Biomolecular Analysis by Dual-Tag Microarrays and Single Molecule Amplification

OLLE ERICSSON
Dissertation presented at Uppsala University to be publicly examined in Rudbeckssalen, Rudbeckslaboratoriet, Daghammarsköldsväg 20, Uppsala, Thursday, March 20, 2008 at 09:15 for the degree of Doctor of Philosophy (Faculty of Medicine). The examination will be conducted in English.

Abstract

Padlock probes and proximity ligation are two powerful molecular tools for detection of nucleic acids and proteins, respectively. Both methods result in the formation of DNA reporter molecules upon recognition of specific target molecules. These reporter molecules can be designed to include tag sequences that can be analyzed by techniques for nucleic acid analysis. Herein, I present a dual-tag microarray (DTM) platform that is suitable for high-performance analyses of DNA reporter molecule libraries, generated by padlock and proximity probing reactions. The DTM platform was applied for analysis of mRNA transcripts using padlock probes, and of cytokines using proximity ligation. The platform drastically improved specificity of detection, and it allowed precise measurements of proteins and nucleic acids over wide dynamic ranges.

The thesis also presents two techniques for multi-probe analyses of biomolecules: the triple-specific proximity ligation assay (3PLA) for protein analyses, and the spliceotyping assay for mRNA analyses. 3PLA allows highly specific measurements of as little as hundreds of target protein molecules by interrogating three target epitopes simultaneously. In spliceotyping the exon composition of individual transcripts are represented as a series of tag sequences in DNA reporter molecules, via a series of target-dependent ligation reactions. Next, the splicing patterns along individual transcripts can be revealed by amplified single molecule detection and step-wise decoding.

Keywords: Microarray, proximity ligation, Padlock probe, Rolling circle amplification, Splicing, single molecule

Olle Ericsson, Department of Genetics and Pathology, Rudbecklabatoriet, Uppsala University, SE-75185 Uppsala, Sweden

© Olle Ericsson 2008

ISSN 1651-6206
urn:nbn:se:uu:diva-8475 (http://urn.kb.se/resolve?urn=urn:nbn:se:uu:diva-8475)
To my family
List of papers

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


II Olle Ericsson*, Rachel Nong*, Katerina Pardali, Ulf Landegren. Parallel protein analysis by proximity ligation with DNA microarray read-out. (Manuscript)


IV Tim Conze, Jenny Göransson, Hamidreza Razzaghian, Ulf Landegren, Mats Nilsson, Olle Ericsson “Single Molecule Analysis of Combinatorial Splicing” (Manuscript)

*Authors contributed equally to the work
Related work by the author

Peer reviewed research


Patent applications


# Contents

Introduction ................................................................................................... 11

Background ................................................................................................... 12
  Microarray analysis .................................................................................. 12
    Microarrays for nucleic acid analysis .................................................. 12
    Microarrays for protein analysis ......................................................... 13
  Microarray probe generation ............................................................... 13
  Microarray production ......................................................................... 14
  Microarrays and specificity ................................................................. 15

Discourse on multiplex nucleic acid amplification .................................. 16
  Non selective target amplification ....................................................... 16
  Selective multiplex amplification ........................................................ 16
  Solid phase DNA amplification........................................................... 17
  On dynamic range and signal amplification ........................................ 19

Single molecule analysis of nucleic acids ................................................ 21
  Probe-based analysis ............................................................................ 22
    Ligation-based nucleic acid assays .................................................... 23
    Polymerase based tag microarray assays .......................................... 24
  Protein detection by nucleic acid reporter molecules .......................... 25
  Protein analysis by DNA ligation ......................................................... 26
  Multiple binder analysis .................................................................... 29

Analysis of splicing patterns .................................................................... 30
  Interaction analysis ............................................................................. 31
    Yeast two hybrid ............................................................................. 31
    Mass spectrometry ........................................................................... 32
    Interaction in action ......................................................................... 32

Present investigations .................................................................................... 34
  Paper I. A dual-tag microarray platform for high-performance nucleic
    acid and protein analyses .................................................................. 34
    Perspectives on paper I ................................................................... 35
  Paper II. Parallel protein analysis by proximity ligation with DNA
    microarray read-out ......................................................................... 37
    Perspectives on paper II .................................................................. 38
  Paper III. Sensitive protein detection via triple-binder proximity ligation
    assay. ............................................................................................... 40
Abbreviations

SNP: Single Nucleotide Polymorphism
CGH: Comparative Genome Hybridization
DTM: Dual Tag Microarray
RNA: Ribonucleic Acid
DNA: Deoxyribonucleic Acid
mRNA: Messenger RNA
C2CA: Circle to Circle Amplification
RCA: Rolling Circle Amplification
Q-PCR: Quantitative (Real Time) PCR
PLA: Proximity Ligation Assay
3PLA: Three binder PLA
VEGF: Vascular Endothelial Growth Factor
PSA: Prostate Specific Antigen
ACT: AntiChymoTrypsin
cDNA: Complementary DNA
HIV: Human Immunodeficiency Virus
Y2H: Yeast Two Hybrid
TAP-MS: Tandem Affinity Pullout coupled with Mass Spectrometry
WGA: Whole Genome Amplification
HRCA: Hyper branched RCA
OLA: Oligonucleotide Ligation Assay
MLPA: Multiplex Ligation Probe Assay
MIP: Molecular Inversion Probe
ASE: Allele Specific Extension
EST: Expressed Sequence Tag
ChIP-chip: Chromatin Immunoprecipitation combined with microarray readout.
LOD: Limit Of Detection
Introduction

The human genome is composed of approximately 3 billion base pairs, encoding roughly 20,000 genes with an average of 10 exons, a majority in multiple splice variants. These genes are transcribed and translated into the proteome complement to the human genome, increasing the molecular diversity further by for example differential expression levels and post translational modification. The accessibility of biomolecular information is inversely proportional to the molecular complexity, and while population-wide genome projects have been initiated the proteome and interactome remains relatively unexplored.

New and improved tools are continuously developed to enhance parallel analysis of biomolecules. The work presented herein is a further contribution to these efforts. Padlock probes and proximity ligation have previously been adapted for highly specific analysis of nucleic acids and proteins by encoding target molecules into reporter nucleic acids which in turn are detected. The current thesis presents a microarray platform enabling high performance readout of the nucleic acid reporter molecules. Furthermore, a proximity ligation assay is presented where three separate protein epitopes are encoded as a single nucleic acid reporter molecule, allowing detection of down to hundreds of target molecule. Finally a strategy is presented for analysis of alternative splicing iso-forms of individual transcripts that enable interrogation of exon composition by single molecule readout.
Background

Microarray analysis

Microarrays for nucleic acid analysis

Arising from the discovery of the double helical DNA molecule and the opportunities for DNA hybridization, technologies have evolved continuously. Solid phase hybridization of nucleic acids like the Southern blot, the dot blot and reverse dot blot, can be seen as logical steps, leading up to the concept of DNA microarrays in the late 80s\textsuperscript{1-5}.

Several technology generations have passed since these initial efforts. Attempts to interrogate numerous nucleic acid sequences in parallel first focused on the development of probe arrays for sequencing by hybridization to enable rapid sequence acquisition. Later, arrays were developed to monitor the expression of many genes in individual samples. Early “macro arrays” were later refined into the current wafers with μm-size features, allowing in the order of a million targets to be interrogated per analysis. Some technological milestones during the microarray technology development are the demonstration of cDNA microarrays and the introduction of light directed \textit{in situ} DNA synthesis and the application thereof for a variety of genetic analyses\textsuperscript{6-8}.

Initially the microarray technology was heterogeneous, and dominated by in-house manufacture of microarrays composed of cloned PCR products, producing data inconsistencies among users\textsuperscript{9}. Introduction of standards for microarray experiments and for data analysis, establishment of standardized RNA controls and community efforts to secure quality control have all increased reproducibility\textsuperscript{10-13}. Commercialization of the production of microarrays by specialized vendors and increased centralization of the technology to dedicated core centers have also contributed to enhanced performance.

Utilization of the microarray technology has evolved from early attempts at gene expression profiling into advanced transcriptional analyses where splicing patterns can be investigated using splice junction arrays and exon microarrays as well as by exon discovery using tiling microarrays\textsuperscript{14-16}. Applications directed at analyses of genomic DNA include studies of copy-
number variation by methods like comparative genome hybridization (CGH), genotyping of single nucleotide polymorphism (SNP), and on-chip analyses of DNA molecules isolated by chromatin immunoprecipitation (ChIP-chip).\textsuperscript{17-20}

The HapMap initiative, establishing a map of human polymorphisms, encouraged development of high-throughput methods for SNP genotyping and today more than a million genetic markers can be screened in one analysis using microarray-based technologies\textsuperscript{21-23}. These high-density, genome-wide screening techniques now allow interrogation of large patient cohorts for quantitative trait loci, revealing the genetic basis of a wide variety of complex traits ranging from hair and skin pigmentation to host response upon HIV infection\textsuperscript{24, 25}.

