Dissertation

Serum auto-antibody testing for early diagnosis of breast cancer

ausgeführt zum Zwecke der Erlangung des akademischen Grades eines
Doktors der Naturwissenschaften unter der Leitung von
Univ.Prof. DI. Dr.techn. Peter Christian Kubicek

eingereicht an der Technischen Universität Wien
bei der Fakultät für Technische Chemie

From
Parvez Syed
Matrikelnummer: 0948411
Goldschlagstraße 21/2
1150 Vienna

For my family

Acknowledgements

I thank Univ.Prof. DI. Dr.techn. Peter Christian Kubicek from the Vienna University of Technology and DI. Priv.-Doz. Dr. Andreas Weinhäusel from the AIT Austrian Institute of Technology GmbH for supervising my thesis.

I also would like to thank Dr. Ulrike Fischer, Dr. Petra Leidinger and Prof. Eckart Meese from Saarland University, Germany for introducing me to the topic of SEREX.

I am thankful to Dr. Albert Kriegner and DI. Klemens Vierlinger for their constant support and helping me in analyzing the complex microarray data. I am grateful to Univ. Prof. Dr. Christian F. Singer and Mag. Christine Rappaport-Fürhauser from the Medical University of Vienna for providing me the tissue and serum samples. Special thanks to all the tissue and blood donors. I express my gratitude to Ronald Kulovics and Silvia Schönthaler from AIT for printing the protein microarrays and to Dr. Johannes Zerweck from JPT Peptide Technologies for providing peptide microarrays. I thank all my colleagues at AIT for creating a great working atmosphere.

I thank my family for the constant love and support.

Contents

Abstract

The aim of this thesis is generate prototype-tests suitable for randomized prospective validation of auto-antibody based diagnostic testing using serum samples. Tumours can stimulate the production of auto-antibodies against autologous cellular proteins known as TAAs (tumour associated antigens). This discovery has lead to a possibility of using the auto-antibodies as serological tools for the early diagnosis and management of breast cancer.

The recombinant proteins expressed by the SEREX clones, identified from screenings of brain and lung tumour, were used for the production of the protein microarrays and macroarrays. The protein microarrays showed better correlation between the replicates of the serum samples used. The optimized protocols were used for the subsequent experiments. A sizable panel of 642 clone-proteins was selected by marker-screening on protein macroarrays with 38000 clones. These 642 clone-proteins were used to generate protein microarrays that differentiated serum samples from breast cancer patients and controls. Antigenic peptide motifs were identified by *in-silico* analysis of 642 clone-proteins and peptide arrays were generated using synthetically generated peptides. Comparative studies between protein microarrays and peptide microarrays were done using breast cancer and healthy control samples.

Simultaneously, SEREX strategy was used for the identification of the immunogenic TAAs. I identified 192 cDNA expression clones derived from breast cancer tissue samples and the selection was done using breast cancer sera. The genes corresponding to these clones were found over-represented for the pathways that are known to be associated with cancers. These genes showed typical features of TAAs, like over-expression, mutations and fusion genes.

General Introduction

Submitted in part in *J Proteomics*. Luna-Coronell J*, **Syed P*** et al. **The current status of cancer biomarker research using tumor-associated antigens for minimal invasive and early cancer diagnostics.** Under review Feb. 2012

*Equal contributions

Background

In 2008 alone an estimated 12.7 million new cancer cases and 7.6 million cancer related deaths were reported worldwide. The most commonly reported cases were lung cancer (1.61 million, 12.7% of the total), breast cancer (1.38 million, 10.9% of the total) and colorectal cancer (1.23 million, 9.7% of the total). Among the most common causes of cancer deaths, lung cancer is the leading form of cancer (1.38 million, 18.2% of the total) followed by stomach (738,000 deaths, 9.7% of the total) and liver cancers (696,000 deaths, 9.2% of the total) [1-3]. The estimated age-standardized worldwide incidence rate of the main four types of cancers, breast, prostate, lung and colorectal cancer are shown in the figure 1 [1]. Breast cancer is the leading tumour type in women in sense of occurrence with an estimated 1 million new cases each year [4,5]. It is well accepted that early diagnosis can improve survival, thus, there is great need and anticipation to identify novel biomarkers for cancer diagnostics at the earliest stage as possible. Auto-antibodies are well known for their pathological role in autoimmune diseases, such as rheumatoid arthritis or lupus erythematosus. Mutated- or aberrantly expressed proteins, acting as antigens, evoke immune response resulting in the production of auto-antibodies. The onset and progression of various cancer types often associated with the appearance of these proteins with unnatural conformations or epitopes. Since auto-antibodies could be detected months or years before the clinical diagnosis of cancer [6-8], tumour-associated antigens (TAAs) and their corresponding auto-antibodies could be used as biomarkers for early diagnosis and prognosis of cancer [9-11]. These tumour-induced antibodies might also give an insight about the host-tumour interactions and the dynamics of carcinogenesis since most of the auto-antibodies are immunological fingerprints of pathological processes [10,12,13].

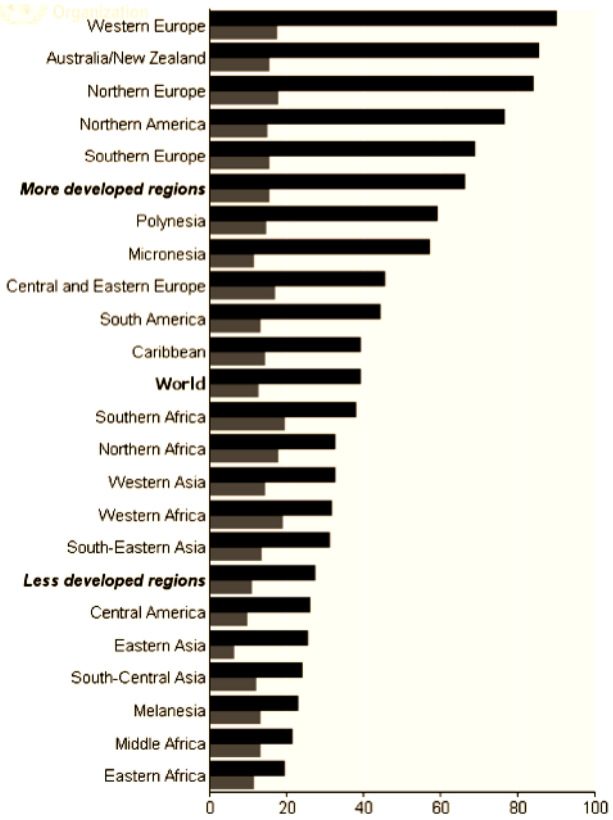

Figure 1: Incidence and mortality rates of breast cancer cases in 2008. The image shows the estimated numbers (thousands) of incidences (black) and mortalities (grey) observed in various parts of the world (Data from GLOBOCAN, IARC, Lyon, France 2010) [14].

Standard diagnostic methods of Breast cancer

Breast cancer diagnosis is proved to improve women's chances for survival and could provide 100% cure if the diagnosis of the breast cancer is done while the cancer is still confined to the breast. The current screening techniques which widely employed for the breast cancer diagnosis are mammography, ultrasound (US) and magnetic resonance imaging (MRI). Although mammography is one of the most common techniques for breast cancer screening, it is associated with low sensitivity problem [15,16], which varies from about 68 to 93%. This may be attributed to variations in both practitioners' skill and experience as well as patient characteristics. This sensitivity is up to 50% less in women who are on hormone replacement therapy, young, Asian, and/or have dense breasts. Additionally, it is also less sensitive for the detection of invasive lobular carcinomas (ILC) and small or diffuse tumours [17]. X-rays are used in mammography, which may lead to radiation-induced mutations. As the image quality from the mammography depends on the

breast structure, this technique may not be suitable for the screening of young women with family history of breast cancer and women with dense breast tissue [18].

Novel protein based analyses

The development of proteomic technologies provides an unprecedented ability to identify novel bio-signatures to diagnose, classify and guide therapeutic decision making in patients with cancer. The novel biomarker candidates and biomarker signatures, which are ready for the use in clinical settings, have proven to require detection steps similarly complex as in their discovery. This complicates their widespread usage in the screening of large populations despite recent improvements in separation techniques based on High-performance liquid chromatography (HPLC) separation, mass spectrometry (MS) and 2D electrophoresis (2DE) [19-24]. Identification of marker molecules that can be targeted in specialized assays using proteomics based techniques relies on antibody or aptamer technology. The development of specific capturing agents for the candidate markers requires expensive production process and thorough validation to ensure high avidity for the target while minimizing the risk of unspecific binding [25-27].

Tumour-associated antigens

It has been found that tumours can stimulate the production of auto-antibodies against autologous cellular proteins known as tumour-associated antigens (TAAs). This discovery has lead to a possibility of using the auto-antibodies as serological tools for the early diagnosis and management of cancer [10]. Tumour-associated antigens have been described in reference to several cancer types, such as lung-, liver-, breast-, prostate-, ovarian-, renal-, head and neck cancer, oesophageal cancer, lymphoma and leukaemia [28-38]. The tumour suppressor gene, p53, is one of the most frequently mutated genes with regards to human tumours and the gene product of it results that this protein could act as TAA [39-41]. Anti-p53 antibodies are known to be associated with many cancer types and were observed prior to the onset of breast -, lung- and prostate cancer [6-8].

The immunogenicity of TAAs is conferred to mutated amino acid sequences, which then exposes an altered non-self epitope [42]. Other explanations are also implicated of this immunogenicity, including alternative splicing, expression of embryonic proteins in

adulthood, deregulation of apoptotic or necrotic processes, abnormal cellular localizations (*e.g.* nuclear proteins being secreted) [43]. There have been found a few examples when epitopes of the tumour-restricted antigens that were encoded by intron sequences (*i.e.* partially unspliced RNA were translated) making the TAA highly immunogenic [44]. The meta-analysis of Backes and colleagues [43] revealed that the protein motif with strongest immunogenic potential were Zinc-finger DNA-binding domains. Post-translational modifications, like proteolytic cleavage, phosphorylation or glycosylation, may also attribute to the immunogenicity of TAAs [45-49].

In the diagnostic point of view, a single auto antigen lacks the adequate sensitivity and specificity which can be overcome by using a panel of TAAs where multiple auto-antibodies being detected simultaneously [10,16,50-53]. For example, antibodies against SOX families B1 and B2 are found in the patients with small cell lung cancer. However, the presence of anti-SOX antibodies in benign diseases leads to an insufficient specificity [54-56]. Koziol *et al*.[57] employed a panel of 7 TAAs namely c-myc, p53, cyclin B, p62, Koc, IMP1 and survivin, which could discriminate between healthy individuals and patients with breast-, colon-, gastric-, liver-, lung- or prostate cancers, with sensitivities within a range of 77 to 92% and specificities ranging from 85 to 91%. An increased antibody response was found to the panel of p53, HER2, MUC1, TOPO2α, IGFBP2, CCND1, and CTSD in breast cancer patients with specificity and sensitivity up to 75% [52]. Antibodies specific to the TAA panel of PIM1, MAPKAPK3, and ACVR2B showed 73.9% and 83.3% of specificity and sensitivity for diagnosing colon cancer [58]. A list of TAA panels used for diagnosis of various cancer types can be found in the table 1.

Table 1. TAA panels used for diagnosis of breast-, prostate-, lung- and colon cancer

TAAs	Cancer type	Sensitivity (%)	Specificity (%)	Reference
SOX1 SOX2 SOX3 SOX21	Small cell lung cancer	67	95	[59,60]
CAGE GBU4–5 NY-ESO-1 MUC1	Small cell and Non-small cell lung cancer	64-92	92-100	[61]
IMPDH PGAM1 UBQN ANXA1 ANXA2 HSP70-9B	Non-small cell lung cancer	94.8	91.1	[62]
BRD2 eIF4G1 RPL22 RPL13a XP_373908	Prostate cancer	81.6	88.2	[63]
Imp1 p62 Koc p53 c-myc	Colon cancer	60.9	89.7	[64]
MAPKAPK3 ACVR2B	Colorectal cancer	83.3	73.9	[65]
ASB-9 SERAC1 RELT	Breast cancer	80	100	[15,53]
p16 p53 c-myc	Breast carcinoma	43.9	97.6	[15]
PPIA PRDX2 FKBP52 MUC1 HSP60	Breast cancer	73	85	[66]

Identification of TAAs

Tumour specific antigens are of high importance with regards to diagnosis since many of them are restrictedly produced against specific tumour; hence act as ideal biomarkers [47]. However, some auto-antibody species are rather generic than specific thus appropriate selection methods and stringent statistical criteria must be applied. The most commonly

employed methods for the discovery of TAAs are protein macro- and microarrays, peptide microarrays, SEREX, phage display and SERPA (Fig. 2).

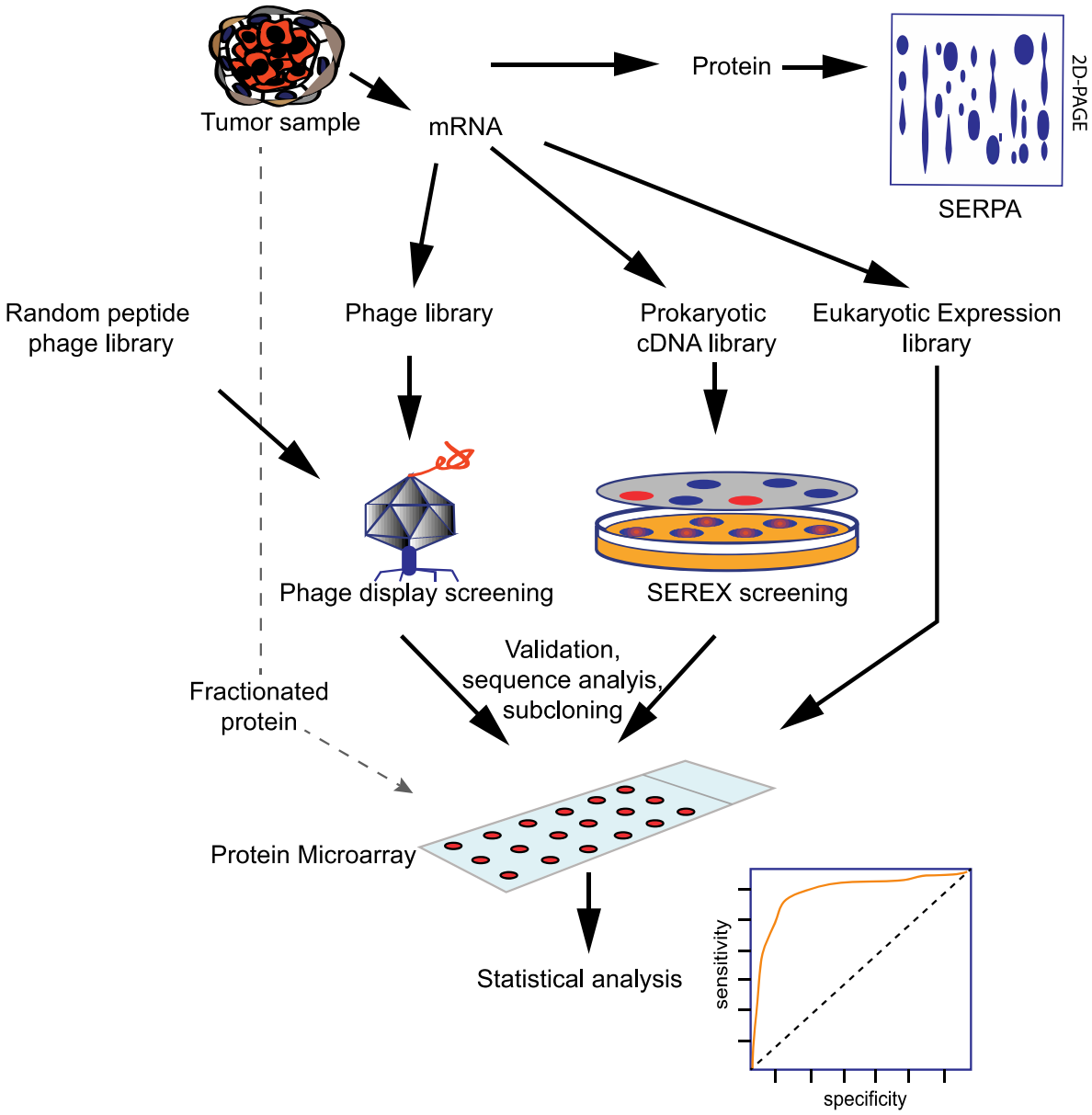

Figure 2: Schematic summary of the methods used for tumour-associated antigen discovery.

Protein macroarrays

Protein macroarrays are the low-density protein arrays which play significant role in identifying auto-antibodies in a various autoimmune diseases and cancers [67]. Protein macroarrays are polyvinylidene fluoride (PVDF) membranes consisting of cDNA expression E. coli clones. Recombinant proteins are expressed on these membranes which are released upon lysis of the cells. These recombinant proteins are recognised by antibodies in the

serum. Protein macroarrays have been used for the identification of highly specific antibody–antigen interactions [68]. Protein macroarrays possess the ability to screen an immune response to a large number of proteins, thus, providing improved diagnosis and identification of auto-antibody signatures that may represent disease subgroups [69]. Protein macroarrays immobilized with disease-associated antigenic proteins obtained from body fluids are used for the detection of circulating antibody repertoires from the patients with autoimmune diseases like chronic obstructive pulmonary disease (COPD) and cancers like ovarian and lung cancer [67,70-73]. The detection of the antibodies reacting to the antigens immobilized on the membranes is done by secondary antibody, anti-human IgG/M/A conjugated with horse radish peroxidase. The images of the processed membranes can be captured by a scanner like Storm 860 (Amersham/GE) or FujiFLA3000 (Raytest).

Protein microarrays

Protein microarrays offer a potent tool to study interactions between proteins and peptides on a large scale. It is the advanced, high-throughput form of immunoassays such as Radioimmunoassay (RIA) described by Yalow *et al* [74] and the enzyme-linked immunosorbant assay (ELISA) by Engvall *et al* [75]. Initially, Ekins *et al* [76] proposed that a-few-micron scale solid support antibody assays could yield high sensitivity comparable to that of macroarrays. Advances in studies on DNA microarray technology, such as spotting machinery, support surface chemistries, detection methods, data analysis and principles, have been successfully applied to protein microarray technology. One of the earliest examples of implementation of DNA microarray experience to protein arrays is the use of standard inkjet printer to apply 200µm diameter monoclonal antibody spots onto a polystyrene film. Silzel *et al.* [77] were able to reduce reagents, capture antibody, analyse molecule per zone with equal sensitivity and specificity as that of ELISA.

At present, protein microarrays denote recombinant-, fractions- or purified proteins. Proteins of interest are immobilized onto a microscope slide coated with various surfaces in a planar or 3D platform or captured by affinity beads in a micro-well plate setting. Planar surface coating chemistries are categorized into 4 main groups based on the binding principle: non-specific covalent and non-covalent, specific covalent and non-covalent. The first group includes nitrocellulose and poly(L-lysine), the second, aldehyde and epoxy, the

third, certain affinity interaction molecules like streptavidin–biotin, His-tag–nickel-chelates, and the last self-assembly monolayers (SAMs) on gold coated surfaces [78]. Coating chemistries are vital in the overall performance of the protein microarray and should offer low background, maintain proper orientation and reduce the effect on the sensitive 3-D structure of the proteins The most commonly used surface chemistries in antibody microarrays are nitrocellulose, amines, aldehyde or epoxy. However, there has been only limited number of studies comparing surface chemistries published so far [78-83]. Due to the highly complex nature of protein interactions, new technologies to improve performance of protein microarrays are developed, like the so-called 3D surface setups based on: agarose hydrogel, hydrogel and more recently nanoparticles and beads in micro-well plate formats. These approaches aim to overcome the difficulty of limited spot density, detecting low abundance proteins, increase signal to noise ratio, avoid non-specific binding and cross-reactivity [83-90]. Furthermore, for TAA profiling or detection of TAA diagnostic signatures, spotted protein arrays are hybridized with cancer patients' and/or control individuals' serum samples and detected using methods that avoid cross-reactivity, enabling high-resolution signals in high throughput format with good reproducibility and are cost effective [78]. The same strategies as in immunoassays (sandwich-, antigen capture- and direct), do also apply to protein microarrays [91-96].

Different approaches to produce the protein analytes for tumour biomarker discovery and validation exist, ranging from immobilization of recombinant proteins to protein tumour tissues fractions. TAA biomarker discovery initially begins with screening, subsequently narrowing down to sets of proteins that best distinguish the tumour type to be diagnosed. Several studies have demonstrated the use of different liquid chromatography fractioning of proteins from tumour [97,98] or tumour-derived cell lines [80,99-101]. Using such native proteins for functional analysis have the advantage of having accurate post-transcriptional modifications but has drawback with the likelihood of manifold proteins present in spots, as well as to depend on the tissue sample [52].

Recombinant proteins from cDNA libraries can be expressed via bacterial, insect or mammalian expression systems avoiding the limitations. An example of proteins produced by insect expression system is the commercial product Human ProtoArray manufactured by Invitrogen. In this system proteins are expressed as N-terminal GST (Glutathione-S-

Transferase)-fusion protein using a baculovirus expression system. The over expressed proteins are then purified from insect cells, and printed in duplicate on a modified glass slide. Compared to bacterial expression systems, protein folding in insect cell expressed proteins is improved [102,103] and is similar to that of mammalian cells. Although mammalian cells are the most suitable expression system for human proteins, in regard to correct folding, the technique lacks in ability to scale up, give good yield and is time and cost effective [104].

Several cancer types have been addressed for biomarker discovery using protein microarray approach. For instance in colon cancer, Nam *et al* [80] identified an antigenic target by using a protein microarray containing 1760 solubilised protein fractions obtained from human colon adenocarcinoma cell line (LoVo). Proteins fractions were immobilized onto nitrocellulose-coated slides and hybridized with 15 plasma samples each from colon and lung cancers and healthy controls. From the total of 1760 fraction, 39 exhibited higher reactivity to colon cancer samples. One fraction was reactive to 9 out of 15 colon cancer sera, which was identified by mass spectrometry as ubiquitin C-terminal hydrolase isozyme 3 (UCH-L3). Antibody to UCH-L3 could be detected in 19/43 sera from patients with colon cancer, but none of the 54 sera of lung cancer, colon adenoma or healthy subjects.

By using commercial protein microarrays containing 8000 human proteins, Babel *et al*. [58] tested sera from colorectal cancer (CRC) patients and healthy subjects. They were able to identify 43 proteins that were recognized by tumoural sera but not the control ones. From these 43 different proteins, 5 immunoreactive antigens, PIM1, MAPKAPK3, STK4, SRC, and FGFR4 were shown to be highly prevalent in cancer samples. By using an ELISA with PIM1, MAPKAPK3, and ACVR2B they showed specificity and sensitivity values of 73.9 and 83.3% (area under the curve, 0.85), respectively. From a 37,830 clone recombinant human protein array, Kijanka *et al* [69] could identify 22 antigens that were able to distinguish between 43 colorectal cancer patients and 40 non cancer patients. Expression of the antigens (p53, high mobility group B1 (HMGB1), TCF3, tripartite motif-containing 28 (TRIM28), longevity assurance gene homologue 5 (LASS5) and zinc finger protein 346 (ZNF346)) were also assessed using quantitative reverse transcription PCR (Q-PCR). Results showed that these proteins are involved various cellular processes.

