer may have partly different etiologies these cases were analyzed separately. For the association analyses, using single loci genotypes, logistic regression analysis was used to calculate odds ratios (OR) and 95 percent confidence intervals (CI).

Disease-haplotype association was tested with likelihood-based approaches using either the software EH plus (Zhao *et al.* 2000) or routines written in the S-PLUS (Insightful) programming language using essentially the method used by Stram and colleagues (Stram *et al.* 2003). The in-house written program is able to estimate models other than multiplicative penetrance and enables adjustment for covariates. Results reported in Study II involving haplotypes are based on our own program. Several models were used, thus multiple testing using a permutation-based approach was performed.

In Study III, the coefficient of variation (CV) for the fluorescent signals from four parallel MDA reactions was calculated to assess possible bias between amplification of the two SNP alleles of a heterozygous sample when small amounts of genomic DNA are MDA amplified. Furthermore, the Pearson correlation coefficient was calculated and used to assess the correlation between the different templates used in minisequencing with tag-microarrays in Study III.
3 RESULTS AND DISCUSSION

3.1 Assay performance
Conclusions on assay performance can be drawn not only from the methodologically oriented Studies I, III, IV and V, but also from Study II where a large number of samples were genotyped simultaneously in the breast cancer and endometrial cancer material. The strength of a method to discriminate between genotypes and the closely related issue of cluster tightness and separation, both influence genotyping performance and were assessed.

3.1.1 Summarized results on assay performance
The main features and results of the specific studies with regard to the genotype assays used are summarized in Table 6 and are discussed in more detail in the following sections.

3.1.2 Homogeneous assays
Homogeneous assays have the benefits of a relatively easy setup and high throughput when a limited number of sequence variations are analyzed in a large sample set. This was the case in Study II where a large sample set was genotyped for two SNPs using the MB assay. A drawback of the homogenous TQ and MB assays is the need for dual labeled probes that are relatively expensive. On the other hand, the assay is performed directly during PCR with DNA polymerase as the only enzyme, reducing the cost of the assays.

TaqMan assay versus Molecular beacon probes
In Study I the genotyping performance of the TaqMan and Molecular beacon assays were evaluated as a pilot study to select the optimal method for the large genotyping effort required in Study II. Ninety DNA samples were analyzed for 3 SNPs with different nucleotide variations in the estrogen receptor (codon 10 TCT>TCC, codon 325 CCC>CCG, and codon 594 ACA>ACG).

Both methods had comparable PoGD ranging from 0.3 to 0.5 and with little variation between samples of the same genotype. For the TQ and MB assays, the lowest proportion of a minority allele that could be distinguished were 1.6% and 3.1% respectively, but varied between the SNPs. Nevertheless, the MB assay was generally found to resolve two-fold differences in allele proportions over a wider range than the TQ assay (Figure 14).

The TQ assay generated five genotype errors by mistyping samples homozygous for the codon 325 C-allele SNPs as heterozygotes. The region around the codon 325 SNP is particularly GC-rich, and the five bases flanking the SNP are all cytosines. The sequence context also influences the fluorescence signal by quenching effects inferred by the secondary structures and the nucleotides in vicinity of the fluorophore (Marras et al. 2002, Nazarenko I. et al. 2002).
<table>
<thead>
<tr>
<th>Study</th>
<th>Main features</th>
<th>Genotype assay results</th>
</tr>
</thead>
<tbody>
<tr>
<td>I</td>
<td>Compares genotyping performance of the TQ and MB assays. 3 SNPs in 90 samples.</td>
<td>Both methods had comparable PoGD. The TQ assay was slightly more sensitive but the MB assay resolved two-fold differences in allele proportions over a wider range and gave more robust genotyping, especially for a SNP in a GC-rich region.</td>
</tr>
<tr>
<td>II</td>
<td>Large genotyping effort using mainly MB assay and minisequencing with FP detection. 4 SNPs in 4,000 samples.</td>
<td>3,000 genotypes produced using the MB assay with success rate &gt; 99%. No conflicts between parallel reactions or methods. 12,200 genotyping assays were performed using minisequencing with FP detection. All samples were genotyped in both polarities and approximately 30 percent of the assays were repeated with no conflicts.</td>
</tr>
<tr>
<td>III</td>
<td>Compares the use of MDA and PEP to genomic DNA as template for multiplex PCR with subsequent genotyping with minisequencing and Tag-arrays. 45 SNPs in 15 samples.</td>
<td>The overall success for genomic samples was 97.0%. Results from the PEP and MDA templates corresponded to the genomic with 88.7% and 99.7% respectively. MDA products correlated well with the genomic results in all aspects studied. The most reproducible results were obtained with 3 ng genomic DNA (~1,000 genome equivalents) subjected to MDA.</td>
</tr>
<tr>
<td>IV</td>
<td>Applied silhouette scores as a quality measure for SNP genotyping clusters. Evaluated the performance of four DNA polymerases in minisequencing with Tag-arrays. 26 SNPs in 16 samples.</td>
<td>The silhouette score is a useful tool to measure genotype cluster quality. Both cluster tightness and distances between clusters are taken into account. The silhouette score could be generally applicable to all genotype assays were genotypes are assigned based on cluster formation. Of the four DNA polymerases, KlenThermase showed the best results over all.</td>
</tr>
<tr>
<td>V</td>
<td>New method for HRV typing applying genotype-specific primer extension on microarrays. 33 genotype-specific capture oligonucleotides and 40 samples.</td>
<td>The final array detected HRV and assigned complete genotypes in all samples. No false genotypes were obtained. All genotypes were in concordance with results obtained by reference methods. A normalized ratio was a good measure for possible automatic genotype scoring.</td>
</tr>
<tr>
<td>MDA-</td>
<td>Investigated MDA products from DNA extracted from buccal cells as a template in minisequencing with FP detection. 4 SNPs in 44 samples.</td>
<td>MDA samples had a better genotyping success rate than the original genomic DNA. 98.8% concordance between genotypes from genomic DNA and MDA product. Low concentration samples were genotyped only after MDA.</td>
</tr>
</tbody>
</table>