Microarrays for protein analysis
Analysis of the protein complement encoded in the human genome allows closer investigation of the actuators executing most cellular function. While whole genome SNP screening arrays allow phenotypes to be associated with specific genomic regions, and gene regulation can be interrogated by expression microarrays, the functional interconnection between gene products so far requires other analytic techniques. Protein analysis enables annotation of genes for example by mapping physical protein-protein interactions or enzyme-substrate pairs, providing functional information.

In parallel to the development of nucleic acid microarrays, methods have also been established for protein analysis in microarray formats. However, it has proven much more difficult to scale-up protein analysis by microarray technology compared to nucleic acid analyses. Microarray-based protein analyses enable investigation of thousands of proteins\textsuperscript{26, 27}. A wide spectrum of microarray embodiments have been presented, ranging from classical affinity-based protein detection to functional characterization of enzymatic activities of proteins, and interactions between proteins and with other proteins or with phospholipids, small molecules and DNA\textsuperscript{27-31}.

Microarray probe generation
Probe synthesis for microarray analysis of nucleic acids primarily depends on bioinformatic analysis of the target nucleic acid to identify one or several sequences unique in the relevant background genome or transcriptome. Probes are then synthesized either on-chip or off-chip followed by dispensing onto an array. On-chip synthesis is generally performed by light directed \textit{in situ} synthesis or combinatorial on-chip synthesis by ink jet dispensing\textsuperscript{7, 32}. Off-chip probe production is typically achieved by PCR amplification or conventional oligonucleotide synthesis, followed by array manufacture using a microarray printer. The bioinformatic design of nucleic acid probes has
been developed extensively and the success rate in for example genotyping applications is high\textsuperscript{20, 33}. Protein probes, on the other hand, generally cannot be synthesized chemically, with the exception of aptamers and peptides, but they require \textit{in vivo} or \textit{in vitro} translation and purification\textsuperscript{30, 34}. Furthermore, construction of protein affinity probes requires target synthesis and purification, followed by \textit{in vitro} or \textit{in vivo} selection to produce suitable affinity binders. Protein probe generation by selection is empirical in contrast to the bioinformatic design of nucleic acid probes. Therefore, unlike for nucleic acid probes it is not possible to evaluate protein-binding reagents by \textit{in silico} analysis. Even though it is possible to select unique protein domains \textit{in silico} for subsequent binder generation, this does not ensure success and the targeted epitopes may be common to other protein molecules. Cross-reactivity among different targets is frequently reported by protein microarray analyses of antibodies\textsuperscript{35, 36}.

\textbf{Microarray production}\n
DNA microarray production techniques have developed rapidly since their introduction, and the parallel is often made between Moore’s law concerning the rapidly increasing transistors density in central processing units and the microarray features density\textsuperscript{37, 38}. This comparison is not as far fetched as may occur at first glance, since the two techniques were spawned from a common ancestor technology. Light-directed oligonucleotide synthesis is an extraordinary example of the impact of interdisciplinary research. The marriage of light-directed, high-precision micro-etching with combinatorial oligonucleotide synthesis via light sensitive chemistry that occurred in Silicon valley during the early 90’s have had dramatic consequences for molecular analyses\textsuperscript{7}. Competing strategies for nucleic acid microarray production also resulted from interdisciplinary research efforts, as exemplified by the combination of the ink-jet printer with oligonucleotide synthesis or self-assembly of beads on top of etched optical fiber bundles\textsuperscript{32, 39}.

Compared to nucleic acid microarrays, protein microarray manufacture poses additional complications, apart from those of protein probe production, due to the relative instability of proteins and the difficulty of on-chip biosynthesis of affinity probes, although efforts are under way to address each of these difficulties. Peptide probes can currently be generated by \textit{in situ} synthesis on chips, and aptamers should accommodate on chip binder synthesis by standard nucleic acid synthesis, but so far this has not been described in the literature\textsuperscript{7, 30}. However for most protein probes on-chip synthesis is not an option. So far protein microarrays are typically manufactured by conventional contact printing or piezoelectric deposition, and microarray density is not considered the major bottleneck in protein microarray production due to the relatively low number of available antibodies and other affinity probes. These manufacturing techniques become limiting when more than 100,000
features are required on a single microscope slide and this number is more than one order of magnitude higher than from the current protein microarray standards.

Recently, mixtures of discretely labeled bead populations have become widely used in the protein analysis community. These “bead arrays” are defined by fluorescent labeling of beads with two different dyes at variable concentrations, allowing two dimensional spectral separation of bead sub-populations. The beads provide unique advantages over planar arrays in regards to protein coupling to solid phases, binding kinetics, and storage, but the number of resolved features is still orders of magnitudes lower than on conventional planar microarrays.

Innovative approaches have been demonstrated where nucleic acid microarrays are converted into protein microarrays in situ, eliminating the upstream protein synthesis and purification steps. Future combination of on-chip nucleic acid synthesis with on-chip protein synthesis could potentially bridge the gap between protein and nucleic acid microarray manufacture. However, current state of the art on-chip nucleic acid synthesis fails to produce DNA strands of the length required to enable protein synthesis by on-chip translation.

Microarrays and specificity

Enhanced specificity of nucleic acid or protein detection is typically achieved by introducing multiple criterions for detection. By way of example, the enhanced specificity of Southern blot compared to hybridization alone, is achieved by combining size discrimination with hybridization mediated specificity, allowing identification of single-copy loci in genomic DNA. In addition, the restriction enzymes used to fragment the DNA are sensitive to even single nucleotide differences, enabling distinction among single nucleotide variants in the human genome.

Several approaches have been developed to allow closely similar targets to be distinguished in microarray analyses. One remarkable example of the possibility to increase specificity by introduction of a separate mechanism to distinguish molecules was published by Gunderson et al. 2005. SNPs were scored in total genomic DNA using a protocol involving the conventional modules of whole-genome amplification prior to hybridization on arrays and allele-specific extension, followed by on-chip signal amplification. The specificity was derived from the combination of hybridization over 40 nucleotides, with the power of a polymerase to selectively extend only correctly matched primers.

The specificity of microarray analyses can also be augmented by utilizing multiple probes per target, both matched and mismatched, and analyzing the hybridization pattern using dedicated algorithms. This approach has also been perfected to allow scoring of SNPs in total genomic DNA.
Discourse on multiplex nucleic acid amplification

An amplification step is often included to enhance the limit of detection (LOD) in techniques for molecular analyses. Two of the most important inventions in molecular biology the last century enabled clonal amplification of nucleic acid sequences, the first by bacterial cloning and the second *in vitro* by PCR amplification. Since the dawn of nucleic acid amplification a wide spectrum of techniques has been developed. Although PCR still dominates *in vitro* DNA amplification, other approaches have filled specific niches.

Non selective target amplification

Whole genome amplification (WGA) approaches are often used in combination with microarrays. These approaches do not add any specificity as they amplify all present nucleic acid and they are a cost-effective alternative to specific primer designs. Examples of amplification procedures are multiple strand displacement and PCR-based adaptor ligation approaches. The latter approach can also allow sequence-independent complexity reduction upon application on genomic DNA, since fragments that are either too long or too short fail to amplify. This reduction can be helpful for subsequent hybridization-based analyses of sequence variation. Another example of target amplification includes RNA transcription approaches often used in combination with expression profiling. One drawback of universal amplification is concomitant amplification of nucleic acids not subject to analysis which nonetheless contribute to amplification saturation and thereby limit the overall amplification level.

Selective multiplex amplification

Specific amplification procedures are used to focus on sets of targets of interest. This is traditionally achieved by PCR, however intrinsic limitations of multiplex PCR restricts the application for multiple target analysis. Selective amplification of multiple loci can instead be achieved by encoding the target information in probe molecules equipped with general amplification motifs as discussed below. For PCR, two primer sites are included in the probe design and used for amplification upon target recognition. Some advantages with PCR are that the amplification is easy to control by a set cycle number and under optimal conditions the exponential amplification allows detection of single molecules.

Circular nucleic acids can be selectively amplified by rolling circle amplification (RCA). For probe molecules this is typically achieved by incorporation of a single primer site in the probe sequence. The Φ29 polymerase has proven to be extraordinarily suitable for RCA, maybe a slightly unexpected
finding since the Φ29 phage genome is linear. Using Φ29 polymerase a 100 nt DNA circle can be amplified 1000-fold in one hour, producing a 100 kb long concatemeric amplification product. Advantages of RCA include the high processivity, iso-thermal amplification, localized product accumulation and high product yield. Compared to PCR where product accumulation results in product re-association and competitive repression of primer hybridization, RCA is not product inhibited and continues until polymerization substrates are depleted or the products precipitate. RCA also seems to introduce less bias among different amplicons compared to PCR. Two drawbacks compared to the PCR are the limited amplification level and the linear level of amplification which is harder to tune into to a desired level compared to PCR for which a fairly exact amplification level can be set by the cycle number. A circle to circle amplification (C2CA) procedure has been developed to enhance the RCA amplification level by additional amplification cycles. This is enabled by restriction enzyme digestion of the concatemeric RCA product and circularizing the monomers to allow consecutive rounds of RCA. Introduction of a general restriction enzyme cassette enables both restriction enzyme digestion and re-circularization. The C2CA protocol provided high precision, reduced amplification bias, and improved product yield compared to PCR. Another approach that enables high amplification levels of circular or linear nucleic acids is the hyperbranched rolling-circle amplification (HRCA) protocol, resulting in up to 10^9-fold amplification in one hour. By addition of a secondary primer of the reverse polarity to the RCA, geometric amplification is achieved upon utilization of a strand displacing polymerase such as the Φ29 polymerase.

