MICROBIAL IDENTIFICATION BY DETECTION OF LIGATION PROBES ON DNA MICROARRAY

Jarmo Ritari

Institute of Biotechnology
and
Department of Biosciences
Faculty of Biological and Environmental Sciences
University of Helsinki

ACADEMIC DISSERTATION IN GENETICS

To be presented for public examination with the permission of the Faculty of Biological and Environmental Sciences of the University of Helsinki in the lecture room 3 at Infocenter Korona (Viikinkaari 11), on February 10th 2012, at 12:30.

Supervisor: Docent Petri Auvinen
Institute of Biotechnology
University of Helsinki
Finland

Thesis committee: Docent Outi Monni
Institute of Biomedicine
University of Helsinki
Finland

Professor Minna Pirhonen
Department of Agricultural Sciences
Faculty of Agriculture and Forestry
University of Helsinki
Finland

Pre-examiners: Professor Jari Valkonen
Department of Agricultural Sciences
Faculty of Agriculture and Forestry
University of Helsinki
Finland

Professor Joakim Lundeberg
Department of Gene Technology
Royal Institute of Technology
Sweden

Custos: Professor Tapio Palva
Department of Biosciences
Faculty of Biological and Environmental Sciences
University of Helsinki
Finland

Opponent: Docent Janna Saarela
Institute for Molecular Medicine Finland
University of Helsinki
Finland

ISNN 1799-7372
ISBN 978-952-10-7599-5 (paperback)
ISBN 978-952-10-7600-8 (PDF; http://ethesis.helsinki.fi)

Helsinki University Print, Helsinki 2012

CONTENTS

ABSTRACT

Microbes in natural and artificial environments as well as in the human body are a key part of the functional properties of these complex systems. The presence or absence of certain microbial taxa is a correlate of functional status like risk of disease or course of metabolic processes of a microbial community. As microbes are highly diverse and mostly not cultivable, molecular markers like gene sequences are a potential basis for detection and identification of key types. The goal of this thesis was to study molecular methods for identification of microbial DNA in order to develop a tool for analysis of environmental and clinical DNA samples. Particular emphasis was placed on specificity of detection which is a major challenge when analyzing complex microbial communities. The approach taken in this study was the application and optimization of enzymatic ligation of DNA probes coupled with microarray read-out for high-throughput microbial profiling.

The results show that fungal phylotypes and human papillomavirus genotypes could be accurately identified from pools of PCR amplicons generated from purified sample DNA. Approximately 1 ng/µl of sample DNA was needed for representative PCR amplification as measured by comparisons between clone sequencing and microarray. A minimum of 0,25 amol/µl of PCR amplicons was detectable from amongst 5 ng/µl of background DNA, suggesting that the detection limit of the test comprising of ligation reaction followed by microarray read-out was approximately 0,04%. Detection from sample DNA directly was shown to be feasible with probes forming a circular molecule upon ligation followed by PCR amplification of the probe. In this approach, the minimum detectable relative amount of target genome was found to be 1% of all genomes in the sample as estimated from 454 deep sequencing results.

Signal-to-noise of contact printed microarrays could be improved by using an internal microarray hybridization control oligonucleotide probe together with a computational algorithm. The algorithm was based on identification of a bias in the microarray data and correction of the bias as shown by simulated and real data. The results further suggest semiquantitative detection to be possible by ligation detection, allowing estimation of target abundance in a sample. However, in practise, comprehensive sequence information of full length rRNA genes is needed to support probe design with complex samples.

This study shows that DNA microarray has the potential for an accurate microbial diagnostic platform to take advantage of increasing sequence data and to replace traditional, less efficient methods that still dominate routine testing in laboratories. The data suggests that ligation reaction based microarray assay can be optimized to a degree that allows good signal-to-noise and semiquantitative detection.

LIST OF ORIGINAL PUBLICATIONS

This thesis is based on the following articles which are referred to in the text by their Roman numerals. The articles are reprinted with the kind permission of the publishers.

I Hultman J, Ritari J, Romantschuk M, Paulin L, Auvinen P. Universal ligation-detection-reaction microarray applied for compost microbes. BMC Microbiol. 2008 Dec 30;8:237.

II Ritari J, Paulin L, Hultman J, Auvinen P. Application of hybridization control probe to increase accuracy on ligation detection or minisequencing diagnostic microarrays. BMC Res Notes. 2009 Dec 14;2:249.

III Ritari J, Hultman J, Fingerroos R, Tarkkanen J, Pullat J, Paulin L, Kivi N, Auvinen P, Auvinen E. Detection of human papillomaviruses by polymerase chain reaction and ligation reaction on universal microarray. *Submitted*

IV Ritari J, Koskinen K, Hultman J, Kurola JM, Kymäläinen M, Romantschuk M, Paulin L, Auvinen P. Molecular analysis of meso- and thermophilic microbiota associated with anaerobic biowaste degradation. *Submitted*

AUTHOR'S CONTRIBUTIONS

I JR participated in conceiving of the study, designing the LDR probes and performing the microarray experiments. JR wrote the scripts and analyzed the microarray data. JR participated in preparing the manuscript.

II JR conceived of the study, analyzed the data and prepared the manuscript.

III JR participated in conceiving of the study, designing the LDR probes and performing the microarray hybridizations. JR analyzed the microarray data. JR participated in preparing the manuscript.

IV JR participated in conceiving of the microarray and sequencing experiments. JR designed the ligation probes and performed the microarray experiments, qPCR experiments and analyzed the data. JR participated in preparing the manuscript.

ABBREVIATIONS

AMP/ATP	adenosine mono/triphosphate
cDNA	complementary DNA
DNA	deoxyribonucleic acid
ITS	internal transcribed spacer
LDR	ligation detection reaction
LOESS	locally weighted quadratic least squares
NTase	nucleotidyl transferase
PCR	polymerase chain reaction
RCA	rolling circle amplification
RNA	ribonucleic acid
rRNA	ribosomal RNA
ssDNA	single stranded DNA
T_m	nucleic acid melting temperature

1 INTRODUCTION

1.1 DNA MICROARRAY TECHNOLOGY

DNA microarrays are composed of distinct microscopic spots containing DNA probes on a solid support like glass slide, silicon wafer or microbead. This architecture enables multiple parallel hybridization reactions to take place between the probes and sample nucleic acids. The immobilized DNA probes in each spot are designed to match a specific complementary sample target sequence. The probes can vary in length from hundreds of bases to approximately twenty bases depending on application and microarray fabrication technique. Hybridized target molecules at each spot are detected in laser scanning by excitation of a fluorescent dye incorporated in the target DNA, followed by recording the emission at a narrow wavelength window to block excitation wavelengths. The fluorescence signal intensity is therefore proportional within a certain intensity range to the bound sample DNA in each spot. Typically scanners record 16 bit images allowing 65563 values per pixel.

In gene expression profiling two cDNA pools, typically labeled with green-fluorescent cyanine 3 (Cy3) and red-fluorescent cyanine 5 (Cy5), are compared through competitive hybridization on the same microarray to identify differentially expressed genes. In practice, the dynamic range is limited by background noise and signal saturation. The background noise, caused by hybridization and cross-hybridization artifacts (Okoniewski & Miller 2006, Casneuf et al. 2007), impedes analysis of low-abundance targets and necessitates complex normalization procedures to overcome technical biases to make different probes and microarrays comparable. The upper bound to the dynamic range is imposed by signal saturation, making differential analysis above a certain intensity threshold difficult. However, in between the boundaries the relationship of signal and target abundance is assumed to be approximately linear because the number of immobilized probe molecules is much higher than the number of target molecules in the hybridization mixture, so that target binding efficiency is not dependent on target abundance under equilibrium. Dynamic range can be extended for instance by scanning the same slide at varying intensities and modeling the distribution of signals according to the obtained data (Gupta et al. 2006).

For about 15 years microarrays have been the essential tool in functional genomics in genome-wide gene expression profiling supported by genome projects that have provided increasing amounts of sequence data of model organisms. Microarrays have also shown significant utility in understanding for instance single nucleotide polymorphisms (Cutler et al. 2001), copy number variations (Iafrate et al. 2004), pathogen analysis (Aittamaa et al.

2008), susceptibility to diseases (Rujescu et al. 2009) and drug response profiles (Shah et al. 2011). In general, despite its versatility, DNA microarray technology is limited by the inability to analyze previously unmapped genes and requires detailed sequence information of targets and references under study.

1.1.1 MICROARRAY PLATFORMS

The first microarrays in research laboratories were manufactured by depositing small droplets of DNA-containing solution on filter membrane (Augenlicht & Kobrin 1982) and thereafter robotically on glass microscope slide surface with special printing pins (Schena et al. 1995). The molecules applied to the spots by contact printing have traditionally been amplicons like shotgun library clones (Hayward et al. 2000) and cDNAs (Schena et al. 1995) or short oligonucleotide probes (Guo et al. 1994) but over the years applications have diversified to include proteins (MacBeath & Schreiber 2000), antibodies (Rivas et al. 2008) and even live cells (Rantala et al. 2011). The contact-printing technique requires dedicated equipment with carefully controlled vibration, temperature, humidity, dust and other environmental factors. After deposition, the DNA can be attached to the microarray surface by electrostatic interactions between the negatively charged sugar-phosphate backbone and positively charged surface groups. Another way is ultraviolet (UV) crosslinking where the thymine bases of DNA are covalently linked onto the surface amine groups induced by UV irradiation. The chemical surface matrix can also contain aldehyde or epoxy groups to which DNA can be bound covalently through a 5' amino group.

High-density microarrays are built using on-chip oligonucleotide synthesis with solid-phase chemistry, photolabile protecting groups and photolithography (Pease et al. 1994). Lower density microarrays are typically produced by contact or inkjet printing of pre-synthesized molecules. Affymetrix (Santa Clara, CA, USA) pioneered the manufacture of in situ synthesized high-density oligo microarrays using photolitography and photochemical synthesis (Chee et al. 1996). A physical photolitography mask allows UV ligth to illuminate specific areas of the chip surface and thus direct the synthesis chemistry of DNA on the surface. Photochemical removal of protective group from deoxynucleosides and adding of nucleotides to the exposed positions is achieved by changing the masking pattern sequentially. Roche Nimblegen (Madison, WI, USA) produces microarrays with similar chemistry but in place of physical photolithography masks, the probes are synthesized using programmable micromirrors to focus light in specific patterns on the surface (Nuwaysir et al. 2002), enabling synthesis of over 60-mer probes and approximately 200000 features.

Oligonucleotide probes on high-density Agilent (Palo Alto, CA, USA) microarrays are produced by non-contact inkjet printing of chemically modified monomers. The synthesis uses phosphoramidite chemistry which

allows high coupling efficiency capable of producing over 100-mer probes (Hughes et al. 2001). A nucleoside phosphoramidite is a modified nucleotide harbouring protective groups at its exocyclic amine and hydroxyl groups to prevent them from reacting. The polymerisation is based on the exposed, reactive N,N-diisopropyl phosphoramidite at 3'-hydroxy position. Phosphoramidite nucleosides are added to the 5'-terminus of the chain in a stepwise fashion. At the end of the synthesis, all the protective groups are removed to release chemically normal DNA strand. The inkjet printing method allows precise deposition of picoliters of coupling and deprotection reactants at each cycle without physical contact, making the resulting array highly consistent and practically devoid of surface anomalies.

Illumina (San Diego, CA, USA) bead array is a high-density microarray platform employing 3 µm silica beads that assemble at random on a microwell substrate. The beads on the random array are mapped using identifiers, i.e. unique oligonucleotides specific for each bead type (Fan et al. 2006). Another kind of bead array is suspension bead array by Luminex (Toronto, Canada) which is based on microscopic polystyrene beads. Each bead type has an internal specific fluorescent label that is used to encode the beads. The beads harbour probe sequences binding to target DNA that is labeled in a separate reaction and hybridised on beads. The beads are measured using standard flow cytometry equipment enabling both bead encoding and hybridized DNA measurement simultaneously for each bead.

1.1.2 FACTORS DETERMINING HYBRIDIZATION SENSITIVITY AND SPECIFICITY

Strand complementarity forms the basis of sequence-specific nucleic acid detection and analysis techniques. As the stability of the double helix is determined by base interactions between the two hybridizing strands, instability caused by mispairing can be utilized to distinguish between sequences, since increasing the number of non-complementary bases lowers the stability of the hybrid. Consequently, in stringent reaction conditions, the number of mispairing bases determines whether the hybrid is formed or not enabling the use of sequence identity as a criterion for target detection. Generally, sequence specificity and binding affinity of DNA and RNA interactions negatively correlate with each other (Lomakin & Frank-Kamenetskii 1998). Longer probes have better affinity and therefore sensitivity of detection, but target specificity is decreased with probe length. Because the thermodynamic free energy gap ($\Delta\Delta G$) between correct and mismatched complexes is not substantially affected by an increase in affinity the mismatched complexes will be more stable in longer probes. In contrast, short oligonucleotides can discriminate even single nucleotide differences but their affinity is lower. A longer stretch of matches has higher affinity than the same number of matches separated by a mismatch (Ohrmalm et al.