Abbreviations: MB, Molecular beacon assay; TQ, TaqMan assay; PoGD, power of genotype discrimination; MDA, multiple displacement amplification; PEP, primer extension preamplification; HRV, human rotavirus; FP, fluorescence polarization.
Figure 14. Results from genotyping mixed samples with the TaqMan (TQ) and Molecular beacon (MB) assays. The results are the mean fluorescence signal ratios from five parallels and the standard deviations are indicated by vertical bars. The percentages of the alleles represented by the different shades are indicated below the diagram.

MB probes can form a hairpin formation, which competes with its hybridization to the target, thus providing greater sensitivity for mismatches and a wider useful temperature range, compared to linear probes such as TQ probes (Tyagi et al. 1998, Bonnet et al. 1999).

Based on the results from Study I, where the MB assay appeared to be more robust in the GC-rich region of the codon 325 SNP, the MB assay was chosen for genotyping the majority of the samples for the two exonic ESR1 SNPs analyzed in Study II. 3,076 genotypes were produced using the MB assay, 2% of the assays were replicated and 3% of them were analyzed with a reference method. No conflicts between parallel reactions or methods were observed. In total we were able to genotype the two exonic SNPs in more than 99% of the DNA samples.

Since the publication of Study I, several modifications and improvements have been made both to the TQ and MB assays. A major improvement to the TQ assay was to conjugate minor groove binding groups to the probe to increase the stability of the probe target hybrid (Kutyavin et al. 2000). The conjugated group folds into the terminal minor groove of the double stranded DNA molecule formed by the probe and target. This stabilizes the hybridization and allows shorter probes to be used, thus enhancing allelic discrimination. Probes used today are normally 13-20 nucleotides long according to Applied Biosystems user bulletin. This is substantially shorter than the 20-25 bp probes used in Study I. The shorter probes increase quenching efficiency and render the oligonucleotide synthesis less error-prone. Furthermore, the TQ assays are now performed with non-fluorescent quenchers enabling some multiplexing (Lee et al. 1999). An advantage of the
TQ assay is the independence of the temperature during fluorescence monitoring which allows for end-point detection and eliminates the need for real-time PCR instruments (de Kok et al. 2002). Additionally, the TQ assay has been successfully combined with FP detection, which eliminates the need for a quencher (Latif et al. 2001).

The stability of MB hybrids has been increases by using peptide nucleic acid (PNA) allowing for shorter probes (Ortiz et al. 1998, Petersen et al. 2004). Since the fluorescent measurement in MB assays is made on an intact probe during hybridization to its target, MB probes, in contrast to TQ probes, can be immobilized on solid-surfaces (Steemers et al. 2000, Wang H. et al. 2002b). Wave-length shifting MB probes have been developed to allow higher multiplexing. A single monochromatic light is used to excite a harvester fluorophore. When the MB probe is in the hairpin formation, the harvester is quenched by a quencher at the opposite end of the probe. When the probe is hybridized, the absorbed energy will be transferred through FRET to an emitter fluorophore attached to the harvester by spacer nucleotides and fluorescence will be emitted. By using a range of emitter fluorophores for different probes, a 10-plex multiplexing capacity is reached (Tyagi et al. 2000).

3.1.3 Semi-homogeneous assays

Semi-homogeneous assays are similar to homogeneous assays in respect to the optimal number of SNPs and samples. Since the two intronic SNPs in Study II were genotyped at a later time than the exonic SNPs, minisequencing with fluorescence polarization detection (Chen et al. 1999) was chosen due to the lower cost combined with its simple laboratory performance and assay design. This project was the first large genotyping study using this assay, performed in our laboratory. Approximately 4,000 samples were genotyped for the two SNPs in both polarities by one person using robotic PCR setup in slightly less than three months including re-runs and control experiments. This method is now one of the routine assays in the genotyping service facility of our research group.