**Solid phase DNA amplification**

For some applications it is desirable to confine the DNA amplification products to the site where the template sequences are located. Examples of such applications include microarrays and massive parallel sequencing. The three most commonly used procedures are versions of solid phase PCR, emulsion PCR and RCA. In comparison to generic localized signal amplification mechanisms like the use of horse radish peroxidase nucleic acid replication confers additional means of increasing the specificity. This is important since amplification alone does not necessarily enhance the assay if amplification artefacts or noise is abundant.

Typically, solid phase nucleic acid amplification is achieved by using immobilized primers followed by subsequent hybridization of a replication template. Replication will only occur upon co-localization of replication template and primer, thereby reducing potential background due to nonspecifically adsorbed replication templates. This mechanism has been exploited for signal amplification on microarrays by RCA in combination with protein detection by sandwich immunoassays and nucleic acid detection by OLA.
These strategies used a generic primer motif and complementary circular DNA template to generate signal amplification from all investigated target molecules. However, to maintain a strict chain of specificity, it is desirable to make amplification conditional on correct target recognition. This ensures specificity by selective amplification upon localization of the target in the correct position. Solid phase PCR has been used to selectively amplify target sequences and primer extension and minisequencing mentioned above also use the polymerase to discriminate targets but do involve nucleic acid amplification. Herein an embodiment of solid phase nucleic acid amplification is presented wherein solid phase RCA is conditional to ligation of the target to the correct microarray probe introducing a very stringent criterion for signal generation.

One often neglected aspect of detection techniques is the impact of specificity on LOD and dynamic range. Securing a strong amplification level will not guarantee a low LOD or enable analysis of rare target molecules. Everyone who has set up a few PCR reactions knows that although the PCR amplification can permit detection of single-copy target molecules, multiple hits in the genome can produce nonspecific amplification and render the assay useless. This applies for all molecular analyses and it is of particular importance in parallel analyses of several targets, and even more so in combination with signal amplification.

Several studies have been described in the literature where inter-feature cross-reactivity has been investigated by e.g. addition of defined target molecules and monitoring appearance of concomitant background signals. Cross-reactive signals, appearing due to non-specific adsorption of targets to irrelevant features, in the order of 1-10% compared to the signal from the correct feature, are often considered insignificant in multiplex analyses of both nucleic acids and proteins. Furthermore, rational design of probes has empirically proven difficult and cross-reactive signals from ~10% of investigated probes have been reported. However, although this level of cross-reactive signal may appear insignificant in comparison with the true signal, the important question is really whether the signal is insignificant for target detection with the affected probe. The consequence of cross-reactive signals may be generation of false positive signals. However, many molecular detection assays compare sets of molecules for differences between states, one frequent comparison is that between cancer and normal tissue. In these instances, cross reactivity will be more likely to generate false negative signals since typically only a minor target subset displays differential expression. Cross reactivity among non-fluctuating targets will lead to an increased noise level that abolishes identification of differential expression for targets below this noise threshold. With this scenario in mind an average cross-reactivity of 1-10% to a few targets can severely impair assay performance for rare target detection. Moreover, widespread analyte cross-reactivity may have adverse effects upon combination with signal amplifica-
tion, generating increased noise levels which effectively prevent investigation of low abundance molecules.

On dynamic range and signal amplification

It is worth pointing out the different effect on the dynamic range for microarray analysis by amplification of nucleic acids before microarray hybridization compared to on-chip signal amplification. The dynamic range of a microarray is set at the lower end by how many detectable units, typically fluorophores, that can be detected over background in one microarray feature with a defined statistical significance. At the upper end the dynamic range is limited by the maximum number of fluorophores per feature. This is in turn is determined by the saturation level when all probes are occupied by targets. A typical detection curve is outlined in Figure 1A. Nucleic amplification before hybridization on arrays does not improve the dynamic range but shifts the window of detectable target concentrations towards lower concentrations. This enables detection of fewer molecules but the arrays are saturated at a lower target concentration (Figure 1B). Solid phase amplification on the other hand facilitates detection of fewer hybridized molecules per feature by increasing signal output from each hybridized target. Compared to solution phase amplification the saturation limit is unaffected by introduction of signal amplification upon array hybridization. Thereby the dynamic range of the analysis is increased since greater numbers of fluorophores can be present per feature at saturation (Figure 1C).
**Figure 1.** Schematic diagrams of the effect of signal amplification on dynamic range and LOD for microarray analyses. A. Characteristics of a detection curve. The dynamic range is determined by the LOD and the point of saturation. LOD is typically defined as two standard deviations above background. B. The effect of on-chip solid phase amplification. Solid phase amplification increases the signal generated from a target concentration, thereby increasing the LOD and dynamic range. C. The effect of solution phase amplification prior to microarray hybridization. The LOD is enhanced by increasing the number of nucleic acid molecules available for hybridization per detected target. The linear dynamic range is shifted towards lower concentrations but it is not expanded since the point of saturation is lowered along with the LOD.

**Single molecule analysis of nucleic acids**

Analysis by direct observation of single molecules introduces unique possibilities to interrogate biomolecular phenomena that may go undetected with averaging measurements. Furthermore, quantification of single molecule events provides high precision and identification of minor subpopulations. The first single biomolecule detection efforts in the 70’s involved identification of individual fluorescence labeled antibodies. Since then, various approaches have been developed for analysis of single biomolecules, some more spectacular than others, including monitoring of enzyme-DNA interactions by attaching single DNA strands to beads. Recent developments of single molecule nucleic acid analysis have focused on DNA sequencing, where single molecule analysis potentially could increase throughput by orders of magnitude. Two of the currently most promising approaches for single molecule sequencing are nanopore based sequencing and fluorescence based sequencing. However, all high throughput sequencing platforms currently on the market utilize different forms of amplified single molecule detection in combination with fluorescence readout. Local clonal amplification of single nucleic acids improve signal to noise ratios by clustering of multiple identical target molecules, compared to identification of single fluorophores. Approaches that enable single molecule amplification for DNA sequencing includes Bert Vogelstein’s emulsion based BEAMing approach (short for beads, emulsion, magnetics and amplification), the polony approach developed in George Church’s laboratory and the on chip bridge PCR used in Illumina’s G1 sequencers. BEAMing involves emulsion based PCR where co-localization of single PCR templates and beads in emulsion vesicles allow clonal amplification and immobilization of the product onto the bead. This allows generation of discrete homogeneous amplification products on each individual bead. Polony amplification involves entrapment of individual PCR templates in a gel matrix followed by contained PCR amplification, generating populations...
of locally immobilized PCR products in the gel using one immobilized primer and one primer in solution phase (Figure 2B)\textsuperscript{59}. This type of gel entrapment was initially demonstrated for \textit{in vitro} cloning by RNA amplification using Q\beta replicase\textsuperscript{74}. Bridge PCR, introduced by Adessi \textit{et al}, use two immobilized primers leading to exclusive solid phase PCR amplification (Figure 2C)\textsuperscript{58}. Finally, localized amplification of single nucleic acid molecules can also be achieved using RCA, as performed in paper I and IV. Polymerization can be initiated using an immobilized primer (Figure 2D) but RCA products generated in solution can also be immobilized and analyzed.

\begin{figure}
\centering
\includegraphics[width=\textwidth]{figure2.png}
\caption{Approaches for single molecule amplification, of a single target sequence (blue). \textbf{A.} Emulsion PCR. An emulsion is used to encapsulate individual templates with PCR primers and a single bead. The template is amplified by PCR using one bead immobilized primer. \textbf{B.} Polony amplification. Single templates are captured using a gel-matrix further comprising immobilized primers. During PCR amplification the products remains localized in the gel. \textbf{C.} Bridge PCR. Two immobilized PCR primers are used to locally amplify a single target. \textbf{D.} RCA. A circular template is amplified by RCA producing a localized amplification product.}
\end{figure}

\textbf{Probe-based analysis}

One intrinsic limitation of analyses on planar solid phases concerns the transport of target molecules to the immobilized probes. Although mixing may enhance binding kinetics, probe-target interactions are more favorably performed in solution, simplifying detection of small numbers or even single molecules\textsuperscript{51}. To enable efficient target-probe interactions in homogenous phase while preserving the advantages of multi-analyte readout provided by planar arrays, the probing and readout steps can be separated into discrete reactions. This can be achieved by encoding information in nucleic acid tags that are subsequently detected by microarray readout. One early demonstration of the use of \textit{in silico} designed tag microarrays to report molecular events was presented by Shoemaker \textit{et al} who established a collection of yeast cell lines, comprising individual gene knock-outs wherein the gene was deleted and replaced by a reporter nucleic acid tag\textsuperscript{75}. The relative impact on
fitness of individual deletions could then be investigated by microarray read-out of the relative tag abundance representing individual clones following PCR amplification of the tag collection using a common primer pair. By allowing nucleic acid tags to represent target molecules, standard microarray can be used to evaluate any set of interrogated targets, avoiding cost and labor associated with array redesign and manufacture. Separation of the probing step from the solid phase further allows insertion of an optional amplification step between target recognition and sorting on microarrays. Amplification of a set of reporter molecules with similar lengths and sequence composition introduces less bias compared to amplification of heterogeneous target DNA pools. Furthermore, amplification increases the reporter molecule concentration, thereby enhancing microarray binding kinetics and allowing detection of minute target amounts.