2010). Thus, the position of a mismatch is an important determinant of probe specificity.

1.1.2.1 Base pairing interactions

Hydrogen bonds contribute strongly to the selectivity of DNA base pairing. In the double helix consecutive nucleotide bases are positioned towards the core of the helix such that hydrogen bonds are formed between geometrically compatible bond donors and acceptors in opposing complementary strands (Rich & Watson 1954). Watson-Crick rule of binding includes adenine-thymine and guanine-cytosine pairs while wobble base pairing, typically within a transfer RNA molecule, can take place between guanine-uracile, inosine-uracile, inosine-adenine and inosine-cytosine. Along with base pairing, base stacking as the sum of various weak electronic forces is recognized to contribute significantly to the double helical stability, being the major force holding the duplex together (Yakovchuk et al. 2006). Adjacent nucleotides in a nucleic acid strand have parallel aromatic rings with partly overlapping π bonds. The total bond energy for a given base pair is thus dependent also on its neighbours in the strand through π bond sharing interactions. The hydrogen bonds and weak interactions can be reversed by energy like heat or a chemical agent. Adenine-thymine pairs are more prone to denaturation than guanine-cytosine pairs, which is reflected in genomic regions that need to be frequently accessible like for instance prokaryotic promoters (Pribnow 1975). Higher salt concentration and longer sequence length increase affinity. Thus, the total affinity in models of nucleic acid interaction is determined by sequence content, length of the oligonucleotide and salt concentration.

Melting temperature (T_m) describes the temperature where hybridization reaction is at dynamical equilibrium, i.e., half of strands are dissociated (in the "random coil" state) at a given time as a result of breaking of bonds between the bases. Since T_m is only defined at equilibrium it is independent of time, but dependent on binding strength. Therefore it can be used as a measure of affinity. There are several ways to estimate the T_m of a nucleic acid sequence even though none of them are accurate with long sequences, most likely because of supercoiled states in the helix. The nearest neighbour model is considered to be the most accurate for short oligonucleotides as it allows the incorporation of sequence-dependent thermodynamic properties taking into account not just base composition but the actual sequence (SantaLucia 1998). The binding affinities of oligonucleotide probes can be estimated with nearest neighbour calculations, either computing T_ms or free energy changes (ΔG).

1.1.2.2 Steric and kinetic effects

The effect of steric interactions, i.e. the spatial arrangement of hybridizing DNA strands, to specificity and sensitivity of microarray probes is still unclear. However, the density of immobilized probes on the microarray surface, as well as spacing from the surface in part determine kinetic and steric accessibility for probe-target hybridizations (Peterson et al. 2001). Typically one end of the probe is attached onto the surface while the other end protrudes to the solution, causing different reaction conditions for the two ends (Hagan & Chakraborty 2004). The surface end of the probe is also less accessible because of neighbouring probes at close proximity on high-density arrays. Availability to hybridization may also be affected by the fact that solid microarray surface imposes sterical restrictions to the duplex formation (Southern et al. 1999). These factors, through efficiency of capturing targets from solution affect the sensitivity of the microarray as has been shown experimentally for example in a 3D matrix microarray platform (Dorris et al. 2003).

1.1.3 MICROARRAY DATA PROCESSING

The goal of microarray data normalization (processing prior to analysis) is to produce uniformity within and between arrays by removing variation caused by technical artifacts while retaining biologically relevant variation. The technical variation in microarray data typically has a systematic and a random component. Systematic variation recurs from experiment to experiment in a predictable way and can be modeled, whereas random error becomes apparent as unpredictable differences between replicate measurements. The biases in data occur both within and between microarrays and can result from differences in sample nucleic acid quality and quantity, labeling efficiency, probe concentrations, fluorescent properties of the dyes, hybridization conditions, array surface effects, data acquisition and other technical sources. In order to make meaningful comparisons between experimental treatments on a given array or between different arrays, signals need to be standardized for all of these parameters. Same kind of techniques can be used for within-array and between-array normalizations.

A typical systematic intra-array source of variation is dye balance as a function of signal intensity or position on the array. For two-colour gene expression microarrays, this is revealed by diagnostic graphs where the difference of log expression values are plotted as a function of average log expression values (M vs. A plot). Methods based on locally weighted quadratic least squares (LOESS) regression curve (Yang et al. 2002) or other non-linear functions like cubic splines (Workman et al. 2002) are commonly used to correct these kinds of artifacts. A LOESS model is fitted into the data showing intensity dependent trend, followed by subtraction of loess function values from data so that the trend is neutralized along the M-axis. The

LOESS method relies on the assumption that the vast majority of genes do not show differential expression or, alternatively, that the expression is symmetrical in terms of up and down regulation. If this does not hold, the method can introduce additional noise into the data. The microarray can be designed to contain control spots with varying intensities through which the LOESS curve is fitted. Ideally, the control spots are mixtures representing the totality of transcripts so that differential expression of any gene can not affect the results. This kind of control method tends to be more reliable at higher signal intensities whereas the LOESS model fitted through all data points is more reliable at lower intensities (Yang et al. 2002). The LOESS model includes a weight function which determines how the data points are weighted according to their distance from the point of estimation (smoothing). This makes the LOESS model non-parametric as it holds no assumptions about which kind of weight function is appropriate. The weight parameter value must therefore be chosen by the researcher which leaves a potentially biased subjective judgment to the model.

Normalization between microarrays is important especially for single-colour platforms like Affymetrix where different samples are quantitatively compared between different microarrays. A similar method, but typically applied for between-array normalizations, is to use an endogenous set of rank invariant genes within the data as a basis normalization (Li & Hung Wong 2001). Rank invariant genes show similar expression level in all samples. For example, Pelz and co-workers proposed using a global set of selected genes and fitting the normalization curve through them (Pelz et al. 2008). Quantile normalization is another powerful method for making data comparable between arrays. It sets intensity distribution similar between all arrays in the dataset by ranking the genes according to signal magnitude and assigning each rank the same value throughout the set of microarrays (Bolstad et al. 2003).

Stochastic error is the sum of many random measurement errors and therefore follows Gaussian distribution. This kind of random noise manifests in microarray measurements as an increasing spread of replicate spot signals as a function of mean intensity, i.e. the higher the signal the higher the variance. A simple log transformation stabilises variation in high signal intensities but fails to work well for low intensity values (Rocke & Durbin 2001). Rocke and coworkers have proposed a model for error estimation where a quadratic function describes the variance vs. mean dependence. The model incorporates a multiplicative error term of the expression value and an additive error term of the signal background. In low signal levels, the multiplicative error of signal is low and the additive term dominates, and vice versa. The values of these two components of error are estimated from data. This method allows data transformation to stabilize the variation making it independent of intensity (Rocke & Durbin 2001, Huber et al. 2002). The prerequisite for using this kind of model is that the model assumptions hold,

in which cases it should perform better than corresponding models with user-defined parameter values.

1.1.4 MICROBIAL PROFILING WITH DNA MICROARRAYS

Over the last decade, advances in DNA sequencing technology have enabled comprehensive metagenomic studies revealing the substantial taxonomic diversity and abundance of microbes in practically all natural and human environments (Torsvik et al. 1990, Lopez-Garcia & Moreira 2008, Singh et al. 2009). Measuring the diversity, changes in composition and activity of microbial communities is needed in various applications ranging from clinical diagnostics to environmental biotechnology and research, thus requiring efficient measuring tools. Even though high-throughput sequencing is necessary to obtain de novo information, it is still not economically feasible for rapid routine monitoring. The more traditional culturing, restriction fragment analysis, antibody labeling and PCR based approaches are generally not capable of detecting a large number of rare microbial types. In this respect, the advantage of DNA microarray technology is the capability of detection and identification and potentially quantification of thousands of distinct DNA molecules in a single experiment relatively rapidly and cost-effectively. Typically microbial microarrays rely on labeled PCR products. However, the sample preparation involves extraction, amplification and labeling steps of the source DNA or RNA that can introduce biases (von Wintzingerode et al. 1997), necessitating careful design and validation of the microarray test. In addition to detection and diagnostics, by interrogation of gene expression and genomic content, DNA microarrays allow comprehensive characterisation of genetic elements attributing to virunlence of pathogens (Aittamaa et al. 2008, Grundmeier et al. 2010, Aguado-Urda et al. 2010).

Microarrays designed for interrogation of microbial communites can be divided broadly into two categories: phylogenetic and functional microarrays. The phylogenetic oligonucleotide microarrays ("phylochips") target the small ribosomal subunit gene regions using hierarchically selective probes for detecting multiple taxonimical levels simultaneously. The rRNA gene has been established as the primary phylogenetic marker since the invention of culture independent sequencing of bacterial genes (Lane et al. 1985), because the gene is present in all organisms and contains both conserved and and taxonomically characteristic variable domains (intergenic transcribed spacer, ITS) making it suitable for phylogenetic classification. Furthermore, currently it is the most characterised marker gene with the highest number of available sequences in public databases, allowing in silico probe design more reliably than other genes. The phylochip rRNA probe hierarchy typically covers groups from species to phylum level, and is therefore potentially capable of exposing the whole taxonomic structure. This also alleviates the inherent inability of microarrays to find novel types;

higher-level probes are likely to capture unknown targets as well, even though they may not indicate their exact taxonomic designation. Phylochips harbouring a large number and a broad taxonomic range of probes are also flexible with regard to usability as the same microarray design can be used to profile microbial populations in many different types of environments. These kind of phylochips contain tens of thousands to hundreds of thousands of probes and have been built on high-density microarray platforms like Affymetrix (Brodie et al. 2006, DeSantis et al. 2007) or Agilent (Palmer et al. 2006, Palmer et al. 2007). High-density phylochips are ideal for analysing complex populations as was demonstrated for example by Palmer and coworkers by studying human colon microflora (Palmer et al. 2006, Palmer et al. 2007). In contrast, smaller scale in-house produced phylochips have probes in the order of hundreds and are primarily designed for a certain type of environment and a restricted selection of taxa (Loy et al. 2002, Rastogi et al. 2010).

In addition to community structure analysis, phylochips have also been used for analysis of community function by using radioisotope labeling as a marker of microbial activity. Radioactive carbon isotope is incorporated into the genomes of replicating microbes so that metabolically active types can be recognized by hybridizing extracted total DNA on phylochip and monitoring signals from the decay of radionuclei in the hybridised DNA (Adamczyk et al. 2003). Thus, functional activity in the studied environment can be inferred from the taxononomic activity profile. A more common approach to characterising community function is the functional gene microarray which targets selected genes or gene families encoding key enzymes of metabolic pathways of interest. For instance, He and coworkers (He et al. 2007) monitored uranium degradation using spotted microarray platform with over 24000 50-mer probes targeting 150 different functional bacterial genomic sequences for biogeochemical cycling, metal resistance and organic contaminant degradation. Relevant genes and microbial groups were recognized by testing correlation in changes in abundance of detected genes with changes in uranium concentration. Besides genomic DNA, environmental RNA (or cDNA) can be targeted for evidence of metabolic gene expression levels (Dennis et al. 2003, Bodrossy et al. 2006). Even though in this approach the abundances of metabolizing groups and mRNA levels are intermixed in the results, in theory it should be possible to study gene expression changes resulting from subtle changes in environmental conditions by monitoring community changes simultaneously using a phylochip.

Pathogen detection by microarrays combines aspects from phylogenetic and functional approaches. Nested probe hierarchy can be used in pathogen detection in a similar way as in phylochips but the target gene does not necessarily have to be a typical phylogenetic marker. Since high qualitative accuracy is desired because of lethality of many common types, rRNA genes may not be ideal for species level identification owing to high conservation of

the sequence. More specific functional genes are commonly used, for instance virulence factors or resistance determinants (Saunders et al. 2004, Sergeev et al. 2004, Cleven et al. 2006), although good discrimating power of infectious bacteria by 23S (Anthony et al. 2000) and 16S (Wang et al. 2002) genes have also been reported. Secondly, as the presence of a pathogen might be relevant only if it exceeds a certain limit of magnitude expected to cause disease, quantitative detection is needed in many cases as well.