Minisequencing using fluorescence polarization detection

For the two intronic SNPs in Study II, a total of 12,200 genotyping assays were performed with an overall success rate of 86%. Both of the SNPs were analyzed in both polarities, approximately 30% of the assays were repeated with identical results and 3% of the genotypes were validated by solid-phase minisequencing.

The assays in Study II were performed on DNA samples that had not been normalized with regard to concentration. This contributed to some difficulties in genotype calling and also resulted in PCR failures as the dominating reason for assay failures. Optimally the concentration of the samples should be measured. However, the same DNA extraction method was used throughout the sample preparation which limits variations in concentration. Thus, it is a less time consuming approach to initially run the whole sample set and then
make the necessary adjustments of the concentrations of samples to be reanalyzed.

The genotypes were scored manually, using the AlleleCaller software provided with the analyze instrument. In a comparison by van den Oord, manual scoring was somewhat more error prone than three tested statistical scoring methods (van den Oord et al. 2003). However, in our study the inclusion of positive controls and analysis of both polarities improve the reliability in genotyping.

It has been proposed that the addition of single strand binding protein prior to FP readout and a reading buffer containing ethanol and glycerol may increase success rate (Hsu et al. 2001a). In Study II the assays worked satisfactorily and no further optimization attempts were made. Hsu et al. also used a universal extension mixture with all four nucleotides labeled enabling any SNP to be studied with the same reagents.

Minisequencing with FP detection is able to detect 40:60 differences and has been applied to quantitatively measure allelic differences in expression levels (Pastinen et al. 2004). A kinetic version of the assay has been developed for determining SNP allele frequencies in pooled DNA samples. A heterozygous sample is needed as a reference, and the detection is performed in real-time during the minisequencing reaction. The monitored rate of increase in FP is proportional to the amount of the two alleles. The smallest detected frequency was approximately 5% and the deviation from the true allele frequencies from individual genotypes was 3.3±0.8% (Xiao et al. 2003). This sensitivity is comparable to our results using the TQ and MB assays in Study I. Minisequencing on solid-phases is able to detect allele frequencies of <1% and 2-5% for microtiter plates using ³H labeled dNTPs and microarrays using fluorescently labeled ddNTPs respectively (Olsson et al. 2000, Lindroos et al. 2002).

**WGA samples and minisequencing with FP detection**

Based on the results of Study III (section 3.1.4), we were interested in evaluating the performance of MDA products in the minisequencing assay with FP detection routinely used by the genotyping facility of the research group. A total of 44 samples were genotyped for up to 4 SNPs. A total of 150 genotypes were produced with a concordance of 98.8% (80/81 genotypes) between genomic DNA and MDA product. However, approximately half of the samples failed to genotype for either or both of the templates (Table 7).

Genotyping was successful for MDA products in the majority of cases, reflecting the ability of the MDA reaction to perform amplification with only minute amounts of DNA as input (Dean et al. 2002, Lovmar et al. 2003b, Rook et al. 2004). One MDA sample was genotyped as homozygous, although the corresponding genomic sample gave a heterozygous genotype. No genotypes derived from the MDA product were in inheritance conflict when compared to known genotypes of family members, indicating an even amplification of the two alleles, see further the discussion in section 3.1.4.
Table 7. Distribution of the obtained genotyping results (+) and failures (-) from genomic (gDNA) and multiple displacement amplification (MDA) sources

<table>
<thead>
<tr>
<th></th>
<th>gDNA +</th>
<th>gDNA -</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>MDA +</td>
<td>81 (54%)</td>
<td>49 (32.7%)</td>
<td>130</td>
</tr>
<tr>
<td>MDA -</td>
<td>6 (4%)</td>
<td>14 (9.3%)</td>
<td>20</td>
</tr>
<tr>
<td>Total</td>
<td>87</td>
<td>63</td>
<td>150</td>
</tr>
</tbody>
</table>

The buccal cell extracted DNA samples analyzed had a concentration range of 2–200 ng/µl. There was no concordance between concentration and failure to genotype MDA samples. All samples (total of 4) with concentrations below 5 ng/µl failed when genotyping was attempted on genomic DNA directly.

The clusters used for genotyping were also analyzed to obtain silhouette scores, see section 3.1.4. Since only four SNPs were analyzed, of which three had low rare allele frequencies, it is difficult to draw a definitive conclusion. However, the MDA clusters gave silhouette scores at least as good as the genomic template. The silhouette scores for the four SNPs, when the failed samples were excluding, were 0.91, 0.86, 0.92, 0.94 and 0.92, 0.92, 0.95, for genomic DNA and MDA product respectively. The better performance of MDA products is also indicated by the higher genotype success rate.

Biobanked material may only be present in sparse amounts. For example, buccal swabs provide only sparse quantities of DNA with a large variation in concentration. Hence, MDA, or other WGA methods, is a promising technique for providing enough material for future genotyping (Hosono et al. 2003).