Ligation-based nucleic acid assays

Several approaches have been demonstrated to translate target molecules into nucleic acid reporter molecules. The approach of probing nucleic acid targets by ligation, introduced 1988 with the oligonucleotide ligation assay (OLA), has become popular in different embodiments combined, for example, with selective PCR amplification of the ligation product\textsuperscript{76}. Examples of techniques using OLA in conjunction with PCR are multiplex ligation-dependent probe amplification (MLPA) using size encoded probes with followed by gel electrophoresis and the cDNA-mediated annealing, selection, extension and ligation (DASL) assay, which also includes a polymerase extension between the probe pair prior to ligation\textsuperscript{77,78}.

In 1994 OLA was successfully refined into the padlock probe assay, by joining the two oligonucleotide probes into one continuous DNA sequence\textsuperscript{79} (Figure 3). The association of the two separate OLA probes into a single molecule provides intramolecular hybridization kinetics upon target recognition. This enhances ligation efficiency and reduces the required probe amount. In addition to PCR amplification the circularized probes can be amplified by RCA, as illustrated in Figure 3\textsuperscript{52,55,80}. Target-dependent probe circularization furthermore enables selective degradation of remaining unreacted, linear probes using exonucleases\textsuperscript{33}. A popular version of the padlock probes, the molecular inversion probes (MIPs), uses a polymerase to perform a single nucleotide gap fill reaction between the probe arms before ligation. The method has been used to interrogate tens of thousands of SNPs in individual reactions using generic tag microarrays to record the outcome for individual loci\textsuperscript{81}. 
Figure 3. Nucleic acid analysis using padlock probes. Padlock circularization is mediated by target templated ligation. The padlock probe comprises two target complementary sequences (black), primer sites for amplification (blue) and a tag sequence for microarray sorting (yellow). **A.** Padlock probe amplification by PCR. Ligated padlock probes are selectively amplified by PCR and subsequently sorted onto a tag microarray. A fluorescence labeled PCR primer is used for detection. **B.** The circularized padlock probes are amplified by RCA with a single primer. Additional rounds of amplification can be introduced by C2CA. The amplification products are monomerized and hybridized to a tag microarray. The primer motif is used to hybridize a general fluorescent oligonucleotide.

Polymerase based tag microarray assays

Tag microarray readout has generally been applied to probes having recognized targets in solution by polymerase-mediated discrimination after amplification of individual loci by PCR, or in combination with ligation as described above. The two major polymerase based nucleic acid analysis techniques are allele specific extension (ASE) and minisequencing. ASE utilizes the capability of DNA polymerases to discriminate primers matched or mismatched at their 3’ ends to their targets, while minisequencing exploits
the ability of polymerases to introduce the correct base in any given position by extending a primer. The target complementary primers can be equipped with nucleic acid tags and following polymerization, typically with a labeled nucleotide mixture, the extended primers are hybridized to the tag microarrays.

Protein detection by nucleic acid reporter molecules

Compared to nucleic acids, proteins are much more complicated to analyze, being composed of 20 amino acids instead of four nucleotides and capable of undergoing far more complex covalent modifications. The classical approach to improve specificity of detection is to also involve a second antibody in the form of a sandwich assay, thus requiring the presence of two epitopes for detection\(^8^4\). However this approach does not scale well due to the rapidly increasing risks of cross reactivity of secondary binders as more pairs of antibodies are combined in the same reactions, and large scale affinity arrays typically depend on single binder analysis\(^8^5\).

In a different approach, photo aptamers increase specificity by requiring first binding between the target and the affinity probe, followed by the formation of a covalent cross link only upon correct binding\(^8^6\). Aptamers are single-stranded nucleic acid selected to exhibit affinity for particular target molecules. Photoaptamers additionally contain residues that allow them to be covalently cross-linked to their target molecules when correctly bound. The combination of binding and cross-linking provides a scalable intramolecular selectivity enhancement. Unspecific binding is unlikely to position the photo-sensitive residue to enable covalent cross-linking, thereby introducing an additional level of specificity. Following covalent cross-linking denaturing washing can be used to remove unspecific background. However, although theoretically very attractive, aptamers have not yet been established as a standard binder type in the proteomic field since their introduction 1990\(^3^4, 8^7\).

One frequently lamented drawback in analyses of rare proteins compared to nucleic acids is the absence of a protein amplification mechanism. The immuno-PCR was introduced as a means of amplifying the signal from antibodies having bound specific proteins via a nucleic acid molecule attached to the antibodies that can be amplified and detected by PCR\(^8^8\). In a similar fashion the immuno-RCA allowed conversion of a protein target into a locally amplified nucleic acid signal used for example in microarray readout\(^5^6, 8^9\). So far, the techniques have not been used to translate sets of proteins into unique nucleic acids sequences for subsequent tag microarray analysis in analogy with the nucleic acid detection.
Protein analysis by DNA ligation

Proximity ligation is an approach to protein detection based on concepts similar to those evolved for high throughput analysis of nucleic acids where enzymatic target discrimination is followed by selective signal amplification. Proximity ligation, outlined in Figure 4, uses two binders targeting the protein of interest, typically antibodies, equipped with an oligonucleotide each. Each of the two antibody-associated oligonucleotides encode one forward or reverse PCR primer site and one motif allowing enzymatic joining of the two oligonucleotides. Binding of the two antibodies to the target of interest brings the respective oligonucleotides in close proximity, and thereby enables joining by ligation. Enzymatic joining is achieved by hybridization of a third oligonucleotide, complementary to the antibody associated oligonucleotides. This forms a bridging hybridization which templates ligation of the oligonucleotides attached to the antibodies. Thereby a new nucleic acid is formed encompassing one forward and one reverse PCR primer site, allowing selective amplification of the ligated probes.

The co-localization criterion introduced by ligation followed by subsequent detection via quantitative real-time PCR (Q-PCR) permits protein detection in solution phase (Figure 4A). Except for the advantage of elimination of all washing steps, thereby reducing labor and providing a simplified assay, solution phase analysis provides further advantages. Similar to nucleic acid detection, analysis in solution phase increases the detection efficiency by enhancing probe-target binding kinetics and allowing detection of fewer target molecules. Furthermore, the amount of probes used can be significantly reduced and the requirement for antibody immobilization is eliminated, significantly reducing antibody consumption. The sample volume required for solution phase analysis by proximity ligation assay (PLA) can be reduced to one micro liter, a significant advantage for analysis of scarce samples like biobank material. The subsequent PCR amplification ensures detection of very low numbers of DNA ligations over background, typically by Q-PCR analysis. The background is set by the probe concentration, since unbound probes occasionally are in proximity. However, this background is predictable and can be adjusted via the probe concentration. Since the background decreases roughly as the square of the probe concentration, a window for protein detection can readily be achieved. Practically the assay setup is very similar to RNA analysis by reverse transcription and Q-PCR. In analogy with reverse transcription of RNA into cDNA for gene expression analyses, a proximity ligation step is used to convert proteins into ligated DNA targets followed by Q-PCR analysis.

Since the introduction of the sandwich assay for protein detection it has become widely accepted that the double specificity inherent to sandwich assays provides significant advantages for analysis of proteins in complex samples. This is in line with the enabling characteristic of PCR, where
specificity is increased by interrogation of nucleic acid samples via two primer motifs in the target sequence. However neither PCR nor sandwich immunoassays scale well when several analyses are performed in the same reaction due to increasing opportunities for pairs of reagents to bind incorrect molecules. For immunoassays the cross-reactivity of solution-phase binders to nonspecific targets gradually erodes the specificity. The cross-reactivity increases upon addition of new binder pairs, which gradually results in diminishing return in specificity approaching that of single-binder assays.

For the purpose of multiplex target analysis PLA introduces unique advantages compared to all other parallel protein analysis platforms since it is possible to constrain the detection reactions to only report reactions involving defined pairs of protein-binding reagents. By using antibody pairs conjugated to oligonucleotides designed to exclusively ligate as cognate pairs background by cross reactivity can be eliminated. This enables scalable addition of new binder pairs without increased background from non-cognate pairs of binders.

The opportunities for multiplex PLA have recently been illustrated by Fredriksson et al who simultaneously detected sets of six proteins using an optimized homogenous PLA. The assays allowed detection of as little as a thousand target proteins in microliter samples over a dynamic range of $10^5$. Following parallel proximity ligation the reactions were split into several reaction tubes where one Q-PCR analysis was performed per target protein. Specific analysis of designated binder pairs was achieved using primers that selectively amplified ligation products for individual proteins.

Proximity ligation can also be adapted for solid phase analysis by further involving an immobilized antibody, that allows detection of the captured target molecule by proximity ligation following a washing step (Figure 4B). This embodiment allows three target molecule epitopes on the detected target protein to be recognized simultaneously, one by the immobilized antibody and two via the added proximity probes.