1.1.5 PROBE DESIGN

The most important criteria for designing microarray probes include minimal crosshybridization with non-targets, ability to bind to low-abundance targets and similarity of duplex formation properties between probes in the probe set. In practice, all of these criteria are not possible to fulfill simultaneously especially when the number of probes is high. Tools utilizing sequence databases for in silico probe design and validation are available, although most programs are for gene expression applications instead of diagnostics. For diagnostics, it would be essential to compare variants of a gene region within a population instead of genes within a single genome.

Generally, conserved genes like rRNA gene and internal transcribed spacer (ITS) regions are targeted for microbial diagnostics in order to cover the majority of related microorganisms in one PCR amplification reaction using consensus primers. The rRNA gene occurs in high copy numbers (Condon et al. 1992), allowing sensitive detection. Large numbers sequences are publicly available, facilitating the validation of the reliability of the assay (Maidak et al. 2000). Further, rRNA gene possesses enough polymorphic information to contain sufficient discriminatory potential to differentiate and characterize even closely related microorganisms. Degenerate primer sets can be designed to increase the coverage of relatively variable common region. However, ribosomal gene is problematic with regard to quantitative detection because the rDNA fragments form easily self-structures causing variation in results. Palmer and co-workers (Palmer et al. 2006) studied human colon microflora with a high-density phylochip and reported that taking into account the predicted hairpin structures increased quantification accuracy. rRNA gene copy numbers vary between taxa and DNA extraction and PCR bias can also introduce errors (Fogel et al. 1999).

Another problem for phylotype-level microarray detection is posed by the complexity of microbial communities. A large number of closely related types can be present in any environment requiring highly specific probes for all detectable types, which seldom is possible. In reality, cross hybridization among probes can not be completely avoided resulting in ever more complex patterns in analyzing highly complex samples. Since the number of probes on any microarray is limited by technological and economic constraints, the pattern to identify a target type must be interpreted. The established concept

in microarray gene expression data analysis is that a given probe represents only one target type. In analyzing environmental data it is more realistic to assume that there may be multiple targets for any given probe. To this end, Urisman and coworkers developed the E-Predict algorithm for making sense of pathogen microarray hybridization patterns (Urisman et al. 2005). The algorithm compares an observed pattern to theoretical patterns predicted by probe binding energies for classification into categories representing targets. A category can contain probes from other genomes as well, so that any given probe could be part of multiple categories. Moreover, Wong and others investigated PCR amplification bias and the ability of microarray probes to capture their intended targets on a custom Nimblegen high-density tiling microarray platform representing 35 pathogen genomes (Wong et al. 2007). They conceived an algorithm for predicting the identity of the pathogen from a complex microarray probe signal pattern. The probe binding to targets was estimated by similarity score and PCR amplification efficiency.

1.2 PRINCIPLE OF LIGATION DETECTION REACTION

Detection of specific DNA molecules by enzymatic ligation was developed by Landegren and coworkers in 1988 to overcome the limitations of oligomeric probes in distinguishing single base mutations associated with genetic diseases (Landegren et al. 1988). Even though oligomeric ssDNA probes typically are quite specific at lengths of 20-25 nucleotides, their sensitivity is compromised due to their relatively low dissociation temperature (T_d). Increasing the length improves sensitivity but specificity then tends to drop considerably (Wetmur 1991, Kane et al. 2000, Tiquia et al. 2004). Ligation based detection takes advantage of both probe hybridization and catalytical selectivity of DNA ligase to improve specificity and sensitivity of detection. Since ligases favour perfect complementarity of a double stranded DNA structure to successfully catalyze the sealing of a nick in the phosphodiester backbone, they can be used to link two adjacently hybridising probes covalently together in the presence of a correct target molecule (Figure 1). The probes constitute a target-specific probe pair which becomes detectable only if the probes are ligated. The so called "discriminating" probe is designed such that the 3'-end matches the target at a unique position which contains a nucleotide that distinguishes the target from other DNA strands. The "common" probe is designed to hybridise next to the discriminating probe. The first implementation of this technique employed the T4 ligase (Landegren et al. 1988), but further improvements have utilized thermostabile ligases enabling thermally cycled reactions with higher ligation product yield and better sensitivity (Barany 1991a).

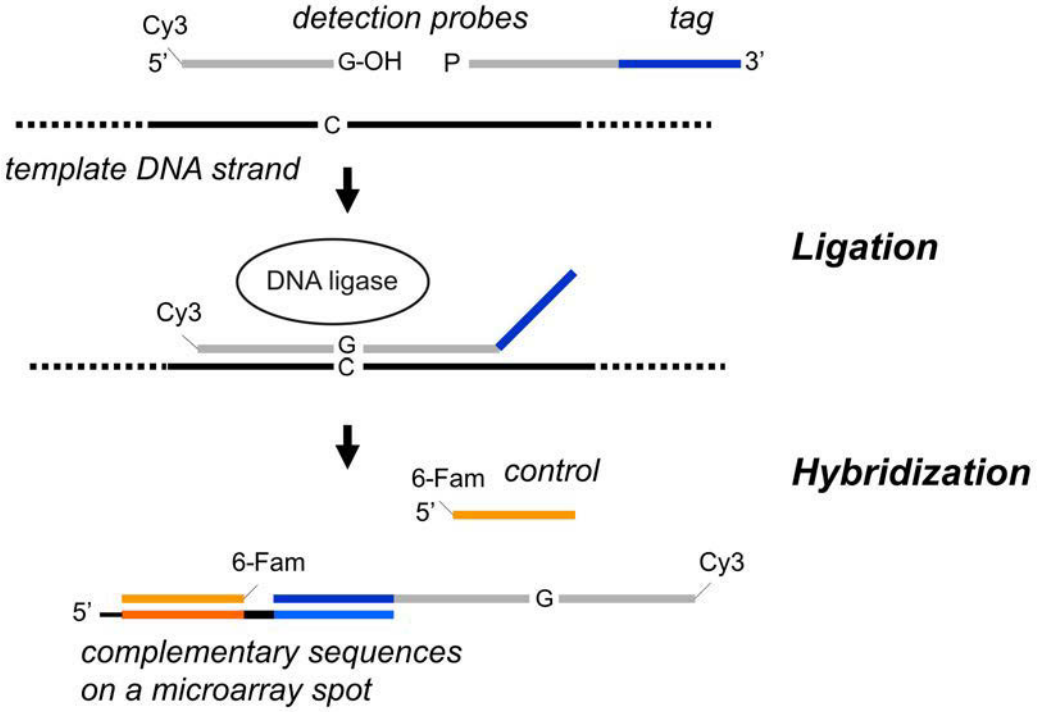

Figure 1. A schematic of ligation detection reaction (LDR). Two probes, one labeled with 5'-fluorescent label (the discriminating probe) and the other harbouring 5-phosphate and a 3'-tag sequence (the common probe), are annealed on target DNA molecule. If there is sequence complementarity at the junction site of the two probes (indicated by G-C pair), they are covalently joined by ligase. The ligated probe construct is then detected by excitation of the 5'-label molecule after hybridization on microarray at locations determined by complementary tag sequence (blue colour). An internal hybridization control probe labeled with 6-Fam (orange colour) is used to measure the quality of each microarray spot.

1.2.1 LIGATION CHEMISTRY AND CATALYSIS

Polynucleotide ligases are a group of enzymes catalyzing the formation of a covalent phosphodiester bond at the site of a single-strand break in DNA or RNA molecule. As sugar-phosphate backbone breaks are a result of both stress and normal manipulation of nucleic acids in replication and recombination, ligases are required ubiquitously and are a part of the basic molecular machinery in all cells. There are multiple genes and proteins of DNA ligases with hypothetical specific cellular functions in mammals and in lower eukaryotic and prokaryotic organisms (Tomkinson et al. 1991, Wilson et al. 1997). In addition, RNA acts a substrate for a specific class of ligases. All the ligases together with the RNA capping enzymes constitute the nucleotidyl transferase superfamily sharing the catalytic NTase domain which interacts with a nucleotide cofactor to form a covalent enzyme-nucleoside monophosphate (Shuman & Schwer 1995). Bacterial ligases require nicotinamide dinucleotide (NAD^+) as a cofactor whereas ligases of viruses, phages and eukaryotes require ATP.

1.2.2 CATALYTIC MECHANISM

The catalytic mechanism of ligation is similar in both eukaryotic and prokaryotic ligases (Shuman, Schwer 1995), reflecting a common evolutionary origin. In the fist reaction step, the nucleotide cofactor is hydrolysed leading to adenylation of the ligase: the α-amino group of lysine residue at the active site of the ligase reacts with the 5'-hydroxyl group of the AMP group of the nucleotide cofactor leading to formation of ligase-AMP complex (Figure 2). In the second step, the monophosphate of the formed complex reacts with the free 5'-phosphate in a nucleotide in the broken strand thus transferring the AMP group and acitvating the phosphate. This intermediate structure leads to the third step whereby the 3'-hydroxyl group performs a nucleophilic attack to the activated 5'-phosphate and upon creating a new phosphodiester bond, AMP and the ligase are released (Lehman 1974).

Protein crystallographic structures have shown that the ligase protein stabilizes the structure necessary for the proceeding of the reaction with hydrogen bond formation between several conserved residues in the NTase domain and the substrate nucleic acid (Shuman & Lima 2004). A diverse collection of N- and C-terminal accessory domains extends from the NTase domain to promote the ligation reaction or to provide special substrate binding properties. For instance, ATP dependent DNA ligases use C-terminal OB domain but NAD^+ dependent ligases use an N-terminal Ia domain to contribute residues that stabilise the structure between active site lysine and the cofactor to facilitate the formation of the ligase-AMP intermediate (Pascal et al. 2006, Gajiwala & Pinko 2004, Mackey et al. 1999). RNA ligases in contrast do not use any additional domain other than the core NTase to form this intermediate (Ho et al. 2004). Mutational studies in Thermos Thermophilus (Tth) ligase suggest that the active site for the phosphodiester bond formation (the third reaction step) is separate from the site of adenynalation and deadenylation as mutations in conserved residues G339 and C433 render only the third reaction step dysfunctional (Luo & Barany 1996).

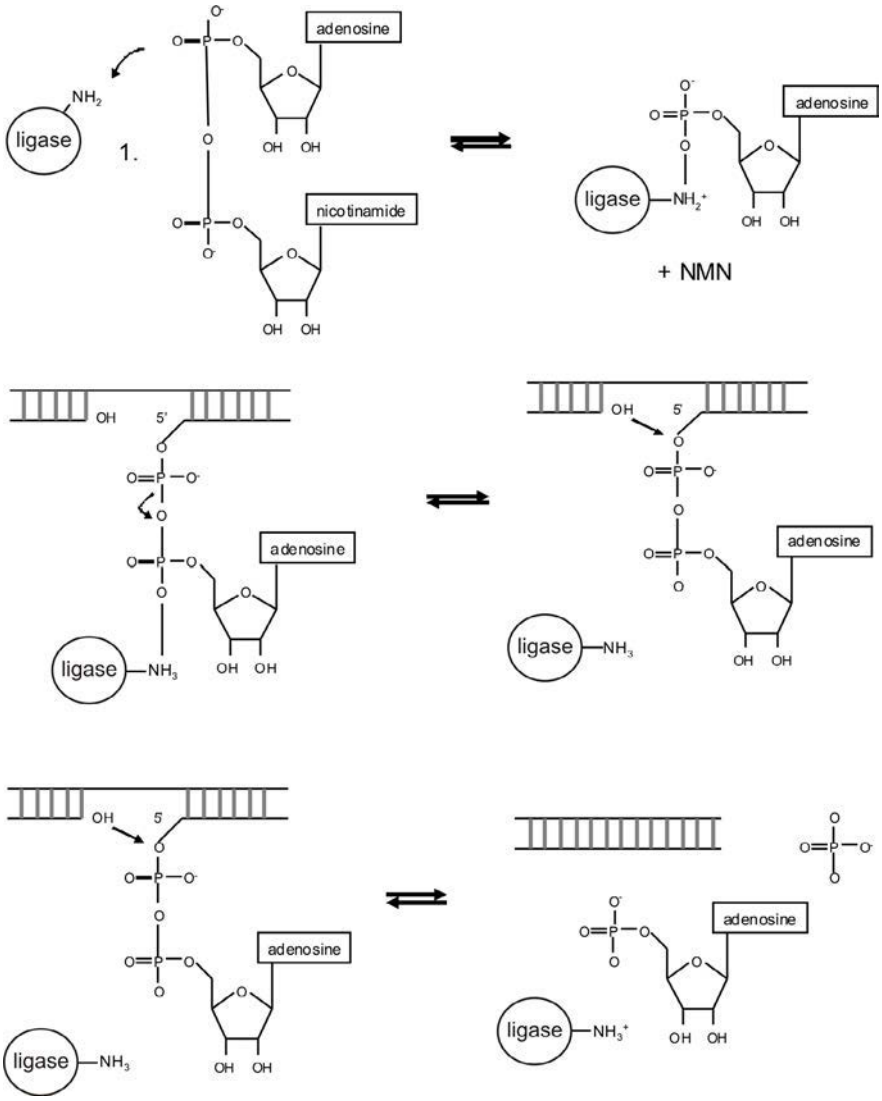

Figure 2. Schematic presentaion of ligation reaction steps. 1: Nicotinamide adenine dinucleotide (or ATP) acts as a cofactor and is hydrolysed by a NH3 group of ligase. Ligase then forms a complex with AMP and a nicotinamide mononucleotide (NMN) or a pyrophosphate is released. 2: The complex reacts with 5'-phosphate of DNA end forming a pyrophosphate linkage. 3: Nucleophilic attack of 3'-OH group to 5'-PO4 displaces AMP. 4: The sealed polynucleotide strand, ligase, AMP and inorganic phosphate are released when the phosphodiester bond is formed. Adapted from (Lehman 1974).