Prostate cancer is one of the most extensively screened cancer potential TAA markers. In one instance, a panel of 12 proteins, including several heat shock proteins, Prostate Specific Antigen (PSA) and alpha-methylacyl-CoA racemase (AMACR), immobilized on a nitrocellulose coated microarray slides were analysed using 48 prostate cancer patient and 28 control subject serum samples. This study by Sreekumar *et al* [105] aimed at elucidating the performance of AMACR in assisting PSA as a diagnostic biomarker for prostate cancer. The AMACR protein showed significantly higher immune-reactivity in patient sera than controls. By immunoblot and ELISA tests, a sensitivity and specificity of 78.1% and 73.1% respectively were found. With the attempt to screen for classifier proteins between prostate cancer and non-cancer samples, Taylor *et al* [106] tested protein microarrays comprising of 2300 protein fractions, obtained by liquid-phase protein fractionation of localized and metastatic prostate cancer tissue lysates and tested their performance with 18 biopsy positive prostate cancer and 16 neoplasm negative prostatic hyperplasia patient sera. Twenty fractions were top predictors having 75% specificity and 78% sensitivity (with 4 samples misclassified) in discriminating between cancer and hyperplasia. Mass spectrometry of the constituents of these 20 fractions revealed 359 unique proteins and 29 of these associated to the humoral immunity. Massoner *et al*. [107] generated microarrays for prostate cancer specific biomarkers by spotting antigenic proteins on nitrocellulose-coated slides, after identifying these antigens on expression clone macroarrays containing more than 37,000 recombinant human proteins. A panel of 15 TAA was found to discriminate between 40 prostate cancer patients and 40 benign disease patients with an ROC curve AUC of 0.71.

In a recent study concerning Breast cancer protein biomarkers, Anderson *et al*. [108] presented a novel protein microarray approach termed Nucleic Acid Programmable Protein Array (NAPPA), containing 4988 candidate tumour antigens. Performance was tested using 53 early stage breast cancer samples, 1,102 benign breast disease samples (group 2 and 3), 140 pre-treatment patient in stage I-III breast cancer and 64 healthy control subject serum samples (group 4), depending on the disease state of patients. After 3 rounds of detection of antibody in samples, and eliminating un-informative clones in each step, finally deduced 28 antigens with 80.8% sensitivity and 61.6% specificity

A protein microarray is a microscopic slide with immobilised proteins and the immobilisation is done using contact or non-contact spotter. The contact spotter (Fig. 3) uses hollow pins for

the deposition of the protein solution on the microarray slides. On the other hand, the non-contact spotter (Fig. 4) uses the same principle as inkjet printing technology. Droplets of the protein solution are fired onto the slides. The spotting is done in a humidity-controlled environment. A variety of slide surfaces, like amines, aldehyde and epoxy are used to prepare a protein microarray. The coating of the slide surface enables the proteins to adsorb covalently to the slide surface [45]. The advantage of protein microarray over macroarray is the miniaturization of the assay which largely reduces the amount of the sample and reagents to be used [52]. DNA microarrays do not provide information about the nature of the protein they are coding for. Protein microarray helps in overcoming this problem. Clinical diagnostics has become an easy task with the advent of protein microarrays. Protein microarrays immobilized with antigens capture and quantify specific antibodies and thus have the capability of serving as a platform for diagnosis [11]. Apart from clinical diagnostics, protein microarrays are extensively used to study biochemical activities of proteins like protein-protein interactions and protein-phospholipid interactions.

Figure 3: Print head of contact spotter. Capillary pins arranged in the allotted slots are used for the deposition of protein solution onto the microarrays. *Image from Arrayit corporation.*

Figure 4: Print head for non-contact spotter. Small droplets of protein solution are fired onto the microarray slide surface. *Image from Eisenstein* [109].

There are various kinds of protein microarrays basing on the function and the immobilized molecule onto the microarray surface. Analytical protein microarrays help in measuring the binding affinities, specificities and expression levels of a protein. It enables the profiling of a complex protein mixture. Protein solution is used as probe on a glass slide arrayed with a library of antibodies [45]. This technique helps in not only diagnosing certain diseases but also monitoring protein expression levels and protein profiling [110]. Another type of protein microarray is the functional protein microarray. It enables the studying of the biochemical activity of the whole proteome in a single experiment. Reverse phase protein microarray is yet another type of protein microarray where the lysates from various tissues are arrayed onto a nitrocellulose slide which is probed with labelled antibodies specific for target proteins. Using reverse phase protein microarray one could detect the altered proteins with different post translational modifications when compared to that of normal proteins [45]. Analysis of multiple proteins and non-soluble proteins are few of the advantages of reverse phase protein microarray [111]. However, if in an experiment, antibodies are immobilized onto the substratum and used as bait molecules then this kind of microarray could be called as forward phase protein microarray [112]. There are few setbacks for forward phase protein microarray. Detection using protein tagged with fluorescent dyes was performed which has been a common practice. If these tags are placed within the epitope regions, they might sterically interfere with the antibody binding [113].

There are mainly two types of assays designed for the protein microarrays. They are label-based assay (direct labelled) and sandwich assay (indirect labelled). Label-based assays involve flourophores such as Cy3 and Cy5. These flourophores are tagged to a protein (antibody or protein antigen) which is targeted onto to the immobilized protein (antibody or protein antigen). Co-incubation of two different samples is possible. Being a competitive assay, this assay has advantages over non-competitive with reference to linearity of response and dynamic range. However, limited sensitivity and specificity are the noticeable drawbacks of this assay. Sandwich assay is non-competitive assay as co-incubation of different samples is not possible. In sandwich assay, an immobilised protein (antibody or protein antigen) captures another protein (antibody or protein antigen) which is unlabelled. This unlabelled protein (antibody or protein antigen) is targeted by a labelled antibody. Sandwich assays have higher sensitivity and specificity when compared to that of competitive assays [11,45]. The fluorescence intensity values from the processed arrays are captured using scanner and are used for the statistical evaluation.

Peptide microarrays

Like protein arrays, peptide arrays enable the screening of large number of binding events in a high throughput way. The information deduced from the peptide microarrays can contribute to the development of substrates and inhibitors which can be used for diagnostic or therapeutic purposes. While the protein arrays give insight to the protein-protein interaction peptide arrays provide the information on characterization of molecular recognition events at the amino acid level. Peptide arrays enable antibody epitope and paratope mapping, protein–protein interaction mapping in general [114,115]. Epitopes are smaller compared to the full proteins and represent the immunogenic portions of the protein. Peptide arrays generated with such epitopes can serve as durable and cheap platform for diagnostic purposes and can be used as alternative to protein microarrays [116]. Apart from the ability to mimic the biological activities of proteins, synthetic peptides are stable and are rather easy to synthesize and manipulate [114,116]. Just like protein macroarrays, peptide macroarrays have low density of 20 spots per cm^2. On the hand, the protein microarrays can accommodate 200 spots per cm^2 [115].

The concept of peptide microarray was first conceived in 1984 by Geysen *et al*. [117] for mapping the B-cell epitopes by immobilizing different peptides using solid-phase peptide synthesis. Later on, a technique called SPOT-synthesis, developed by Frank [118], made possible for the synthesis of large number of peptides directly on cellulose membranes. The same principle has been applied for the generation of peptide microarrays. Glass slides coated with amino, epoxide, aldehyde and sulfhydryl groups are used for the immobilization of the peptides onto the glass surface [115]. The immobilization is done using a non-contact printer. Peptide microarrays have been used for the identification of auto-antibodies against autoimmune diseases and cancers [119,120]. The processing of the peptide microarray is rather similar to that of protein microarray. There are not so many studies published describing the usage of peptide microarrays in the field of cancer diagnosis. This gives one an opportunity to explore and develop a platform for early diagnosis of cancer using peptide microarrays.

SEREX

The mostly used technique, which allows exploring tumour's antigen repertoire, is SEREX (Serological Analysis of Recombinant cDNA Expression Libraries). The SEREX method was originally developed by Michael Pfreundschuh and his colleagues in 1995 [121]. The first step of this technique is the construction of lambda phage cDNA library from fresh tumour samples. Using primary tumour samples instead of cancer cell lines also helps to eliminate the *in vitro* artefacts associated (e.g. altered ploidity and DNA methylation status) with the numerous passaging of the cells upon a DNA expression library generation. Subsequent steps are include plating phages, transfer onto nitrocellulose membrane and induction of protein expression. In the screening step membranes must be blocked with phage- and bacterial lysates to reduce reactivity against host antigens. Immune-reactive clones with antibodies from autologous- or allogeneic sera of cancer patients can be identified with a secondary antibody-coupled enzymatic (colorimetric) reaction. Identified clones are sequenced and can be validated with an independent method, such as ELISA or protein macro- or microarray. Though this method is quite laborious, more than 2700 immunogenic antigens were identified so far as listed in the Cancer Immunome Database developed by the Ludwig Institute of Cancer Research (www.Ludwig-sun5.unil.ch/CancerImmunomeDB).

SEREX defined antigens can be classified as Cancer-testis antigens, differentiation antigens, mutated genes derived antigens, chimeric proteins as a result of chromosomal translocations, products of over-expressed genes, spliced variants' products and products having a retroviral origin [122,123]. Cancer-testis (CT) antigens have salient features, which make these antigens an interesting group of immunogenic proteins. The CT antigens show tissue-specific expression in normal testis and embryonic ovary and highly expressed in various human tumours. Increased expression levels of CT antigens in other than tumour tissues can be a good indication of malignant transformation. More than 40 CT antigens or antigen families have been discovered using the well-established SEREX method. CT antigens can be ideal for the development of cancer vaccines as they have high immunogenicity and restricted expression [122]. Differentiation antigens are those, which are showing lineage-specific expression pattern in tumours. These antigens are also expressed in normal cells that have the same origin, like melanocyte-specific tyrosinase and glial fibrillary acidic protein GFAP. Both of these proteins are found to be antigenic in malignant melanoma and glioma besides their expressions in melanocytes and brain cells, respectively [121,124]. Identification of over-expressed HER-2/neu by SEREX approach demonstrates the immunogenicity of over expressed gene products [125]. Another example which can be cited under this category is HOM-RCC-3.1.3 found in renal cancer cells [126].

Mutated gene products have also been identified by SEREX and these are frequently relevant in cancer. The tumour suppressor protein p53 isolated from a case of colon cancer is an example for this class of SEREX-defined antigens [127]. Hodgkin's disease-associated splice variant of restin (CLIP1) and gastric cancer-associated splice variants of TACC1 serve as example for splice variants of known genes which become immunogenic in cancer patients [121,128].

Several SEREX studies have been conducted for the identification of TAAs against sera from breast-, prostate-, colon- and lung cancers [33,127-136]. Jäger and colleagues [130] analysed a breast cancer library and a normal testicular library using autologous and allogeneic breast cancer sera and identified 3 novel antigens, two CT antigens namely NY-ESO-1 and SSX2 and a candidate breast cancer suppressor gene, ING1. In another study, Jäger *et al*. [129] isolated p33ING1, which is an immunogenic breast cancer antigen and is encoded by a putative tumour suppressor gene. During a SEREX study performed on prostate cancer, Fossa and

colleagues [131] isolated 3 CT antigens, namely NY-ESO-1, LAGE-1, and XAGE-1. Zhou *et al*. [137] identified an antigen PARIS-1, reacting to the sera from prostate cancer patients, which might play a role in regulation of cell differentiation and growth. Scanlan *et al*. [127] identified 48 antigens (NY-CO-1 to NY-CO-48), which were found reactive to the sera from colon cancer patients. Of the 48 antigens, one of them (NY-CO-13) was found to be a mutated form of the p53 tumour suppressor gene. Three out of 48 antigens were found to have a differential mRNA expression pattern. One such antigen is NY-CO-27 (galectin-4), which is usually expressed in gastrointestinal tract and the other two (NY-CO-37 and NY-CO-38) exhibited a pattern of tissue-specific isoforms. In another study, anti-CDX2 serum antibodies against CDX2 with a frame-shift mutation in the repetitive G sequences (microsatellite) in the coding region were identified by Ishikawa and colleagues [134]. Line *et al*. [128] isolated 8 antigens reacting to the serum from colon cancer patients. Of these 8 antigens, RHAMM and AD034 have a differential tissue distribution. AD034, which carries a frame-shift insertion, along with NAP1L1 and RHAMM showed over expression in tumours when compared to the neighbouring non-cancerous tissue. Güre *et al*. [135] identified Aldolase 1 (ALDOA) in a SEREX screen, which gene is known to be over expressed in lung cancer [138,139]. In another study, Güre *et al*. [136] isolated SOX1, SOX2, SOX3, SOX21 and ZIC2 which elicited immune response in sera from small cell lung cancer (SCLC) patients.

Although SEREX has been widely employed for the identification of TAAs, there are few practical problems associated with SEREX. An inherent bias can be observed in the selection of highly expressed antigens in the particular tumour that was used for the construction of cDNA library, as these antigens are overrepresented in the library It is also possible that majority of those auto-antibodies detected by SEREX might not have anything to do with cancer. It is therefore crucial to investigate the functionality of each of the found genes thoroughly [123].

Phage display

Phage display was established back in 1985 by Smith [140] as a method that displays polypeptides on the surfaces of phage particles. The most frequently employed phage strains are usually the members of Ff filamentous phage family, such as M13, f1, Fd and ft. Alternatively, lytic phages like T7 are also used for phage display [141]. Practically,

polypeptides from cells lines or tumour tissues are expressed as fusion to the phage coat proteins and hence exposed or displayed on their external surface [63]. Alternatively, random peptide libraries are utilized to map antigenicity of large number of epitopes. During this method, phages are immobilized and a process of affinity purification is used to identify and isolate phages carrying peptides with high affinity and specificity to a given target molecule like IgG. Unbound phage particles are washed away while the bound phages are eluted, and this process is repeated several times (biopanning rounds). In this way, the library is enriched for phage-peptides that bind to IgGs specifically associated with the tested cancer serum samples and not those found in control serum [142]. The resulting phages clones are sequenced for identification [120] and then arrayed onto membranes or glass slides were screening/validation takes place.

Phage display has been used for the identification of cancer-associated antibodies. Panels of TAA markers for colon-, prostate-, ovarian- and non-small cell lung cancer (NSCLC) have been found by phage display technology [63,71,143-145]. For instance, Ran *et al* [146] built a phage cDNA expression library of colon cancer, based on a modified SEREX screening. They achieved with a training set combination of 6 markers using the logistic regression model leave-one-out validation, 91.7% sensitivity and 91.7% specificity. They also found that the 6 antigen sequences in their phage display system are relatively short peptides, with only 2 of them showing homology to known protein sequences.

With a phage-display library derived from prostate-cancer tissue, Wang *et al*. [63] developed and used phage protein microarrays. . Using a 22 phage-display derived peptide detector, obtained 88.2 percent specificity (95 percent confidence interval, 0.78 to 0.95) and 81.6 percent sensitivity (95 percent confidence interval, 0.70 to 0.90) in discriminating between patients with prostate cancer and the control group.

Another study exploiting a breast cancer cDNA T7 phage library for the identification of tumour-associated proteins using biopanning enrichment yielded 100 putative tumour-associated phage clones. Sequencing of clones revealed that 6 phage proteins were in-frame and unique, and phage protein ELISAs were developed to measure the reactivity of these proteins. ELISA results demonstrated that 3 of the phage clones had proved statistical significance in discriminating patients from healthy individuals, and Basic Local Alignment

Search Tool (BLAST) search gave high similarity scores to proteins like ankyrin repeat and SOCS box-containing 9 (ASB-9), serine active site containing 1 (SERAC1), and RELT tumour necrosis factor receptor . Measurements of the 3 predictive phage proteins were combined in a leave-one-out validation achieved 77.0% sensitivity and 82.8% specificity [53].

Similar work was performed by Zhong *et al*. [145], starting from a T7-phage NSCLC (non-small lung cancer) cDNA library. The screening procedure was performed with patient plasma to identify phage-expressed NSCLC derived proteins recognized by TAAs. Altogether 212 immunoreactive phage-expressed proteins were selected and spotted in duplicates onto FAST slides for generation of targeted protein microarray. Combining the measurement of the 5 most predictive TAA, a logistic regression model yielded an area under the receiver operating characteristics curve (ROC) of 0.99, while leave-one-out validation achieved 91.3% sensitivity and 91.3% specificity respectively. Also working with NSCLC, Wu and collaborators achieved with a logistic regression model (leave one-out cross validation) that the sensitivity and specificity of a panel of a 6 phage peptide clone detector were 92.2% and 92.2%, respectably [147].

Although panels of TAAs have been found using this technique, the technique does have some limitations. For example, a limitation associated with phage display technique includes the need to sequence each and every immunoreactive phage clone. Additionally, only proteins that can be displayed on the surface of the phages are expressed, which lack mammalian post-translational modifications and may not be in the native configuration [9].

SERPA

Another proteomic based approach which allows the exploitation of the B-cell repertoire of patients with cancer is the serological proteomics analysis (SERPA), also known as Proteomex [148] or serological and proteomic evaluation of antibody response (SPEAR) [149]. This technique takes advantage of the classical 2-D PAGE. A further step involves a MALDI–TOF MS, which analyses and identifies the nature and abundance of total proteins in tissues [150]. This method was developed by Klade *et al.* [151], where they compared the proteomes of non-tumourous kidney and renal cell carcinoma (RCC). Using this technique, they were able to identify two antigens in kidney cancer patients, SM22-alpha and CAI.

The 2-D electrophoresis comprise separation of the proteins in the gel by their isoelectric point (IP) then a further separation based on the molecular mass [152]. Usually three gels are run simultaneously. Two of the gels are electro-blotted onto membranes and then probed with patient or control samples. The third gel is stained with Coomassie. Comparison between the spots obtained in the patient blot against the control blot is performed. The differentially expressed protein spots are identified on the blot and cut out from the Coomassie blue stained gel. Proteins are then extracted from the gel and identified by MALDI–TOF MS [150].

In 2003, Unwin *et al.* [149] used this method to screen RCC patients for naturally occurring antitumour antibody responses and thus, identified six immunogenic candidates, namely ANXA1, ANXA4, TIMP, CA1, SOD2 and MVP in the patients with high-grade disease. A related work was performed with the aim to identify tumour antigens that commonly induce a humoral immune response in patients with infiltrating ductal breast carcinomas [153]. Sera from 40 patients with invasive breast cancer and 42 healthy controls were screened individually for the presence of IgG antibodies to MCF-7 cell line proteins. Immunoreactive proteins were isolated and subsequently identified by MALDI-TOF mass spectrometry. A total of 26 antigens were identified, from which a significantly higher frequency was observed against the molecular chaperone HSP60, beta-tubulin, HPR and PRDX2. Additionally, PHB was also included, which is an oestrogen-regulated gene and frequently mutated in sporadic breast cancer [154]. Using SERPA, Suzuki *et al.* [155] were able to identify five proteins, namely EEF2, ENO1, ALDOA, GAPDH and HNRNP from the patients with melanoma. Tumour proteins that elicit humoral response in colorectal cancer (CRC) were also identified by SERPA. In a study, the protein source for the 2-DE and subsequent Western blot analysis came from the CRC cell line HCT116. An auto-antibody against HSP60 identified by MS was detected in patients with CRC and from one control individual. Results showed that both the expressions of HSP60 in tumour tissue and serum antibody titre to HSP60 were significantly higher in patients with CRC than in healthy controls [156].

One of the advantages of the SERPA method, in comparison to the SEREX method is the reduced time required to complete an experiment. The construction of a representative cDNA library in phage requires numerous days of work, while using SERPA, proteins can be prepared from tumour cryosections in matter of hours [152]. Also, 2DE is a very good

method to achieve separation of a complex mixture of proteins. Additionally, the proteins used for the 2-D gel contain their post-translational modifications intact, allowing a comprehensive view of the antibody-TAA interaction [66].

On the other hand, drawbacks of this technique are mainly due to the use of 2-DE. Excising proteins from the 2-DE gels can be a cumbersome task, and the reproducibility of the gel is quite weak [113]. Another disadvantage is that proteins which have the same molecular weight but different post translational modifications may not be able to separate adequately, which has consequences on the quantification of the visualized spot [157].

Objectives

There is great anticipation to identify novel biomarkers for further diagnostic improvements. The development of proteomic technologies provides an unprecedented ability to identify novel bio-signatures to diagnose, classify and guide therapeutic decision making in patients with cancer. The aim of this thesis is to develop a platform for the early diagnosis of breast cancer using protein biomarkers. Proteomic approaches like protein and peptide arrays and SEREX were used for the identification of potential protein biomarkers. Protein arrays help in the profiling of the autoantigens presented by the cDNA expression clones by treating the arrays with patient sera. Antigenic peptides sequences were identified within the proteins expressed by the cDNA expression clones. These antigenic peptide sequences were synthesized and immobilized onto glass surfaces and treated with patient sera. A comparison is done between the performance of protein and peptide arrays. This comparison helped in drawing the conclusion as to which platform suits better for the autoantigen profiling. On the other hand, SEREX technique was used to deduce set of sero-reactive cDNA expression clones derived from the cDNA library made from the RNA of breast cancer tissue samples. These clones were selected on the basis of immunoscreening done with autologous serum sample. To find if the genes expressed by the clones cluster into certain biological categories, gene set enrichment analysis was performed. This analysis helped in deducing the information on the categories in which the sets of genes/proteins cluster and their relevance with cancer.

Reference list

1. Ferlay J, Shin HR, Bray F, Forman D, Mathers C, Parkin DM: **Estimates of worldwide burden of cancer in 2008: GLOBOCAN 2008.** *Int J Cancer* 2010, **127:** 2893-2917.

2. Pecorelli S, Favalli G, Zigliani L, Odicino F: **Cancer in women.** *Int J Gynaecol Obstet* 2003, **82:** 369-379.

3. Ferlay J, Parkin DM, Steliarova-Foucher E: **Estimates of cancer incidence and mortality in Europe in 2008.** *Eur J Cancer* 2010, **46:** 765-781.

4. Pisani P, Bray F, Parkin DM: **Estimates of the world-wide prevalence of cancer for 25 sites in the adult population.** *Int J Cancer* 2002, **97:** 72-81.

5. Sturgeon CM, Duffy MJ, Stenman UH, Lilja H, Brunner N, Chan DW *et al*.: **National Academy of Clinical Biochemistry laboratory medicine practice guidelines for use of tumor markers in testicular, prostate, colorectal, breast, and ovarian cancers.** *Clin Chem* 2008, **54:** e11-e79.

6. Lubin R, Zalcman G, Bouchet L, Tredanel J, Legros Y, Cazals D *et al*.: **Serum p53 antibodies as early markers of lung cancer.** *Nat Med* 1995, **1:** 701-702.

7. Trivers GE, De B, V, Cawley HL, Caron G, Harrington AM, Bennett WP *et al*.: **Anti-p53 antibodies in sera from patients with chronic obstructive pulmonary disease can predate a diagnosis of cancer.** *Clin Cancer Res* 1996, **2:** 1767-1775.

8. Abendstein B, Marth C, Muller-Holzner E, Widschwendter M, Daxenbichler G, Zeimet AG: **Clinical significance of serum and ascitic p53 autoantibodies in epithelial ovarian carcinoma.** *Cancer* 2000, **88:** 1432-1437.

9. Anderson KS, LaBaer J: **The sentinel within: exploiting the immune system for cancer biomarkers.** *J Proteome Res* 2005, **4:** 1123-1133.