In comparison with solution phase PLA solid phase PLA is less sensitive to proximity probe-generated background, since unbound probes can be readily eliminated by washes. Furthermore, the dynamic range in homogenous PLA is limited at the upper end by the proximity probe concentration. When the target concentration exceeds that of the proximity probes, then the probes are increasingly separated instead of co-localized as the target concentration is increased, giving rise to a so-called hook or prozone effect at higher concentrations of the analyzed proteins. By contrast, in solid phase assays detection signals reach a plateau at higher target concentrations. Solid phase analyses are also less sensitive to antibody-oligonucleotide conjugates that include unbound oligonucleotides since these can be eliminated by washing following probe incubation. Furthermore, antibodies having poor affinity impair performance of the solution phase assay more compared to
solid phase analysis since in solution-phase assays the background is set by the binder concentration independently of the binder affinity but the degree of target mediated co-localization of binders depends on their affinity. In solid phase assays it is possible to compensate for poor affinities by increasing the concentration of binders.

In comparison with standard solid phase sandwich immunoassays, background signals can be reduced significantly using solid phase PLA with one immobilized capture antibody and two oligonucleotide-conjugated probes added in solution. In contrast to sandwich immunoassays where both non-specifically adsorbed and target bound secondary probes produce signals, background does not arise from single matrix-adsorbed PLA probes. The solid phase assays further permit larger sample volumes to be interrogated to search for rare molecules, and it is possible to remove substances that could inhibit binding, the enzymatic reactions steps or optical detection.

PLA has been applied for analysis of individual or interacting proteins, for protein-DNA interactions, and for detection of single pathogens\textsuperscript{93, 94}. The assays have also been configured to reveal the presence of three epitopes on individual target proteins or protein complexes simultaneously in solution\textsuperscript{95} (Figure 4C). These analyses have been enabled by design of dedicated probes and corresponding oligonucleotides for each application, followed by readout by Q-PCR. A separate readout format has been applied for analysis of protein-protein interactions and post translational modifications in fixed cells\textsuperscript{96, 97} (Figure 4D). If the oligonucleotides are designed to give rise to circular DNA strands upon proximity-dependent ligation, then the reaction products can be locally amplified by RCA for visualization of individual and co-localized proteins \textit{in situ}, providing spatial target information in samples.
Multiple binder analysis

Since the introduction of the sandwich immunoassay it has become clear that analysis of multiple target epitopes can enhance immunoassays dramatically. The approach efficiently eliminates background signals from single binder
cross reactivity and thereby facilitates analysis of complex samples. Multiple affinity epitope targeting has also been successful using serial purifications for mass spectrometry with tandem affinity purification (MS-TAP) although in contrast to conventional immunoassays genetically introduced epitopes are exploited. Multiple epitope analysis also enables interaction analysis by targeting different proteins, or post translational modifications targeting e.g. phosphorylation sites.

Proximity ligation has been demonstrated using an immobilized antibody for antigen capture followed by detection of the immobilized antigen using proximity ligation, thereby interrogating three target epitopes simultaneously. The requirement for three epitopes for detection enables analysis of even more complex structures, introduction of even higher detection stringency or combinations thereof like sandwich detection using two binders in combination with a third binder targeting a post translational modification. Co-localization of complexes of three target proteins has also been demonstrated using in situ PLA.

Analysis of splicing patterns

Most human transcripts are estimated to be alternatively spliced and 15% of the single base pair mutations that cause human disease have been estimated to affect splicing. Splicing has also been put forward as a regulator of protein interactions. Splicing analysis is traditionally performed by capillary sequencing of gel electrophoresis and blotting. However, more recently several different microarray platforms have been introduced for high throughput analysis of alternative splicing. There are three major microarray approaches for analysis of alternative splicing, exon probes, exon junction probes and tiling arrays. Analysis using exon probes identifies differential exon expression, in turn indicating alternative splicing. Exon junction probes are designed to target known or predicted splice junctions, thereby reporting association of exons. This approach typically requires a priori knowledge of splice variants to reduce probe numbers. Tiling arrays comprising overlapping probes that cover genomic regions of interest can be used for both exon discovery and analysis of alternative splicing, similar to exon arrays. Splice junction arrays provides more information about the mRNA iso-forms present compared to exon arrays and tiling arrays, since information of which exons are joined is retrieved. However even splice junction arrays can not resolve mRNA iso-forms comprising two variant positions separated by homologous sequence common to both subtypes.

Splicing patterns are also mapped by expressed sequence tag (EST) sequencing of cDNA libraries, followed by mapping of the sequence on to the chromosomal DNA sequence. Analysis of gene expression and alternative splicing by sequencing will be further extended in combination with the
recently introduced sequencing platforms. Generation of a comprehensive map over alternative splicing is essential for probe design when using techniques based on oligonucleotide probes like splice junction microarrays or the splicing analysis approach presented in paper IV.

Interaction analysis

The desire to characterize the link between genotype and phenotype often prompts investigators to wander off into the maze of cellular interaction networks. With only a rough blueprint at hand this journey often reaches dead ends. Due to the chaotic appearance of the maps that have been established, they are popularly referred to as “hairballs” or “ridiculograms”. Maps of different interactomes are currently assembled by methods such as yeast two hybrid (Y2H) assays or tandem affinity purification coupled to mass spectrometry (TAP-MS). However, a complete interactome map is difficult to establish, due to the intrinsic difficulty of defining an interactome. Mark Vidal and colleagues define the interactome as a “complete collection of binary protein-protein interactions detectable in one or more exogenous assays”100. The definition recognizes the difficulty introduced by splice variation, which ultimately should be included, significantly increasing the complexity. However functional and dynamic effects are purposely ignored since the aim is to create an information scaffold of all potential interactions on which additional information can be superimposed101. In analogy, the human genome project did not involve analysis of e.g. genetic variation, DNA derivatizations like methylation, or functional characterization of sequence elements but merely provided a basis for further studies.

Yeast two hybrid

The Y2H assay is one of the major techniques used in genome wide approaches to the analysis of interactions. By fusing one protein with the DNA binding domain (forming a bait) and one with the activation domain of a transcription factor (forming a prey), a protein pair can be screened for interaction by mating yeast cells carrying the respective plasmids, thereby colocalizing the protein expression102. Upon protein interaction, reconstitution of the transcription factor activates a reporter gene, typically lacZ, allowing interaction to be scored using a β-galactosidase assay. Compared to mass spectrometry where proteins can be identified de novo, exhaustive screening by Y2H requires pair-wise mating of all ORFs. Two studies have been published so far aiming to map the human interactome, or at least significant parts thereof, by Y2H. Stelzl and colleagues used a matrix approach where ~5,500 single prey clones spotted in a 384 format were mated with ~5,500
baits in pools of 8, forming a total of ~400,000 yeast fusions interrogating over 25 million potential interactions\textsuperscript{103}. Scored interactions were subsequently verified by mating single yeast clones in a second run. Rual et al exhaustively screened 8300 ORFs, constituting a total space of ~70 million interactions, by investigating ~400,000 pools of single baits mated with 188 preys and subsequently verifying positive interactions by Sanger sequencing\textsuperscript{100}. Strategies using larger libraries enhance throughput but typically generate lower numbers of interaction partners per interrogated protein compared to when smaller pools are used\textsuperscript{100, 104, 105} presumably due to domination of certain interactions. This renders shotgun approaches difficult and screening of large libraries interrogating millions of potential interactions have resulted in recovery of the single interaction between α-globin and β-globin\textsuperscript{106}. Although the efforts to investigate the human interactome are quite ambitious, Rual et al estimates that the ~2,500 identified protein-protein interactions in their study only represents ~1% of the human interactome, leaving much of the road ahead unexplored.

Mass spectrometry
Mass spectrometry-based analysis of interactomes has been performed by investigating proteins fused to a tandem affinity purification tag that includes two distinct affinity tags. The two tags are used to serially capture affinity complexes, in order to eliminate concomitant purification artefacts. Using this approach protein complexes have been identified in genome wide screens by individually tagging bait proteins and subsequently affinity purifying interacting prey partners. Affinity-purified complexes are typically separated by gel electrophoresis before analysis by mass spectrometry to identify components of protein-complexes. Using TAP-MS, high-throughput analyses of the yeast interactome have been published, covering a majority of all protein-protein interactions\textsuperscript{107, 108}. Mass spectrometry avoids the need to individually interrogate all pair-wise interactions due to the possibility of de novo protein detection, however the ~2,000 yeast proteins investigated by Gavin et al required analysis of ~50,000 purified complexes identifying 2,760 distinct, interacting proteins\textsuperscript{108}.