1.2.3 SUBSTRATE SELECTIVITY

NTase domain N- and C-terminal appendages provide the selective RNA/DNA binding properties for RNA and DNA specific ligases (Shuman & Lima 2004), although some DNA ligases can accept RNA as a substrate to some extent which is probably due to the fact that DNA ligases first modify the substrate DNA into a RNA-like helical conformation to carry out the catalysis (Shuman & Lima 2004). Another important aspect of ligation fidelity is the strong discrimination against mismatched base pairs at the 3'-OH end of the nick and in upstream positions. As a general rule, mismatched base pairs are poorly tolerated upstream of the 3'-OH DNA terminus and

better tolerated downstream of the 5'-PO_4 end of a nicked DNA substrate (Sriskanda & Shuman 1998, Luo et al. 1996). Crystallographic studies on DNA ligase structures have shown that the 5'-phosphorylated nick terminus is positioned through multiple contacts with the attached AMP group (Riballo et al. 1999, Pascal et al. 2004) and is therefore less likely to be affected by mispaired bases. In contrast, there are relatively fewer contacts with the 3'-OH end of the nick. Therefore the correct positioning of 3'-OH is more likely to depend on base pairing interactions with the intact template strand to correctly position the 3'-OH for end joining. Mutational analysis studies on Tth ligase have shown that the fidelity is lower against thymine-guanine mismatches at the 3'-end of template (Luo et al. 1996). Both bacterial and mammalian DNA ligases have a very high fidelity against all purine–purine mismatches (Luo et al. 1996, Husain et al. 1995), suggesting that base pair stability is not the predominant factor influencing ligase fidelity because then other mismatches should have a similar effect as well. Rather, the fidelity of DNA ligases is likely to be influenced by both the base stacking on nucleic acid helix and the structural properties of the enzyme protein. Also, the work by Ginya and others suggests that the fidelity of ligation is dependent on sequence at the 3' terminus (Ginya et al. 2010).

1.2.4 APPLICATIONS OF LIGATION TECHNIQUES

LDR in its various forms has been usually coupled to microarray detection. In the "universal microarray" approach, the common probe has a 3'-tag sequence which directs it to a microarray spot harbouring a complementary tag sequence, while the discriminating probe is fluorescently labelled (Gerry et al. 1999). The advantages of the universal array lie in the uniform conditions of all tag sequences, and in flexibility as the same array platform can be used with multiple ligation probe sets. The potential for relatively high discriminatory accuracy, sensitivity and throughput has made ligation based microarray detection techniques a candidate tool for mutation screening (Gerry et al. 1999, Hogervorst et al. 2003b, Hardenbol et al. 2005) and for characterising complex and extremely diverse microbial populations (Busti et al. 2002, Rantala et al. 2008, Szemes et al. 2005, Candela et al. 2010). For example, Candela and coworkers developed a LDR-based phylogenetic microarray consisting of 30 probes for interrogation of human intestinal microbiota (Candela et al. 2010), a similar approach that had been previously taken by Palmer and others using a high density phylochip (Palmer et al. 2007). The LDR microarray identified the same taxa as the phylochip with comparable sensitivity of 0,02 %.

A variety of DNA amplification strategies can be used to enhance sensitivity by amplifying the ligated probes. The majority of approaches use either PCR or isothermal DNA polymerase based method in some form as a separate step from ligation. Multiple ligation-dependent probe amplification is a PCR based approach to amplifying ligation products. In this method, the

ligation probes are flanked by PCR primer binding sequences enabling selective amplification of ligated probes (Schouten et al. 2002). The method is capable of relative quantification of targets and it has been widely used in research in medical diagnostics e.g. cancer mutations (Hogervorst et al. 2003a) and copy number variations (Armour et al. 2007) and recently also in detecting the pathogenic bacterium M. tuberculosis (Bergval et al. 2008). PCR amplification produces linear double stranded DNA fragments that can be detected using rudimentary equipment and methods like electrophoresis, making it practical to be used in most laboratories. However, for microarray hybridizations, the double stranded PCR product should be made single stranded by endonuclease which requires extra work steps. In order to create hybridizable products directly from PCR, linear-after-the-exponential PCR (Sanchez et al. 2004) has been applied to ligation probe amplification (Szemes et al. 2005). Linear-after-the-exponential PCR is based on differing concentrations of PCR primers so that after the low copy primer is consumed, the amplification continues linearly with the other primer producing fragments from only one of the template strands. Because of the concentration difference, the lengths of the primers must be adjusted in order to achieve similar theoretical Tm for both primers. Ligation chain reaction is an equally simple way with exponential kinetics (Barany 1991b). It involves using an additional probe pair per each actual probe pair in ligation. The extra pair recognises the same sequence as the actual pair but instead in the complementary target strand. As a result, ligation takes place on both the probe pairs creating an additional ligation template in each cycle. One of the probes can have a tag sequence making the ligation chain reaction products ready for hybridisation on microarray without further processing.

A somewhat more elaborate approach involves using padlock probes, that is, approximately 100-mer ssDNA oligos with the 3' and 5' termini representing the discriminating and common probes, respectively (Nilsson et al. 1994). The probe is circularised upon ligation. The advantage of padlock probes is that as the two recognition sequences are in the same molecule, the reaction kinetics is much faster when the correct partners of a probe pair are constantly in close proximity and independent of random diffusion. This advantage becomes more distinct when multiplexing ligation probes because of the lack of combinatorial explosion of possible ligationable probe pairs. Another advantage is the circular form which allows using exonucleases to remove all linear DNA from mixture to increase detection accuracy. In addition, amplification using either PCR as in multiple ligation-dependent probe amplification or isothermal rolling circle amplification (RCA) is possible. In the rolling circle method, tandem copies of the circular ssDNA template are produced isothermally by a strand displacing DNA polymerase, for instance ϕ29 (Fire & Xu 1995). RCA can be performed with linear or exponentially branching kinetics, making it suitable for quantitatively accurate (Antson et al. 2003) and highly sensitive (Lizardi et al. 1998) detection. In the basic single primed configuration, the polymerase produces

complementary repeats of the original molecule with linear kinetics. The double primed version takes advantage of additional primer which is targeted to the primary product leading to multi branched double stranded structure. The benefit of RCA over conventional PCR is its lower error rate and much higher product yield which enables detecting a very small number of molecules. RCA coupled with padlock probes has been mainly applied to medical diagnostics in detecting SNPs (Faruqi et al. 2001) while the PCR or non amplified versions have been applied in environmental research (Szemes et al. 2005, van Doorn et al. 2009).

2 AIMS OF THE STUDY

To study methods based on ligation reaction of ssDNA probes in combination with microarray platform for detection and identification of microbial target types from environmental and clinical samples. To further develop existing methodology.

3 MATERIALS AND METHODS

The key methods used in this study are summarized in Table 1 and described in detail in the indicated articles.

Method	**Described in article**	**Used in article**
DNA extractions:		
MasterPure Yeast kit		
Extraction genomic DNA from yeast pure cultures.	I	I
Fast prep soil		
Extraction of genomic DNA from compost samples.	I	I
Multiscreen PCR kit		
Purification of fungal 18S PCR products.	I	I
MagNAPure LC instrument		
DNA extraction from human patient samples.	III	III
FastDNA Spin for Soil kit		
DNA extraction from methane tank samples.	IV	IV
Ligation reaction for padlock probes		
Protocol for circular padlock probe ligation reaction.	IV	IV
Ligation reaction for LDR		
Protocol for ligation reaction of linear probes.	I, III	I, III
Padlock probe design	IV	IV
Description of synthesis by Agilent		
LDR probe synthesis and design	I, III	I, III
Microarray fabrication and testing		
Print solution comparisons and description of printing for contact printed arrays	I, III, IV	I, III, IV
Microarray hybridization		
Hybridization protocols for contact printed and Agilent microarrays.	I,III,IV	I, III,IV
Clone library sequencing		
Preparation of clone library and sequencing of clones of compost sample DNA.	I	I
454 pyrosequencing		
Description of primers and protocols of methane tank sample deep sequencing.	IV	IV
Microarray normalization		
Computational bias correction using hybridization control signals.	II	III, IV
18S universal PCR		
PCR for fungal ITS1 and ITS2 regions. Amplicons were used as LDR templates.	I	I

HPV multiplex PCR *Primers and protocol for multiplex HPV PCR.*	III	III
Padlock probe PCR *Primers and protocol for amplification of padlock ligation reactions.*	IV	IV

Table 1. *Key methods used in the study.*

4 RESULTS AND DISCUSSION

The objective in study **I** was to investigate the composition of fungal community of composting process and correlate the taxonomic profile to composting status. To this end, different kinds of composts were characterized by clone sequencing; four research composts with controlled temperature and oxygen levels to minimize acidic phase to increase efficiency of decomposition and four full-scale composts with different temperatures and locations. The starting point for study **IV** was very similar, employing 454 pyrosequencing to characterize the microbiota of different stages of anaerobic decomposition in methane tanks. Both studies aimed at developing a microarray test for monitoring the status of these processes by identifying fungal or other microbial taxa present in the system on basis of recovered sequence information. In study **III**, in contrast, the microarray was designed to identify and genotype infections of clinically characterized, previously known HPVs to monitor cervical cancer risk status of patients.

4.1 SPECIFICITY TESTS OF LIGATION PROBE POOLS

Functionality of the ligation probe pool in studies **I** and **III** was assessed in a series of microarray hybridization experiments where the entire probe pool was tested against probe-specific templates, one template type at a time for each probe. This experimental setup allowed identifying false positive signals individually for all probes as well as identifying false negatives on corresponding specific targets for each probe. However, the experiments did not include probe signal responses to different concentrations of templates which would ideally be needed to determine sensitivity and specificity of each probe; the lower the template concentration that can be detected at some predetermined level of accuracy, the better the sensitivity, and, the higher the non-specific template concentration that does not produce false positives, the better the specificity. Despite this limitation, the relatively high constant template concentrations of 0,5 fmol/µl and 2 fmol/µl in studies **I** and **III**, respectively, allowed a reasonable estimate for specificity because these concentrations were high enough to produce strong signals for true positive probes (**I**:Figure 4, **III**:Figure 2) and consequently could be expected to yield detectable false positive signals as well should they exist in the probe set. The templates used in the specificity tests in **I** and **III** were PCR products generated from genomic DNA. PCR primers were the same as with real biological samples, but the following LDR test conditions differed from real samples in that instead of an unknown amount and composition of sample DNA as a non-specific background, the reaction mixture contained 100 ng of herring sperm DNA as background DNA. Herring DNA does not accurately

represent a real biological sample where competition for the probe with highly similar genomic sequences is likely to take place. On the other hand, this simplified test setting allowed determination of probeset functionality without confounding factors of biological samples.

In **I**, the probeset of 16 probes did not produce any false positives for any of the fungal phylotypes tested (**I**:Figure 4) and in **III**, the probeset of 38 probes produced one clear and six ambiguous false positives (**III**:Figure 2). However, in **III** a second probe targeting the same genotype was in all cases accurate meaning that an alternative specific probe for each of the seven HPV genotypes can be selected. The data provides proof that ligation probes are accurate in recognizing genotypes of closely related targets. The data generated by the specificity experiments, taken together, is of importance in estimating the limit between true positives and true negatives where the experimental values constitute the background signal distribution representing non-specific signals.