10. Casiano CA, Mediavilla-Varela M, Tan EM: **Tumor-associated antigen arrays for the serological diagnosis of cancer.** *Mol Cell Proteomics* 2006, **5:** 1745-1759.

11. Sanchez-Carbayo M: **Antibody arrays: technical considerations and clinical applications in cancer.** *Clin Chem* 2006, **52:** 1651-1659.

12. Tomkiel JE, Alansari H, Tang N, Virgin JB, Yang X, VandeVord P *et al*.: **Autoimmunity to the M(r) 32,000 subunit of replication protein A in breast cancer.** *Clin Cancer Res* 2002, **8:** 752-758.

13. Frenkel K, Karkoszka J, Glassman T, Dubin N, Toniolo P, Taioli E *et al*.: **Serum autoantibodies recognizing 5-hydroxymethyl-2'-deoxyuridine, an oxidized DNA base, as biomarkers of cancer risk in women.** *Cancer Epidemiol Biomarkers Prev* 1998, **7:** 49-57.

14. Ferlay J, Shin HR, Bray F, Forman D, Mathers C and Parkin DM. GLOBOCAN 2008 v1.2, Cancer Incidence and Mortality Worldwide: IARC CancerBase No. 10 [Internet]. Lyon, France: International Agency for Research on Cancer; 2010. Available from: http://globocan.iarc.fr, accessed on 21/04/2012.

15. Piura E, Piura B: **Autoantibodies to tailor-made panels of tumor-associated antigens in breast carcinoma.** *J Oncol* 2011, **2011:** 982425.

16. Chapman C, Murray A, Chakrabarti J, Thorpe A, Woolston C, Sahin U *et al*.: **Autoantibodies in breast cancer: their use as an aid to early diagnosis.** *Ann Oncol* 2007, **18:** 868-873.

17. Brooks J, Cairns P, Zeleniuch-Jacquotte A: **Promoter methylation and the detection of breast cancer.** *Cancer Causes Control* 2009, **20:** 1539-1550.

18. Peng G, Hakim M, Broza YY, Billan S, bdah-Bortnyak R, Kuten A *et al*.: **Detection of lung, breast, colorectal, and prostate cancers from exhaled breath using a single array of nanosensors.** *Br J Cancer* 2010, **103:** 542-551.

19. Gangadharan B, Antrobus R, Dwek RA, Zitzmann N: **Novel serum biomarker candidates for liver fibrosis in hepatitis C patients.** *Clin Chem* 2007, **53:** 1792-1799.

20. Pan S, Shi M, Jin J, Albin RL, Lieberman A, Gearing M *et al*.: **Proteomics identification of proteins in human cortex using multidimensional separations and MALDI tandem mass spectrometer.** *Mol Cell Proteomics* 2007, **6:** 1818-1823.

21. Meyer HE, Stuhler K: **High-performance proteomics as a tool in biomarker discovery.** *Proteomics* 2007, **7 Suppl 1:** 18-26.

22. Clarke CH, Buckley JA, Fung ET: **SELDI-TOF-MS proteomics of breast cancer.** *Clin Chem Lab Med* 2005, **43:** 1314-1320.

23. Diamandis EP: **Mass spectrometry as a diagnostic and a cancer biomarker discovery tool: opportunities and potential limitations.** *Mol Cell Proteomics* 2004, **3:** 367-378.

24. Mathelin C, Cromer A, Wendling C, Tomasetto C, Rio MC: **Serum biomarkers for detection of breast cancers: A prospective study.** *Breast Cancer Res Treat* 2006, **96:** 83-90.

25. Borrebaeck CA, Wingren C: **High-throughput proteomics using antibody microarrays: an update.** *Expert Rev Mol Diagn* 2007, **7:** 673-686.

26. Hu S, Li Y, Liu G, Song Q, Wang L, Han Y *et al*.: **A protein chip approach for high-throughput antigen identification and characterization.** *Proteomics* 2007, **7:** 2151-2161.

27. Li Y, Lee HJ, Corn RM: **Detection of protein biomarkers using RNA aptamer microarrays and enzymatically amplified surface plasmon resonance imaging.** *Anal Chem* 2007, **79:** 1082-1088.

28. Diesinger I, Bauer C, Brass N, Schaefers HJ, Comtesse N, Sybrecht G *et al*.: **Toward a more complete recognition of immunoreactive antigens in squamous cell lung carcinoma.** *Int J Cancer* 2002, **102:** 372-378.

29. Krackhardt AM, Witzens M, Harig S, Hodi FS, Zauls AJ, Chessia M *et al*.: **Identification of tumor-associated antigens in chronic lymphocytic leukemia by SEREX.** *Blood* 2002, **100:** 2123-2131.

30. Tureci O, Usener D, Schneider S, Sahin U: **Identification of tumor-associated autoantigens with SEREX.** *Methods Mol Med* 2005, **109:** 137-154.

31. Wang Y, Han KJ, Pang XW, Vaughan HA, Qu W, Dong XY *et al*.: **Large scale identification of human hepatocellular carcinoma-associated antigens by autoantibodies.** *J Immunol* 2002, **169:** 1102-1109.

32. Qian F, Odunsi K, Blatt LM, Scanlan MJ, Mannan M, Shah N *et al*.: **Tumor associated antigen recognition by autologous serum in patients with breast cancer.** *Int J Mol Med* 2005, **15:** 137-144.

33. Fossa A, Siebert R, Aasheim HC, Maelandsmo GM, Berner A, Fossa SD *et al*.: **Identification of nucleolar protein No55 as a tumour-associated autoantigen in patients with prostate cancer.** *Br J Cancer* 2000, **83:** 743-749.

34. Stone B, Schummer M, Paley PJ, Thompson L, Stewart J, Ford M *et al*.: **Serologic analysis of ovarian tumor antigens reveals a bias toward antigens encoded on 17q.** *Int J Cancer* 2003, **104:** 73-84.

35. Devitt G, Meyer C, Wiedemann N, Eichmuller S, Kopp-Schneider A, Haferkamp A *et al*.: **Serological analysis of human renal cell carcinoma.** *Int J Cancer* 2006, **118:** 2210-2219.

36. Vaughan HA, St CF, Scanlan MJ, Chen YT, Maraskovsky E, Sizeland A *et al*.: **The humoral immune response to head and neck cancer antigens as defined by the serological analysis of tumor antigens by recombinant cDNA expression cloning.** *Cancer Immun* 2004, **4:** 5.

37. Shimada H, Nakashima K, Ochiai T, Nabeya Y, Takiguchi M, Nomura F *et al*.: **Serological identification of tumor antigens of esophageal squamous cell carcinoma.** *Int J Oncol* 2005, **26:** 77-86.

38. Liggins AP, Guinn BA, Banham AH: **Identification of lymphoma-associated antigens using SEREX.** *Methods Mol Med* 2005, **115:** 109-128.

39. Umano Y, Tsunoda T, Tanaka H, Matsuda K, Yamaue H, Tanimura H: **Generation of cytotoxic T cell responses to an HLA-A24 restricted epitope peptide derived from wild-type p53.** *Br J Cancer* 2001, **84:** 1052-1057.

40. Wurtzen PA, Pedersen LO, Poulsen HS, Claesson MH: **Specific killing of P53 mutated tumor cell lines by a cross-reactive human HLA-A2-restricted P53-specific CTL line.** *Int J Cancer* 2001, **93:** 855-861.

41. Azuma K, Shichijo S, Maeda Y, Nakatsura T, Nonaka Y, Fujii T *et al*.: **Mutated p53 gene encodes a nonmutated epitope recognized by HLA-B*4601-restricted and tumor cell-reactive CTLs at tumor site.** *Cancer Res* 2003, **63:** 854-858.

42. Somasundaram R, Swoboda R, Caputo L, Otvos L, Weber B, Volpe P *et al*.: **Human leukocyte antigen-A2-restricted CTL responses to mutated BRAF peptides in melanoma patients.** *Cancer Res* 2006, **66:** 3287-3293.

43. Backes C, Ludwig N, Leidinger P, Harz C, Hoffmann J, Keller A *et al*.: **Immunogenicity of autoantigens.** *BMC Genomics* 2011, **12:** 340.

44. Baurain JF, Colau D, van BN, Landry C, Martelange V, Vikkula M *et al*.: **High frequency of autologous anti-melanoma CTL directed against an antigen generated by a point mutation in a new helicase gene.** *J Immunol* 2000, **164:** 6057-6066.

45. Hall DA, Ptacek J, Snyder M: **Protein microarray technology.** *Mech Ageing Dev* 2007, **128:** 161-167.

46. Pollard KM, Lee DK, Casiano CA, Bluthner M, Johnston MM, Tan EM: **The autoimmunity-inducing xenobiotic mercury interacts with the autoantigen fibrillarin and modifies its molecular and antigenic properties.** *J Immunol* 1997, **158:** 3521-3528.

47. Utz PJ, Anderson P: **Posttranslational protein modifications, apoptosis, and the bypass of tolerance to autoantigens.** *Arthritis Rheum* 1998, **41:** 1152-1160.

48. Rosen A, Casciola-Rosen L, Wigley F: **Role of metal-catalyzed oxidation reactions in the early pathogenesis of scleroderma.** *Curr Opin Rheumatol* 1997, **9:** 538-543.

49. Wu X, Molinaro C, Johnson N, Casiano CA: **Secondary necrosis is a source of proteolytically modified forms of specific intracellular autoantigens: implications for systemic autoimmunity.** *Arthritis Rheum* 2001, **44:** 2642-2652.

50. Tan EM, Zhang J: **Autoantibodies to tumor-associated antigens: reporters from the immune system.** *Immunol Rev* 2008, **222:** 328-340.

51. Zhang JY: **Tumor-associated antigen arrays to enhance antibody detection for cancer diagnosis.** *Cancer Detect Prev* 2004, **28:** 114-118.

52. Lu H, Goodell V, Disis ML: **Humoral immunity directed against tumor-associated antigens as potential biomarkers for the early diagnosis of cancer.** *J Proteome Res* 2008, **7:** 1388-1394.

53. Zhong L, Ge K, Zu JC, Zhao LH, Shen WK, Wang JF *et al*.: **Autoantibodies as potential biomarkers for breast cancer.** *Breast Cancer Res* 2008, **10:** R40.

54. Tschernatsch M, Gross O, Kneifel N, Kaps M, Blaes F: **SOX-1 autoantibodies in patients with paraneoplastic neurological syndromes.** *Autoimmun Rev* 2009, **8:** 549-551.

55. Titulaer MJ, Klooster R, Potman M, Sabater L, Graus F, Hegeman IM *et al*.: **SOX antibodies in small-cell lung cancer and Lambert-Eaton myasthenic syndrome: frequency and relation with survival.** *J Clin Oncol* 2009, **27:** 4260-4267.

56. Davis TM, Mehta Z, Mackay IR, Cull CA, Bruce DG, Fida S *et al*.: **Autoantibodies to the islet cell antigen SOX-13 are associated with duration but not type of diabetes.** *Diabet Med* 2003, **20:** 198-204.

57. Koziol JA, Zhang JY, Casiano CA, Peng XX, Shi FD, Feng AC *et al*.: **Recursive partitioning as an approach to selection of immune markers for tumor diagnosis.** *Clin Cancer Res* 2003, **9:** 5120-5126.

58. Babel I, Barderas R, az-Uriarte R, Martinez-Torrecuadrada JL, Sanchez-Carbayo M, Casal JI: **Identification of tumor-associated autoantigens for the diagnosis of colorectal cancer in serum using high density protein microarrays.** *Mol Cell Proteomics* 2009, **8:** 2382-2395.

59. Tschernatsch M, Gross O, Kneifel N, Kaps M, Blaes F: **SOX-1 autoantibodies in patients with paraneoplastic neurological syndromes.** *Autoimmun Rev* 2009, **8:** 549-551.

60. Titulaer MJ, Klooster R, Potman M, Sabater L, Graus F, Hegeman IM *et al*.: **SOX antibodies in small-cell lung cancer and Lambert-Eaton myasthenic syndrome: frequency and relation with survival.** *J Clin Oncol* 2009, **27:** 4260-4267.

61. Chapman CJ, Murray A, McElveen JE, Sahin U, Luxemburger U, Tureci O *et al*.: **Autoantibodies in lung cancer: possibilities for early detection and subsequent cure.** *Thorax* 2008, **63:** 228-233.

62. Farlow EC, Patel K, Basu S, Lee BS, Kim AW, Coon JS *et al*.: **Development of a multiplexed tumor-associated autoantibody-based blood test for the detection of non-small cell lung cancer.** *Clin Cancer Res* 2010, **16:** 3452-3462.

63. Wang X, Yu J, Sreekumar A, Varambally S, Shen R, Giacherio D *et al*.: **Autoantibody signatures in prostate cancer.** *N Engl J Med* 2005, **353:** 1224-1235.

64. Liu W, Wang P, Li Z, Xu W, Dai L, Wang K *et al*.: **Evaluation of tumour-associated antigen (TAA) miniarray in immunodiagnosis of colon cancer.** *Scand J Immunol* 2009, **69:** 57-63.

65. Babel I, Barderas R, az-Uriarte R, Martinez-Torrecuadrada JL, Sanchez-Carbayo M, Casal JI: **Identification of tumor-associated autoantigens for the diagnosis of colorectal cancer in serum using high density protein microarrays.** *Mol Cell Proteomics* 2009, **8:** 2382-2395.

66. Desmetz C, Bascoul-Mollevi C, Rochaix P, Lamy PJ, Kramar A, Rouanet P *et al*.: **Identification of a new panel of serum autoantibodies associated with the presence of in situ carcinoma of the breast in younger women.** *Clin Cancer Res* 2009, **15:** 4733-4741.

67. Huang RP: **Protein arrays, an excellent tool in biomedical research.** *Front Biosci* 2003, **8:** d559-d576.

68. Holt LJ, Bussow K, Walter G, Tomlinson IM: **By-passing selection: direct screening for antibody-antigen interactions using protein arrays.** *Nucleic Acids Res* 2000, **28:** E72.

69. Kijanka G, Murphy D: **Protein arrays as tools for serum autoantibody marker discovery in cancer.** *J Proteomics* 2009, **72:** 936-944.

70. Pedersen SK, Sloane AJ, Prasad SS, Sebastian LT, Lindner RA, Hsu M *et al*.: **An immunoproteomic approach for identification of clinical biomarkers for monitoring disease: application to cystic fibrosis.** *Mol Cell Proteomics* 2005, **4:** 1052-1060.

71. Chatterjee M, Mohapatra S, Ionan A, Bawa G, li-Fehmi R, Wang X *et al*.: **Diagnostic markers of ovarian cancer by high-throughput antigen cloning and detection on arrays.** *Cancer Res* 2006, **66:** 1181-1190.

72. Leidinger P, Keller A, Heisel S, Ludwig N, Rheinheimer S, Klein V *et al*.: **Novel autoantigens immunogenic in COPD patients.** *Respir Res* 2009, **10:** 20.

73. Leidinger P, Keller A, Heisel S, Ludwig N, Rheinheimer S, Klein V *et al*.: **Identification of lung cancer with high sensitivity and specificity by blood testing.** *Respir Res* 2010, **11:** 18.

74. YALOW RS, BERSON SA: **Assay of plasma insulin in human subjects by immunological methods.** *Nature* 1959, **184 (Suppl 21):** 1648-1649.

75. Engvall E, Perlmann P: **Enzyme-linked immunosorbent assay (ELISA). Quantitative assay of immunoglobulin G.** *Immunochemistry* 1971, **8:** 871-874.

76. Ekins R, Chu F, Biggart E: **Multispot, multianalyte, immunoassay.** *Ann Biol Clin (Paris)* 1990, **48:** 655-666.

77. Silzel JW, Cercek B, Dodson C, Tsay T, Obremski RJ: **Mass-sensing, multianalyte microarray immunoassay with imaging detection.** *Clin Chem* 1998, **44:** 2036-2043.

78. Tomizaki KY, Usui K, Mihara H: **Protein-detecting microarrays: current accomplishments and requirements.** *Chembiochem* 2005, **6:** 782-799.

79. Huang RP, Huang R, Fan Y, Lin Y: **Simultaneous detection of multiple cytokines from conditioned media and patient's sera by an antibody-based protein array system.** *Anal Biochem* 2001, **294:** 55-62.

80. Nam MJ, Madoz-Gurpide J, Wang H, Lescure P, Schmalbach CE, Zhao R *et al*.: **Molecular profiling of the immune response in colon cancer using protein microarrays: occurrence of autoantibodies to ubiquitin C-terminal hydrolase L3.** *Proteomics* 2003, **3:** 2108-2115.

81. Olle EW, Messamore J, Deogracias MP, McClintock SD, Anderson TD, Johnson KJ: **Comparison of antibody array substrates and the use of glycerol to normalize spot morphology.** *Exp Mol Pathol* 2005, **79:** 206-209.

82. Fici DA, McCormick W, Brown DW, Herrmann JE, Kumar V, Awdeh ZL: **A protein multiplex microarray substrate with high sensitivity and specificity.** *J Immunol Methods* 2010, **363:** 60-66.

83. Lee DW, Kim HJ, Choi CH, Shin JH, Kim EK: **Development of a protein chip to measure PKCbeta activity.** *Appl Biochem Biotechnol* 2011, **163:** 803-812.

84. Kang M, Trofin L, Mota MO, Martin CR: **Protein capture in silica nanotube membrane 3-D microwell arrays.** *Anal Chem* 2005, **77:** 6243-6249.

85. Guilleaume B, Buness A, Schmidt C, Klimek F, Moldenhauer G, Huber W *et al*.: **Systematic comparison of surface coatings for protein microarrays.** *Proteomics* 2005, **5:** 4705-4712.

86. Zubtsov DA, Savvateeva EN, Rubina AY, Pan'kov SV, Konovalova EV, Moiseeva OV *et al*.: **Comparison of surface and hydrogel-based protein microchips.** *Anal Biochem* 2007, **368:** 205-213.

87. Tanaka H, Hanasaki M, Isojima T, Takeuchi H, Shiroya T, Kawaguchi H: **Enhancement of sensitivity of SPR protein microarray using a novel 3D protein immobilization.** *Colloids Surf B Biointerfaces* 2009, **70:** 259-265.

88. Chen H, Jiang C, Yu C, Zhang S, Liu B, Kong J: **Protein chips and nanomaterials for application in tumor marker immunoassays.** *Biosens Bioelectron* 2009, **24:** 3399-3411.

89. Tsarfati-Barad I, Sauer U, Preininger C, Gheber LA: **Miniaturized protein arrays: Model and experiment.** *Biosens Bioelectron* 2011.

90. Chandra H, Reddy PJ, Srivastava S: **Protein microarrays and novel detection platforms.** *Expert Rev Proteomics* 2011, **8:** 61-79.

91. Haab BB, Dunham MJ, Brown PO: **Protein microarrays for highly parallel detection and quantitation of specific proteins and antibodies in complex solutions.** *Genome Biol* 2001, **2:** RESEARCH0004.

92. Sreekumar A, Nyati MK, Varambally S, Barrette TR, Ghosh D, Lawrence TS *et al*.: **Profiling of cancer cells using protein microarrays: discovery of novel radiation-regulated proteins.** *Cancer Res* 2001, **61:** 7585-7593.

93. Schweitzer B, Roberts S, Grimwade B, Shao W, Wang M, Fu Q *et al*.: **Multiplexed protein profiling on microarrays by rolling-circle amplification.** *Nat Biotechnol* 2002, **20:** 359-365.

94. Nolen BM, Marks JR, Ta'san S, Rand A, Luong TM, Wang Y *et al*.: **Serum biomarker profiles and response to neoadjuvant chemotherapy for locally advanced breast cancer.** *Breast Cancer Res* 2008, **10:** R45.

95. Carlsson A, Wingren C, Kristensson M, Rose C, Ferno M, Olsson H *et al*.: **Molecular serum portraits in patients with primary breast cancer predict the development of distant metastases.** *Proc Natl Acad Sci U S A* 2011, **108:** 14252-14257.

96. Pla-Roca M, Leulmi RF, Tourekhanova S, Bergeron S, Laforte V, Moreau E *et al*.: **Antibody colocalization microarray: a scalable technology for multiplex protein analysis in complex samples.** *Mol Cell Proteomics* 2012, **11:** M111.

97. Bouwman K, Qiu J, Zhou H, Schotanus M, Mangold LA, Vogt R *et al*.: **Microarrays of tumor cell derived proteins uncover a distinct pattern of prostate cancer serum immunoreactivity.** *Proteomics* 2003, **3:** 2200-2207.

98. Yan F, Sreekumar A, Laxman B, Chinnaiyan AM, Lubman DM, Barder TJ: **Protein microarrays using liquid phase fractionation of cell lysates.** *Proteomics* 2003, **3:** 1228-1235.

99. Qiu J, Madoz-Gurpide J, Misek DE, Kuick R, Brenner DE, Michailidis G *et al*.: **Development of natural protein microarrays for diagnosing cancer based on an antibody response to tumor antigens.** *J Proteome Res* 2004, **3:** 261-267.

100. Pereira-Faca SR, Kuick R, Puravs E, Zhang Q, Krasnoselsky AL, Phanstiel D *et al*.: **Identification of 14-3-3 theta as an antigen that induces a humoral response in lung cancer.** *Cancer Res* 2007, **67:** 12000-12006.

101. Madoz-Gurpide J, Kuick R, Wang H, Misek DE, Hanash SM: **Integral protein microarrays for the identification of lung cancer antigens in sera that induce a humoral immune response.** *Mol Cell Proteomics* 2008, **7:** 268-281.

102. Albala JS, Franke K, McConnell IR, Pak KL, Folta PA, Rubinfeld B *et al*.: **From genes to proteins: high-throughput expression and purification of the human proteome.** *J Cell Biochem* 2000, **80:** 187-191.

103. Hollister J, Grabenhorst E, Nimtz M, Conradt H, Jarvis DL: **Engineering the protein N-glycosylation pathway in insect cells for production of biantennary, complex N-glycans.** *Biochemistry* 2002, **41:** 15093-15104.

104. Schena M. Protein microarrays. Boston, Mass: Jones and Bartlett; 2005.

105. Sreekumar A, Laxman B, Rhodes DR, Bhagavathula S, Harwood J, Giacherio D *et al*.: **Humoral immune response to alpha-methylacyl-CoA racemase and prostate cancer.** *J Natl Cancer Inst* 2004, **96:** 834-843.

106. Taylor BS, Pal M, Yu J, Laxman B, Kalyana-Sundaram S, Zhao R *et al*.: **Humoral response profiling reveals pathways to prostate cancer progression.** *Mol Cell Proteomics* 2008, **7:** 600-611.

107. Massoner P, Lueking A, Goehler H, Hopfner A, Kowald A, Kugler KG *et al*.: **Serum-autoantibodies for discovery of prostate cancer specific biomarkers.** *Prostate* 2011.

108. Anderson KS, Ramachandran N, Wong J, Raphael JV, Hainsworth E, Demirkan G *et al*.: **Application of protein microarrays for multiplexed detection of antibodies to tumor antigens in breast cancer.** *J Proteome Res* 2008, **7:** 1490-1499.

109. Eisenstein M: **Protein arrays: growing pains.** *Nature* 2006, **444:** 959-962.

110. Phizicky E, Bastiaens PI, Zhu H, Snyder M, Fields S: **Protein analysis on a proteomic scale.** *Nature* 2003, **422:** 208-215.