Interaction in action
Interactions play a key role in human physiology and pathology, and it is therefore of utmost importance to measure such interactions in response to cellular reactions, and in health and disease. Genetic interaction analyses can provide further, more indirect information about interactions between gene product\textsuperscript{109}. Genome-wide association screens of patient cohorts and transcription profiling of pathologically relevant material may further assist studies of mechanisms involved in pathology. However, all these approaches:
Y2H, genetic interactions, genome wide screens and transcription profiling, merely implicate genes in cellular functions in an indirect manner. Ultimately, it will be important to detect functional interactions, not as potential affinities but actual complexes directly in pathological tissues. Genetically modified model systems can only provide a rough estimate of human molecular pathology, considering the complexity that results from mechanisms like tissue-specific expression and splicing of most genes, interaction regulation by site specific protein modifications, tissue heterogeneity and subcellular organization of proteins. It will therefore be essential to interrogate endogenous molecular interactions directly in genetically unmodified cells in order to characterize disease-specific interaction maps. Furthermore, the ability to simultaneously analyze many proteins, and their respective interactions, instead of performing parallel analysis of multiple pairs in different experiments, will provide information about temporal and spatial relations of molecular processes.
Present investigations

Paper I. A dual-tag microarray platform for high-performance nucleic acid and protein analyses

Even though microarrays are successfully used for analysis a wide variety of nucleic acid targets, ranging from SNP analysis and CGH to gene expression and ChIP-chip analyses, there is room for improvements. Gene transcript expression has been demonstrated to span over 5-6 orders of magnitude in homogenous cell populations, while microarrays typically only identify transcripts over a thousand fold concentration range. Furthermore, even though the understanding of hybridization is continuously expanding it is clear that cross-hybridization is difficult to predict and it is continuously reported in the microarray community, even for in silico designed tag microarrays. Finally, due to technical and biological factors, transcript expression and protein expression correlate only to a moderate degree and integrated analysis would therefore be very attractive.

Herein we present a dual tag microarray (DTM) platform, which essentially abolishes cross-hybridization while extending the microarray dynamic range towards five orders of magnitude. Upon combination of the DTM read-out and padlock probe based target analysis, ~100,000 fold lower target concentrations could be detected compared to direct target hybridization. Furthermore, DTM read-out of proximity ligation for the first time demonstrates detection of proteins using a DNA tag microarray, potentially also allowing combined microarray analysis of mRNA and protein levels in the future. Finally, we also demonstrated two additional useful properties resulting from on-chip RCA, namely on-chip real-time monitoring of RCA and digital quantification of single molecule RCA products for increased precision.

Our group has previously demonstrated the advantages of RCA for padlock probe analyses by C2CA resulting in reduced amplification bias, improved LOD, product yield, and enhanced precision. In C2CA RCA products are monomerized using an oligonucleotide to direct restriction digestion and thereafter circularization by ligation and priming of a new generation of RCA. The current protocol comprises a further development of this procedure, which allows ligation of monomers of RCA products to specific tag microarray sequences, thereby practically eliminating the risk of cross-hybridization between features. Furthermore the on-chip circularization by
ligation enables on-chip RCA initiated from the 3’ ends of tag oligonucleotides for a further signal amplification. Circularization to specific microarray tags is achieved by using the type IIS restriction enzyme MlyI that cleaves the target uni-directionally, 5 bp away from the recognition site. This allows the RCA product to be cleaved next to a tag sequence, allowing formation of short reporter molecules having a tag sequence in the 3’ and 5’ ends. Compared to standard hybridization of fluorescent nucleic acids, no signal is generated upon non-specific adsorption or cross-hybridization when ligation is used for reporter molecule localization, as illustrated in Figure 5. RCA enhanced LOD and the dynamic range, and the specificity introduced by on-chip ligation ensures that microarray cross-hybridization does not drown weak signals.

Figure 5. Schematic diagram of potential sources of background. A. Correct hybridization localizes fluorophores to the reporting feature (i). Background may arise upon cross-hybridization (ii) or non-specific adsorption (iii). B. Ligation to the correct microarray probe forms a circular DNA substrate required for signal generation by RCA (i). Cross-hybridization (ii) or non-specific adsorption (iii) does not generate circular DNA polymerization substrates and therefore does not contribute to background.

Perspectives on paper I

Microarrays are used quite successfully for analysis of gene regulation, and recent reports validating microarray platforms by Q-PCR are encouraging. However, there is need for improvements of e.g. false negative rates, and the dynamic range obtained in molecular analysis by microarrays is typically 3 logs even though RNA expression in homogenous cell populations has been shown to span over six orders of magnitude. Specificity problems resulting from cross-hybridization that amount to less than 1-10% are often considered insignificant, although this level may well contribute to the frequently reported reduced correlation between weak microarray signals and Q-PCR measurements of transcript levels.

There is no doubt that the transcriptome represents massive amounts of information and the current analysis platforms typically overwhelm users with information, an effect of the microarray platform scalability.
quently, a high density microarray experiment leaves few asking for more data, but important information may remain buried at lower concentration levels. Finally, even if it would be possible to correctly assess gene expression, RNA levels typically correlate only to a moderate extent with protein expression patterns, due both to biological and technical factors, further emphasizing the importance of correct measurements to minimize validation efforts

The DTM presented in this paper provides a potential solution to several important issues of microarray analysis by enhancing LOD, dynamic range and specificity. One unique feature of the DTM is the possibility to interrogate reporter molecules for the presence of two distinct tag motifs. This will enable combinatorial analysis of tag pair combinations. Finally, protein detection by tag microarray readout was demonstrated herein for the first time, an application that is further expanded in paper II.
Paper II. Parallel protein analysis by proximity ligation with DNA microarray read-out.

This investigation reports ongoing work to establish multiplex protein analysis by proximity ligation with dual-tag microarray readout. The proximity ligation assays were performed either in liquid phase samples or in fixed cells.

Proximity ligation has previously been demonstrated for simultaneous analysis of several proteins in individual reactions with excellent performance. However, the Q-PCR assay readout format used in this study is not suitable for large numbers of analyses since at least one Q-PCR reaction must be performed per target analyte. In analogy with high throughput analysis of gene expression by mRNA detection, where microarray analysis provides larger-scale analyses than Q-PCR, microarray analysis could also prove of great value for multiplex protein analysis by proximity ligation, provided that adequate dynamic range and detection specificity can be maintained.

One unique feature of the dual tag microarray platform presented in this thesis is the possibility to interrogate nucleic acid reporter molecules for the presence of two distinct nucleic acid tag motifs. In combination with proximity ligation this offers important advantages in multiplex analyses. Proximity ligation utilizes two oligonucleotide-coupled protein binders to interrogate each target protein. By encoding one unique tag sequence motif in each of the two antibody-oligonucleotide conjugates, protein detection can give rise to a nucleic acid reporter molecule comprising a unique combination of tag motifs. Consequently, upon analysis of multiple analytes each target is represented by a distinct pair of nucleic acid tags. Following PCR amplification, where the PCR products represent the individual proteins as a pair of tag sequences, each tag pair was interrogated using the dual tag microarray. One important advantage of this approach, where specific tag pairs are interrogated in each microarray feature, is the ability to distinguish correct signals from cross-reactive events involving binder pairs specific for different proteins as outlined in Figure 6.

Multiplex protein analysis of liquid samples was set up using capture antibodies immobilized on distinct beads, enabling facile combination of bead types for analysis of several targets. As a start, three separate oligonucleotide systems were designed and benchmarked with one common analyte, and they were subsequently used to interrogate three separate cytokines. Protein detection by PLA was demonstrated over a dynamic range of $10^5$-$10^7$ and with a LOD in the sub pM range as measured by Q-PCR. Microarray readout of the PLA maintained the LOD although the dynamic range was reduced to $\sim10^4$.

Microarray readout of proximity ligation performed in fixed cells was demonstrated by comparing the results from analyses of cells transfected
with the Smad4 gene with those from untransfected *wt* cells. Additionally, Smad4 was visualized by microscopy using both *in situ* PLA and *in situ* immunofluorescence. Microarray readout was performed using a two-color detection scheme, very similar to those often used in standard gene expression microarrays. Using the two color setup the protein level from Smad4 transfected cells was represented on the microarray by Cy5 and the *wt* protein level was represented by Cy3. Beta actin was interrogated in parallel in all reactions and used for normalization. The results were confirmed by performing a dye switch experiment between *wt* and transfected cells.

The demonstrated results are promising for scaling PLA to large numbers of parallel analyses with microarray readout. The two color assay demonstrated herein by Smad4 is also advantageous for protein detection in liquid-phase samples to preserve the dynamic range of the underlying PLA. In this manner the relative proportion of target proteins in the two samples is reflected by the ratio of the two fluorophores over a wide dynamic range.

**Perspectives on paper II**

Compared to nucleic acid analysis protein analysis imposes even more stringent requirements on analysis platforms due to wider range of target concentrations and increased molecular complexity. The protein analysis platform presented herein enables scalable protein detection using a sandwich immunoassays involving binding of individual target proteins by two or even three affinity reagents. By exploiting PLA proteins can be detected over a wide dynamic range with excellent specificity and LOD. Sandwich immunoassays are typically not multiplexed beyond 20-40 antibody pairs. Further scaling of antibody mediated detection can be achieved using arrays of single binders for each protein 40, 56, 85. Compared to single binder assays PLA requires two separate affinity reagents providing enhanced specificity by dual epitope recognition. However, as demonstrated herein, this can be accommodated using a single batch of polyclonal antibodies which is split and conjugated to two separate oligonucleotides. This obviates the need for matched pairs of antibodies, but still enables sandwich immunoassay specificity.