In study **IV**, the specificity was estimated by dividing the templates into four different sub pools and testing each sub pool separately with the entire probe pool. Compared to **I** and **III**, the experimental setup in **IV** was limited in that true and false negatives and positives could not be identified individually for each probe, but only in relation to the sub pool of templates each probe belonged to, leaving a degree of uncertainty to the functionality assessment of individual probes in the pool. For example, within each template sub pool, a given expected positive probe could be a false positive on one or more templates present because different templates in the corresponding sub pool were not tested separately. Thus, each probe could be ascertained reliably for non-specific binding on 75% of the targets and for specific binding as part of a group constituting 25% of targets. Despite this limitation, the probe pool response to template sub pools was highly predictable in terms of signal presence or absence, with only a few non-functional probes (**IV**:Supplemental File 7), suggesting good accuracy overall. However, the signal levels of different probes at the same template concentration varied substantially which can be due to combination of differing ligation and PCR efficiencies. Varying signal levels can be also seen in studies **I** and **III** using ordinary ligation probes. Further, the templates were synthetic 80-mer oligos not representing the total sequence content in an actual biological sample. Mispriming to possible genomic sequences present in the biological sample DNA material can not be tested with synthetic templates, even though the most probable mispriming sites are in the ribosomal target region shared by many orthologous genes in the sample where priming accuracy can be most effectively tested in silico. Another limiting factor in **IV** is that the 80-mer oligos are not a good model for possible secondary structures (Dixon & Hillis 1993) that might occur in actual, longer genomic ribosomal target sequences.

4.2 SENSITIVITY TESTS OF LIGATION PROBE POOLS

In study **I**, sensitivity was assessed for the entire ligation probe pool instead of determining it at probe level individually. Approximately 5 amol/µl was determined to be the absolute limit and 0,04% from total DNA content to be the relative limit of detection (**I**:Figure 2). This was better than in previous reports for LDR (Busti et al. 2002, Rantala et al. 2008, Castiglioni et al. 2004), the reason probably being in optimized ligation reaction and microarray hybridization conditions. Different microarray slide coatings and printing buffers were evaluated (data not shown) and tetramethylammonium chloride (TMAC) was used as an additive in ligation reactions. TMA derivatives have been shown to enhance the formation of specific products in PCR (Chevet et al. 1995, Kovarova & Draber 2000) and this effect is likely to take place in ligation as well. In study **III**, sensitivity was estimated through PGMY-t HPV multiplex PCR amplification of 0,04 pg/µl and 0,04 ng/µl of template plasmid pools, and then using a maximal amount of resultant PCR product as a template for LDR. All correct templates were detected by LDR from both 0,04 pg/µl and 0,04 ng/µl template samples after 40 cycles of PCR (**III**:Figure 3) suggesting that multiplex PGMY-t PCR is robust for template amount and thus hardly presents a limiting factor for the sensitivity of the method. Further, sensitivity of LDR in **III** can be assumed to be in similar range as in I because the LDR methodology is effectively identical in both studies even though sequences differ. However, as amplification and detection efficiency of varying amounts of different target templates present in the same PCR reaction were not tested in these studies, the effect of possible PCR bias to the final detection sensitivity could not be accurately estimated. PCR bias occurs when targets have different efficiencies in primer binding, polymerization initiation and elongation (Suzuki & Giovannoni 1996, Polz & Cavanaugh 1998), and in addition, random effects in initial conditions that cause variation in end-point values. This concerns especially settings where varying amounts of different genotypes present in the sample should be detected, but low-copy genotypes may not be amplified by a specific or common primer. The final sensitivity is naturally the outcome of sample handling, PCR and ligation, and failure in any of these would result in non-functional test.

Sensitivity in **IV** was tested with a dilution series of templates (**IV**:Figure 3) in a similar manner as in **I**. Template concentration 0,1 fmol/µl was still detectable for the majority of probes, being higher than in **I**. This result was unexpected because the padlock probes are amplified by PCR and were therefore presupposed to give stronger signals on low template concentrations than ordinary LDR probes which are amplified with linear kinetics. It is not clear why sensitivity was not improved. One reason could be the low concentration of probes. The amount of LDR padlock probes was 200 amol/probe (c.a. 10^8 molecules) per reaction, of which an unknown proportion is eventually ligated into amplifiable constructs. Despite that 10^8

probe molecules is lower than the amount of templates ($6x10^8$ - $6x10^{10}$) in all dilutions which gave a positive signal, the template amount in these experiments correlated positively with signal strength (**IV**:Supplemental Figure 8). The prerequisite for this kind of concentration dependent response is that the signal from the probes does not saturate, i.e. the probes are not the limiting factor, but instead the template amount limits the signal. If target recognition of padlock probes was maximally efficient, or close to it, saturated signal would be expected in all aforementioned dilutions because template molecule concentration exceeds probe concentration. As saturation was not observed, ligation and other factors determining probe target recognition efficiency must be relatively low. Because about $6x10^8$ copies of template was the lowest detectable concentration, and optimized PCR should be able to amplify from 1 - 10^2 copies of target in 30-40 cycles (Li et al. 1988, Skakni et al. 1992, Palmer et al. 2003), target recognition efficiency is likely to be below 10^{-6}.

4.3 PCR AMPLIFICATION OF PADLOCK PROBES

Since ultimately sensitivity is dependent on both PCR and ligation, a major contributing factor can be that PCR from a padlock template is not efficient enough to produce amplicons from few ligated probes. The circular shape and short length of the ligated padlock might impede amplification, even though Taq polymerase has been shown to be able to generate product from short circular molecule (Liu et al. 1996), implying that template shape itself is probably not restricting. Results from PCR protocol optimization tests done for study IV suggest that DNA polymerases in general do not readily amplify a padlock template, since only one enzyme (Paq5000, Agilent Technologies) out of several tested, including Taq (Biotools), produced relatively clear and detectable agarose gel electrophoresis bands from 250 fmol of ligated probes (data not shown). Activity of 5'-3' exonuclease of the polymerase is probably crucial, since the synthesized template will be easily degraded rather than displaced with high exonuclease activity. Moreover, high helicase activity should favour amplification because ligated probe forms a double helical structure with template. Paq5000 probably fulfills these criteria better than other tested thermostabile polymerases, although explicit data on the properties of the enzymes is not available. The Vent exo- DNA polymerase lacking all 5'-3' exonuclease activity has been succesfully used for PCR amplification of padlock probes (Prins et al. 2008) although its advantage over Taq was not rigorously established. Sensitivity after 80 cycles of asymmetrical linear-after-the-exponential PCR (Sanchez et al. 2004) was found to be in the order of 0,1 pmol (Prins et al. 2008). On the other hand, van Doorn and co-workers detected just 1000 copies of target molecules by real-time quantitative PCR approach (van Doorn et al. 2007). The differences between these two similar approaches suggest that implementation of

detection is crucial for maximal sensitivity. In this regard, the sensitivity of 0,1 fmol/µl attained in study **IV** is relatively good but could likely be optimized further.

Contrary to PCR amplification, padlock probes amplified with isothermal strand displacement methods can reach much higher sensitivity, up to the level of a few copies of target molecule per reaction. For instance, Zhang and co-workers were able to detect viral transcripts from a dilution of less than 1 infected lysed cell per reaction (Zhang et al. 1998). The efficiency of their amplification, a combination of PCR and strand displacement using Taq polymerase and two primers (RAM; ramification-extension amplification method), is emphasized by the fact that ligation was done without thermal cycling. Similar approach has been used to detect 10 target molecules of C. trachomatis per reaction using Bst polymerase (Zhang et al. 2002). Further, single-primer rolling circle amplified (RCA) padlock probes have been employed in detection of individual transcripts from human cells in situ (Larsson et al. 2010). While both PCR and isothermal strand-displacing amplification methods can achieve 109 fold amplification from a few copies, RCA and RAM and their variants are more sensitive in practice probably because one priming event is enough to produce maximally amplified signal. With PCR, in contrast, priming multiple times over multiple subsequent rounds to achieve amplification can lead to amplification of background signal. This implies that for low-copy PCR templates the risk of amplifying background increases.

Other reports of padlock-based applications have employed exonucleases to degrade linear DNA prior to amplification (Zhang et al. 1998, Szemes et al. 2005, Akhras et al. 2007), but this procedure was left out in **IV** because while testing samples no difference was found between exonuclease treatment and non-treated control (data not shown). Because non-specific background amplification did not present a problem as illustrated by the specificity experiments, exonuclease was not applied. Increasing the concentration of probes and number of PCR cycles and similar further optimizations were not pursued in study **IV**. It is likely that the method could still benefit from optimization in that area in the future.

4.4 ANALYSIS OF BIOLOGICAL SAMPLES

Studies **I** and **III** employed PCR amplification of target gene region from sample DNA prior to LDR, whereas in **IV** the sample DNA served directly as a template for ligation of padlock probes which were subsequently PCR amplified. This difference in approaches to target detection is also reflected in results of analysis of biological samples: in **I** and **III**, the method was able to detect the majority of targets present in samples verified by sequencing. In **I**, LDR detected 21 out of 24 sequenced phylotypes (**I**:Table 3) and in **III**, LDR detected all HPV genotypes verified by sequencing in patient samples

(**III**:Supplemental file 3). These results are in agreement with other studies of applying LDR for microbial identification; for instance, the majority of target human intestinal microbial groups of validated LDR probes could be detected from biological samples (Candela et al. 2009) and in another study, all of the clone library sequencing verified cyanobacteria from environmental samples (Castiglioni et al. 2004). In contrast in **IV**, only a few probes that were shown to be functional with artificial templates could detect their target from sample DNA. The probes in **IV** were initially designed to match certain phylotypes or phylotype-level OTUs (97% sequence similarity), but as these can typically correspond to relatively few sequences in the sample material, the sequence abundances were likely to be below detection limit of the method. Consequently, the sensitivity of detection between the approaches represented by **I** and **III**, and on the other hand **IV**, differed markedly. In **I**, 1 ng/µl of sample DNA as a template for fungal ITS PCR provided more consistent detection compared to 0,2 ng/µl (**I**:Table 3). In **IV**, approximately 20 ng/µl of sample DNA was used per ligation reaction and detection sensitivity directly from sample DNA was approximately 1%, calculated as a proportion of total pyrosequencing reads corresponding to the probe target (**IV**:Figure 3) in the samples. Assuming that one read corresponds to one genome in the sample, the actual proportion of probe target site from total non-target sequence content in ligation is lower in **IV** than in **I** and **III** where the total sequence content is constituted by relatively short PCR amplicons instead of full genomes. Thus, the fact that in **IV** there is no reduction of complexity of the target pool (like with initial PCR there would) is likely to affect the sensitivity and efficiency of target recognition by ligation. The padlock probe approach has not been applied to complex environmental samples before, but studies in genotyping in a highly multiplexed setting have shown that ratio of target to non-target signals is relatively low, most likely due to competition (Hardenbol et al. 2005). Same kind of effect could take place in **IV**, since the high sequence background can cause unspecific competitive binding with probes.

Even though LDR sensitivity in I was better than the padlock approach in **IV** with regard to artificial templates, the padlock method seems to work to some extend with non-amplified environmental sample DNA. As an indirect comparison to **IV**, the LDR probes could not detect anything when applied directly to 0,5 ng/µl of non-amplified sample DNA (**I**:Table 3). The LDR probe pair has two separate probes that are required to anneal simultaneously in order to be ligated, whereas in the padlock approach each pair is in the same molecule, meaning that there is much less possible combinations of all pairs that occurs in LDR. This could in part explain the result that the padlock probe approach works better in an environment with a large amount of complex background DNA.

Another factor in **IV** likely to contribute to probe-target recognition is the sequence mismatch between probes and targets. Specific microarray probes typically could not be designed merely on the basis of trimmed 454 sequence

reads due to their limited length of about 150 bp, which necessitated to retrieve full-length rRNA genes matching to OTUs from the NCBI nucleotide database. The closest matching gene to an OTU was typically only 94% similar, leaving considerable uncertainty regarding the estimated target specificity of the probes in the context of the sample DNA. The probes were designed to variable sites in the gene which are also most likely to differ from known database sequences. As expected under the probe-target sequence mismatch hypothesis, probes that could be aligned with mismatches to the rRNA genes were less accurate quantitively (**IV**:Supplemental File 9) than 100% matching probes (**IV**:Figure 3). Since the probes in the initial specificity tests responded highly accurately to their cognate target oligo pools, it is reasonable to assume that at least some missing signals and those not correlating well with the sequence read numbers are explained by unknown sequence differences in the rRNA genes in the samples. This conclusion is supported by studies on padlock probes for pathogen detection, where polymorphisms in the template has been reported to cause attenuation of probe signals (Szemes et al. 2005).