111. Haab BB: **Antibody arrays in cancer research.** *Mol Cell Proteomics* 2005, **4:** 377-383.

112. Sheehan KM, Calvert VS, Kay EW, Lu Y, Fishman D, Espina V *et al*.: **Use of reverse phase protein microarrays and reference standard development for molecular network analysis of metastatic ovarian carcinoma.** *Mol Cell Proteomics* 2005, **4:** 346-355.

113. Balboni I, Chan SM, Kattah M, Tenenbaum JD, Butte AJ, Utz PJ: **Multiplexed protein array platforms for analysis of autoimmune diseases.** *Annu Rev Immunol* 2006, **24:** 391-418.

114. Cretich M, Damin F, Pirri G, Chiari M: **Protein and peptide arrays: recent trends and new directions.** *Biomol Eng* 2006, **23:** 77-88.

115. Reimer U, Reineke U, Schneider-Mergener J: **Peptide arrays: from macro to micro.** *Curr Opin Biotechnol* 2002, **13:** 315-320.

116. Stafford P, Johnston S: **Microarray technology displays the complexities of the humoral immune response.** *Expert Rev Mol Diagn* 2011, **11:** 5-8.

117. Geysen HM, Meloen RH, Barteling SJ: **Use of peptide synthesis to probe viral antigens for epitopes to a resolution of a single amino acid.** *Proc Natl Acad Sci U S A* 1984, **81:** 3998-4002.

118. Frank R, Overwin H: **SPOT synthesis. Epitope analysis with arrays of synthetic peptides prepared on cellulose membranes.** *Methods Mol Biol* 1996, **66:** 149-169.

119. Robinson WH, Steinman L, Utz PJ: **Protein and peptide array analysis of autoimmune disease.** *Biotechniques* 2002, **Suppl:** 66-69.

120. Bradford TJ, Wang X, Chinnaiyan AM: **Cancer immunomics: using autoantibody signatures in the early detection of prostate cancer.** *Urol Oncol* 2006, **24:** 237-242.

121. Sahin U, Tureci O, Schmitt H, Cochlovius B, Johannes T, Schmits R *et al*.: **Human neoplasms elicit multiple specific immune responses in the autologous host.** *Proc Natl Acad Sci U S A* 1995, **92:** 11810-11813.

122. Wang Y, Gu Q, Liu B, Zhu Z: **Perspectives of SEREX-defined antigens in diagnosis and immunotherapy for gastric cancer.** *Cancer Biol Ther* 2004, **3:** 806-811.

123. Lee SY, Jeoung D: **The reverse proteomics for identification of tumor antigens.** *J Microbiol Biotechnol* 2007, **17:** 879-890.

124. Schmits R, Cochlovius B, Treitz G, Regitz E, Ketter R, Preuss KD *et al*.: **Analysis of the antibody repertoire of astrocytoma patients against antigens expressed by gliomas.** *Int J Cancer* 2002, **98:** 73-77.

125. Cheever MA, Disis ML, Bernhard H, Gralow JR, Hand SL, Huseby ES *et al*.: **Immunity to oncogenic proteins.** *Immunol Rev* 1995, **145:** 33-59.

126. Tureci O, Sahin U, Vollmar E, Siemer S, Gottert E, Seitz G *et al*.: **Human carbonic anhydrase XII: cDNA cloning, expression, and chromosomal localization of a carbonic anhydrase gene that is overexpressed in some renal cell cancers.** *Proc Natl Acad Sci U S A* 1998, **95:** 7608-7613.

127. Scanlan MJ, Chen YT, Williamson B, Gure AO, Stockert E, Gordan JD *et al*.: **Characterization of human colon cancer antigens recognized by autologous antibodies.** *Int J Cancer* 1998, **76:** 652-658.

128. Line A, Slucka Z, Stengrevics A, Silina K, Li G, Rees RC: **Characterisation of tumour-associated antigens in colon cancer.** *Cancer Immunol Immunother* 2002, **51:** 574-582.

129. Jager D, Stockert E, Gure AO, Scanlan MJ, Karbach J, Jager E *et al*.: **Identification of a tissue-specific putative transcription factor in breast tissue by serological screening of a breast cancer library.** *Cancer Res* 2001, **61:** 2055-2061.

130. Jager D, Stockert E, Scanlan MJ, Gure AO, Jager E, Knuth A *et al*.: **Cancer-testis antigens and ING1 tumor suppressor gene product are breast cancer antigens: characterization of tissue-specific ING1 transcripts and a homologue gene.** *Cancer Res* 1999, **59:** 6197-6204.

131. Fossa A, Alsoe L, Crameri R, Funderud S, Gaudernack G, Smeland EB: **Serological cloning of cancer/testis antigens expressed in prostate cancer using cDNA phage surface display.** *Cancer Immunol Immunother* 2004, **53:** 431-438.

132. Scanlan MJ, Welt S, Gordon CM, Chen YT, Gure AO, Stockert E *et al*.: **Cancer-related serological recognition of human colon cancer: identification of potential diagnostic and immunotherapeutic targets.** *Cancer Res* 2002, **62:** 4041-4047.

133. Scanlan MJ, Gout I, Gordon CM, Williamson B, Stockert E, Gure AO *et al*.: **Humoral immunity to human breast cancer: antigen definition and quantitative analysis of mRNA expression.** *Cancer Immun* 2001, **1:** 4.

134. Ishikawa T, Fujita T, Suzuki Y, Okabe S, Yuasa Y, Iwai T *et al*.: **Tumor-specific immunological recognition of frameshift-mutated peptides in colon cancer with microsatellite instability.** *Cancer Res* 2003, **63:** 5564-5572.

135. Gure AO, Altorki NK, Stockert E, Scanlan MJ, Old LJ, Chen YT: **Human lung cancer antigens recognized by autologous antibodies: definition of a novel cDNA derived from the tumor suppressor gene locus on chromosome 3p21.3.** *Cancer Res* 1998, **58:** 1034-1041.

136. Gure AO, Stockert E, Scanlan MJ, Keresztes RS, Jager D, Altorki NK *et al*.: **Serological identification of embryonic neural proteins as highly immunogenic tumor antigens in small cell lung cancer.** *Proc Natl Acad Sci U S A* 2000, **97:** 4198-4203.

137. Zhou Y, Toth M, Hamman MS, Monahan SJ, Lodge PA, Boynton AL *et al*.: **Serological cloning of PARIS-1: a new TBC domain-containing, immunogenic tumor antigen from a prostate cancer cell line.** *Biochem Biophys Res Commun* 2002, **290:** 830-838.

138. Ojika T, Imaizumi M, Abe T, Kato K: **Immunochemical and immunohistochemical studies on three aldolase isozymes in human lung cancer.** *Cancer* 1991, **67:** 2153-2158.

139. Asaka M, Kimura T, Meguro T, Kato M, Kudo M, Miyazaki T *et al*.: **Alteration of aldolase isozymes in serum and tissues of patients with cancer and other diseases.** *J Clin Lab Anal* 1994, **8:** 144-148.

140. Smith GP: **Filamentous fusion phage: novel expression vectors that display cloned antigens on the virion surface.** *Science* 1985, **228:** 1315-1317.

141. Danner S, Belasco JG: **T7 phage display: a novel genetic selection system for cloning RNA-binding proteins from cDNA libraries.** *Proc Natl Acad Sci U S A* 2001, **98:** 12954-12959.

142. Mersmann M, Meier D, Mersmann J, Helmsing S, Nilsson P, Graslund S *et al*.: **Towards proteome scale antibody selections using phage display.** *N Biotechnol* 2010, **27:** 118-128.

143. Kelly KA, Jones DA: **Isolation of a colon tumor specific binding peptide using phage display selection.** *Neoplasia* 2003, **5:** 437-444.

144. Somers VA, Brandwijk RJ, Joosten B, Moerkerk PT, Arends JW, Menheere P *et al*.: **A panel of candidate tumor antigens in colorectal cancer revealed by the serological selection of a phage displayed cDNA expression library.** *J Immunol* 2002, **169:** 2772-2780.

145. Zhong L, Coe SP, Stromberg AJ, Khattar NH, Jett JR, Hirschowitz EA: **Profiling tumor-associated antibodies for early detection of non-small cell lung cancer.** *J Thorac Oncol* 2006, **1:** 513-519.

146. Ran Y, Hu H, Zhou Z, Yu L, Sun L, Pan J *et al*.: **Profiling tumor-associated autoantibodies for the detection of colon cancer.** *Clin Cancer Res* 2008, **14:** 2696-2700.

147. Wu L, Chang W, Zhao J, Yu Y, Tan X, Su T *et al*.: **Development of autoantibody signatures as novel diagnostic biomarkers of non-small cell lung cancer.** *Clin Cancer Res* 2010, **16:** 3760-3768.

148. Lichtenfels R, Kellner R, Atkins D, Bukur J, Ackermann A, Beck J *et al*.: **Identification of metabolic enzymes in renal cell carcinoma utilizing PROTEOMEX analyses.** *Biochim Biophys Acta* 2003, **1646:** 21-31.

149. Unwin RD, Harnden P, Pappin D, Rahman D, Whelan P, Craven RA *et al*.: **Serological and proteomic evaluation of antibody responses in the identification of tumor antigens in renal cell carcinoma.** *Proteomics* 2003, **3:** 45-55.

150. Tan HT, Low J, Lim SG, Chung MC: **Serum autoantibodies as biomarkers for early cancer detection.** *FEBS J* 2009, **276:** 6880-6904.

151. Klade CS, Voss T, Krystek E, Ahorn H, Zatloukal K, Pummer K *et al*.: **Identification of tumor antigens in renal cell carcinoma by serological proteome analysis.** *Proteomics* 2001, **1:** 890-898.

152. Gunawardana CG, Diamandis EP: **High throughput proteomic strategies for identifying tumour-associated antigens.** *Cancer Lett* 2007, **249:** 110-119.

153. Hamrita B, Chahed K, Kabbage M, Guillier CL, Trimeche M, Chaieb A *et al*.: **Identification of tumor antigens that elicit a humoral immune response in breast cancer patients' sera by serological proteome analysis (SERPA).** *Clin Chim Acta* 2008, **393:** 95-102.

154. Sato T, Saito H, Swensen J, Olifant A, Wood C, Danner D *et al*.: **The human prohibitin gene located on chromosome 17q21 is mutated in sporadic breast cancer.** *Cancer Res* 1992, **52:** 1643-1646.

155. Suzuki A, Iizuka A, Komiyama M, Takikawa M, Kume A, Tai S *et al*.: **Identification of melanoma antigens using a Serological Proteome Approach (SERPA).** *Cancer Genomics Proteomics* 2010, **7:** 17-23.

156. He Y, Wu Y, Mou Z, Li W, Zou L, Fu T *et al*.: **Proteomics-based identification of HSP60 as a tumor-associated antigen in colorectal cancer.** *Proteomics Clin Appl* 2007, **1:** 336-342.

157. Gygi SP, Corthals GL, Zhang Y, Rochon Y, Aebersold R: **Evaluation of two-dimensional gel electrophoresis-based proteome analysis technology.** *Proc Natl Acad Sci U S A* 2000, **97:** 9390-9395.

Chapter 1: Tumour auto-antibody screening: performance of protein microarrays using SEREX derived antigens

---

Published in *BMC Cancer* as Stempfer R*, **Syed P*** *et al*.: **Tumour auto-antibody screening: performance of protein microarrays using SEREX derived antigens.** *BMC Cancer* 2010, **10:** 627.

*Equal contribution

Keywords: macroarrays, microarrays, SEREX, recombinant protein expression and purification

Abstract

The simplicity and potential of minimal invasive testing using serum from patients make auto-antibody based biomarkers a very promising tool for use in diagnostics of cancer and auto-immune disease. Although, several methods exist for elucidating candidate-protein markers, immobilizing these onto membranes and generating so called macroarrays is of limited use for marker validation. Especially when several hundred samples have to be analysed, microarrays could serve as a good alternative since processing macro membranes is cumbersome and reproducibility of results is moderate.

Candidate markers identified by SEREX (serological identification of antigens by recombinant expression cloning) screenings of brain and lung tumour were used for macroarray and microarray production. For microarray production recombinant proteins were expressed in *E. coli* by autoinduction and purified His-tag (histidine-tagged) proteins were then used for the production of protein microarrays. Protein arrays were hybridized with the serum samples from brain and lung tumour patients.

Methods for the generation of microarrays were successfully established when using antigens derived from membrane-based selection. Signal patterns obtained by microarrays analysis of brain and lung tumour patients' sera were highly reproducible (R=0.92-0.96). This provides the technical foundation for diagnostic applications on the basis of auto-antibody patterns. In this limited test set, the assay provided high reproducibility and a broad dynamic range to classify all brain and lung samples correctly.

Protein microarray is an efficient means for auto-antibody -based detection when using SEREX-derived clones expressing antigenic proteins. Protein microarrays are preferred to macroarrays due to the easier handling and the high reproducibility of auto-antibody testing. Especially when using only a few microliters of patient samples protein microarrays are ideally suited for validation of auto-antibody signatures for diagnostic purposes.

Background

The idea of early diagnosis of the onset of a disease via biomarkers has inspired several molecular biological approaches. In the past decade, since the unravelling of the human genome to a large extent, genomics technologies have been used to identify disease biomarkers. For cancerous diseases recently the most promising results were obtained by gene expression profiling. Excellent results have been achieved with these techniques in terms of improved patient stratification and increased potential of a clearer prognosis by a more detailed initial diagnosis. However, the true challenge is to develop techniques which are suitable for early diagnosis and prophylactic screening. These techniques should be minimally invasive, cost effective and ideally they indicate several diseases of the screened patient [1].

Proteomics techniques have shifted biomarker identification and validation research to the level of the main actual biological agents of health and disease, the proteins. Despite recent improvements in separation techniques based on HPLC (High-performance liquid chromatography) separation, mass spectrometry and 2D electrophoresis, so far the novel biomarker candidates and biomarker signatures which are ready for the use in clinical settings have proven to require detection steps similarly complex as in their discovery and thus complicate their widespread use in the screening of large populations [1].

Ideally the proteomics based techniques result in the identification of marker molecules that can be targeted in specialized assays relying on antibodies or aptamers [2-4]. The development of specific capturing agents for the candidate markers requires a costly production process and thorough validation. This ensures high avidity for the target while minimizing the risk of unspecific binding.

In the auto-antibody approach these issues are sophisticatedly avoided. The need to identify aberrant nucleic acid sequences, disease related biochemical compounds, disease affected cells or their debris is reduced by making use of a highly sensitive detection system closest to the patient, the human immune system. Mutated, modified and aberrantly expressed proteins evoke an immunological response leading to the production of auto-antibodies [5,6]. The auto-antibody based biomarkers could be used as serological tool for the early

diagnosis and prognosis of cancer as auto-antibodies are specific to each kind of cancer [5,7,8]. Most of the auto-antibodies are immunological finger prints of pathological processes which are involved in the development of autoimmunity [5]. Such a molecular finger print of auto-antibodies which is produced against certain disease states can be called auto-antibody signature [7]. Assays for the detection of auto-antibodies at present are mainly ELISA (enzyme linked immunosorbent assay) and fluorescence immunoassays. However, protein microarrays have great potential to characterize auto-antibodies [9].

Strategies like the SEREX have been developed for the serological definition of immunogenic tumour antigens [10-13]. A similar approach has been used successfully for the identification of tumour endothelium associated antigen genes from human liver cancer vascular endothelial cells by generating a cDNA expression library. For the identification of auto-antibodies against pancreatic ductal adenocarcinoma-associated antigens that could be useful for early cancer diagnosis and therapy, a proteomics approach was followed up [14,15]. Proteins from pancreatic ductal adenocarcinoma cell lines were separated by 2D electrophoresis, and the serum IgG (immunoglobulin G) reactivity was tested by Western blot analysis. Spots specifically reacting with auto-antibodies from the sera of pancreatic ductal adenocarcinoma patients, which were analyzed by mass spectrometry, corresponded to metabolic enzymes or cytoskeletal proteins which proved to be specific targets of the humoral response to pancreatic ductal adenocarcinoma.

Over recent years most approaches have used so called macroarrays for autoantigen-profiling. These macroarrays are generated by spotting cDNA expression clones on membranes. Expression clones are grown on these membranes and recombinant proteins over-expressed upon induction are directly immobilized on the reactive membrane surfaces. Because entire colonies are lysed directly on the membranes and proteins of interest are immobilized in the background of the proteins of the expression-host bacteria, the targeted proteins are accessible for detection only after removal of the reactive anti-*E. coli*-Ig (immunoglobulin) from the analyte. This can be achieved either by masking of the anti-*E. coli* antibodies in human sera by addition of saturating concentrations of *E. coli* crude protein extracts and by blocking unwanted reactivity against *E. coli* by repeated incubation of membranes with the human serum. In this latter approach reactive Ig's from sera are captured by the macroarrays, sera are collected upon this primary incubation and after

washing/stripping the membranes the sera are applied again onto the membranes obtaining then the signal from the Ig's specific for the reactive over-expressed antigens. Handling membranes and processing sera is cumbersome, and sensitivity and reproducibility of these macroarrays are limiting. Signals derived from membranes are not dynamic. In analogy to western blotting different strategies exist to enhance sensitivities and to extend the dynamic range of membrane-based measures, but are rather limiting compared to the 16 bit ($0\text{-}2^{16}$) dynamic range of standard microarrays.

In this study, the methods for the generation of protein microarrays were optimized. This experiment provided an optimized protocol for generation of biomarker profiles with high reproducibility using 10µL amounts of patient serum samples. The provided data do also confirm that clones derived from SEREX membrane screens can be successfully transferred onto microarray slides retaining reactivity and gaining dynamic signal measures suitable for class-comparison to elucidate and validate protein-biomarkers.

Methods

Candidate marker screening

The candidate markers were identified by previous SEREX screenings of brain and lung cancer, and screening macroarrays of a foetal brain cDNA expression library. Potential tumour associated antigens derived from SEREX screens were isolated and sub cloned in the expression vector pQE30NST for production of His-tag (histidine-tagged) fusion proteins [16-20]. Marker candidate screening involved testing of serum from lung tumour patients and from brain tumour patients as well as control sera under the patients' informed consent. The local ethics committee (Ärztekammer des Saarlandes, Kenn-Nr. 213/08) approved the study and the research was carried out in compliance with the Helsinki Declaration.

E. coli culturing and induction

E. coli culturing techniques were adapted and modified after [19,20]. Deep-96well micro titre plates were filled with 1200 µL 2xYT medium (per L: Bactotryptone 16 g, yeast extract 10 g, NaCl 5 g) supplemented with 2% glucose, 100 µg/mL ampicillin, and 15 µg/mL kanamycin. Plates were inoculated with 3µL from master plates of 96 *E. coli* cultures (using

vector constructs of the SEREX library) each, sealed with gas permeable film and incubated overnight at 37°C while shaking at 1000 rpm. For the glycerol daughter plates, in each well 20 µL glycerol were combined with 50 µL of the overnight cultures and stored at -80°C.

Deep-96well microtiter plates were filled with 100 µL 2xYT medium with the aforementioned supplements. Media were inoculated with 3 µL of the glycerol stocks. Plates were sealed with gas permeable film and shaken overnight at 37°C. In the morning, 900µL SB (Sabouraud broth) medium pre-warmed to 37°C were added to each well. One litre of SB medium consisted of: bactotryptone 12 g, yeast extract 24 g, glycerol 4 mL, potassium phosphate buffer 50 mL (consisting of 2.4 g KH_2PO_4 and 12.15 g K_2HPO_4 in 50 mL) were added per litre medium; upon autoclaving the medium was supplemented with stock solutions obtaining final concentrations of 100 µg/mL ampicillin, 15 µg/mL kanamycin, and 20 µg/mL thiamine.) Plates were sealed with gas permeable film and shaken at 37°C at 1000rpm until an OD_{600} of 0.5-0.8 was reached for the IPTG (isopropyl-β-D-thiogalactopyranosid) induction. Then, 100 µL 11 mM IPTG in distilled water was added and cultures were shaken for 3h at 37°C at 1000rpm. *E. coli* cells were pelleted for 10min at 3500rpm, and the supernatants were discarded. Pellets were washed in PBS (phosphate buffered saline, pH 7.0) and the washed pellets were frozen at -80°C pending protein extraction. Alternatively, for the autoinduction protocol, the SB medium was additionally supplemented with 0.5% glycerol, 0.05% glucose, and 0.2% lactose. Upon addition of SB medium, plates were shaken overnight at 30°C at 1000rpm and then pelleted.

Comment on induction-strategies for protein-expression

In order to determine the ideal time-point for induction of the *E. coli* culture with IPTG, the OD had to be controlled tightly with a set of test clones over a course of 0.5 to 3h. Since these test clones may not ideally represent all the clones, and measuring the OD of all clones simultaneously is not feasible, the optimum time-point of induction may not be met by several clones. The autoinduction principle of William F. Studier [21] has been modified and adapted for the use in complex media to overcome this issue of having to choose the ideal time for the addition of an inducer. The minute amounts of glucose in the medium prevent untimely induction by the present lactose until high cell densities are reached. Then, when all of the glucose is metabolized, and with no other nutrients limiting, the less attractive

nutrient lactose is taken up by the cells and then causes induction. With IPTG induction, protein concentrations after the Ni-NTA affinity purification on average reached 0.26 mg/mL while with the auto-induction concentrations of 0.21 mg/mL were achieved. These similar yields in protein concentration point towards similar effectiveness in protein expression. Considering the rate of clones expressing His-tagged protein of 41% of the IPTG-induced clones as determined by the Penta-His antibody detection on chip and the expression rate of 40% of the auto-induced clones as determined by the His-tag ELISA, it becomes clear that both methods of induction of recombinant protein expression were equally successful.

Protein extraction

Bacterial pellets were thawed and resuspended in 100 µL lysis buffer (50 mM Tris-HCl, pH 8.0, 0.3 M NaCl, 0.1 mM EDTA, 0.02 % NaN_3). Lysozyme was added to a final concentration of 50µg/well in a 1% Brij58 solution. Pellets were resuspended and incubated on ice for 30min to yield the whole cellular extract. To each well 25 µL of benzonase mixture were added containing 0.3 µL 1 M $MgCl_2$, 0.1 µL benzonase grade 11, and 24.6 µL 50 mM Tris-HCl. Lysates were mixed, incubated at RT (room temperature) for 30min and centrifuged for 30min at 6200rpm at 4°C for removal of cell debris. The supernatant was transferred to a small-pore filter plate (Millipore GmbH) to remove any remaining particles. Filter plates were placed on top of capture plates and supernatants were filtered by centrifugation at 1000g obtaining the clarified protein extracts.

Protein purification

The filtered protein extracts were transferred to small-pore filter plates (Millipore GmbH) placed on top of a capture plate. Then, 15 µL of 0.1 M imidazole were added to each well to reach an end concentration of 10 mM. Upon addition of 25 µL of 20% Ni-NTA agarose (Qiagen) equilibrated in 50 mM Tris-HCl, pH 8.0, samples were shaken for 30min at RT. Liquid was removed by centrifugation of the filter plates at 1000 g followed by three washing steps with sodium phosphate wash buffer (50 mM monobasic and 50 mM dibasic sodium phosphate mixed together to produce a solution with pH 8.0), 0.3 M NaCl, and 20 mM imidazole. Upon the last wash protein was eluted with 50 µL of the elution buffer (sodium

phosphate buffer as abovementioned, pH 8.0 containing 250 mM or 500 mM imidazole, 0.01% SDS and 0.01% NaN_3).