3.1.4 Microarrays
In all three microarray studies of the thesis the "array of arrays" format was applied. In Studies III and IV universal microarrays with immobilized cTags were used in combination with cyclic minisequencing reactions performed in solution. In Study V, microarray technology was applied to genotype human rotavirus group A using genotype-specific extension of immobilized primers. The “array of arrays” format is especially well suited for performing comparisons, e.g. between different reaction conditions, since up to 80 samples can be processed in parallel at the same conditions. The four-color minisequencing system used in Study III and IV is also beneficial because signal and noise are measured from the same spot as described above.

When analyzing many SNPs in parallel on microarrays, the different clustering properties between SNPs become obvious. The reason for the differences are that the flanking sequence as well as the fluorophores attached to the dideoxynucleotides affect the efficiency and sequence specificity of nucleotide incorporation by the DNA polymerase (Lindroos et al. 2002). Additionally, different properties of the fluorophore, and the sequence context influence the fluorescence signal (Marras et al. 2002, Nazarenko I. et al. 2002).
Silhouette scores for cluster quality assessment

To perform a quantitative assessment of the genotyping clusters quality we developed a script to assign silhouette scores. The silhouette score is dependent on the distance between clusters and on how tight the clusters are; both features are relevant for robust genotype assignment. This implies that the quality of the assignment to clusters, i.e., how well the genotyping is performed, is assessed and condensed into one single measure in the range of -1.0 to 1.0. Optimally a SNP should give correct genotypes and high silhouette scores as a result of the cluster validation. Silhouettes were originally introduced in 1987 as a general graphical aid in the interpretation and validation of cluster analysis (Rousseeuw 1987).

Silhouette scores were applied to the genotype data in Studies III, IV and the MDA-FP study and can be calculated by using distances measured either in one dimension or as vectors (Figure 13). For the silhouette scores calculated in Studies III and IV, distances measured only on the x-axis, i.e., the differences in fluorescence signal ratios, were used. In Figure 15, the silhouette scores for a selection of genotype clusters from minisequencing on Tag-arrays are shown. In minisequencing with FP detection, genotypes are assigned based on clusters with the same measure on both axis (see Figure 12, panel D), thus vectors were chosen to calculate the silhouette scores in the MDA-FP study.

Figure 15. Example of silhouette scores for SNP genotyping assays using minisequencing on Tag-arrays. The three genotype clusters are indicated by black squares, grey triangles and black filled circles and the silhouette score is given in the top right corner of each graph.

Our “array of arrays” microarray system is a useful tool for comparative analysis, as was done in Study III and IV, especially in combination with the silhouette score to assess cluster quality. The silhouette score has the potential to be useful as quality measure in any genotyping system where the genotypes...
are assigned by cluster analysis and is included as part of the complex quality system used for interpretation of genotype data from Affymetrix hybridization arrays (Liu et al. 2003).

**WGA samples and minisequencing on Tag-arrays**

In Study III, the “array of arrays” microarray system was applied to investigate the performance of templates from two whole genome amplification procedures, primer extension preamplification and multiple displacement amplification compared to genomic DNA. The three different types of template were used in multiplex PCR with subsequent genotyping by cyclic minisequencing in solution, followed by capture of the extended primers on Tag-arrays. The assay procedure is illustrated in the flowchart of Figure 16.

![Figure 16. Procedure for assay set up and genotyping of SNPs by cyclic minisequencing with capture on Tag-arrays. Required preparations and assay development are shown to the left whereas the genotyping procedure is given to the right. © Humana Press](image)

A 14x14 array of immobilized cTags was used and 45 SNPs in both polarities were analyzed. This is the highest spot-density applied thus far in our laboratory. In total 15,000 spots and 60,000 fluorescent signals were analyzed in one array experiment.

Based on initial titrating experiments, 3 ng of DNA, equivalent to 1,000 copies of the genome, were used as starting material for MDA and PEP. The overall success for genomic samples was 97.0% and no conflicts of inheritance were observed. The results from the PEP and MDA templates corresponded to the genomic with 88.7% and 99.7%, respectively. Overall, the PEP template generated lower genotype success rate, silhouette scores, PoGD and S/N ratios (Table 8). The PoGD values and silhouette scores emphasized the differences in cluster quality between the templates.
Table 8. Genotyping call rate, success rate and average silhouette score, power of genotype discrimination (PoGD), and signal to noise ratios (S/N) for the three templates