4.5 NORMALIZATION OF MICROARRAY SIGNALS

In order to remove technical variance from microarray data, the source of variation should be identified and measured. Normalization methods for gene expression microarrays are typically based on correcting the non-linear relationship between channel ratios and intensity or spatial location of a large number of spot values (Yang et al. 2002). Same kind of non-linear smoothing is used for single-channel normalizations to correct spatial and between-array effects (Edwards 2003). These methods assume that only a small minority of genes are differentially expressed, but this condition does not generally hold for diagnostic microarray platforms where the number of spots may be much lower and the number of positive and negative probes cannot be known beforehand. Unlike gene expression microarrays, there are no established normalization methods for microbial detection or diagnostic microarray applications in general. For instance, phylochips used for microbial profiling on high-density Affymetrix platform utilize only the mismatch probes as an indicator of background hybridization signal threshold to identify positive probes (DeSantis et al. 2005). In another study using Agilent platform, a hybridization control in one channel was used to normalize the signals from probes in another channel (Palmer et al. 2007). Similar internal control has been proposed for lower density diagnostic microarrays as well (Peterson et al. 2009, Yin et al. 2008). In these studies, a given spot control signal was first compared with the mean control signal and the obtained value was then used to divide the detection signal of the same spot. Common to all these methods is that they do not identify or measure any kind of error in the data but implicitly presume that spot-wise technical

variation exists, and that the procedure is sufficient to correct it. However, while it is a reasonable assumption that technical variation in spot quality exists in microarrays, the methods to efficiently use an internal control to correct errors should be studied separately as was done in **II**.

Microarray normalization in **III** and **IV** was based on signal from internal hybridization control probe and an algorithm which takes into account a bias caused by the low signal intensities from spots not binding any LDR probe but still binding control probe showing high intensities in control channel. This may result in large variances in control signals causing aberrant probe-control ratios and outliers (**II**:Figure 3). On the other hand, the control signal is necessary to quantify spot-to-spot variance in quality in contact printed microarrays. The algorithm does not contain any explicit model on how variances affect the signal of ligation probes, but is rather a heuristic solution based on the bias described above. The ability of the algorithm to correct the bias was demonstrated by empirical data from a few microarray experiments in **II**. To provide further proof for the utility of this normalization concept, results from a simulation experiment are presented here in Figure 2. The bias effect was simulated by random sampling from normal distributions with $\mu=1$, $\sigma=0,5$ for empty probe channel spots (n=900), $\mu=10$, $\sigma=10$ for spots harbouring control probes (n=1000) and $\mu=10$, $\sigma=10$ for spots harbouring positive ligation probes (n=100). The distribution parameter values were intended to be representative of situations with small mean and variation in probe channel for negative spots and higher mean and variation in probe or control channel for spots with bound control or ligation probes, respectively. Figure 3A shows how computing the probe-control ratio can not easily differentiate the majority of true positives (filled circles) from the background distribution. In addition, there are some false positive outliers among the negative population with significantly higher values than the bulk of true positives. Normalization algorithm presented in **II** can correct the bias, differentiate positives from the background and remove outliers in the simulated data (Figure 3B). Examples of reanalysis of data in **I** is presented in the supplements of **II**.

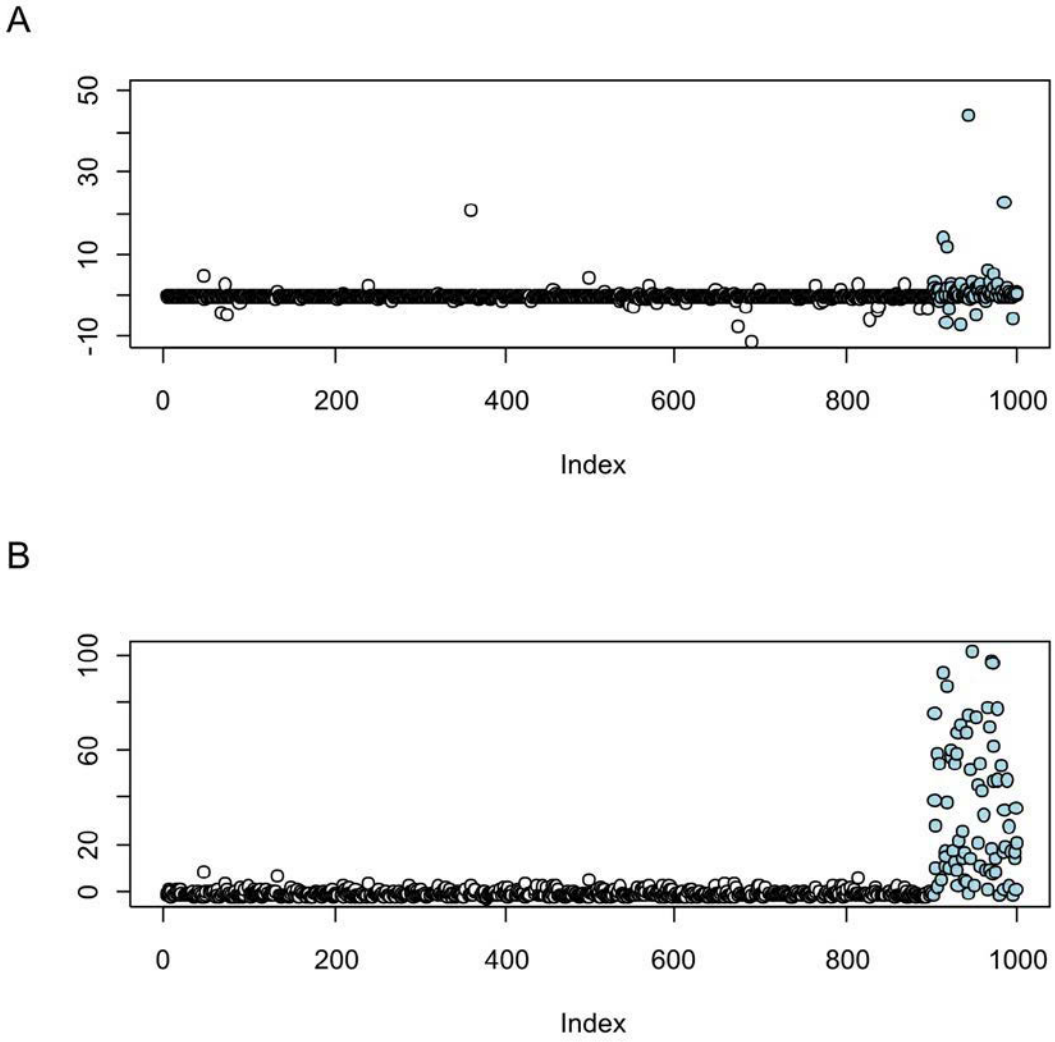

Figure 3. Effect of normalization on simulated microarray data. Blue-filled circles represent positive LDR channel signals (n=100) and white circles represent negative LDR channel signals (n=900). All spots are simulated to produce a control probe signal with same mean and variation as positive LDR signals. See text for details. (A) The simulated data normalized by LDR-control ratio. Extreme outliers are not shown in the plot. (B) The simulated data processed by normalization algorithm described in **II**.

CONCLUSIONS

The results presented in this thesis show the capability of ligation detection based microarray methods to identify microbial groups from environmental and human samples. In addition to appropriate sample handling and DNA extraction, initial PCR of the target gene region is required to achieve high enough sensitivity and robust results using linear amplification of ligation probes. With PCR amplifiable probes, source DNA can be used as a template without prior amplification, although in this approach the target gene is not enriched in the sample DNA and overall sensitivity is about 1%. PCR as an amplification method incorporates biases, but the results regardless show that obtained microarray signals correlate with target sequence number. This suggests that these methods could be used for semiquantitave assaying despite that detection of PCR amplicons as an end-point assay is not likely to provide as high a dynamical range as qPCR. However, achieved quantitative range was at least three orders of magnitude, which is possibly enough for practical pathogen risk level estimation. Furthermore, results from computational method utilizing internal control probe signals suggest that appropriate normalization can significantly improve signal to noise. In summary, the data presented in this thesis suggests that ligation-based microarray assay can be optimized to a degree that allows good signal-to-noise and semiquantitative detection, making it a potential microbial diagnostic platform to take advantage of increasing sequence data and to replace traditional, less efficient culturing based methods that still dominate routine testing. Furthermore, as the methodology is not dependent on any particular sequence, it can in principle be used for detection of DNA or RNA variants in other applications as well.

ACKNOWLEDGEMENTS

This study was carried out at the Institute of Biotechnology of the University of Helsinki during 2007-2011. The financial support by the Technology Agency of Finland (TEKES), the Finnish Cultural Foundation, Maj and Tor Nessling Foundation and the University of Helsinki are gratefully acknowledged. I would like to express my gratitude to my supervisor, head of the DNA sequencing and genomics laboratory docent Petri Auvinen and my inofficial co-supervisor Lars Paulin for their significant scientific support, advice and guidance throughout the project.

I am also grateful to all my coauthors and collaborators in Viikki, Meilahti and Lahti campuses, especially Jenni Hultman, Kaisa Koskinen, Eeva Auvinen, Miia Pitkäranta, Martin Romantschuk and Jukka Kurola. I likewise thank all the personnel and scientists of the DNA sequencing and genomics laboratory for their contribution in providing a professional research environment.

I would like to thank Outi Monni and Minna Pirhonen for taking part in my thesis advisory committee. I am also most thankful to Joakim Lundeberg and Jari Valkonen for the critical review of my thesis.

Finally, I would like to thank my family for their support at all times.

Helsinki, January 2012

Jarmo Ritari

REFERENCES

Adamczyk J, Hesselsoe M, Iversen N, Horn M, Lehner A, Nielsen PH, et al. The isotope array, a new tool that employs substrate-mediated labeling of rRNA for determination of microbial community structure and function. Appl Environ Microbiol. 2003 Nov;69(11):6875-87.

Aguado-Urda M, Lopez-Campos GH, Fernandez-Garayzabal JF, Martin-Sanchez F, Gibello A, Dominguez L, et al. Analysis of the genome content of lactococcus garvieae by genomic interspecies microarray hybridization. BMC Microbiol. 2010 Mar 16;10(1):79.

Aittamaa M, Somervuo P, Pirhonen M, Mattinen L, Nissinen R, Auvinen P, et al. Distinguishing bacterial pathogens of potato using a genome-wide microarray approach. Mol Plant Pathol. 2008 Sep;9(5):705-17.

Akhras MS, Thiyagarajan S, Villablanca AC, Davis RW, Nyren P, Pourmand N. PathogenMip assay: A multiplex pathogen detection assay. PLoS ONE. 2007 Feb 21;2(2):e223.

Anthony RM, Brown TJ, French GL. Rapid diagnosis of bacteremia by universal amplification of 23S ribosomal DNA followed by hybridization to an oligonucleotide array. J Clin Microbiol. 2000 Feb;38(2):781-8.

Antson DO, Mendel-Hartvig M, Landegren U, Nilsson M. PCR-generated padlock probes distinguish homologous chromosomes through quantitative fluorescence analysis. Eur J Hum Genet. 2003 May;11(5):357-63.

Augenlicht LH, Kobrin D. Cloning and screening of sequences expressed in a mouse colon tumor. Cancer Res. 1982 Mar;42(3):1088-93.

Barany F. Genetic disease detection and DNA amplification using cloned thermostable ligase. Proc Natl Acad Sci U S A. 1991 Jan 1;88(1):189-93.

Bergval IL, Vijzelaar RN, Dalla Costa ER, Schuitema AR, Oskam L, Kritski AL, et al. Development of multiplex assay for rapid characterization of mycobacterium tuberculosis. J Clin Microbiol. 2008 Feb;46(2):689-99.

Bodrossy L, Stralis-Pavese N, Konrad-Koszler M, Weilharter A, Reichenauer TG, Schofer D, et al. mRNA-based parallel detection of active methanotroph populations by use of a diagnostic microarray. Appl Environ Microbiol. 2006 Feb;72(2):1672-6.

Bolstad BM, Irizarry RA, Astrand M, Speed TP. A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. Bioinformatics. 2003 Jan 22;19(2):185-93.

Brodie EL, Desantis TZ, Joyner DC, Baek SM, Larsen JT, Andersen GL, et al. Application of a high-density oligonucleotide microarray approach to study bacterial population dynamics during uranium reduction and reoxidation. Appl Environ Microbiol. 2006 Sep;72(9):6288-98.