Electrophoresis

Purified protein samples of 5 µL were resolved on NuPAGE Novex 4-12% Bis-Tris gels (Invitrogen) in the MES SDS running buffer system (Invitrogen). PageRuler prestained protein ladder (Fermentas GmbH) was used as a standard. Gels were stained first with the 6xHis Protein Tag staining kit (Pierce Biotechnology) according to manufacturer's instructions, and then according to the conventional protocol of the PageBlue protein staining solution (Fermentas GmbH).

Determination of protein-concentrations in microtiter plate

The DC Protein Assay from kit II (Bio-Rad) was adapted for the use in microtiter plates. BSA (bovine serum albumin) standards were prepared in elution buffer. Pipetting volumes of the protein assay were scaled according the manufacturer's instruction for using 5 µL of sample-volumes for the measurements. Absorbance at 630nm was read on an ELx800 automated microplate reader (Bio-Tek Instruments, Inc.) and protein-concentrations calculated based on the BSA standard measures.

ELISA-measurement of His-tagged proteins

For the relative quantification of His-tagged proteins an ELISA assay was established employing the Penta-His HRP (horseradish-peroxidase) conjugate detection antibody. Samples were diluted in immobilization buffer (50 mM sodium carbonate buffer; 100 mM sodium carbonate and 100 mM sodium bicarbonate mixed together to produce a solution with pH 9.6) to a total protein content of 10 µg/mL. 100 µL of the samples were transferred to high-binding Nunc MaxiSorp 96well plates (Nunc GmbH & Co. KG, Thermo Fisher Scientific) and incubated overnight at 4°C. The solution was removed and 200µL blocking solution (50 mM Tris, pH 8.0, 0.14 M NaCl, 1% BSA) was added for an incubation of 30min at RT. The solution was removed and 100µL HRP-conjugated antibody were added to each well in the required dilution of 1:100,000 in conjugate diluent (50 mM Tris, pH 8.0, 0.14 M NaCl, 1% BSA, 0.05% Tween 20) for an incubation of 60min at RT. Plates were washed 5 times with

ELISA wash buffer (50 mM Tris, pH 8.0, 0.14 M NaCl, 0.05% Tween20). Equal volumes of the two-substrate reagents TMB peroxidase substrate (Kierkegaard & Perry Laboratories, Gaithersburg, MD) and of solution B (Kierkegaard & Perry Laboratories, Gaithersburg, MD) were mixed and 100 µL of this mixture were added to each well. Plates were incubated at RT for 5-30min until colour development and stopped by the addition of 100 µL 2 M H_2SO_4 to each well, added in the same order as the enzyme substrate. Plates were read at 450nm on the ELx800 automated microplate reader (Bio-Tek Instruments, Inc.).

Protein Microarray production

E. coli culturing and induction was performed in 96well format with slight modifications [19]. Recombinant protein expression was induced either by IPTG (Isopropyl β-D-1-thiogalactopyranoside) or by cultivation of bacterial clones in autoinduction medium (1mL) [21]. Upon cultivation His-tagged recombinant proteins were purified using Ni-NTA (nickel immobilized onto agarose resin via nitrilo triacetic acid) agarose and chosen elution conditions were adopted warranting protein-binding onto ARChip Epoxy coated slides [22]. Elution of His-tag protein was done using 500 mM imidazole.

Purified proteins were electrophoresed and analyzed using standard procedures. Protein eluates from Ni-metal-chelate purification were controlled for specificity via a His-Tag antibody ELISA. Protein antigens were printed in triplicates on ARChip Epoxy glass slides. Crude clarified protein extracts of the *E. coli* host was used for positive control spots, plain buffer spots were used as negative controls.

Protein arraying

Protein antigens were spotted using an Omnigrid arrayer (GeneMachines) with SMP 3 pins (TeleChem International Inc.) under adjusted air humidity of between 55% and 60%. Spots were printed in triplicates on ARChip Epoxy [22] glass slides. Crude clarified protein extract of the *E.coli* host was used for positive control spots, plain buffer spots were used as negative controls.

Quality control of the spotting process

The recombinant protein expression, the spotting process and the immobilization of protein on the chip surface were controlled by binding of an anti-Penta-His Alexa Fluor 647 conjugate antibody recognizing groups of five histidines in the Hexa-His-tag. To the blocked slides, detection antibody was added in a 1:50,000 dilution in blocking buffer. Slides were washed twice in wash buffer and blown dry with filtered air.

Assay protocols

Reactive groups on the slide surface were blocked for 2h in PBST (Phosphate buffered saline with 0.1% Tween 20) blocking buffer with 5% non-fat milk powder. Slides were washed 2 times 5min in PBST wash buffer, rinsed with distilled water and blown dry with filtered air. Arrays were incubated for 1h with patients' sera and control sera diluted 1:10 in blocking buffer. Upon washing twice for 5min in wash buffer, slides were rinsed with distilled water and blown dry with filtered air. Arrays were incubated for 1h with goat anti human IgG detection antibody fluorescently labelled with *Alexa647* dye (Invitrogen), diluted 1:500 in blocking buffer. Following the final washing steps of twice 5min washing in wash buffer, arrays were rinsed with distilled water and dry-blown. Array images were captured using an Axon Genepix 4000A microarray scanner (Molecular Devices).

Data analysis

Fluorescence intensities - medians after subtraction of the local background - were calculated from the scanned array images with the Genepix software (Molecular Devices). Statistical data analysis was performed using R version 2.6.2 [23], BRB-ArrayTools Version: 3.6.0 - Stable Release, limma software package [24] and nearest shrunken centroid algorithm. The nearest shrunken centroid algorithm is used to find out the clusters in the samples using hierarchical clustering methods on expression arrays [25].

Results

Methods optimization

The bacterial wet biomass (30 mg/mL culture) obtained by autoinduction was twice when compared to that of obtained from IPTG cultures. Using Ni-NTA-metal chelate purification the amount of purified protein from 1mL of bacterial culture (autoinduced) was 7-70 ng at an average concentration of about 0.2-0.25 mg in 75 µL of elution-buffer. Protein yields were similar with both methods which points towards similar effectiveness in protein expression. The expression rate of recombinant proteins was 40% for both IPTG and autoinduction as determined by Penta-His antibody ELISA. Although there were some minor differences between different batches of 1 mL bacterial cultures grown in 96well plates and distinct runs of protein-purification it becomes clear that both induction methods of recombinant protein expression were equally successful. The repeated experiments did not show a great difference between the induction strategies. As autoinduction method was easier for technical handling, this method was chosen for the continuation of the experiments. It was found out that the amount of His-tag protein yielded upon elution using 250 and 500mM imidazole to be more or less the same. Eventually, 500mM imidazole was used for elution of all proteins which were used for microarray printing.

Optimizations for processing the protein-arrays covered the 1) blocking-reagent, 2) serum-incubation time, and 3) detection of serum-auto-antibodies using anti-humanIg-Alexa647 conjugate. Addition of 5% non-fat- dry milk into PBST was efficient when blocking slides for 30 min at room temperature. Prolonged blocking did not significantly increase signal to noise ratios. Although bovine serum albumin has been described for blocking, milk powder is an efficient and an inexpensive alternative. Omission of the milk powder, however led to strong unspecific binding of serum-proteins to the microarray and thereby to high background signals.

The serum-incubation time was tested with respect to signal intensities of microarray spots. Using a 1:10 dilution of sera from healthy controls in PBS (Phosphate buffered saline) signal intensities reached a plateau after 2h incubation at room temperature. Signal intensities upon 4h incubation were comparable to intensities after 2h incubation. Therefore, a 2h

serum-incubation step was used for all further tests and found to give sufficiently high signals to identify clear and distinct auto-antibody patterns from controls as well as from patient's samples. Negative control experiments conducted without serum excluded unspecific binding of the anti-human Ig-Alexa647 detection antibody and indicate that the detection step is specific to human auto-antibodies. Thus, spot signals are derived from specific serum antibody-binding with the antigens presented on the chip. No direct correlation of the amounts of spotted protein with the yielded signal strengths could be detected. Hence, the yielded signals are due to the presence of antibodies specific to the target proteins and not to unspecific binding which would clearly correlate with protein mass. Positive control spots of *E. coli* crude protein extracts showed high signals indicating the presence of high levels of antibodies against *E. coli* proteins in the sera of all donors and patients, whereas buffer spots serve as controls were clearly negative. Thus, false positive-signals derived from carry-over of reactive proteins from printing spots with the same set of pins during microarray fabrication can be excluded. Here, it has to be mentioned that recombinant proteins derived from single step Ni-His(6)-affinity protein purification are not pure and will contain several percentages of *E. coli* proteins, which could be problematic when covering specific signals. This, however, might be especially true for primary-screens to identify specific antigens from clone libraries. In this experimental setting, all clones used for antigen-purification and microarray fabrication were selected via several pre-screens within the SEREX procedure. Reactivity of the spotted proteins was not covered by serum-reactivity against remaining *E. coli* proteins, therefore, the microarray enabled a specific and clear differentiation between sera derived from lung cancer and brain cancer patients.

Serum samples have been initially tested during the SEREX-membrane screen. Some subsets of clones which have been tested positive with several patient sera were used for evaluation of protein microarrays. On the chip, binding of auto-antibodies to the candidate marker proteins was observed as demonstrated in the initial SEREX screens (Fig. 1). Binding events detected in addition to the marker candidates that were expected from macro-membrane screens (in analogy to SEREX screens) indicate greater detection limit/signal intensity of the microarray when compared to the membrane method. This might be due to the smaller reaction surfaces and better distribution of the serum-sample over the array (Fig. 1). Moreover, this greater detection limit/signal intensity is achieved with a few microliters of

analyte i.e., 1:10 diluted sera (about 75 µL is sufficient for wetting the entire area of a standard slide).

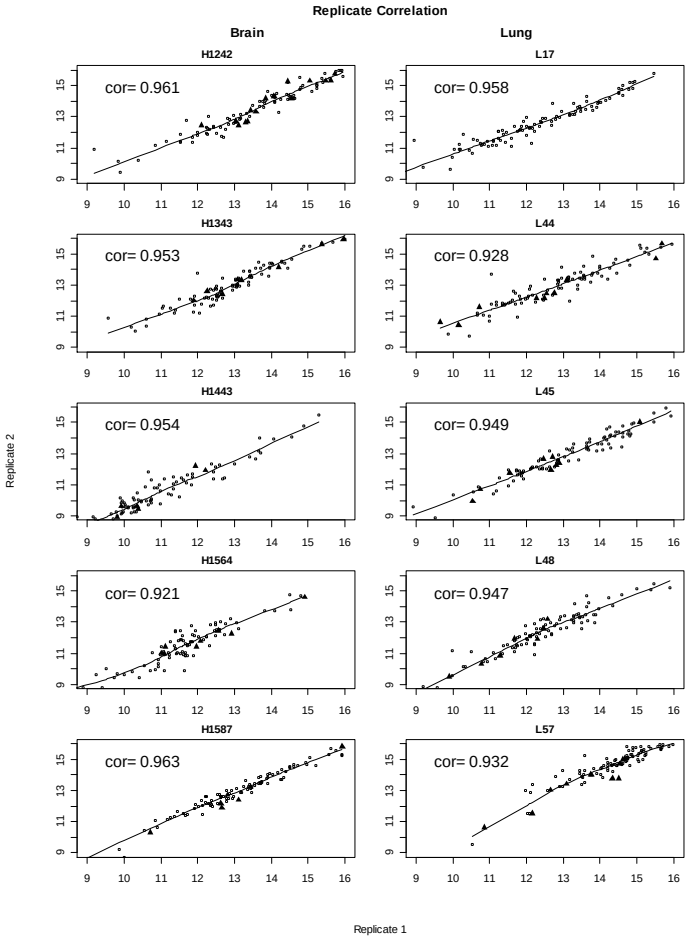

Figure 1: Pair wise correlation of repeated protein microarray analyses. Pair wise correlation of repeated analyses of serum samples. Log2 transformed unnormalised intensities, with a threshold set to 512 intensity-units (derived from Genepix .gpr files) were used for analyses. Correlation coefficients are given in the paired scatterplots and were above 0.92 upon repetitive analyses. The "filled triangles" represent reactive clones from each individual serum found within membrane-based macroarray testing. Data from repetitive analyses (replicate-1 on x-axes; replicate-2 on y-axes) microarray analyses using serum from brain (left) and lung (right) tumour patients (identifiers of different patient sera on top of each scatter plot) are plotted.

Performance of microarray based serum-auto-antibody testing

Detection using anti-human Ig-Alexa647 conjugate upon application of patients' sera to the blocked arrays yielded clearly visible binding patterns that already at an optical level displayed almost identical patterns. Also the control sera yielded specific patterns, yet different to the ones of patient's sera. Pair-wise correlation plots of repetitive serum-testing

on different slides confirmed the high reproducibility of the signal patterns and results in correlation coefficients ranging from 0.92 to 0.96 of (Figure 1). Statistical data analyses of brain and lung cancer serum-microarray data was performed in analogy to gene-expression microarray data analysis using the limma software package. Figure 2 shows the normalised signal intensities of the three most differentially reactive clones between brain and lung. It shows that across replicate measurements the assay is capable of distinguishing these two biological classes. For these genes, both technical variances and within-group variances are small, compared to the between-group variances. Therefore it is not surprising, that when attempting to build a classification rule using the nearest shrunken centroid algorithm on replicate-1 and testing this rule on replicate-2, all samples are classified correctly. In the reverse case (building the rule on replicate-2 and testing it on replicate-1), one sample is misclassified. The good separation between the two classes is also visualised in Figure 3.

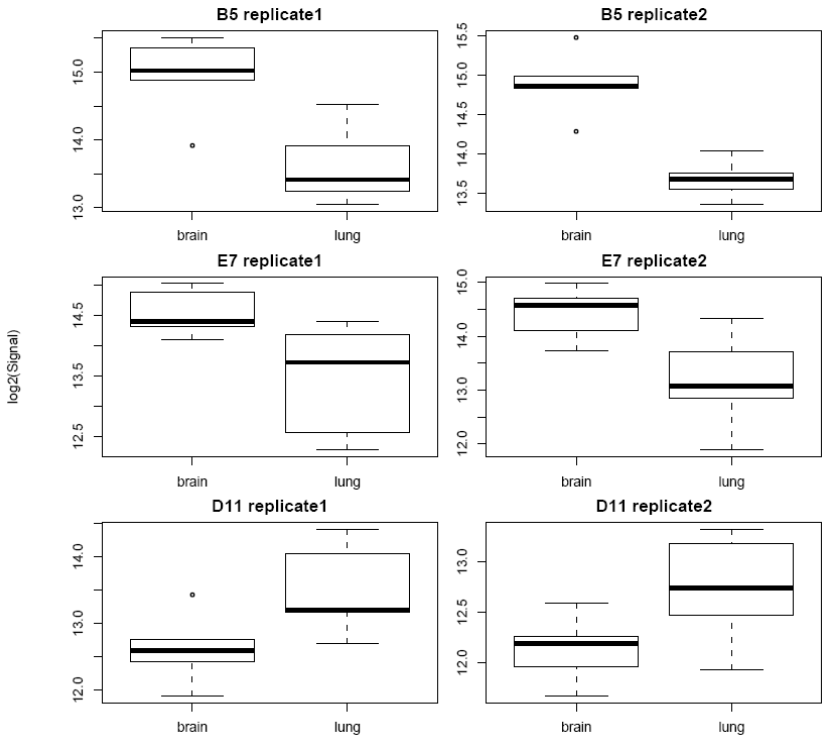

Figure 2: Replicate measurements of the top three differentially reactive clones. Performance of the top three differentially reactive clones (B5, E7 and D11 with p-values less than 0.002, 0.01, 0.05, respectively) in replicate experiments. The normalized signal intensity values of these reactive clones across the replicate measurements (replicate-1 and -2) distinguishes between brain (n=5) and lung (n=5) cancer serum samples.

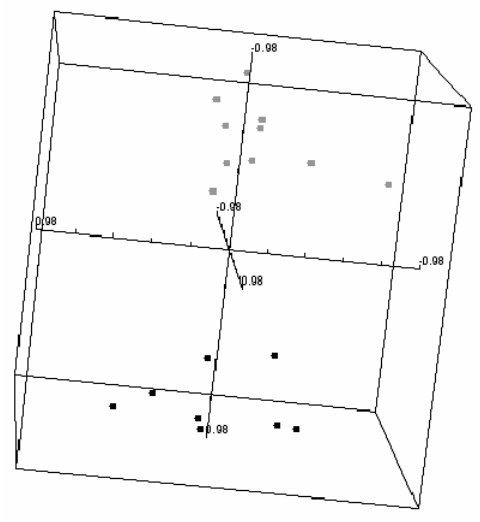

Figure 3: Multidimensional scaling of protein microarray data of brain and lung tumour patients' sera. Multidimensional scaling using centered correlation of significant antigens derived from class comparison (using 10 most significant antigens). Microarray data as depicted in scatterplots (figure 1) of duplicate analysis of brain (black) and lung (gray) cancer patients' serum samples were used for class comparison.

Discussion

Auto-antibodies are very potent biomarkers which would be useful for minimal invasive testing for early diagnosis of autoimmune and cancerous disease. Besides SEREX-based screening using immobilized expression clones on membranes, macroarrays with several thousand expression-clones derived from human cDNA libraries are suitable platforms for screening for determining reactive clones over-expressing proteins which are biomarker-candidates [20]. However, the membrane based clones are not a versatile tool for validation of those candidate makers. Drawbacks of membrane based screening are low reproducibility, low dynamic range of signal intensities, and difficulties in handling membranes. In addition to these technically problems, several hundred microliters of patient serum for processing the membranes or macroarrays are required. Because sample size of clinically well documented samples is always limited, miniaturization of assays using microarrays would be a great option to save samples (about 75 µL of sample is sufficient to cover an entire 1x2 inch standard slide). As known from the performance of DNA-microarrays, obtaining high reproducibility, high dynamic range of intensity-measures (usually in the range of 4-6 orders of magnitudes; derived form 16 or 20 bit microarray

scanners) and easy handling microarrays, protein microarrays would be a potential alternative for validation of disease specific serum-auto-antibody profiles.

The aim of this experiment was to generate protein microarrays and evaluate their performance with respect to technical aspects like reproducibility and suitability using patient serum samples already used for candidate biomarker screening. Therefore, the establishment of techniques and the optimizations were done respect to 1) recombinant protein expression from candidate clones, 2) protein-purification in a 96well standard plate format and microarray printing, and 3) finding best conditions of serum-testing on antigen-microarrays. It was found during optimization of protein microarray production that proteins concentrations of up to 0.5 mg/mL are well suited for spotting using a contact spotter. At that protein-/antigen-concentrations microarrays perform well (with respect to signal intensities and spot morphology) and at that concentration clogging of microarrayer-pins is also avoided.

The protein concentrations were measured upon purification using His-tag/Ni-affinity and the determination of specific recombinant proteins was done using a His-tag-ELISA. Although, the comparison between the microarrays generated from different batches of protein-purification was not done, the ratio of His-tag-ELISA signals and protein-concentration would be a practicable measure of “purity” which should be taken into consideration when using different protein-batches for microarray generation. The “different slide batch effect” is known also from DNA-chips, this would be clearly more critical using proteins derived from different batches of clone-cultivation, expression and purification. Therefore, while using (protein) microarrays for screening purposes defining biomarkers would be done best when using the same batch of microarrays avoiding these effects. When not avoidable that must be considered by proper experimental planning. It was observed that the membrane-blot derived classifiers (which enabled distinction of brain and lung tumour serum antibody profiles) did perform well also on the microarray-derived data set, confirming the reliability of the reactive markers. This is true even when data were derived from two entirely distinct methodologies. Membrane blots are generated by fixation of proteins upon growing *E. coli* clones on membranes and microarrays are spotted using proteins from distinct 1ml culturing of clones. Optimized conditions for obtaining maximum signal intensities on microarrays were achieved with 1:10 serum dilutions after 2h incubation

at room temperature. Thus arrays covering an entire standard slide can be processed with only 10µL of serum. This would enable paralleled detection of about 20000 different spots, a spot density usually achieved with standard microarray printing techniques. This experiment also elucidated a high reproducibility of protein microarray-data. Correlation coefficients of repeated analyses using patient sera were in the range of 0.919-0.971 (median 0.957). While the differentially reactive clones identified in this study need more independent testing to prove their usefulness as clinical markers, this experiment has shown that the assay is capable of detecting differences between biological groups which are stable and reproducible and are therefore suitable for class comparison and class prediction. Thus, this kind of microarrays has several advantages over macroarrays and microarray based testing of patient samples is the method of choice for highly paralleled auto-antibody testing. Especially when many different samples have to be processed for validation of biomarker candidates, handling many microarrays is much easier and also for screening approaches microarrays are best suited and will replace membrane-based macroarray screens. Biostatistical analysis of high-dimensional data derived from microarray-feature intensities is also well established and can be used in the analysis of auto-antibody data.

Conclusion

In conclusion, this study successfully demonstrated the feasibility for auto-antibody identification technology by means of recombinant protein expression and arraying the proteins on microarray solid supports. Because panels of auto-antigens rather than individual antigens enhance the likelihood of detecting cancer antigens with diagnostic potential [26-28], highly paralleled detection of auto-antibody signatures yielded from this platform will be aiding disease diagnosis and improve patient stratification [20,29].

Acknowledgements

I want to thank Michael Stierschneider, Karin Fohringer and Agnes Burg (AIT) for their help with the pipetting robot during the *E. coli* culturing and protein isolation.

Reference list

1. Committee on Developing Biomarker-Based Tools for Cancer Screening DaT: *Cancer Biomarkers: The Promises and Challenges of Improving Detection and Treatment*. The National Academic Press; 2009.

2. Li Y, Lee HJ, Corn RM: **Detection of protein biomarkers using RNA aptamer microarrays and enzymatically amplified surface plasmon resonance imaging.** *Anal Chem* 2007, **79:** 1082-1088.

3. Ferreira CS, Matthews CS, Missailidis S: **DNA aptamers that bind to MUC1 tumour marker: design and characterization of MUC1-binding single-stranded DNA aptamers.** *Tumour Biol* 2006, **27:** 289-301.

4. Vicens MC, Sen A, Vanderlaan A, Drake TJ, Tan W: **Investigation of molecular beacon aptamer-based bioassay for platelet-derived growth factor detection.** *Chembiochem* 2005, **6:** 900-907.

5. Casiano CA, Mediavilla-Varela M, Tan EM: **Tumor-associated antigen arrays for the serological diagnosis of cancer.** *Mol Cell Proteomics* 2006, **5:** 1745-1759.

6. Anderson KS, LaBaer J: **The sentinel within: exploiting the immune system for cancer biomarkers.** *J Proteome Res* 2005, **4:** 1123-1133.

7. Wang X, Yu J, Sreekumar A, Varambally S, Shen R, Giacherio D *et al*.: **Autoantibody signatures in prostate cancer.** *N Engl J Med* 2005, **353:** 1224-1235.

8. Sanchez-Carbayo M: **Antibody arrays: technical considerations and clinical applications in cancer.** *Clin Chem* 2006, **52:** 1651-1659.

9. Robinson WH, Steinman L, Utz PJ: **Protein arrays for autoantibody profiling and fine-specificity mapping.** *Proteomics* 2003, **3:** 2077-2084.