<table>
<thead>
<tr>
<th>Template</th>
<th>Call rate</th>
<th>Success rate</th>
<th>Silhouette score</th>
<th>PoGD</th>
<th>S/N</th>
</tr>
</thead>
<tbody>
<tr>
<td>Genomic</td>
<td>97.0%</td>
<td>100%</td>
<td>0.88 ± 0.11</td>
<td>0.44 ± 0.18</td>
<td>36</td>
</tr>
<tr>
<td>PEP</td>
<td>90.2%</td>
<td>88.7%</td>
<td>0.81 ± 0.14</td>
<td>0.36 ± 0.14</td>
<td>14</td>
</tr>
<tr>
<td>MDA</td>
<td>97.6%</td>
<td>99.7%</td>
<td>0.85 ± 0.12</td>
<td>0.44 ± 0.17</td>
<td>34</td>
</tr>
</tbody>
</table>

1 Abbreviations: PoGD, power of genotype discrimination; S/N, signal to noise ratio; PEP, primer extension preamplification; MDA, multiple displacement amplification.
2 The genotyping results from the genomic DNA were regarded as correct and the results from MDA and PEP templates were compared to the results from genomic DNA.
3 The standard deviations of the averages for all SNPs and template types are indicated.

The MDA products correlated well with the results from genomic DNA (Table 8). The Pearson correlation coefficients were 0.997 for fluorescence signal ratios from all samples in the duplicate reaction with genomic DNA and 0.963 and 0.991 for the PEP and MDA templates respectively, as compared to the use of genomic template (Figure 17).

Figure 17. Correlation between fluorescence signal ratios for the three templates in a single microarray experiment. The signal ratios from the reference genomic DNA are plotted on the x-axis and the parallel genomic DNA (black circles), MDA (light gray diamonds), and PEP (dark gray triangles) are plotted on the y-axis. The linear trend lines are shown and their coefficients of determination (R²) are indicated top left next to the legends.

Most previous studies evaluating WGA protocols have focused on the representation of different genomic loci in the WGA products; however, when performing SNP genotyping un-biased amplification of the two alleles is the most crucial feature. Therefore, MDA templates were analyzed to assess
possible bias between amplification of the alleles and DNA polarities when small amounts of genomic DNA are MDA amplified. Independently of the input of genomic template (3 ng, 0.3 ng, 0.03 ng or 3 pg) we obtained a yield of 200–400 ng/µl per MDA reaction when quantified using the PicoGreen assay. This is in agreement with the reported plateau reached in a 100 µl MDA reaction at about 20–30 µg of DNA independently of the amount of genomic template (Dean et al. 2002). Minisequencing results from the different amounts of input genomic DNA to MDA are summarized in Table 9. Four parallel MDA reactions were performed on genomic DNA from an individual selected to be heterozygous for the four assayed SNPs that were interrogated in both polarities and in triplicate reactions. The results show that the most reproducible fluorescent signals were obtained with 3 ng (~1,000 genome equivalents) of genomic DNA subjected to MDA. This result is concordant with a study by Stenman et al., who observed increased stochastic variation in signal ratios after competitive PCR with less than 1,000 molecules of mRNA (Stenman et al. 1999).

Table 9. Different genomic input in MDA

<table>
<thead>
<tr>
<th>Amount of DNA</th>
<th>Signals 1</th>
<th>Coefficient of variation 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>3 ng</td>
<td>32 (100%)</td>
<td>Acceptable (&lt; 20%)</td>
</tr>
<tr>
<td>0.3 ng</td>
<td>32 (100%)</td>
<td>High (mean &gt; 20%)</td>
</tr>
<tr>
<td>0.03 ng</td>
<td>22 (69%)</td>
<td>High (mean &gt; 40%)</td>
</tr>
<tr>
<td>0.003 ng</td>
<td>8 (25%)</td>
<td>n.d.</td>
</tr>
</tbody>
</table>

1 Reactions that yielded detectable fluorescence signals (maximum of 32).
2 Obtained from signal ratios from four heterozygous SNPs in four parallel MDA reactions analyzed in triplicate and assayed in both polarities, n.d. = not determined

Interestingly, when a low input of DNA was used, we detected unequal fluorescent signals from the two polarities analyzed. This is possible if considering the high processivity and hyperbranched but linear amplification mediated by the Φ29 DNA polymerase. A strategy to avoid these problems when low inputs of genomic DNA are used is to perform parallel amplification reactions and to pool the WGA products prior to genotyping (Rook et al. 2004).

In the absence of templates MDA may produce spurious products occasionally visible on agarose gel. These products have not yielded any PCR-products and have not interfered with genotyping. Introduction of two 5’-terminal nitroindole residues in the random primers has been shown to limit spurious amplification although the mechanism for this phenomenon is not understood (Lage et al. 2003).

Extension enzymes in minisequencing on Tag-arrays

The DNA polymerase is the major factor determining the efficiency and specificity of a minisequencing reaction, irrespective of the assay format. Recently, new thermostable enzymes compatible with fluorescent ddNTPs

64
have become available. In Study IV, we compared the performance of three commercially available DNA polymerases: TERMIPol (Solis BioDyne), Therminator (New England BioLabs, Inc.) and KlenThermase (Gene Craft) to ThermoSequenase (Amersham Biosciences), which is routinely used in minisequencing assays in many laboratories, including our own. A subset (26) of the SNPs used in Study IV was analyzed using the same four-color fluorescence minisequencing system in a microarray format. The “array of arrays” format was used, allowing up to 80 samples to be analyzed in parallel. The genotyping performance was evaluated with respect to fluorescent signals, clustering properties, signal-to-noise ratios and genotyping success.