Brodie EL, DeSantis TZ, Parker JP, Zubietta IX, Piceno YM, Andersen GL. Urban aerosols harbor diverse and dynamic bacterial populations. Proc Natl Acad Sci U S A. 2007 Jan 2;104(1):299-304.

Busti E, Bordoni R, Castiglioni B, Monciardini P, Sosio M, Donadio S, et al. Bacterial discrimination by means of a universal array approach mediated by LDR (ligase detection reaction). BMC Microbiol. 2002 Sep 20;2:27.

Candela M, Consolandi C, Severgnini M, Biagi E, Castiglioni B, Vitali B, et al. High taxonomic level fingerprint of the human intestinal microbiota by ligase detection reaction--universal array approach. BMC Microbiol. 2010 Apr 19;10:116.

Casneuf T, Van de Peer Y, Huber W. In situ analysis of cross-hybridisation on microarrays and the inference of expression correlation. BMC Bioinformatics. 2007 Nov 26;8:461.

Castiglioni B, Rizzi E, Frosini A, Sivonen K, Rajaniemi P, Rantala A, et al. Development of a universal microarray based on the ligation detection reaction and 16S rrna gene polymorphism to target diversity of cyanobacteria. Appl Environ Microbiol. 2004 Dec;70(12):7161-72.

Chee M, Yang R, Hubbell E, Berno A, Huang XC, Stern D, et al. Accessing genetic information with high-density DNA arrays. Science. 1996 Oct 25;274(5287):610-4.

Chevet E, Lemaitre G, Katinka MD. Low concentrations of tetramethylammonium chloride increase yield and specificity of PCR. Nucleic Acids Res. 1995 Aug 25;23(16):3343-4.

Cleven BE, Palka-Santini M, Gielen J, Meembor S, Kronke M, Krut O. Identification and characterization of bacterial pathogens causing bloodstream infections by DNA microarray. J Clin Microbiol. 2006 Jul;44(7):2389-97.

Condon C, Philips J, Fu ZY, Squires C, Squires CL. Comparison of the expression of the seven ribosomal RNA operons in escherichia coli. EMBO J. 1992 Nov;11(11):4175-85.

Cutler DJ, Zwick ME, Carrasquillo MM, Yohn CT, Tobin KP, Kashuk C, et al. High-throughput variation detection and genotyping using microarrays. Genome Res. 2001 Nov;11(11):1913-25.

Dennis P, Edwards EA, Liss SN, Fulthorpe R. Monitoring gene expression in mixed microbial communities by using DNA microarrays. Appl Environ Microbiol. 2003 Feb;69(2):769-78.

DeSantis TZ, Brodie EL, Moberg JP, Zubieta IX, Piceno YM, Andersen GL. High-density universal 16S rRNA microarray analysis reveals broader diversity than typical clone library when sampling the environment. Microb Ecol. 2007 Apr;53(3):371-83.

Dixon MT, Hillis DM. Ribosomal RNA secondary structure: Compensatory mutations and implications for phylogenetic analysis. Mol Biol Evol. 1993 Jan;10(1):256-67.

Dorris DR, Nguyen A, Gieser L, Lockner R, Lublinsky A, Patterson M, et al. Oligodeoxyribonucleotide probe accessibility on a three-dimensional DNA microarray surface and the effect of hybridization time on the accuracy of expression ratios. BMC Biotechnol. 2003 Jun 11;3:6.

Edwards D. Non-linear normalization and background correction in one-channel cDNA microarray studies. Bioinformatics. 2003 May 1;19(7):825-33.

Eid J, Fehr A, Gray J, Luong K, Lyle J, Otto G, et al. Real-time DNA sequencing from single polymerase molecules. Science. 2008 Nov 20.

El Fantroussi S, Urakawa H, Bernhard AE, Kelly JJ, Noble PA, Smidt H, et al. Direct profiling of environmental microbial populations by thermal dissociation analysis of native rRNAs hybridized to oligonucleotide microarrays. Appl Environ Microbiol. 2003 Apr;69(4):2377-82.

Fan JB, Gunderson KL, Bibikova M, Yeakley JM, Chen J, Wickham Garcia E, et al. Illumina universal bead arrays. Methods Enzymol. 2006;410:57-73.

Faruqi AF, Hosono S, Driscoll MD, Dean FB, Alsmadi O, Bandaru R, et al. High-throughput genotyping of single nucleotide polymorphisms with rolling circle amplification. BMC Genomics. 2001;2(1):4.

Fire A, Xu SQ. Rolling replication of short DNA circles. Proc Natl Acad Sci U S A. 1995 May 9;92(10):4641-5.

Fogel GB, Collins CR, Li J, Brunk CF. Prokaryotic genome size and SSU rDNA copy number: Estimation of microbial relative abundance from a mixed population. Microb Ecol. 1999 Aug;38(2):93-113.

Gerry NP, Witowski NE, Day J, Hammer RP, Barany G, Barany F. Universal DNA microarray method for multiplex detection of low abundance point mutations. J Mol Biol. 1999 Sep 17;292(2):251-62.

Ginya H, Matsushita R, Yohda M. Quantification and improvement of error rate during ligase detection reaction. J Biosci Bioeng. 2010 Feb;109(2):202-4.

Grundmeier M, Tuchscherr L, Bruck M, Viemann D, Roth J, Willscher E, et al. Staphylococcal strains vary greatly in their ability to induce an inflammatory response in endothelial cells. J Infect Dis. 2010 Mar 15;201(6):871-80.

Guo Z, Guilfoyle RA, Thiel AJ, Wang R, Smith LM. Direct fluorescence analysis of genetic polymorphisms by hybridization with oligonucleotide arrays on glass supports. Nucleic Acids Res. 1994 Dec 11;22(24):5456-65.

Gupta R, Auvinen P, Thomas A, Arjas E. Bayesian hierarchical model for correcting signal saturation in microarrays using pixel intensities. Stat Appl Genet Mol Biol. 2006;5:Article20.

Hagan MF, Chakraborty AK. Hybridization dynamics of surface immobilized DNA. J Chem Phys. 2004 Mar 8;120(10):4958-68.

Hardenbol P, Yu F, Belmont J, Mackenzie J, Bruckner C, Brundage T, et al. Highly multiplexed molecular inversion probe genotyping: Over 10,000 targeted SNPs genotyped in a single tube assay. Genome Res. 2005 Feb;15(2):269-75.

Hayward RE, Derisi JL, Alfadhli S, Kaslow DC, Brown PO, Rathod PK. Shotgun DNA microarrays and stage-specific gene expression in plasmodium falciparum malaria. Mol Microbiol. 2000 Jan;35(1):6-14.

He Z, Gentry TJ, Schadt CW, Wu L, Liebich J, Chong SC, et al. GeoChip: A comprehensive microarray for investigating biogeochemical, ecological and environmental processes. ISME J. 2007 May;1(1):67-77.

Hogervorst FB, Nederlof PM, Gille JJ, McElgunn CJ, Grippeling M, Pruntel R, et al. Large genomic deletions and duplications in the BRCA1 gene identified by a novel quantitative method. Cancer Res. 2003 Apr 1;63(7):1449-53.

Huber W, von Heydebreck A, Sultmann H, Poustka A, Vingron M. Variance stabilization applied to microarray data calibration and to the quantification of differential expression. Bioinformatics. 2002;18 Suppl 1:S96-104.

Huber W, von Heydebreck A, Sultmann H, Poustka A, Vingron M. Variance stabilization applied to microarray data calibration and to the quantification of differential expression. Bioinformatics. 2002;18 Suppl 1:S96-104.

Hughes TR, Mao M, Jones AR, Burchard J, Marton MJ, Shannon KW, et al. Expression profiling using microarrays fabricated by an ink-jet oligonucleotide synthesizer. Nat Biotechnol. 2001 Apr;19(4):342-7.

Hultman J, Ritari J, Romantschuk M, Paulin L, Auvinen P. Universal ligation-detection-reaction microarray applied for compost microbes. BMC Microbiol. 2008 Dec 30;8:237.

Iafrate AJ, Feuk L, Rivera MN, Listewnik ML, Donahoe PK, Qi Y, et al. Detection of large-scale variation in the human genome. Nat Genet. 2004 Sep;36(9):949-51.

Kane MD, Jatkoe TA, Stumpf CR, Lu J, Thomas JD, Madore SJ. Assessment of the sensitivity and specificity of oligonucleotide (50mer) microarrays. Nucleic Acids Res. 2000 Nov 15;28(22):4552-7.

Kostic T, Weilharter A, Rubino S, Delogu G, Uzzau S, Rudi K, et al. A microbial diagnostic microarray technique for the sensitive detection and identification of pathogenic bacteria in a background of nonpathogens. Anal Biochem. 2007 Jan 15;360(2):244-54.

Kovarova M, Draber P. New specificity and yield enhancer of polymerase chain reactions. Nucleic Acids Res. 2000 Jul 1;28(13):E70.

Landegren U, Kaiser R, Sanders J, Hood L. A ligase-mediated gene detection technique. Science. 1988 Aug 26;241(4869):1077-80.

Lane DJ, Pace B, Olsen GJ, Stahl DA, Sogin ML, Pace NR. Rapid determination of 16S ribosomal RNA sequences for phylogenetic analyses. Proc Natl Acad Sci U S A. 1985 Oct;82(20):6955-9.

Larsson C, Grundberg I, Soderberg O, Nilsson M. In situ detection and genotyping of individual mRNA molecules. Nat Methods. 2010 May;7(5):395-7.

Lehman IR. DNA ligase: Structure, mechanism, and function. Science. 1974 Nov 29;186(4166):790-7.

Li C, Hung Wong W. Model-based analysis of oligonucleotide arrays: Model validation, design issues and standard error application. Genome Biol. 2001;2(8):RESEARCH0032.

Li HH, Gyllensten UB, Cui XF, Saiki RK, Erlich HA, Arnheim N. Amplification and analysis of DNA sequences in single human sperm and diploid cells. Nature. 1988 Sep 29;335(6189):414-7.

Liu D, Daubendiek SL, Zillman MA, Ryan K, Kool ET. Rolling circle DNA synthesis: Small circular oligonucleotides as efficient templates for DNA polymerases. J Am Chem Soc. 1996 Feb 21;118(7):1587-94.

Lizardi PM, Huang X, Zhu Z, Bray-Ward P, Thomas DC, Ward DC. Mutation detection and single-molecule counting using isothermal rolling-circle amplification. Nat Genet. 1998 Jul;19(3):225-32.

Lomakin A, Frank-Kamenetskii MD. A theoretical analysis of specificity of nucleic acid interactions with oligonucleotides and peptide nucleic acids (PNAs). J Mol Biol. 1998 Feb 13;276(1):57-70.

Lopez-Garcia P, Moreira D. Tracking microbial biodiversity through molecular and genomic ecology. Res Microbiol. 2008 Jan-Feb;159(1):67-73.

Loy A, Lehner A, Lee N, Adamczyk J, Meier H, Ernst J, et al. Oligonucleotide microarray for 16S rRNA gene-based detection of all recognized lineages of sulfate-reducing prokaryotes in the environment. Appl Environ Microbiol. 2002 Oct;68(10):5064-81.

Luo J, Barany F. Identification of essential residues in thermus thermophilus DNA ligase. Nucleic Acids Res. 1996 Aug 1;24(15):3079-85.

Luo J, Bergstrom DE, Barany F. Improving the fidelity of thermus thermophilus DNA ligase. Nucleic Acids Res. 1996 Aug 1;24(15):3071-8.

MacBeath G, Schreiber SL. Printing proteins as microarrays for high-throughput function determination. Science. 2000 Sep 8;289(5485):1760-3.

Maidak BL, Cole JR, Lilburn TG, Parker CT,Jr, Saxman PR, Stredwick JM, et al. The RDP (ribosomal database project) continues. Nucleic Acids Res. 2000 Jan 1;28(1):173-4.

Nilsson M, Malmgren H, Samiotaki M, Kwiatkowski M, Chowdhary BP, Landegren U. Padlock probes: Circularizing oligonucleotides for localized DNA detection. Science. 1994 Sep 30;265(5181):2085-8.

Nuwaysir EF, Huang W, Albert TJ, Singh J, Nuwaysir K, Pitas A, et al. Gene expression analysis using oligonucleotide arrays produced by maskless photolithography. Genome Res. 2002 Nov;12(11):1749-55.