It is a high priority to now expand the assay to interrogating interactions between investigated target proteins. This will be achieved by adding microarray features that can detect all nucleic acid tag pairs that represent all possible combinations of protein pairs as illustrated in *Figure 6D*. Classical microarray analysis of protein interactions outlined in *Figure 6C* use immobilized proteins to investigate interactions upon addition of a single labeled protein. In the proposed strategy however, all interactions among sets of proteins can be investigated in parallel. Since target labeling is not required the analysis could allow interaction screening of clinical material. This setup will allow analysis of protein-protein interactions by microarray analysis in a
fashion very similar to that of gene expression analysis by microarrays. Thereby different samples can be readily investigated for altered patterns of interactions. Any identified changes can then be validated by Q-PCR readout of PLA or by detecting protein interactions by *in situ* PLA, as demonstrated herein\(^9\).

**Figure 6.** Cross reactivity in multiplex sandwich immunoassays and proximity ligation. **A.** The possible sources of cross-reactive events of a multiplexed sandwich immunoassay using labeled secondary antibodies are identified by thin lines. **B.** Layout of multiplex proximity ligation with dual tag microarray readout. Only correct tag pair combinations are detected in each microarray feature, thereby eliminating signals from cross reactive events between antibody pairs. **C.** Conventional microarray-based interaction screening. A single protein is labeled and interrogated for affinity to members on a protein microarray. **D.** Protein-protein interactions are encoded as pairs of DNA tags that can be analyzed by constructing a microarray interrogating all possible DNA tag pairs. Each microarray feature represents one specific interaction between two proteins.
Paper III. Sensitive protein detection via triple-binder proximity ligation assay.

The analysis of proteins by affinity probes was revolutionized by the introduction of the sandwich assay allowing two epitopes to be interrogated simultaneously. Proximity ligation allows protein analysis by sandwich assay specificity in solution phase without any washing steps. Herein we present an approach for analysis of three epitopes simultaneously using the solution phase proximity ligation assay. The prerequisite of three recognition events further decreases non-specific background, allowing very low target protein LOD.

In contrast to standard proximity ligation, where a PCR template is formed upon co-localization of two binders, the current protocol is designed to only form amplifiable nucleic acid strains if three affinity probes are being co-localized by the target. In standard PLA two oligonucleotides are brought together and ligated using a complementary non-conjugated oligonucleotide. The current embodiment however, uses a third affinity probe-conjugated oligonucleotide that is complementary to the two other oligonucleotides allowing them to be ligated. Thereby all three binders are required to bind the target to generate signal. The background is reduced further by addition of blocking oligonucleotides that hybridize competitively with the third affinity probe conjugated oligonucleotide. This reduces background due to random co-localization and ligation events. However, upon target mediated co-localization of binders the local concentration of the conjugates exceeds that of the blocking oligonucleotides and the blocking is reversed.

Using the described protocol, detection of less then 100 recombinant vascular endothelial growth factor (VEGF) molecules could be demonstrated. Furthermore, to illustrate the flexibility of the assay design the complex formed between prostate specific antigen (PSA) and the carrier protein antichymotrypsin (ACT) was detected using separate binders for PSA and ACT, respectively. The application for biomarker detection is also exemplified by analysis of the myocardial infarction biomarker troponin I in both heparin plasma and citrate plasma.

Perspectives on paper III

The ability to analyze three epitopes simultaneously is a unique feature of proximity ligation. The current design encodes sequence motifs from three separate binders into one single nucleic acid molecule synthesized by ligation upon detection. The approach could potentially be used to decode combinatorial interactions involving three participants and to investigate how proteins enter and exit protein complexes to elucidate macro molecular dynamics. In addition to analysis of heterogeneous molecule complexes homomeric polymers can be efficiently differentiated from monomers using one
single monoclonal antibody to detect several identical epitopes. This application could be valuable in, for example, analysis of prion protein complexes or discrimination of Amyloid beta monomers from polymers appearing as a result of Alzheimer’s disease.
Most human transcripts are expressed in multiple splicing isoforms and 10-30% are differentially expressed among tissues. If alternative splicing occurs at two or more distinct sites of a transcript in parallel, analysis of individual transcript compositions using e.g. microarrays is difficult in heterogeneous cell populations. One classical example of a transcript exhibiting this type of combinatorial splicing is the *Drosophila melanogaster* transcript Dscam, displaying remarkable transcript diversity. The Dscam gene has the potential to generate up to 38,000 transcript isoforms by combining two alternatively spliced positions of mutually exclusive exons.

The current project presents an alternative approach for parallel analysis of alternative splicing. The technique allows association of alternatively spliced sites on the same transcript molecule. The approach is named spliceotyping, in analogy with haplotyping, where genetic variation located on the same chromosomal DNA molecule is identified. In spliceotyping approach the exon compositions of interrogated transcripts are translated into nucleic acid reporter molecules. Each transcript is converted into a chain of generic nucleic acid tag sequences that represent the exon composition of the original transcript. This is achieved using probes complementary to the exon boundaries combined by ligation to reflect the splicing pattern. The 5’- and 3’- most probes are equipped with PCR primer sites, allowing amplification of the probe library following ligation. Each DNA molecule in the probe library will after ligation comprise a set of generic exon tags that represent the splicing pattern and one transcript. To enable identification of individual transcript patterns the amplified library is circularized and amplified by RCA, forming a long concatemeric RCA product. The RCA product, representing the exon composition of a single transcript, can be readily interrogated by sequential hybridization of fluorescent probes that target the generic tag sequences, thereby decoding transcript identity.

We used the beta actin transcript as a reference gene and the adenovirus gene E1A as a model system for splicing. Probes targeting the five exons and single intron of E1A were designed and converted into a library, amplified and decoded successfully. Probe stability and temperature dependence were optimized and a ligation efficiency of ~70% was achieved for each probe junction. Decoding was performed by hybridization of three tag positions in each sequential round of hybridization visualized by three separate fluorophores. Upon analysis of the adenovirus transcript E1A, all expected isoforms were identified. However, the absolute exon composition was biased towards short isoforms, most likely due to unbalanced efficiencies during ligation and PCR.
Perspectives on paper IV

Combinatorial splicing is impossible to analyze using microarrays or next generation sequencing platforms and no truly parallel platform is available for analysis of transcript compositions. Previously published approaches for single-molecule-based splicing analysis are based on specific PCR amplification of transcripts and only analyze one transcript per analysis\textsuperscript{119}. In contrast, probe ligation approaches have previously enabled analysis of thousands of transcripts in parallel\textsuperscript{120}. By using nucleic acid probes, the presented platform can be further applied to facilitate analysis of alternative promotor and poly adenylation sites.
Future perspectives

Considering the central dogma of molecular genetics, the modest correlation between level of transcripts and proteins is perhaps unexpected and stresses the importance of undertaking both protein and nucleic acid analyses. The techniques presented herein could be used to interrogate both nucleic acids and protein components of the same sample, either in parallel but possibly also by serial analysis of the same sample. This can be done by encoding the protein content as tag DNA-sequences before sample denaturation, followed by padlock probe ligation and readout on microarray to decode both proteins and nucleic acids.

One potential synergy of the work presented in paper I and IV is the combination of microarray-based analysis of single RCA products and of splicing pattern analysis by spliceotyping. The potential to enumerate reporter molecules with single molecule precision in separate array features and in the same process decoding the individual exon composition of the targets is attractive. Similar to standard microarrays, the microarray position would then reveal the gene identity and the transcript splicing pattern could be sequentially decoded using generic exon probes.

Furthermore, alternative splicing primarily appears at the protein surface and may be a potent regulator of protein interactions by differential inclusion of protein-protein affinity domains. The techniques presented in paper II and paper IV could be used to combine analyses of protein interaction patterns and alternative splicing.

The probing problem

For nucleic acid probe libraries the situation is bright compared to protein affinity probes. Highly parallel in situ synthesis of oligonucleotide libraries enables high throughput production of microarrays for solid phase assays. Recently successful assembly of probe libraries have also been demonstrated for use in solution phase. This could be useful for construction of large probe sets for both padlock probing and spliceotyping analyses.

The concept of microarray analyses has proven much more difficult to implement for proteins than for nucleic acids. Compared to nucleic acid applications the feature density of protein microarrays lags three orders of magnitude behind, and so far the technology has had much less impact in the research community compared to nucleic acid microarrays. This is a result of
the increased difficulty of producing protein-binding reagents and also applying them in microarray formats, further complicated by the increased complexity of the proteome compared to that of the genome and transcriptome. Community efforts to assemble open reading frame library resources and produce protein products of the so called ORFeome will benefit the whole proteome field by providing antigens for affinity binder selections, microarray probes, yeast two hybrid (Y2H) libraries and tandem affinity pullout mass spectrometry (TAP-MS) approaches. Furthermore, efforts have been initiated to generate and quality-control large sets of affinity binders. However, significant further investments will be required to generate a renewable proteome resource, including affinity reagents and targets, and to make this available to the scientific community.

To enable parallel analysis of proteins using PLA, sets of protein binders coupled to DNA are required. Once a clonal repertoire of affinity reagents is available, synthesis of probes for PLA could be achieved by conjugating each affinity reagent to the respective DNA strand. However, potentially this can be performed in parallel, thereby avoiding individual conjugation of each binder to DNA. For the purpose of affinity binder selection several methods have been developed that enable association of protein libraries with their encoding DNA upon expression. This serves to maintain the link between genotype and phenotype during the selection procedure. These methods have been used for construction of DNA associated protein libraries using DNA libraries and *in vitro* expression. In this manner a DNA library encoding recombinant binders, could be expressed and associated to DNA. The DNA associated to the protein binders could then be used in PLA for ligation, amplification and identification instead of synthetic oligonucleotides.