In Figure 18 the average fluorescent signal intensities for each ddNTP with respect to the genotype of the samples are shown for the four DNA polymerases. The signal intensities were similar, with the exception of the T-signals that were lower using Therminator and ThermoSequenase. Generally KlenThermase generated high levels of correct signals together with little mis-incorporation, and ThermoSequenase generated lower signals, but also had a lower level of mis-incorporation.

![Figure 18. Average fluorescent signal intensities with respect to SNP genotype. The average signals for each ddNTP type for homozygotes (AA, CC, GG and TT), heterozygotes (AX, CX, GX and TX) and signals from mis-incorporated ddNTPs from samples not carrying the nucleotide type (XX) are given on the y-axis for each nucleotide and DNA polymerase.](image)

All four enzymes performed satisfactorily in our minisequencing assay. However, the performance varied between the evaluated features as summarized in Table 10. The overall results, especially if considering the cost, would be in favor for KlenThermase.
Table 10. *Silhouette score, genotyping performance and signal to noise ratios for the four enzymes*<sup>1</sup>

<table>
<thead>
<tr>
<th>Genotype calls&lt;sup&gt;2&lt;/sup&gt;</th>
<th>Average silhouette score</th>
<th>Highest silhouette score</th>
<th>Genotype calls&lt;sup&gt;2&lt;/sup&gt;</th>
<th>Average S/N</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>n</td>
<td>%</td>
<td>n</td>
<td>%</td>
</tr>
<tr>
<td>TERMIPol</td>
<td>0.75</td>
<td>25</td>
<td>31.6</td>
<td>2337</td>
</tr>
<tr>
<td>Therminator</td>
<td>0.73</td>
<td>15</td>
<td>19.0</td>
<td>2323</td>
</tr>
<tr>
<td>KlenThermase</td>
<td>0.77</td>
<td>22</td>
<td>27.8</td>
<td>2346</td>
</tr>
<tr>
<td>ThermoSequenase</td>
<td>0.73</td>
<td>17</td>
<td>21.5</td>
<td>2324</td>
</tr>
</tbody>
</table>

<sup>1</sup> The average for all SNPs with results in all four enzymes is given for each enzyme.

<sup>2</sup> Number of genotype calls (n) and call rate (%). The genotype obtained from a majority of the assays was considered to be correct.

As discussed in the introduction, the polymerase is crucial for the performance of minisequencing independent of format and assay set-up. The use of a 3´-5´ proofreading DNA polymerase in combination with protected extension primers may further improve the performance by reducing noise (Di Giusto and King 2003).

**Genotype-specific primer extension on arrays for HRV genotyping**

The assay used in Study V combines the G- and P-type specific capture of RT-PCR amplified viral sequences on microarrays with genotype-specific DNA-polymerase mediated extension of the capture oligonucleotides with fluorescent dNTPs. Thus, the specificity of the primer extension reaction depends firstly, on the presence of a viral sequence that is sufficiently similar for hybridization to the capture oligonucleotide, and secondly, on the presence of a shorter 3´-genotype specific sequence with an obligatory requirement of a 3'-terminal nucleotide that is perfectly matched to the viral template. A total of 33 capture oligonucleotides were immobilized in each subarray. Seven oligonucleotides served as controls as described in section 2.5.1. Two or three genotype specific oligonucleotides were included per genotype.

All genotypes obtained in Study V were concordant with results obtained by sequence analysis and genotype-specific multiplex RT-PCR. However, several versions of the arrays were tested before the final version was successful with sequences representing all 40 samples. The possibility to easily expand and alter the arrays is one of the advantages with the applied array system.

A normalized ratio used to assign genotypes was obtained by dividing the mean signal for the duplicate capture oligonucleotide spots for each sample by the mean signals from the quadruplicate MS2 spots. In all samples the capture oligonucleotides scoring the highest ratio corresponded to the correct genotype rendering the ratio a good measure for automatic genotype scoring. The signal from corresponding capture oligonucleotides in a negative reaction well without DNA template was subtracted as background. The signals from the negative control capture oligonucleotides with a mismatched nucleotide in
the 3’-end were generally very low, but when slightly higher levels of background were observed this was accompanied with higher signals and thus eliminated by normalization of the signals.

HRV genotyping on microarrays has also been described using hybridization of PCR products to immobilized ASO probes (Chizhikov et al. 2002). This approach, in analogy with our assay, has several advantages to a strictly PCR based assay. Several probes are used for each genotype, making the assay less sensitive to new mutations. It is also easy to expand the arrays without additional optimization procedures. However, the procedure presented by Chizhikov et al. is relatively laborious and may also be vulnerable to unspecific signals from closely related genotypes. In contrast, the hybridization in Study V was combined with a genotype-specific extension providing an increased specificity through the action of the DNA polymerase.