Ohrmalm C, Jobs M, Eriksson R, Golbob S, Elfaitouri A, Benachenhou F, et al. Hybridization properties of long nucleic acid probes for detection of variable target sequences, and development of a hybridization prediction algorithm. Nucleic Acids Res. 2010 Nov;38(21):e195.

Okoniewski MJ, Miller CJ. Hybridization interactions between probesets in short oligo microarrays lead to spurious correlations. BMC Bioinformatics. 2006 Jun 2;7:276.

Palmer C, Bik EM, DiGiulio DB, Relman DA, Brown PO. Development of the human infant intestinal microbiota. PLoS Biol. 2007 Jul;5(7):e177.

Palmer C, Bik EM, Eisen MB, Eckburg PB, Sana TR, Wolber PK, et al. Rapid quantitative profiling of complex microbial populations. Nucleic Acids Res. 2006 Jan 10;34(1):e5.

Palmer S, Wiegand AP, Maldarelli F, Bazmi H, Mican JM, Polis M, et al. New real-time reverse transcriptase-initiated PCR assay with single-copy sensitivity for human immunodeficiency virus type 1 RNA in plasma. J Clin Microbiol. 2003 Oct;41(10):4531-6.

Pease AC, Solas D, Sullivan EJ, Cronin MT, Holmes CP, Fodor SP. Light-generated oligonucleotide arrays for rapid DNA sequence analysis. Proc Natl Acad Sci U S A. 1994 May 24;91(11):5022-6.

Pelz CR, Kulesz-Martin M, Bagby G, Sears RC. Global rank-invariant set normalization (GRSN) to reduce systematic distortions in microarray data. BMC Bioinformatics. 2008 Dec 4;9:520.

Peterson AW, Heaton RJ, Georgiadis RM. The effect of surface probe density on DNA hybridization. Nucleic Acids Res. 2001 Dec 15;29(24):5163-8.

Peterson G, Bai J, Narayanan S. A co-printed oligomer to enhance reliability of spotted microarrays. J Microbiol Methods. 2009 Jun;77(3):261-6.

Pingle MR, Granger K, Feinberg P, Shatsky R, Sterling B, Rundell M, et al. Multiplexed identification of blood-borne bacterial pathogens by use of a novel 16S rRNA gene PCR-ligase detection reaction-capillary electrophoresis assay. J Clin Microbiol. 2007 Jun;45(6):1927-35.

Polz MF, Cavanaugh CM. Bias in template-to-product ratios in multitemplate PCR. Appl Environ Microbiol. 1998 Oct;64(10):3724-30.

Pribnow D. Nucleotide sequence of an RNA polymerase binding site at an early T7 promoter. Proc Natl Acad Sci U S A. 1975 Mar;72(3):784-8.

Prins TW, van Dijk JP, Beenen HG, Van Hoef AA, Voorhuijzen MM, Schoen CD, et al. Optimised padlock probe ligation and microarray detection of multiple (non-authorised) GMOs in a single reaction. BMC Genomics. 2008 Dec 4;9:584.

Rantala A, Rizzi E, Castiglioni B, de Bellis G, Sivonen K. Identification of hepatotoxin-producing cyanobacteria by DNA-chip. Environ Microbiol. 2008 Mar;10(3):653-64.

Rantala JK, Makela R, Aaltola AR, Laasola P, Mpindi JP, Nees M, et al. A cell spot microarray method for production of high density siRNA transfection microarrays. BMC Genomics. 2011 Mar 28;12:162.

Rastogi G, Osman S, Vaishampayan PA, Andersen GL, Stetler LD, Sani RK. Microbial diversity in uranium mining-impacted soils as revealed by high-density 16S microarray and clone library. Microb Ecol. 2010 Jan;59(1):94-108.

Reich M, Kohler A, Martin F, Buee M. Development and validation of an oligonucleotide microarray to characterise ectomycorrhizal fungal communities. BMC Microbiol. 2009 Nov 24;9:241.

Reijans M, Dingemans G, Klaassen CH, Meis JF, Keijdener J, Mulders B, et al. RespiFinder: A new multiparameter test to differentially identify fifteen respiratory viruses. J Clin Microbiol. 2008 Apr;46(4):1232-40.

Rivas LA, Garcia-Villadangos M, Moreno-Paz M, Cruz-Gil P, Gomez-Elvira J, Parro V. A 200-antibody microarray biochip for environmental monitoring: Searching for universal microbial biomarkers through immunoprofiling. Anal Chem. 2008 Nov 1;80(21):7970-9.

Rocke DM, Durbin B. A model for measurement error for gene expression arrays. J Comput Biol. 2001;8(6):557-69.

Rujescu D, Ingason A, Cichon S, Pietilainen OP, Barnes MR, Toulopoulou T, et al. Disruption of the neurexin 1 gene is associated with schizophrenia. Hum Mol Genet. 2009 Mar 1;18(5):988-96.

Sanchez JA, Pierce KE, Rice JE, Wangh LJ. Linear-after-the-exponential (LATE)-PCR: An advanced method of asymmetric PCR and its uses in quantitative real-time analysis. Proc Natl Acad Sci U S A. 2004 Feb 17;101(7):1933-8.

SantaLucia J,Jr. A unified view of polymer, dumbbell, and oligonucleotide DNA nearest-neighbor thermodynamics. Proc Natl Acad Sci U S A. 1998 Feb 17;95(4):1460-5.

Saunders NA, Underwood A, Kearns AM, Hallas G. A virulence-associated gene microarray: A tool for investigation of the evolution and pathogenic potential of staphylococcus aureus. Microbiology. 2004 Nov;150(Pt 11):3763-71.

Schena M, Shalon D, Davis RW, Brown PO. Quantitative monitoring of gene expression patterns with a complementary DNA microarray. Science. 1995 Oct 20;270(5235):467-70.

Schouten JP, McElgunn CJ, Waaijer R, Zwijnenburg D, Diepvens F, Pals G. Relative quantification of 40 nucleic acid sequences by multiplex ligation-dependent probe amplification. Nucleic Acids Res. 2002 Jun 15;30(12):e57.

Sergeev N, Volokhov D, Chizhikov V, Rasooly A. Simultaneous analysis of multiple staphylococcal enterotoxin genes by an oligonucleotide microarray assay. J Clin Microbiol. 2004 May;42(5):2134-43.

Shah MY, Pan X, Fix LN, Farwell MA, Zhang B. 5-fluorouracil drug alters the microRNA expression profiles in MCF-7 breast cancer cells. J Cell Physiol. 2011 Jul;226(7):1868-78.

Singh J, Behal A, Singla N, Joshi A, Birbian N, Singh S, et al. Metagenomics: Concept, methodology, ecological inference and recent advances. Biotechnol J. 2009 Apr;4(4):480-94.

Skakni L, Sardet A, Just J, Landman-Parker J, Costil J, Moniot-Ville N, et al. Detection of mycoplasma pneumoniae in clinical samples from pediatric patients by polymerase chain reaction. J Clin Microbiol. 1992 Oct;30(10):2638-43.

Southern E, Mir K, Shchepinov M. Molecular interactions on microarrays. Nat Genet. 1999 Jan;21(1 Suppl):5-9.

Suzuki MT, Giovannoni SJ. Bias caused by template annealing in the amplification of mixtures of 16S rRNA genes by PCR. Appl Environ Microbiol. 1996 Feb;62(2):625-30.

Szemes M, Bonants P, de Weerdt M, Baner J, Landegren U, Schoen CD. Diagnostic application of padlock probes--multiplex detection of plant pathogens using universal microarrays. Nucleic Acids Res. 2005 Apr 28;33(8):e70.

Taroncher-Oldenburg G, Griner EM, Francis CA, Ward BB. Oligonucleotide microarray for the study of functional gene diversity in the nitrogen cycle in the environment. Appl Environ Microbiol. 2003 Feb;69(2):1159-71.

Thomas DC, Nardone GA, Randall SK. Amplification of padlock probes for DNA diagnostics by cascade rolling circle amplification or the polymerase chain reaction. Arch Pathol Lab Med. 1999 Dec;123(12):1170-6.

Tiquia SM, Wu L, Chong SC, Passovets S, Xu D, Xu Y, et al. Evaluation of 50-mer oligonucleotide arrays for detecting microbial populations in environmental samples. BioTechniques. 2004 Apr;36(4):664,70, 672, 674-5.

Tomkinson AE, Vijayakumar S, Pascal JM, Ellenberger T. DNA ligases: Structure, reaction mechanism, and function. Chem Rev. 2006 Feb;106(2):687-99.

Torsvik V, Goksoyr J, Daae FL. High diversity in DNA of soil bacteria. Appl Environ Microbiol. 1990 Mar;56(3):782-7.

Turner DH. Thermodynamics of base pairing. Curr Opin Struct Biol. 1996 Jun;6(3):299-304.

Urisman A, Fischer KF, Chiu CY, Kistler AL, Beck S, Wang D, et al. E-predict: A computational strategy for species identification based on observed DNA microarray hybridization patterns. Genome Biol. 2005;6(9):R78.

van Doorn R, Slawiak M, Szemes M, Dullemans AM, Bonants P, Kowalchuk GA, et al. Robust detection and identification of multiple oomycetes and fungi in environmental samples by using a novel cleavable padlock probe-based ligation detection assay. Appl Environ Microbiol. 2009 Jun;75(12):4185-93.

van Doorn R, Szemes M, Bonants P, Kowalchuk GA, Salles JF, Ortenberg E, et al. Quantitative multiplex detection of plant pathogens using a novel ligation probe-based system coupled with universal, high-throughput real-time PCR on OpenArrays. BMC Genomics. 2007 Aug 14;8:276.

von Wintzingerode F, Gobel UB, Stackebrandt E. Determination of microbial diversity in environmental samples: Pitfalls of PCR-based rRNA analysis. FEMS Microbiol Rev. 1997 Nov;21(3):213-29.

Wang F, Zhou H, Meng J, Peng X, Jiang L, Sun P, et al. GeoChip-based analysis of metabolic diversity of microbial communities at the juan de fuca ridge hydrothermal vent. Proc Natl Acad Sci U S A. 2009 Mar 24;106(12):4840-5.

Wang Q, Wang S, Beutin L, Cao B, Feng L, Wang L. Development of a DNA microarray for the detection and serotyping of enterotoxigenic escherichia coli. J Clin Microbiol. 2010 Mar 29.

Wang RF, Beggs ML, Robertson LH, Cerniglia CE. Design and evaluation of oligonucleotide-microarray method for the detection of human intestinal bacteria in fecal samples. FEMS Microbiol Lett. 2002 Aug 6;213(2):175-82.

Wetmur JG. DNA probes: Applications of the principles of nucleic acid hybridization. Crit Rev Biochem Mol Biol. 1991;26(3-4):227-59.

Wetmur JG. DNA probes: Applications of the principles of nucleic acid hybridization. Crit Rev Biochem Mol Biol. 1991;26(3-4):227-59.

Wilson WJ, Strout CL, DeSantis TZ, Stilwell JL, Carrano AV, Andersen GL. Sequence-specific identification of 18 pathogenic microorganisms using microarray technology. Mol Cell Probes. 2002 Apr;16(2):119-27.

Wong CW, Heng CL, Wan Yee L, Soh SW, Kartasasmita CB, Simoes EA, et al. Optimization and clinical validation of a pathogen detection microarray. Genome Biol. 2007;8(5):R93.

Workman C, Jensen LJ, Jarmer H, Berka R, Gautier L, Nielser HB, et al. A new non-linear normalization method for reducing variability in DNA microarray experiments. Genome Biol. 2002 Aug 30;3(9):research0048.

Yakovchuk P, Protozanova E, Frank-Kamenetskii MD. Base-stacking and base-pairing contributions into thermal stability of the DNA double helix. Nucleic Acids Res. 2006 Jan 31;34(2):564-74.

Yang YH, Dudoit S, Luu P, Lin DM, Peng V, Ngai J, et al. Normalization for cDNA microarray data: A robust composite method addressing single and multiple slide systematic variation. Nucleic Acids Res. 2002 Feb 15;30(4):e15.

Yin BC, Li H, Ye BC. A dual-probe hybridization method for reducing variability in single nucleotide polymorphism analysis with oligonucleotide microarrays. Anal Biochem. 2008 Dec 15;383(2):270-8.

Zhang DY, Brandwein M, Hsuih TC, Li H. Amplification of target-specific, ligation-dependent circular probe. Gene. 1998 May 12;211(2):277-85.

Zhang W, Cohenford M, Lentrichia B, Isenberg HD, Simson E, Li H, et al. Detection of chlamydia trachomatis by isothermal ramification amplification method: A feasibility study. J Clin Microbiol. 2002 Jan;40(1):128-32.