On high-throughput sequencing and microarray analyses

During the work on this thesis the high-throughput sequencing platforms have evolved from startup ventures to fully developed instruments that are currently entering the market. Very high expectations have been placed on the new sequencing platforms for e.g. routine sequencing of human genomes. Projects have been launched targeting both technologies for the $100,000 and $1000 genome along with the “1000 genomes project”, aiming to completely sequence the genomes of 1000 individuals.

In the context of the technology developments presented in the thesis, sequencing can also be used for highly specific and precise measurement of DNA reporter molecules. Identification of the target molecule sequence provides a very high level of specificity, and sequencing also increase precision by single molecule enumeration. Utilization of paired-reads also enables analysis of distal sequences, such as in our designs for proximity ligation. Furthermore, tag identification by sequencing also allows for shorter tags to
be used, and a tag complexity of six nucleotides is sufficient to encode thousands of tag identities.

Although promising, sequencing is still associated with some drawbacks. The current sequencing platforms provide unbiased analyses of collections of nucleic acids, which may be a feasible strategy for analysis of nucleic acid populations where all sequences are present at roughly the same levels, as in genome sequencing. However, for sequencing of nucleic acid populations with unbalanced target levels abundant targets will be sequenced repeatedly. Identification of rare transcripts at increased dynamic range requires a considerable sequencing depth. In order to analyze 10,000 transcripts with a dynamic range of 1000, in the order of 10,000,000 reads may be required depending on the average transcript frequency. For microarray analyses abundant transcripts saturate the microarray features but do not affect the analysis of remaining transcript, if potential cross-hybridizations are neglected.

Future combination of spatial microarray localization and on-chip sequencing would ensure specificity and avoid excessive resequencing of abundant targets. Possibly these efforts are not too distant since microarray platforms are available that can sequence a single base\textsuperscript{129}. However, so far high throughput re-sequencing remains an expensive solution for expression analyses or enumeration of tag sequences, and microarray solutions can be expected to remain competitive for years to come.

The future of protein microarray analysis

Protein microarrays are expected to pick up the race with DNA microarrays for analyses of gene expression\textsuperscript{130}. Currently microarrays are sold that encompass thousands of proteins\textsuperscript{131}. However it is clear that for protein analyses the dual-binder assay design is an essential key for high-performance protein detection in clinical samples. Although protein microarrays have been scaled for analysis of thousands of proteins, analysis of target proteins by sandwich dual-binder assays has not been successfully scaled past a few tens of proteins. One of the largest published sandwich detection assays demonstrated by Schweitzer et al. interrogated in the order of 80 proteins, but the antibody pairs were split into two separate reactions to minimize cross reactivity. The combination of PLA and DTM holds promise to allow parallel protein analyses by combination of individually optimized immunoassays without reduced performance. The approach also avoids the challenge of protein microarray manufacture. This would enable scaling sandwich assays to the same level as single binder assays and allow high-throughput analysis of unlabeled samples from biobanks and clinical material.
Protein networks as potential biomarkers

So far, it has been difficult to identify individual biomarkers with sufficient prognostic power to predict clinical outcome. To overcome this problem, profiles derived from multiple biomarkers are being investigated, thereby collectively increasing the predictive power\cite{85}. It has recently been shown that cancer genes have a higher average number of protein-protein interactions compared to non-disease genes and that they are more likely to participate in interaction network hubs compared to other genes\cite{101}. It is also likely that many of the post translational modifications and splice variants involved in disease will affect protein-protein interaction networks, as reviewed in \cite{132,133}. These network modulations do not necessarily affect protein abundance and may remain undetected with traditional biomarker analysis. The combination of DTM and PLA presented herein could be used to acquire snapshots of interaction networks for use as biomarkers. Potentially these networks may reveal a new dimension of predictive clinical information.
Acknowledgements

This work has been performed at the department of genetics and pathology at the Rudbeck laboratory. This work has been supported by the EU framework program 6 project Moltools, the Wallenberg Consortium North for Functional Genomics and the Swedish Research Council. I would like to thank everyone who has contributed to the successful assembly of this thesis both by direct and indirect support.

I would like to express my gratitude to:

**Ulf Landegren**, who has been my supervisor and mentor during my PhD and corrected this thesis during Christmas. Thanks for your confidence in me and for providing me with resources and possibilities to pursue my own projects. I could not have received a better scientific scholarship.

My co-supervisor **Mats Nilsson**, totally unrivaled in his extraordinary talent to see patterns in chaotic data sets, and always has a valuable comment on everything.

People at the department; especially **Ulla Steimer** who always enthusiastically have time for my annoying questions along with **Pirkko, Lena, Frida, Elisabeth, BM, Gunilla Åberg, Tommy, Eddie, Helena Jernberg Wiklund, Ulf Gyllensten** and everyone else who makes the department spin.

My office mates **Magnus Isaksson** who introduced me to pink cocktails, **Jenny** the “Blobologist” **Göransson** for heavy metal sessions in the lab, **Spyros “Malaka” Darmanis** and **Malin Jarvius**.

My main co-authors; **Rachel Nong** for continuously injecting a lot of crazy jumping energy into the project and **Jonas Jarvius** for introducing me to the lab, showing me how to draw smileys with the $4 million confocal microscope and taking care of my bike, which still does
not have any brakes btw. Tim Conze for staying organized when I can’t. Edith Shallmeiner for interesting discussions and keeping the office door closed.

Henrik Johansson for introducing me to the intriguing life cycle of the kebab animal and Johan Stenberg for strict moral & linguistic guidance. All the in situ people, Ola, Sara, Catharina, Kalle, Agatha, Ida, Irene (party Piff & Puff…). Lena & Carla for keeping the lab running. Anders for maintaining a happy (and organized) atmosphere and Masood for adventures among aptamers and Mojitos. Katerina Pardali for bringing some Mediterranean action into the project and for enthusiastic application writing night-shifts together with Mathias Howell. Yuki, Jonas Melin, Reza, Carolina W, Ji-juan, Mia, Maritha, Johan Olerich and all other current and former members of the group for maintaining a great place to work at.

Sigrun, Fredrik Dahl and Johan Banér who remembers when we still were allowed to have group meetings on Finland cruises… Marek, for interesting discussions about science and other things, and for not killing Andrey during their vivid discussions.

People at Olink, Björn, Mats, Göran, Simon and everyone else for fun interactions. Andrea Reyes for punishing me in the running track.

My present and former students; Samer Yammine, Erika Groth, Daniel Seisdedos, 3 x Sara, David Eklund, Maria Hammond, Hamidreza Razzaghian and Johan Vänelid for enthusiastically trying to overcome my pedagogic attempts.

My friends; Marcus Svedberg who showed me how to steal a room using a piano, Karin Hellström, Aaron “sleeping skavfötters“ Maltais, Anki Roos, Mattias Bergström for dating exhibitionists and various adventures with Gerda, Anders Eriksson for standing up for Anneberg at Stadskällaren, no matter what!, Helena Jacob, Jonatan Dahlkvist for living in my apartment, Ellinor Åslin Hägg, Calle Wolf for sprints in Cadiz, Emelie Torehall for showing the boys at Stallet how to be men, Tomas Adolfsson, Marigo Oulis, Peter Wallenberg for adventures in Africa, Jesper Göthe for fun parties and the new carpet, Tobias Bengtzén and Christoffer Eriksson. Without you all this thesis work would have been finished so much faster, thanks for all the fun.
Old, current and new members of the Wednesday beer club including **Håkan Zeffer, Martin Karlsson, Pär Matsson, Per Engström, Clas Linnman** among others, along with all the fun people at my second home Stallet.

The “Marielundsgatan crew”; **Jonatan Jacobsson, Jessica Berg, Lisa Johansson, Henry Virtanen & Co** for diverse adventures, often together with a couple of fisherman’s friends...

Those I forgot here, thanks and please do not be offended; I am an absent-minded scientist after all.

My family, for everything.
References


84. Wide, L. & Juhlin, L. Detection of penicillin allergy of the immediate type by radioimmunoassay of reagins (IgE) to penicilloyl conjugates. *Clin Allergy* 1, 171-177 (1971).


Acta Universitatis Upsaliensis

Digital Comprehensive Summaries of Uppsala Dissertations from the Faculty of Medicine 315

Editor: The Dean of the Faculty of Medicine

A doctoral dissertation from the Faculty of Medicine, Uppsala University, is usually a summary of a number of papers. A few copies of the complete dissertation are kept at major Swedish research libraries, while the summary alone is distributed internationally through the series Digital Comprehensive Summaries of Uppsala Dissertations from the Faculty of Medicine. (Prior to January, 2005, the series was published under the title “Comprehensive Summaries of Uppsala Dissertations from the Faculty of Medicine”.)

Distribution: publications.uu.se
urn:nbn:se:uu:diva-8475

ACTA UNIVERSITATIS UPSALIENSIS
UPPSALA
2008