3.2 Applications
The investigated genotyping assays has to some extent been applied in an epidemiological context. Both breast cancer and endometrial cancer development is estrogen dependent; thus, a possible association between variants of the estrogen receptor alpha gene (ESR1) and cancer risk were investigated, see sections 3.2.1 (Study IIA) and 3.2.2 (Study IIb). The HRV genotyping method developed in Study V were used to genotype part of the material in an epidemiological study of infectious HRV strains in Nicaragua, see section 3.2.3.

3.2.1 ESR1 polymorphisms in breast cancer
In Study IIA, using logistic regression models, odds ratios and 95% confidence intervals no association between the ESR1 SNPs and breast cancer risk was found, when analyzed separately. However, we found suggestions of an association between the risk for postmenopausal ductal breast cancer of mild to moderate severity and ESR1 haplotypes formed by the alleles of either of the intronic SNPs and the codon 325 SNP. The association appeared stronger for women with higher body mass index (BMI). One copy of the intron 1 -351A and codon 325C haplotype gave an OR of 1.19 (CI 1.06-1.33) and two copies of this haplotype gave an OR of 1.42 (CI 1.15-1.77). The association with the intron 1 -397T and codon 325C haplotype was similar. The associations were not statistically significant after adjustment for multiple comparisons using a permutation based approach. Despite the moderate strength of the association, the high-risk haplotypes have the potential to play a substantial role in breast cancer etiology overall because of their high prevalence.

The studied ESR1 polymorphisms were not specifically selected to define haplotypes. The ability to capture the haplotype diversity of the ESR1 gene is limited using only four markers chosen without regard to LD patterns. However, the strength of Study IIA is that it is population-based and has a large sample size compared to previous studies. Additionally, the results in
the previous literature on ESR1 polymorphism, in relation to breast cancer risk, have been inconclusive (see Table 1 of the publication).

The efforts to search for genes contributing to complex disease have grown dramatically over the past years. Nevertheless, as discussed earlier, the findings have been sparse. It is likely that the genetic alterations we are looking for have low penetrance or that disease arises due to several genetically distinct causal mechanisms. This genetic heterogeneity would distort gene-disease associations.

3.2.2 ESR1 polymorphisms in endometrial cancer

Study IIb is at present the largest study investigating the relation between ESR1 and endometrial cancer. We found an association between the microsatellite (TAn) and the two intronic ESR1 SNPs that are in considerable LD with each other, and endometrial cancer. The association with intron 1 -351A>G was the strongest, OR 0.75 (CI 0.60-0.93) for heterozygotes and OR 0.53 (CI 0.37-0.77) for GG homozygotes compared to AA homozygotes. The remaining markers lacked significant explanatory value once this SNP was taken into account. The results replicate the findings in a separate but equivalent Swedish study population with observed associations of similar strength and magnitude (Weiderpass et al. 2000).

In contrast to the findings in Study IIa, haplotype models did not provide any additional information. The association between ESR1 and disease was weaker for breast cancer than for endometrial cancer, most likely due to the more complex nature of breast cancer etiology, which leaves room for random variation in the data.

In conclusion, we found that intronic variation in ESR1 was associated with endometrial cancer risk with GG homozygotes for the intron 1 -351 SNP, having an almost halved risk compared to AA homozygotes.

3.2.3 Rotavirus genotyping

In Study V, fecal specimens collected in Leon (Nicaragua) and in Uppsala (Sweden) and determined to be rotavirus antigen positive were analyzed with the microarray-based method for genotyping HRV. All 40 isolates were amplified successfully and all samples were shown to contain HRV by the generic capture oligonucleotides on the array. The initial set of probe-primers assigned both G- and P-genotypes in 35 of the 40 strains analyzed and it was subsequently expanded to detect all analyzed sequences. No false genotypes were obtained. The results of genotyping using the microarray were in full agreement with results obtained by nucleotide sequence analysis and genotype specific multiplex RT-PCR (Gouvea et al. 1990, Gentsch et al. 1992). All G- and P-types represented by capture oligonucleotides in the array were present among the 35 strains that were fully typed: 6 samples [P8]G1, 22 samples [P4]G2, 1 sample [P9]G3, 3 samples [P6]G4, 1 sample [P8]G4, and 2 samples [P8]G9.

The technique established in Study VI has been applied to genotype samples from an epidemiological survey of infectious HRV strains in Nicaragua that is currently being completed. Samples were collected over
three years from 2001 to 2003 from children younger than three years that had acute diarrhea as part of the NeTropica collaborative research program. A total of 269 rotavirus positive samples were complete or partially G- and P-characterized, by using either the microarray assay, sequencing or multiplex RT-PCR.
EPILOGUE

Carl von Linnaeus was appointed professor of Medicine at Uppsala University in 1741. The knowledge of Linnaeus spanned a large area; although he was a physician by profession, his most famous work was the Systema Naturae – the celebrated classification of nature. Since the days of Linnaeus, the amount of knowledge in the academic world has increased. We do not only have one professor of Medicine at Uppsala University, but a range of professors in specialized subjects, which reflect the field of research of today.