10. Gunawardana CG, Diamandis EP: **High throughput proteomic strategies for identifying tumour-associated antigens.** *Cancer Lett* 2007, **249:** 110-119.

11. Heubeck B, Wendler O, Bumm K, Schafer R, Muller-Vogt U, Hausler M *et al*.: **Tumor-associated antigenic pattern in squamous cell carcinomas of the head and neck - Analysed by SEREX.** *Eur J Cancer* 2006.

12. Tureci O, Sahin U, Zwick C, Neumann F, Pfreundschuh M: **Exploitation of the antibody repertoire of cancer patients for the identification of human tumor antigens.** *Hybridoma* 1999, **18:** 23-28.

13. Krebs P, Kurrer M, Sahin U, Tureci O, Ludewig B: **Autoimmunity seen through the SEREX-scope.** *Autoimmun Rev* 2003, **2:** 339-345.

14. Zhong X, Ran YL, Lou JN, Hu D, Yu L, Zhang YS *et al*.: **Construction of human liver cancer vascular endothelium cDNA expression library and screening of the endothelium-associated antigen genes.** *World J Gastroenterol* 2004, **10:** 1402-1408.

15. Tomaino B, Cappello P, Capello M, Fredolini C, Ponzetto A, Novarino A *et al*.: **Autoantibody signature in human ductal pancreatic adenocarcinoma.** *J Proteome Res* 2007, **6:** 4025-4031.

16. Weiner H, Faupel T, Bussow K: **Protein arrays from cDNA expression libraries.** *Methods Mol Biol* 2004, **264:** 1-13.

17. Diesinger I, Bauer C, Brass N, Schaefers HJ, Comtesse N, Sybrecht G *et al*.: **Toward a more complete recognition of immunoreactive antigens in squamous cell lung carcinoma.** *Int J Cancer* 2002, **102:** 372-378.

18. Comtesse N, Zippel A, Walle S, Monz D, Backes C, Fischer U *et al*.: **Complex humoral immune response against a benign tumor: frequent antibody response against specific antigens as diagnostic targets.** *Proc Natl Acad Sci U S A* 2005, **102:** 9601-9606.

19. Büssow K, Cahill D, Nietfeld W, Bancroft D, Scherzinger E, Lehrach H *et al*.: **A method for global protein expression and antibody screening on high-density filters of an arrayed cDNA library.** *Nucleic Acids Res* 1998, **26:** 5007-5008.

20. Büssow K, Nordhoff E, Lübbert C, Lehrach H, Walter G: **A human cDNA library for high-throughput protein expression screening.** *Genomics* 2000, **65:** 1-8.

21. Studier FW: **Protein production by auto-induction in high density shaking cultures.** *Protein Expr Purif* 2005, **41:** 207-234.

22. Preininger C, Bodrossy L, Sauer U, Pichler R, Weilharter A: **ARChip epoxy and ARChip UV for covalent on-chip immobilization of pmoA gene-specific oligonucleotides.** *Anal Biochem* 2004, **330:** 29-36.

23. R Development Core Team: R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing 2005

24. Simon R, Lam A: BRB-ArrayTools User Guide, version 3.2. Biometric Research Branch, National Cancer Institute. http://linus.nci.nih.gov/brb. Biometric Research Branch, National Cancer Institute. 2009.

25. Tibshirani R, Hastie T, Narasimhan B, Chu G: **Diagnosis of multiple cancer types by shrunken centroids of gene expression.** *Proc Natl Acad Sci U S A* 2002, **99:** 6567-6572.

26. Shoshan SH, Admon A: **Novel technologies for cancer biomarker discovery: humoral proteomics.** *Cancer Biomark* 2007, **3:** 141-152.

27. Fernandez MF: **Autoantibodies in breast cancer sera: candidate biomarkers and reporters of tumorigenesis.** *Cancer Lett* 2005, **230:** 187-198.

28. Amos MD, Bridges AJ: **Standards for autoantibody testing; addressing future needs for autoimmune disease and cancer diagnosis.** *Cancer Biomark* 2005, **1:** 221-227.

29. Caron M, Choquet-Kastylevsky G, Joubert-Caron R: **Cancer immunomics using autoantibody signatures for biomarker discovery.** *Mol Cell Proteomics* 2007, **6:** 1115-1122.

Chapter 2: Evaluation of auto-antibody serum biomarkers for breast cancer screening and *in silico* analysis of sero-reactive proteins

---

Submitted in *J Mol BioChem*. Syed P *et al*. **Evaluation of auto-antibody serum biomarkers for breast cancer screening and *in silico* analysis of sero-reactive proteins**. Under review April 2012.

Keywords: auto-antibody biomarker, breast cancer, protein microarray, tumour-associated antigens

Abstract

Aberrantly expressed proteins in tumours evoke an immunological response. These immunogenic proteins can serve as potential biomarkers for the early diagnosis of cancers. A candidate marker screening was performed on macroarrays containing 38,016 human proteins, derived from a human foetal-brain expression library, with the pools of sera from breast cancer patients (1 pool of benign samples and 3 pools of ductal carcinoma and 2 pools of lobular carcinoma) and 1 pool of sera from healthy women. A panel of 642 sero-reactive clones were deduced from these macroarray experiments which include 284 in-frame clones. Over-representation analyses of the sero-reactive in-frame clones enabled the identification of the sets of genes over-expressed in various pathways of the functional categories (KEGG, Transpath, Pfam and GO). Protein microarrays, generated using the His-tag proteins derived from the macroarray experiments, were used to evaluate the sera from breast cancer patients (24 malignant, 16 benign) and 20 control individuals. The PAM algorithm elucidated a panel of 50 clones which enabled correct classification prediction of 93% of the breast-nodule positive group (benign & malignant) sera from healthy individuals' sera with 100% sensitivity and 85% specificity. This was followed by over-representation analysis of the significant clones derived from class prediction.

Background

Within the European countries, in 2008, there were estimated 3.2 million new cases of cancer and 1.7 million cancer related deaths. Out of the 1.7 million cancer cases, 129,000 cases (7.5% of all forms of cancer) were diagnosed with breast cancer [1]. Therefore, there is a great anticipation to identify novel biomarkers for diagnosing breast cancer.

An immunological response can be evoked by a mutated or an aberrantly expressed protein resulting in the production of auto-antibodies. In the context of cancer, these immunogenic proteins are known as tumour-associated antigens (TAA). The corresponding tumour-auto-antibodies could be used as biomarkers for early diagnosis and prognosis of cancer [2-4]. Proteins like ANXA11, p53, HIP1 and ECPKA are known to serve as TAA biomarkers for various cancers [5-8]. Tomaino *et al*. [9] used Western blot analysis to identify auto-antibodies against pancreatic ductal adenocarcinoma (PDAC) associated antigens from the PDAC sera. Various studies elucidated a range of TAAs in breast cancer, such as MUC1, HSP90, HER2/neu, c-myc, NY-ESO-1/LAGE-1 and Lipophilin B [10-13]. However, it has been shown that measurement of a single TAA is neither sensitive nor specific enough to be used as a diagnostic biomarker. Assessment of auto-antibodies to a tailor-made panel of TAAs may have a promising diagnostic potential [14]. Various studies have reported the panels of TAAs which differentiated the breast cancer patients from healthy controls with higher specificity but low sensitivity (Tab. 1).

Table 1. TAA panels identified in breast cancer patients identified from various studies.

TAA/panel of TAA	Sensitivity (%)	Specificity (%)	Study size	Ages (Mean average in years)		Method use	Ref.
ASB-9SERAC1 RELT	80	100	87 patients & 87 controls	n.a		cDNA T7 phage library protein screening with ELISA	[15]
p16 p53 c-myc	43.9	97.6	41 patients & 82 controls	n.a		ELISA	[15,16]
PPIA PRDX2 FKBP52 MUC1 HSP60	73	85	60 primary breast cancer patients, 82 carcinoma in situ patients & 93 controls	55 (Patients)		ELISA	[17]
p53 c-myc HER2, NY-ESO-1 BRCA2 MUC1	64	85	97 patients & 94 controls	Patients (59)	Controls (54)	ELISA	[18]
IMP1 p62 Koc p53 c-MYC cyclin B1 survivin	70	95	64 Chinese patients, 82 healthy Chinese controls & 264 healthy USA controls	n.a		ELISA	[19,20]

For TAA profiling both macro- and microarrays are used. Macroarrays, polyvinylidene fluoride (PVFD) membranes, are spotted with *E. coli* clones expressing recombinant proteins. Using macroarrays spotted with *E. coli* clones (hEx1 library), Ludwig *et al*. [21], could differentiate glioma sera from healthy controls with a specificity and sensitivity of 90.3% and 87.3%, respectively. On the other hand, the microarrays are spotted with purified recombinant proteins. Babel *et al.* [22], used protein microarrays, containing 8000 human GST-tagged proteins, to differentiate sera from 20 colorectal cancer (CRC) patients and healthy individuals. They found that antibodies against PIM1, MAPKAPK3, STK4, SRC, and FGFR4 were found in high abundance in cancer samples and antibodies against ACVR2B were found in abundance in healthy controls [23].

In this chapter, the identification of a panel of 642 sero-reactive clones from a collection of 38,016 recombinant protein expressing clones (hEx1 library [24]) using macroarrays and sera from the breast cancer patients and healthy controls is discussed. After the identification of the panel of sero-reactive clones, “GeneTrail” gene set analysis toolkit was used to find the genes which are significantly over-represented and are accumulated into certain functional categories (Transpath, Pfam and GO). GeneTrail is an efficient software tool which enables a

statistical evaluation of high-throughput genomic or proteomic data sets with regards to the enrichment of functional categories. Furthermore, the genes expressed by the 642 sero-reactive clones were compared to the SEREX (serological expression of cDNA expression libraries) database and their role in cancer is discussed. Using the recombinant proteins derived from the 642 sero-reactive clones, protein microarrays were generated which enabled distinguishing serum samples from breast-nodule positive patients (benign and malignant) and healthy controls.

Methods

Serum samples

Serum samples were obtained after the approval from the patients and healthy women and were stored at -80°C. The study was approved by the Ethics Committee of the Medical University of Vienna and the General Hospital of Vienna (study number: 143/2007). For macroarray experiments, an aliquot (80 µL) of each serum sample was used for the generation of 7 serum pools. For microarray experiments, 60 serum samples (malignant *n*=24; benign *n*=16; healthy *n*=20) were used. The pathological and clinical cohort characteristics of the breast cancer samples can be found in the table 2.

Table 2. Clinical and pathological charecteristics of the sera used in macro- and microarray screenings. Pools 1-7 were used for the macroarray experiments. Pools 1 and 2 consist of sera from patients with benign fibroadenoma and healthy controls, respectively. Pools 3-4 comprise sera from patients with ductal carcinoma, while pools 6 and 7 contain sera from patients with lobular carcinoma. The data enlisted in the columns, Control, Benign and Malignant, are the samples used for microarray experiments.

	Pool 1	**Pool 2**	**Pool 3**	**Pool 4**	**Pool 5**	**Pool 6**	**Pool 7**	**Control**	**Benign**	**Malignant**
Number of samples	10	10	10	10	10	10	10	20	16	24
Median age (years)	43	73	71	57.5	65.5	54	63	77	45[a]	60
Grading[b]										
G1			5			3				6
G2			5	10		7	9			11
G3					10		1			5
P53 Positive			1	1	9	2	1			
Hormone receptor positive										
Her2/neu				2	3		2			8
Oestrogen			10	10	1	10	10			18
Progesterone			10	8		10	9			
pT stage[c](%)										
Tx, Tis, T1; T1a, T1b, T1c, T1mic, T2, T3; T4b			0; 0; 0; 20; 30; 40; 10; 0; 0; 0.	0; 10; 0; 0; 0; 10; 0; 60; 0; 10.	0; 0; 0; 10; 0; 30; 10; 40; 0; 0.	0; 0; 0; 0; 20; 50; 0; 10; 10; 0.	0; 0; 0; 0; 0; 20; 0; 60; 20; 0.			4.17; 4.17; 16.67; 29.17; 4.17; 12.50; 0; 4.17
pN stage[d](%)										
Nx; N0; N1; N1a, N1biv; N1mi; N2a; N3			0; 90; 0; 0; 0; 0; 0; 0.	10; 0; 10; 20; 10; 0; 30; 10	0; 60; 0; 0; 0; 0; 20; 10	0; 90; 0; 0; 0; 0; 0; 0.	0; 0; 10; 50; 0; 10; 0; 0.			20; 50; 10; 10; 10
Menopause status[e]										
Pre-menopause	5		3	2	2	3	1		5	3
Peri-menopause						1				
Post-menopause	1		7	7	7	5	9			18

[a]Data available for 14 patients. [b]Data available for 22 malignant patients used in microarray experiments. G1 (low-grade), G2 (intermediate grade) and G3 (high-grade). Low-grade tumours are usually slow growing and are less likely to spread. High-grade tumours are likely to grow more quickly and are more likely to spread. [c]Data available for 24 malignant patients used in microarray experiments. [d]Data available for all patient of samples (40, Pools 3-6) samples and 9 samples from Pool 7; used in macroarray experiments and data available for 20 Malignant patients. [e]Data available for 47 patient (Pools 3-7) and 6 benign samples (Pool 1); used in macroarray experiments and data available for 26 patients; used in microarray experiments.

Candidate marker screening

Protein macroarrays, containing duplicates of 38,016 clones (hEx1 library) were purchased from RZPD (now Source Bioscience), Germany. The protein features were generated by expression of spotted *E. coli* clones, which harbour an expression vector, pQE30NST. The expressed recombinant proteins are His-Tagged. Duplicate clones are present on a set of 2 macroarrays and the macroarrays were processed according the detailed protocol for membrane processing which can be found on the Source Bioscience homepage (http://www.lifesciences.sourcebioscience.com/media/290406/sbs_ig_manual_proteinarray_v1.pdf).

In a pre-test, the reliability of auto-antibody screening on PVFD membranes containing 38,016 foetal brain proteins was evaluated using the native-serum samples and the IgG purified serum fraction isolated by affinity purification of immunoglobulins. The purification of IgG from the serum was done using Melon™ Gel IgG Purification Kit (Thermo Scientific) and the procedure was followed as per manufacturer's instructions. In this pre-test, an individual serum sample was tested against a pool of 10 healthy control serum samples (including also the single individual sample) with and without Melon™ Gel IgG Purification for making a decision for applying either serum or the affinity enriched Ig-fraction onto the macroarrays.

Based on the results derived from the pre-test it was decided to use the pools of native serum samples to perform a candidate marker screen on PVFD membranes containing 38,016 human proteins derived from hEx1, a human foetal-brain expression library. For having a measure of the reproducibility of the macroarrays, all the membranes were hybridized with a male-serum sample (without any individual or familial breast cancer history). Then the membranes were stripped and blinded duplicates of each pool of patient sera (Pool 3-7) and non-malignant sera (Pool 1 & 2) were applied onto the macroarrays and the data was generated upon signal detection according to the protocol from RZPD, Germany. The selection of the clones was done on the basis of sero-reactivity in all the experiments. A total of 642 sero-reactive clones (after excluding the duplicates) from different screening experiments, were considered for the production of microarrays.

GeneTrail analysis

GeneTrail analysis was done for 284 in-frame clones among the panel of 642 sero-reactive clones. A statistical approach of Over-Representation Analysis (ORA) was followed for the comparison of test set with the reference set ("Heidelberg human foetal brain"), provided by the gene set analysis tool [25,26]. The analyses were performed with the following parameters: Multiple testing adjustment method: false discovery rate (FDR), significance level threshold (α-level): 0.05.

Protein microarray production and processing

E. coli clones were cultured using the autoinduction protocol according to Stempfer *et al.* [27]. Recombinant protein expression was induced by cultivation of *E. coli* clones in autoinduction medium (SB medium) and purified using Ni-NTA agarose (Qiagen) Elution of the His-Tag proteins was done using elution buffer (50 mM KH_2PO_4 and 50 mM K_2HPO_4, pH 8.0, 500 mM imidazole, 0.01% SDS and 0.01% NaN3). Purified His-Tag proteins were then spotted on ARChip Epoxy slides [28]. Each microarray consisted of 4 sub-arrays with protein antigens printed in duplicates. Clarified *E. coli* lysate with a concentration of 0.5 mg/mL was used as positive control and plain buffer spots as negative control. Processing of the protein microarrays was performed as described previously [27]. The processed microarray images were captured using an Axon Genepix 4000A microarray scanner (Molecular Devices, Union City, CA). Fluorescence intensity-medians after subtraction of local background were calculated from the scanned array images and used for the data analysis.

Statistical analysis

The statistical data analysis of the data from the scanned images of macroarrays was performed using R version 2.10.0 [29]. For microarray data analyses in addition to R, BRB-ArrayTools Version: 3.6.0 - Stable Release [30] were also used.

For class prediction, Prediction Analysis for Microarrays (PAM) algorithm was used. PAM algorithm uses "nearest shrunken centroid" method which identifies a subset of significant genes/clones for the best classification of the samples [31]. Cross-validation of the predicted class and the true class was done.

Results

In brief, from the collection of 38,016 cDNA expression clones 642 clones were selected based on their sero-reactivity. Over-representation analysis was performed using 284 in-frame clones. Protein microarrays were generated using the purified proteins from the 642 sero-reactive clones. Using these protein microarrays breast-nodule positive samples could be differentiated from healthy controls. A schematic over-view of the results obtained during the course of the study is shown in the figure 1.

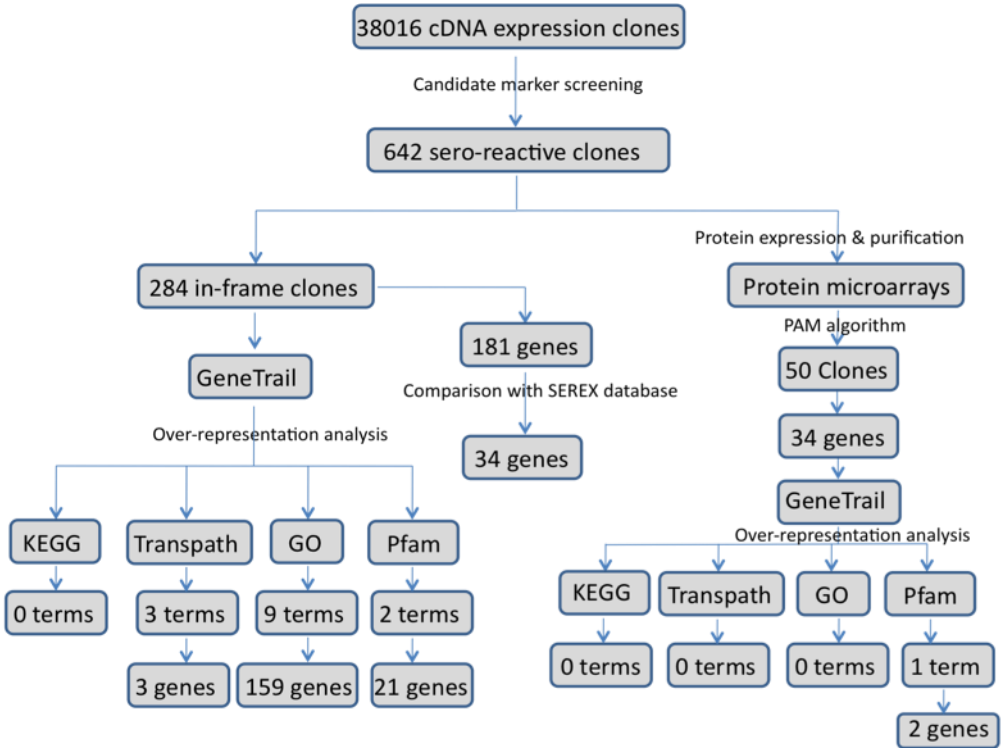

Figure 1. An over-view of the number of clones and genes identified in this study.

Evaluation of purified IgG versus serum for membrane screening

Clones on the membranes which were reactive to native serum samples (Pooled serum samples and single serum sample) and purified IgG (Pooled serum samples and single serum sample) were compared. Signals of duplicate spots were counted as positive signals within the colour-range of 0-4 based on the staining intensity of the spots (Fig. 2). A total of 170 sero-reactive clones were found during this experiment. 32 and 67 clones reacted positively

to pooled purified IgG and native pooled serum samples, respectively. Whereas, 22 and 125 clones were observed reacting positively to purified IgG and native serum sample, respectively (Tab. 3). Based on the number of the clones showing positive reaction, the decision was made to use native sera for the membrane screening.

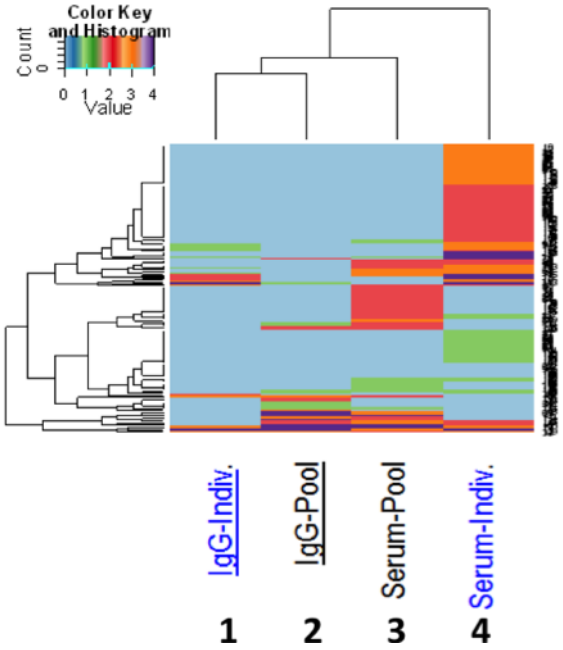

Figure 2. Purified IgG vs native serum Hierarchical clustering of the comparison between native serum ("Serum"- 3&4) versus the purified IgG ("IgG"- 1&2) derived either from a serum-pool (3: "Pool"; from 10 individual sera) or a single sample (4, "Indiv.") also included in the pool. Number of clones showing positive reactivity to purified IgG and native serum sample of an individual were found to be 22 and 125, While, for the purified IgG and sera of the pooled serum samples were 32 and 67, respectively.

Table 3. Number of clones with overlapping reactivity within different samples analysed.

	Purified IgG- Single (1)	Purified IgG- Pool (2)	Native sera- Pool (3)	Native serum- Single (4)
Purified IgG- Single (1)	22	7	11	19
Purified IgG- Pool (2)	7	32	21	11
Native sera- Pool (3)	11	21	67	31
Native serum- Single (4)	19	11	31	125

The numbers (1-4) in the brackets correspond to the lanes in the Fig. 1

Antigen Identification on Macroarrays

Macroarrays were hybridized with pooled samples (pools 1-7) after being processed with single serum-control (reference) and then stripped. Hierarchical clustering results of reference serum sample on different membranes used for sample analysis are shown in Figure 3 (right part of the figure) and the number of the sero-reactive clone from each membrane can be found in the Table 4. The correlation coefficient values derived from the processed membranes with the same reference serum range from 0.68 to 0.98.Analysis of signal intensities derived from the membranes, processed with blinded duplicates (Pools 1-7) was done and sero-reactive clones were identified (Fig. 3 (left part of the figure)). The correlation coefficients of the two runs of each serum pool (Pool 1-7) on macroarrays were found to be ranging from 0.12 to 0.89.