One main focus of Molecular Medicine during the last decades has been development of new methods and technology to analyze genomes, as reflected in this thesis. Several of these techniques focus on analyzing a selected part, most often only single nucleotides, of the genome. It is expected that the analysis of genomic variations will lead to the dissection of the genetics of common diseases and add knowledge usable in several other research fields. For example, the genetic information could be channeled into diagnostics, pharmaceutical development and individualized treatment, or even cures for genetic disorders.

However, as techniques improve and information accumulates, it becomes increasingly obvious that one single person cannot grasp the whole subject, as was the situation at the days of Linnaeus. I have been dependent on my collaborators, without whom the results presented in this thesis would not have been possible. In addition to further technology development, one of the largest challenges for the future may be handling the abundance of collected data and analyzing and interpreting the results into findings which benefit humanity and ultimately improve life quality. To achieve this, co-operation between research groups and research fields will become increasingly important.
I have been a member of the Molecular Medicine research group headed by Ann-Christine Syvänen since 1998, when she first hired me to do some simple laboratory work in parallel with my studies. During all these years I have met and co-operated with many wonderful people. I will not mention all by name, some of you will receive a copy of the thesis, some will not, but I would like to thank all of you. Without you I would never have tried my wings in research.

Chrisse, Professor Ann-Christine Syvänen, you are the best supervisor one could ever wish for. Without you, your support and your trust in me, all from the beginning, I would never have done this. I am so happy that I did! You will forever be an important role model for me when it comes to science and how supervision should be done.

All the present and former members in the Molecular Medicine group, thank you for all the time spent. Especially Ulrika, Lotta R, Mona, Snaevar and Lotta O, for all discussions and knowledge-sharing. Katarina, thank you for being a wonderful friend and also setting the standard for how a PhD degree should be made. Thank you Anki, for running all the gels during my pregnancy and Lillebil for being helpful on all occasions.

My collaborators in the different projects definitely deserve a warm thank you. Especially Sara and Kåre, for also reading my thesis and answering my numerous questions, despite my lack of knowledge in epidemiology and virology.

“My” students, especially Kerstin, Annika and Johanna, thank you for all of your hard work that is partly included in the thesis, for helping me to learn how to teach and for being terrifically fun co-workers.

All office-sharers, you have become quite many, and the members of the research department, thank you for listening to all of my joys and problems whether you wished to or not.

This work has been financially supported by grants mainly from the Swedish Research Council and the instrumentation partly by K&A Wallenbergs Stiftelse. My participation in conferences and courses has been made possible by Upsala Läkareförbundet, Anna Maria Lundins Stiftelse (Smålands Nation)
and the European Genetics Foundation. Anders Wall Stiftelse made my pause from the medical studies feasible and encouraged me to continue with research.

Additionally, some friends have been especially important during my “research years”: Thank you: Karin, without whom I would never have joined the group; Malin, who has listened to “everything” and still always supports me; Johan and Kristin for your scientific encouragement and friendship; Anna and Fredrik for being there and keeping the connection to medicine; Holly for nice “fika” and for taking precious time to proofread my thesis.

I would also like to thank my family in Skåne: Dad, Anders, Magnus, your families and grandma, for supporting me and never tiring of asking questions about “how it is going with the research” even though I never could answer properly. Maybe now you have the answer? I especially thank you Mom for your endless support from the very beginning of life and throughout everything.

At last I would like to thank my “new” family. Martin, for being the best of friends and a great support. You always pick up the pieces when everything breaks, you have shared all the work during the last year and you even managed to read through and comment on the whole thesis. You are my most important critic and I love you for it, even though I would never admit it! Hugo, for being such a lovely and joyful child and for not complaining about all these hours at my office. Being with both of you has kept me from ruining my self during the thesis work!

Lovisa Lovmar, Uppsala, 22nd of August 2004
REFERENCES


Strachan T & Read A P. Human molecular genetics, Edn. 3. (Garland Science, New York, 2004).


Acta Universitatis Upsaliensis

Comprehensive Summaries of Uppsala Dissertations from the Faculty of Medicine

Editor: The Dean of the Faculty of Medicine

A doctoral dissertation from the Faculty of Medicine, Uppsala University, is usually a summary of a number of papers. A few copies of the complete dissertation are kept at major Swedish research libraries, while the summary alone is distributed internationally through the series Comprehensive Summaries of Uppsala Dissertations from the Faculty of Medicine. (Prior to October, 1985, the series was published under the title “Abstracts of Uppsala Dissertations from the Faculty of Medicine”.)