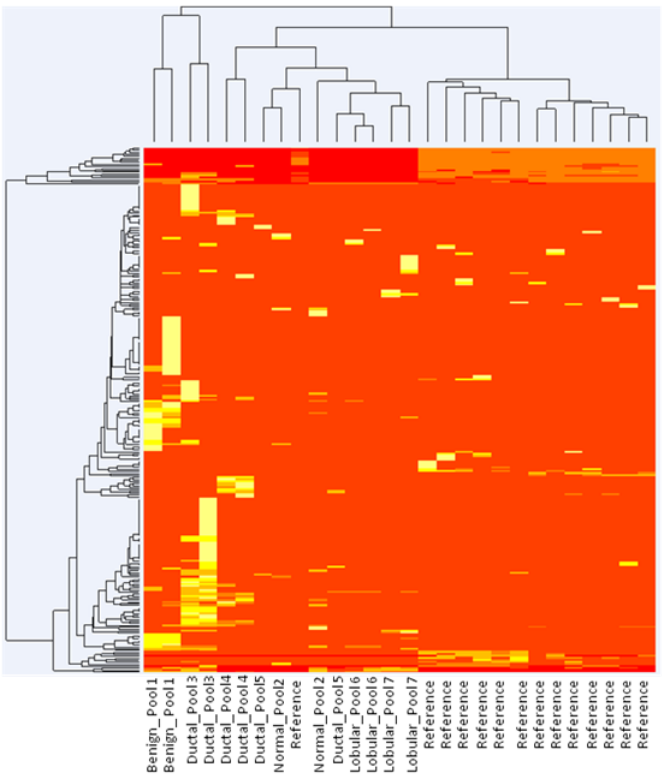

Figure 3 Heat map derived from hierarchical clustering. Macroarrays were treated with a single serum-control (reference) and then stripped and treated with patient serum sample pools. Signal intensities derived from membranes hybridized with sample pools and single serum-control hybridization are depicted in the heat map derived from hierarchical clustering. Duplicate analyses of serum samples and reference are clustering together and are depicted in the heat map as neighbouring lanes.

Table 4: Comparison of sum of positive clones identified from macroarrays. Each membrane was processed with reference sample and these membranes, after stripping, were treated with patient and control serum pools.

Serum	Run	Macroarray 1	Macroarray 2
Reference	Run 1	10	4
	Run 2	23	7
Pool 1	Run 1	21	31
	Run 2	36	42
Reference	Run 1	20	7
	Run 2	18	11
Pool 2	Run 1	16	4
	Run 2	1	8
Reference	Run 1	17	11
	Run 2	20	7
Pool 3	Run 1	63	25
	Run 2	64	17
Reference	Run 1	21	10
	Run 2	18	13
Pool 4	Run 1	22	3
	Run 2	26	3
Reference	Run 1	14	16
	Run 2	15	24
Pool 5	Run 1	6	1
	Run 2	2	2
Reference	Run 1	17	19
	Run 2	16	21
Pool 6	Run 1	8	2
	Run 2	5	2
Reference	Run 1	19	14
	Run 2	16	18
Pool 7	Run 1	6	7
	Run 2	10	15

A total of 1691 sero-reactive clones were found, including the clones identified from the "IgG versus serum" pre-test. Of these 1691 clones, 642 clones were identified as unique clones showing sero-reactivity in all the macroarray experiments. 284 clones out of 642 clones were confirmed (based on DNA sequences of the clones) to be cloned in-frame. Out of the 284 in-frame clones, 71 clones reacted positively to the serum samples from benign breast cancer patients, while 41 and 133 clones showed positive reaction to the serum samples from health control and malignant breast cancer patients, respectively.

All the 642 clones found positive within all the experiments for protein expression were used and thereby using the subsequent proteins for the production of protein microarrays.

In silico analysis of sero-reactive clones

Out of 284 in-frame clones, 181 code for unique proteins. Upon comparison of the 181 genes with 1545 genes from the SEREX database [32], 34 genes were found which were over-lapping between the lists. These 34 genes were reported in the SEREX database from a variety of cancer studies. Among these 34 genes, 7 genes (ALDOA, CENBP, EEF2, GAPDH, MAZ, PRDX1 and TP53) are reported in various cancer studies as TAAs (Tab. 5).

Table 5. Comparison with the SEREX database. The genes encoded by the in-frame clones were compared to genes enlisted in the SEREX database and TAA related published literature

Gene symbol	Cancer study in SEREX db	TAA study-Cancer	Gene symbol	Cancer study in SEREX db	TAA study-Cancer
ACTG1	Colon, Fibrosarcoma		***MARK3***	Prostate	
*ALDOA**	Breast, Lung	Melanoma [33]	***MAZ****	Squamous cell carcinoma, Colon adenomacarcinoma,	Hodgkin's disease [34]
ANKHD1	Renal cell carcinoma, Glioma, Prostate		***MRPS24***	Prostate,	
ATP5B	Malignant fibrous histiocytoma		***PDAP1***	Fibrosarcoma	
BAG5	Melanoma		***PRDX1****	Melanoma	Oesophageal squamous cell carcinoma [35]
CD320	Prostate cancer		***PRKRA***	Testis	
CDC42BPB	Renal carcinoma RCC, thyroid		***RBM5***	Renal cancer	
*CENPB**	Melanoma	Breast cancer [36,37] Small cell lung cancer [38]	***RPL5***	Colon cancer	
CKB	Colon adenocarcinoma		***RPS12***	Renal cell carcinoma	
*EEF2**	Head neck cancer	Melanoma Hepatocellular carcinoma [39]	***RPS13***	Testis	
FDFT1	Fibrosarcoma		***RUFY1***	Prostate cancer, Stomach cancer,	
FKBP3	Stomach cancer, melanoma		***SMARCA4***	Melanoma, Prostate	
*GAPDH**	Breast cancer	Melanoma	***STUB1***	Colorectal adenocarcinoma, Breast carcinoma, Prostate cancer, Ovarian cancer, Glioma	
HIST1H1C	Testis		***TP53****	Colorectal adenocarcinoma, Breast cancer, Colon cancer,	Hepatocellular carcinoma [40] Ovarian cancer [41] Lymphocytic leukaemia [42] Breast cancer [43]
HSPH1	Colorectal adenocarcinoma, Melanoma, Glioma, Lung, pancreas adenocarcinoma,		***TRIM21***	Breast cancer	
IDH2	breast		***TTC3***	Stomach cancer, Glioma, Prostate cancer	
IK	Testis		***ZNF232***	Breast carcinoma	

*antigens against which auto-antibodies have been reported through various cancer studies

Using GeneTrail, *in silico* analysis of the 284 in-frame clone-protein sequences (test set) was done to get the information about functional categories (KEGG, Transpath, Pfam and GO) and their sub-categories, protein families, domains and pathways. The number of genes annotated in the test set to the selected functional categories was found to be 168, out of 284 sequences. While the number of genes annotated in the "Heidelberg human foetal brain" reference set were 3527 (out of 3553). It was found that the observed number of genes involved in cellular processes, various pathways was higher when compared to the expected number of genes. For example, the expected numbers of genes involved in the sub-category "cellular process" were 121 and the observed number of genes was found to be 139 when compared to the reference set, with p-value 0.03. This indicates the over-representation of the genes involved in the respective functional categories in breast cancer. Some of the sub-categories which were enriched in the test set when compared to the reference set are cellular process (GO), wnt pathway (Transpath) and R3H domain (Pfam). Sum of the genes found over-represented in all the enriched subcategories of Transpath, Pfam and GO were found to be 3, 21 and 159, respectively. No sub-category pertaining to KEGG was found enriched in the test set compared to the reference set. A detailed list of sub-categories, the genes encoded by the sero-reactive clones and the number of expected and observed genes are shown in the table 6.

Table 6: Over-represented genes corresponding to the enriched functional categories.

Functional category	Subcategory	P-value	Expected no. Of genes	Observed no. Of genes	Gene symbols
Transpath	CH000000251	0.004	0.11	2	PLK3, TP53
	CH000000255	0.004	0.11	2	PLK3,TP53
	wnt pathway	0.023	0.27	2	CUL1, TP53
Pfam	Zinc finger C2H2 type	0.0005	6.7	19	ZNF33B, ZNF232, ZNF133, ZNF618, ZNF410, ZNF768, ZFP64, ZNF502, ZNF436, ZNF358, ZNF44, ZNF761, BCL11A, ZNF214, ZXDC, BCL11B, ZNF836, MAZ, ZNF238
	Iron only hydrogenase large subunit C-terminal domain	0.02	0.09	2	NARF, NARFL
Gene ontology	Small ribosomal subunit	0.002	1	8	RPS12, RPS3, MRPS24, RPS4X, MRPS11, RPS8, RPS3A, RPS13
	Cell proliferation	0.009	8	21	PDAP1, CUL1, CRIP2, PRDX1, TP53, COL18A1, SRRT, NOP2, AZGP1, HDAC2, RPS4X, RBM5, TP53I11, CCND2, FTH1, ING4, PRKRA, CDC25A, PTN, DCTN2, C19orf10
	Cytosolic small ribosomal subunit	0.009	1	6	RPS12, RPS3, RPS4X, RPS8, RPS3A, RPS13
	Regulation of cell proliferation	0.009	5	16	CUL1, CRIP2, TP53, COL18A1, NOP2, AZGP1, HDAC2, RPS4X, RBM5, TP53I11, CCND2, FTH1, ING4, PRKRA, PTN, C19orf10
	Regulation of cellular process	0.028	63	85	ZNF33B, ZNF232, ZNF133, ZNF618, PDAP1, TNC, CUL1, LPHN1, POU6F1, CRIP2, CDC42BPB, AHSG, TRAF4, ZNF410, C14orf153, SPTBN1, HES5, ARPC2, STUB1, NPTN, ZNF768, ZFP64, ZNF502, SARNP, ZNF436, ZNF358, ZNF44, EIF4EBP3, PRDX1, TP53, ARF1, ZNF761, TPT1, COL18A1, MED8, GLRX3, NOP2, PRKAG1, RPS3, BCL11A, PSIP1, SMARCA1, AZGP1, STK25, HNRPDL, HDAC2,ZNF214, FUT8, RPS4X, TTC3, SRP14, CD320, RBM5, TP53I11, CCND2, CAP1, ATP5B, ZXDC, HDAC6, CHMP5, NARFL, FTH1, SMARCA4, CENPB, YEATS2, ALDOA, ING4, PRKRA, BCL11B, TRIOBP, ZNF207, STMN4, SRI, GIT1, ZNF836, DBN1, CDC25A, PTN, MAZ, TANK, ZNF238, RPS3A, RPS13, JPH3, C19orf10
	Ribosomal subunit	0.03	2	9	RPL7A, RPS12, RPS3, MRPS24, RPS4X, MRPS11, RPS8, RPS3A, RPS13
	Regulation of biological process	0.03	65	87	ZNF33B, ZNF232, ZNF133, ZNF618, PDAP1, TNC, CUL1, LPHN1, POU6F1, CRIP2, CDC42BPB, AHSG, TRAF4, ZNF410, C14orf153, SPTBN1,HES5, ARPC2, STUB1, NPTN, ZNF768, ZFP64, ZNF502, SARNP, ZNF436, ZNF358, ZNF44, EIF4EBP3, PRDX1, TP53, PEX6, ARF1, ZNF761, TPT1,COL18A1, MED8, SRRT, GLRX3, NOP2, PRKAG1, RPS3, BCL11A, PSIP1, SMARCA1, AZGP1, STK25, HNRPDL, HDAC2, ZNF214, FUT8, RPS4X, TTC3, SRP14, CD320, RBM5, TP53I11, CCND2, CAP1, ATP5B, ZXDC, HDAC6, CHMP5, NARFL, FTH1, SMARCA4, CENPB, YEATS2, ALDOA, ING4, PRKRA, BCL11B, TRIOBP, ZNF207, STMN4, SRI, GIT1, ZNF836, DBN1, CDC25A, PTN, MAZ, TANK, ZNF238, RPS3A, RPS13, JPH3, C19orf10
	Negative regulation of cell proliferation	0.05	2	9	CUL1, TP53, COL18A1, AZGP1, RBM5, TP53I11, FTH1, ING4, PRKRA
	Cellular process	0.028	121	139	ZNF33B, PIN4, ZNF232, NDEL1, RPL7A, STX16, EEF2, ZNF133, TTLL1, ZNF618, PDAP1, TNC, CUL1, LPHN1, POU6F1, STK32C, RPS12, CRIP2, CDC42BPB, KIF5B, AHSG, TRAF4, UFC1, ZNF410, C14orf153, SPTBN1, HES5, ARPC2, STUB1, NPTN, ZNF768, ZFP64, IK, ZNF502, PLK3, SARNP, EXOSC10, EBNA1BP2, RUFY1, ZNF436, ZNF358, ZNF44, KIF18B, CLSTN1, NUMA1, EIF4EBP3, PRDX1, GPD1, TP53, MARK3, ZFYVE27, PEX6, ARF1, COQ4, ACTG1, HIST1H1C, ZNF761, PDRG1, TPT1, COL18A1, MED8, SRRT, GLRX3, NOP2, PRKAG1, TUBA1A, RPS3, BCL11A, PSIP1, SMARCA1, AZGP1, STK25, MRPS24, HNRPDL, PSMD10, IDH2, ISY1, HDAC2, FKBP3, ZNF214, FUT8, RPS4X, TTC3, SRP14, EGFL6, EHD1, SNX5, CD320, CKB, RBM5, TP53I11, CCND2, FDFT1, CAP1, ATP5B, LMO7, ZXDC, HDAC6, CHMP5, NARFL, FTH1, SMARCA4, MRPS11, CENPB, YEATS2, PODXL2, TSPAN7, ALDOA, ING4, PRKRA, TRAPPC2L, SFRS4,BCL11B, BIRC5, TRIOBP, ZNF207, BAG5, RNF130, STMN4, SRI, GAPDH, GIT1, ZNF836, DBN1, RPS8, ELAVL3, CDC25A, PTN, MAZ, CCT4, TANK, GRINL1A, DCTN2, GTF3C1, ZNF238, RPS3A, RPS13, JPH3, C19orf10

Protein microarray analysis

BRB-ArrayTools was used to analyze the data derived from the microarrays processed with patient and healthy control sera. Using the PAM algorithm, a panel of 45 significant clones was identified which enabled the classification of benign, malignant and the control samples (Tab. 7). Out of 16 benign breast cancer samples 13 were predicted as malignant and 1 as control samples. Out of 24 malignant samples, 15 were correctly identified as malignant and out of 20 control samples, 15 were identified as healthy controls (Tab. 8).

Table 7. Significant clones which gave 53% correct classification of benign, malignant and control samples.

Clone	Mean of intensities-Benign	Mean of intensities-Malignant	Mean of intensities-Control	Clone	Mean of intensities-Benign	Mean of intensities-Malignant	Mean of intensities-Control
MPMGp800O19569	771	873	2350	**MPMGp800C12577**	4612	6308	9372
MPMGp800A04578	661	988	2089	**MPMGp800D07572**	2779	3842	5680
MPMGp800B23591	3413	6361	8182	**MPMGp800J03569**	2577	3938	5595
MPMGp800C05534	2620	3276	6120	**MPMGp800K23566**	1159	1388	2249
MPMGp800J12588	2811	4769	7704	**MPMGp800K09596**	486	803	931
MPMGp800P03549	713	1251	1824	**MPMGp800B14547**	4080	6496	7714
MPMGp800E06562	1821	3146	4255	**MPMGp800E07573**	1070	1455	2322
MPMGp800N08514	2117	2684	5100	**MPMGp800J14559**	1241	1816	2795
MPMGp800C13512	6892	11924	14827	**MPMGp800M16590**	633	987	1310
MPMGp800M18568	3546	5284	10265	**MPMGp800B07542**	3237	4954	6160
MPMGp800G21543	582	1036	1178	**MPMGp800H10585**	2021	3191	3776
MPMGp800P10579	2721	5109	5867	**MPMGp800A04595**	962	1231	1891
MPMGp800D20603	5286	7078	10918	**MPMGp800C23586**	2583	3625	5031
MPMGp800P20514	1056	1685	2472	**MPMGp800L24584**	1432	2119	3074
MPMGp800C17586	2674	3354	5335	**MPMGp800M08589**	2304	3599	4731
MPMGp800B14528	3876	5321	8507	**MPMGp800D21597**	1291	1624	2542
MPMGp800D08553	5203	5688	9919	**MPMGp800C08590**	3648	5139	7301
MPMGp800C09514	3077	4065	6247	**MPMGp800C23548**	5102	7799	9416
MPMGp800P21572	529	908	1009	**MPMGp800C06602**	9398	9864	16845
MPMGp800A09563	5505	8596	11100	**MPMGp800C18590**	2327	3270	4706
MPMGp800O17527	470	790	882	**MPMGp800E01587**	1156	1817	2295
MPMGp800G21537	708	1016	1513	**MPMGp800L04581**	2651	2609	4436
MPMGp800H01584	650	1130	1258				

Table 8. Prediction of classes (Benign, Malignant and Control) using the classifier from PAM algorithm. A cross-tabulation of the classes in rows (true) versus columns (predicted) and the corresponding sensitivity, specificity, positive predictive value (PPV) and negative predictive value (NPV) is shown in the table

Class	Benign	Malignant	Normal	Sensitivity	Specificity	PPV	NPV
Benign	2	13	1	0.125	0.932	0.4	0.745
Malignant	2	15	7	0.625	0.556	0.484	0.69
Normal	1	4	15	0.75	0.8	0.652	0.865

Since majority of the benign samples were classified as malignant samples, the decision was made to compare the breast-nodule positive samples with the healthy controls. This comparison yielded 50 significant clones which enabled classification of breast-nodule positive samples and healthy controls (Tab.9). These significant 50 clones gave 93% correct classification prediction of breast-nodule positive sera from normal sera with 100% sensitivity and 85% specificity. 4 out of 16 control samples were predicted as breast-nodule positive, while all the 40 breast-nodule positive samples were correctly predicted as breast-nodule positive (Tab. 10).

Table 9. Classifier clones derived from PAM which correctly classified breast-nodule samples from healthy controls.

Clone	Gene	E value	Control-Intensities	Breast-nodule intensities	Ratio of intensities	Clone	Gene	E value	Control-Intensities	Breast-nodule intensities	Ratio of intensities
MPMGp800O19569	*YBX1*	0	1793	853	2.1	**MPMGp800E07573**	*NUBP2*	0	1772	1321	1.3
MPMGp800A04578	*HIP1R*	0	1594	865	1.8	**MPMGp800K01579**	*YBX1*	0	1626	2150	0.8
MPMGp800I19548	*YBX1*	0	1747	2831	0.6	**MPMGp800M08528**[#]	*PODXL2*	0	983	1353	0.7
MPMGp800F12540	*PRPF19*	0	975	1405	0.7	**MPMGp800F17571**	*RDBP*	0	1500	1940	0.8
MPMGp800H22523	*RBM10*	0	684	1043	0.7	**MPMGp800G17568**	*RNF187*	5E-119	818	1040	0.8
MPMGp800P06511	*YBX1*	0	1303	2006	0.6	**MPMGp800K23566**	*H2AFY*	0	1716	1326	1.3
MPMGp800I18557	*CPLX2*	0	898	1524	0.6	**MPMGp800L15517***			680	837	0.8
MPMGp800C05534*			4670	3077	1.5	**MPMGp800B14528**[#]	*LPHN1*	0	6491	4815	1.3
MPMGp800N23548	*YBX1*	0	1684	2591	0.6	**MPMGp800D08553**	*EIF3C*	0	7569	5638	1.3
MPMGp800P01595*			514	744	0.7	**MPMGp800K10577**[#]	*JUP*	0	1280	1624	0.8
MPMGp800J06581	*YBX1*	0	2020	2966	0.7	**MPMGp800F05518**[#]	*SPAG7*	0	3939	5041	0.8
MPMGp800H22512[#]	*CENPB*	0	2616	3923	0.7	**MPMGp800C17586**	*EEF2*	0	4071	3146	1.3
MPMGp800K07565	*YBX1*	0	1993	2936	0.7	**MPMGp800K22574**[#]	*SNX5*	0	979	1175	0.8
MPMGp800H05540	*OSBPL7*	0	6652	9692	0.7	**MPMGp800H22541**[#]	*HIST1H1C*	0	1910	2449	0.8
MPMGp800I15594	*ARPP21*	0	1873	2618	0.7	**MPMGp800M24582**	*PRDX1*	0	1323	1004	1.3
MPMGp800I07520[#]	*SRRT*	0	1743	2575	0.7	**MPMGp800O02506**[#]	*SPAG7*	0	5536	6937	0.8
MPMGp800P13536	*H2AFY*	0	4062	6036	0.7	**MPMGp800O13595**[#]	*PKM2*	0	725	608	1.2
MPMGp800P08541[#]	*TANK*	0	846	1187	0.7	**MPMGp800C06602**	*CKB*	0	12854	9937	1.3
MPMGp800N08514	*MAZ*	0	3892	2507	1.6	**MPMGp800H07541**	*MAZ*	0	6275	4741	1.3
MPMGp800L16562[#]	*SPAG7*	0	6575	9168	0.7	**MPMGp800M05558**	*PIM3*	0	4082	3033	1.3
MPMGp800G16536*			968	1283	0.8	**MPMGp800J24571**	*CBLL1*	0	1175	1354	0.9
MPMGp800M18568[#]	*MAZ*	0	7833	4627	1.7	**MPMGp800N11538**	*C16orf13*	1.63E-135	1043	1211	0.9
MPMGp800N14581[#]	*RPS3A*	0	576	774	0.7	**MPMGp800E06542**[#]	*MAZ*	0	16454	13877	1.2
MPMGp800J12588[#]	*SMARCA1*		5879	3965	1.5	**MPMGp800M08567**	*EPB41L3*	0	1042	920	1.1
MPMGp800G05508	*AKR7A2*	0	1315	1736	0.8	**MPMGp800K16540***			7281	6380	1.1

[#]In-frame clones. *****Clones whose sequences were not available

Table 10. Prediction of classes (Breast-nodule positive and Control) using the classifier from PAM algorithm. A cross-tabulation of the classes in rows (true) versus columns (predicted) and the corresponding sensitivity, specificity, positive predictive value (PPV) and negative predictive value (NPV) is shown in the table

Class	Control	Breast-nodule positive	Sensitivity	Specificity	PPV	NPV
Control	16	4	0.8	1	1	0.909
Breast-nodule positive	0	40	1	0.85	0.93	1

From the lists derived from PAM algorithm, 12 clones were found significant in both the lists. These lists of significant clones were compared to the list of positively reacting clones to breast-nodule positive sera and healthy control sera. 40 clones were found reacting positively to the breast-nodule positive sera and 9 clones were reacting positively to the healthy control sera, exclusively. 14 clones reacted positively to the sera from the patients and the controls.

GeneTrail was used to find the set of genes, among the 34 genes encoded by the 50 significant clones (which gave 93% correct classification prediction), which are over-represented in the functional categories like KEGG, Transpath, Pfam and GO, with "Heidelberg human foetal brain" as reference sets. The parameters for the analyses were the same as used for the analysis of the 284 in-frame clones. The number of genes found annotated within the test set of 43 genes for KEGG, Transpath, GO and Pfam were found to be 7, 1, 27 and 26, respectively. However, no genes related to any of the KEGG, Transpath and GO were found to be over-represented in the test set when compared to the reference set. 2 genes, ARPP21 and SPAG7 were found to be over-represented in R3H domain sub-category of Pfam (p-value 0.001). The expected number of genes was 0.05 while the observed number of genes was 2.

Discussion

Over the years, macroarrays spotted with cDNA expression clones, have been used for TAA profiling. Macroarrays, spotted with hEx1 cDNA expression library clones, have been used

for the identification of auto-antibodies patients with glioma, chronic obstructive pulmonary disease (COPD) and Wilm's tumour [44,45]. Auto-antibodies are known to be observed in the serum prior to the onset of breast cancer, lung cancer and prostate cancer [46-48]. This gives possibility of using the auto-antibodies as serological tools for the early diagnosis and management of cancer.

These macroarrays were used for identifying a panel of 642 sero-reactive clones from a collection of 38,016 cDNA expression clones. An initial experiment was conducted to check the performance of the macroarrays when hybridized with purified IgG and native serum. It was observed that the number of positive clones was higher when using native sera, compared to purified IgG. Based on the number of clones on the macroarrays, reacting positively to the sera/serum and purified IgG the decision was made to use native serum samples for TAA profiling.

To check the reproducibility of the macroarrays, a reference serum was hybridized on the macroarrays which were then stripped and hybridized with blinded duplicates of serum pools from breast cancer patients and healthy controls (Pools 1-7). Blinded duplicates of the serum pools were used to avoid the experimental biasness. Signal intensities derived from the sero-reactive clones were used for hierarchical clustering. Although, the results from the single control serum analysed on every single membrane did cluster in a distinct tree, the sum of the positive-clones detected from each pool in both of the repeated analyses did not cluster with respect to the sample groups "normal", "benign", and "5 different pools of ductal and lobular breast tumour" (Pools 1-7) (Fig. 2). A total of 642 clones were found positive within all the macroarray experiments (including positive clones detected along the pre-test).

Out of the panel of 642 sero-reactive clones identified from macroarray experiments, 284 clones are cloned in-frame. 181 proteins were found to be encoded by the 284 clones, out of which 34 protein encoding genes were found to be enlisted in the SEREX database. These genes were reported in the database basing on various cancer studies. Through literature search, 7 (ALDOA, CENPB, EEF2, GAPDH, MAZ, PRDX1, and TP53) out of 34 genes were found to be reported as TAAs against variety of cancers (Tab. 5). In a study, conducted by Suzuki *et al.*, on melanoma antigen identification by serological proteome approach found that 5

genes ALDOA, EEF2, GAPDH, ENO1 and HNRNP showed high reactivity in patient sera with G361 cell line protein spots compared to melanocytes [33]. In another study antibodies against ALDOA were identified in the sera from patients with Hepatocellular carcinoma [49]. Genes CENPB were mentioned to be significantly expressed in autoimmune diseases [50] and several studies have shown CENPB along with TP53 to be markedly associated with breast cancer survival and prognosis [51]. Over-expression of the genes CENBP, MAZ and PRDX1 was postulated to be linked to regulation of tumour progression, proliferation and metastasis [52,53]. PRDX1 was over-expressed in human oesophagus squamous cell carcinoma and MAZ protein isolated from cerebellar expression library showed significant reactivity against Hodgkin's disease patient sera [54].

Information of the molecular mechanisms is important to understand cellular behaviour and to predict the reasons for dysregulation, which may lead to cancer [55]. *In silico* analysis was done with the aim to find any sets of genes, among the genes expressed by the sero-reactive clones, which cluster together in accordance with certain functional categories like Transpath, Pfam and GO and are over-represented in breast cancer. Transpath is a database which provides information on signalling molecules, their reactions and the pathways these molecules are involved in [56]. KEGG is a collection of databases related to genomes, enzymatic pathways and biological chemicals of a cell [57]. Pfam is a database of protein families based on multiple sequence alignments and profile hidden Markov models [58,59]. GO is an initiative which helps standardizing the representation of a gene and gene product attributes across species and databases [60]. GO provides structured ontologies which classify the gene products with regards to biological process, cellular components and molecular functions irrespective of species [61]. In a meta-analysis study, conducted by Chopra, global cancer maps for KEGG, GO and Pfam were created based on 23 breast cancer microarray expression data sets. These maps revealed "hotspots" of activation/de-activation of breast cancer [62].

For having a better understanding of the genes/proteins, encoded by the sero-reactive clones, and their over-expression in various pathways, a web based toolkit called GeneTrail was employed. The 284 in-frame clones (test set) were compared with a reference set, "Heidelberg human foetal brain". No genes were found to be over-represented in any of the

KEGG pathways in the test. A significant over-representation of the genes involved in various enriched sub-categories of Pfam, Transpath and GO were observed.

Although the reproducibility of the macroarrays was not good enough to draw conclusions, this study enabled the identification of a sizable panel of clones which was used for recombinant protein expression and purification. Protein microarrays serve as very good alternative to protein macroarrays. Protein microarrays do have certain advantages over protein macroarrays. One of them being the signals derived from macroarrays are not as dynamic as compared to 16 bit (0-2^{16}) dynamic range of standard microarrays. Only few microliters (approximately 10 µL) of serum sample are enough for the validation of auto-antibody signatures. In another experiment, it was observed that the signal patterns obtained by microarrays analysis of brain and lung tumour patients' sera were highly reproducible (R=0.92-0.96) [27]. The panel of 642 sero-reactive clones obtained from macroarray screenings were used for the expression of His-tag proteins. These recombinant proteins were used for the production of targeted protein microarrays for TAA profiling using serum samples from breast cancer patients (n=24), females with benign fibroadenomas (n=16) and control individuals (n=20). Upon statistical evaluation of the signal intensities derived from the processed microarrays, using the PAM algorithm, the healthy control serum samples were differentiated from breast-nodule positive patients' sera with 100% sensitivity 85% specificity. However, the attempt to differentiate all the three classes (benign, malignant and healthy controls) yielded only 53% correct classification. Furthermore, GeneTrail analysis of the genes expressed by the classifier clones showed enrichment of R3H domain.

Conclusion

The macroarrays were used for a broad screening and deduced a panel of 642 sero-reactive clones from an expression library consisting 38,016 recombinant protein expressing clones. *In silico* analysis of the in-frame clones revealed enrichment of functional categories, like Transpath, Pfam and GO, in breast cancer. Using the recombinant proteins derived from 642 sero-reactive clones targeted array was generated for TAA profiling using patient sera and controls. With these protein microarrays, breast-nodule positive (benign and malignant) sera

could be differentiated from healthy control sera using 50 clones derived from PAM algorithm.

Acknowledgements

I want to thank Michael Stierschneider, Agnes Burg, Rudolf Pichler, Silvia Schönthaler and Ronald Kulovics (AIT) for their help with the *E. coli* cultivation, protein isolation and printing microarrays. Above all, I thank Jubiläumsfonds der Österreichischen National bank for funding this study (project no. 12551).

Reference list

1. Ferlay J, Parkin DM, Steliarova-Foucher E: **Estimates of cancer incidence and mortality in Europe in 2008.** *Eur J Cancer* 2010, **46:** 765-781.

2. Casiano CA, Mediavilla-Varela M, Tan EM: **Tumor-associated antigen arrays for the serological diagnosis of cancer.** *Mol Cell Proteomics* 2006, **5:** 1745-1759.

3. Anderson KS, LaBaer J: **The sentinel within: exploiting the immune system for cancer biomarkers.** *J Proteome Res* 2005, **4:** 1123-1133.

4. Sanchez-Carbayo M: **Antibody arrays: technical considerations and clinical applications in cancer.** *Clin Chem* 2006, **52:** 1651-1659.

5. Fernandez-Madrid F, Tang N, Alansari H, Granda JL, Tait L, Amirikia KC *et al.*: **Autoantibodies to Annexin XI-A and Other Autoantigens in the Diagnosis of Breast Cancer.** *Cancer Res* 2004, **64:** 5089-5096.

6. Soussi T: **p53 Antibodies in the sera of patients with various types of cancer: a review.** *Cancer Res* 2000, **60:** 1777-1788.

7. Bradley SV, Oravecz-Wilson KI, Bougeard G, Mizukami I, Li L, Munaco AJ *et al.*: **Serum antibodies to huntingtin interacting protein-1: a new blood test for prostate cancer.** *Cancer Res* 2005, **65:** 4126-4133.

8. Nesterova MV, Johnson N, Cheadle C, Bates SE, Mani S, Stratakis CA *et al.*: **Autoantibody cancer biomarker: extracellular protein kinase A.** *Cancer Res* 2006, **66:** 8971-8974.

9. Tomaino B, Cappello P, Capello M, Fredolini C, Ponzetto A, Novarino A *et al.*: **Autoantibody signature in human ductal pancreatic adenocarcinoma.** *J Proteome Res* 2007, **6:** 4025-4031.

10. Chapman C, Murray A, Chakrabarti J, Thorpe A, Woolston C, Sahin U *et al*.: **Autoantibodies in breast cancer: their use as an aid to early diagnosis.** *Ann Oncol* 2007, **18:** 868-873.

11. Disis ML, Calenoff E, McLaughlin G, Murphy AE, Chen W, Groner B *et al*.: **Existent T-cell and antibody immunity to HER-2/neu protein in patients with breast cancer.** *Cancer Res* 1994, **54:** 16-20.

12. Conroy SE, Gibson SL, Brunstrom G, Isenberg D, Luqmani Y, Latchman DS: **Autoantibodies to 90 kD heat-shock protein in sera of breast cancer patients.** *Lancet* 1995, **345:** 126.

13. Carter D, Dillon DC, Reynolds LD, Retter MW, Fanger G, Molesh DA *et al*.: **Serum antibodies to lipophilin B detected in late stage breast cancer patients.** *Clin Cancer Res* 2003, **9:** 749-754.

14. Piura E, Piura B: **Autoantibodies to tailor-made panels of tumor-associated antigens in breast carcinoma.** *J Oncol* 2011, **2011:** 982425.

15. Zhong L, Ge K, Zu JC, Zhao LH, Shen WK, Wang JF *et al*.: **Autoantibodies as potential biomarkers for breast cancer.** *Breast Cancer Res* 2008, **10:** R40.

16. Looi K, Megliorino R, Shi FD, Peng XX, Chen Y, Zhang JY: **Humoral immune response to p16, a cyclin-dependent kinase inhibitor in human malignancies.** *Oncol Rep* 2006, **16:** 1105-1110.

17. Desmetz C, Bascoul-Mollevi C, Rochaix P, Lamy PJ, Kramar A, Rouanet P *et al*.: **Identification of a new panel of serum autoantibodies associated with the presence of in situ carcinoma of the breast in younger women.** *Clin Cancer Res* 2009, **15:** 4733-4741.

18. Chapman C, Murray A, Chakrabarti J, Thorpe A, Woolston C, Sahin U *et al*.: **Autoantibodies in breast cancer: their use as an aid to early diagnosis.** *Ann Oncol* 2007, **18:** 868-873.

19. Zhang JY, Casiano CA, Peng XX, Koziol JA, Chan EK, Tan EM: **Enhancement of antibody detection in cancer using panel of recombinant tumor-associated antigens.** *Cancer Epidemiol Biomarkers Prev* 2003, **12:** 136-143.

20. Koziol JA, Zhang JY, Casiano CA, Peng XX, Shi FD, Feng AC *et al*.: **Recursive partitioning as an approach to selection of immune markers for tumor diagnosis.** *Clin Cancer Res* 2003, **9:** 5120-5126.

21. Ludwig N, Keller A, Heisel S, Leidinger P, Klein V, Rheinheimer S *et al*.: **Improving seroreactivity-based detection of glioma.** *Neoplasia* 2009, **11:** 1383-1389.

22. Babel I, Barderas R, az-Uriarte R, Martinez-Torrecuadrada JL, Sanchez-Carbayo M, Casal JI: **Identification of tumor-associated autoantigens for the diagnosis of colorectal cancer in serum using high density protein microarrays.** *Mol Cell Proteomics* 2009, **8:** 2382-2395.

23. Babel I, Barderas R, az-Uriarte R, Martinez-Torrecuadrada JL, Sanchez-Carbayo M, Casal JI: **Identification of tumor-associated autoantigens for the diagnosis of colorectal cancer in serum using high density protein microarrays.** *Mol Cell Proteomics* 2009, **8:** 2382-2395.

24. Bussow K, Nordhoff E, Lubbert C, Lehrach H, Walter G: **A human cDNA library for high-throughput protein expression screening.** *Genomics* 2000, **65:** 1-8.

25. Backes C, Keller A, Kuentzer J, Kneissl B, Comtesse N, Elnakady YA *et al*.: **GeneTrail--advanced gene set enrichment analysis.** *Nucleic Acids Res* 2007, **35:** W186-W192.

26. Keller A, Backes C, Lenhof HP: **Computation of significance scores of unweighted Gene Set Enrichment Analyses.** *BMC Bioinformatics* 2007, **8:** 290.

27. Stempfer R, Syed P, Vierlinger K, Pichler R, Meese E, Leidinger P *et al*.: **Tumour auto-antibody screening: performance of protein microarrays using SEREX derived antigens.** *BMC Cancer* 2010, **10:** 627.

28. Preininger C, Bodrossy L, Sauer U, Pichler R, Weilharter A: **ARChip epoxy and ARChip UV for covalent on-chip immobilization of pmoA gene-specific oligonucleotides.** *Anal Biochem* 2004, **330:** 29-36.

29. R Development Core Team: **R: A Language and Environment for Statistical Computing.** *R Foundation for Statistical Computing* 2005.

30. Simon R, Lam A: *BRB-ArrayTools User Guide, version 3.2. Biometric Research Branch, National Cancer Institute. http://linus.nci.nih.gov/brb*. Biometric Research Branch, National Cancer Institute. http://linus.nci.nih.gov/brb; 2009.

31. Tibshirani R, Hastie T, Narasimhan B, Chu G: **Diagnosis of multiple cancer types by shrunken centroids of gene expression.** *Proc Natl Acad Sci U S A* 2002, **99:** 6567-6572.

32. The SEREX database [http://www.licr.org/D_programs/d4a1i_SEREX.php]

33. Suzuki A, Iizuka A, Komiyama M, Takikawa M, Kume A, Tai S *et al*.: **Identification of melanoma antigens using a Serological Proteome Approach (SERPA).** *Cancer Genomics Proteomics* 2010, **7:** 17-23.

34. Bataller L, Wade DF, Graus F, Rosenfeld MR, Dalmau J: **The MAZ protein is an autoantigen of Hodgkin's disease and paraneoplastic cerebellar dysfunction.** *Ann Neurol* 2003, **53:** 123-127.

35. Zhang J, Wang K, Zhang J, Liu SS, Dai L, Zhang JY: **Using proteomic approach to identify tumor-associated proteins as biomarkers in human esophageal squamous cell carcinoma.** *J Proteome Res* 2011, **10:** 2863-2872.

36. Atalay C, Dogan L, Atalay G: **Anti-CENP-B antibodies are associated with prolonged survival in breast cancer.** *Future Oncol* 2010, **6:** 471-477.

37. Atalay C, Atalay G, Yilmaz KB, Altinok M: **The role of anti-CENP-B and anti-SS-B antibodies in breast cancer.** *Neoplasma* 2005, **52:** 32-35.

38. Briasoulis E, Kamposioras K, Tzovaras V, Pafitanis G, Kostoula A, Mavridis A *et al*.: **CENP-B specific anti-centromere autoantibodies heralding small-cell lung cancer. A case study and review of the literature.** *Lung Cancer* 2008, **60:** 302-306.

39. Li L, Chen SH, Yu CH, Li YM, Wang SQ: **Identification of hepatocellular-carcinoma-associated antigens and autoantibodies by serological proteome analysis combined with protein microarray.** *J Proteome Res* 2008, **7:** 611-620.

40. Liu H, Zhang J, Wang S, Pang Z, Wang Z, Zhou W *et al*.: **Screening of autoantibodies as potential biomarkers for hepatocellular carcinoma by using T7 phase display system.** *Cancer Epidemiol* 2011.

41. Lu D, Kuhn E, Bristow RE, Giuntoli RL, Kjaer SK, Shih I *et al*.: **Comparison of candidate serologic markers for type I and type II ovarian cancer.** *Gynecol Oncol* 2011, **122:** 560-566.

42. Messmer BT, Nour-Omid TS, Ghia E, Sanchez AB, Kipps TJ: **Autoantibodies against p53 are associated with chromosome 17p deletions in chronic lymphocytic leukemia.** *Leuk Res* 2011, **35:** 965-967.

43. Dalifard I, Daver A, Larra F: **Cytosolic p53 protein and serum p53 autoantibody evaluation in breast cancer. Comparison with prognostic factors.** *Anticancer Res* 1999, **19:** 5015-5022.

44. Leidinger P, Keller A, Heisel S, Ludwig N, Rheinheimer S, Klein V *et al*.: **Novel autoantigens immunogenic in COPD patients.** *Respir Res* 2009, **10:** 20.

45. Schmitt J, Heisel S, Keller A, Leidinger P, Ludwig N, Habel N *et al*.: **Multicenter study identified molecular blood-born protein signatures for Wilms Tumor.** *Int J Cancer* 2011.

46. Lubin R, Zalcman G, Bouchet L, Tredanel J, Legros Y, Cazals D *et al*.: **Serum p53 antibodies as early markers of lung cancer.** *Nat Med* 1995, **1:** 701-702.

47. Trivers GE, De B, V, Cawley HL, Caron G, Harrington AM, Bennett WP *et al*.: **Anti-p53 antibodies in sera from patients with chronic obstructive pulmonary disease can predate a diagnosis of cancer.** *Clin Cancer Res* 1996, **2:** 1767-1775.

48. Abendstein B, Marth C, Muller-Holzner E, Widschwendter M, Daxenbichler G, Zeimet AG: **Clinical significance of serum and ascitic p53 autoantibodies in epithelial ovarian carcinoma.** *Cancer* 2000, **88:** 1432-1437.

49. Looi KS, Nakayasu ES, Diaz RA, Tan EM, Almeida IC, Zhang JY: **Using proteomic approach to identify tumor-associated antigens as markers in hepatocellular carcinoma.** *J Proteome Res* 2008, **7:** 4004-4012.

50. Nakano M, Ohuchi Y, Hasegawa H, Kuroda T, Ito S, Gejyo F: **Clinical significance of anticentromere antibodies in patients with systemic lupus erythematosus.** *J Rheumatol* 2000, **27:** 1403-1407.

51. Kulic A, Sirotkovic-Skerlev M, Jelisavac-Cosic S, Herceg D, Kovac Z, Vrbanec D: **Anti-p53 antibodies in serum: relationship to tumor biology and prognosis of breast cancer patients.** *Med Oncol* 2010, **27:** 887-893.

52. Liang QJ, Lu XF, Cheng XL, Luo S, He DC, Wang YC: **[The active expression of CenpB, a constitutive protein in the centromeres of chromosomes, in breast cancer tissues].** *Yi Chuan Xue Bao* 2004, **31:** 236-240.

53. Zaytseva YY, Wang X, Southard RC, Wallis NK, Kilgore MW: **Down-regulation of PPARgamma1 suppresses cell growth and induces apoptosis in MCF-7 breast cancer cells.** *Mol Cancer* 2008, **7:** 90.

54. Bataller L, Wade DF, Graus F, Rosenfeld MR, Dalmau J: **The MAZ protein is an autoantigen of Hodgkin's disease and paraneoplastic cerebellar dysfunction.** *Ann Neurol* 2003, **53:** 123-127.

55. Krull M, Pistor S, Voss N, Kel A, Reuter I, Kronenberg D *et al.*: **TRANSPATH: an information resource for storing and visualizing signaling pathways and their pathological aberrations.** *Nucleic Acids Res* 2006, **34:** D546-D551.

56. Schacherer F, Choi C, Gotze U, Krull M, Pistor S, Wingender E: **The TRANSPATH signal transduction database: a knowledge base on signal transduction networks.** *Bioinformatics* 2001, **17:** 1053-1057.

57. Kanehisa M, Goto S, Kawashima S, Okuno Y, Hattori M: **The KEGG resource for deciphering the genome.** *Nucleic Acids Res* 2004, **32:** D277-D280.

58. Liu X, Lv B, Guo W: **The size distribution of protein families within different types of folds.** *Biochem Biophys Res Commun* 2011, **406:** 218-222.

59. Bateman A, Coin L, Durbin R, Finn RD, Hollich V, Griffiths-Jones S *et al.*: **The Pfam protein families database.** *Nucleic Acids Res* 2004, **32:** D138-D141.

60. The Gene Ontology Consortium: Gene ontology: tool for the unification of biology. Nat. Genet 2000, 25: 25-29.

61. Lee D, Redfern O, Orengo C: **Predicting protein function from sequence and structure.** *Nat Rev Mol Cell Biol* 2007, **8:** 995-1005.

62. Chopra P: Data Mining Techniques to Enable Large-scale Exploratory Analysis of Heterogeneous Scientific Data. PhD Thesis. North Carolina State University, Deapartment of Computer Science; 2009

Chapter 3*: In silico* design and performance of peptide microarrays for breast cancer tumour-auto-antibody testing

---

Submitted in *J Mol BioChem*. Syed P *et al*. ***In silico*** **design and performance of peptide microarrays for breast cancer tumour-auto-antibody testing**. Under review April 2012.

Keywords: peptide microarrays, auto-antibody, protein microarray, antigenic peptide design, Antigenic motif enrichment

Abstract

The simplicity and potential of minimal invasive testing using serum from patients make auto-antibody based biomarkers a very promising tool for use in diagnostics of cancer. Protein microarrays have been used for the identification of such auto-antibody signatures. On the other hand, peptide microarrays have proven to be a very useful means to study protein-protein interaction. This property can be exploited in the field of cancer diagnostics. In this study, 1185 antigenic peptides were designed which were deduced from proteins expressed by 642 cDNA expression clones found sero-reactive in breast cancer patients and controls. The subsequent peptides and proteins were used for the production of peptide and protein microarrays and serum samples from females with benign and malignant breast tumours and healthy control sera (*n=16* per group) were analysed. Correct classification of the serum samples on peptide microarrays 78% for discrimination of malignant versus healthy controls, 72% for benign versus malignant and 94% for benign vs controls; on protein arrays correct classification for these contrasts was 69%, 59% and 59%, respectively. The over-representation analysis of the classifiers derived from the class prediction showed enrichment of genes associated with ribosomes, spliceosomes, endocytosis and pentose phosphate pathway. Sequence analyses of the peptides with highest sero-reactivity showed enrichment of the zinc-finger domain as well as peptide reactivities are particularly negatively correlated with hydrophobicity while conversely they are positively correlated with positive charge, high inter-residue protein contact energies and possibly a secondary structure propensity bias. This study hints the possibility of using *in silico* designed antigenic peptide microarrays as an alternative to protein microarrays for improving tumour-auto-antibody based diagnostics.

Background

Breast cancer is the leading tumour type in women in sense of occurrence with an estimated 1 million new cases each year [1,2]. Survival rate is highly correlated with the stage of cancer when diagnosis take place thus early detection of this malignancy would be essential. Over the past decades several new diagnostic tools were developed, such as mammography and magnetic resonance imaging (MRI), although definitive answers still require biopsy and histopathological examination. Biopsy is highly invasive and usually applied only with a more advanced stage of the disease.

Blood-based biomarker discovery is an emerging field of cancer research with the hope of identification specific and sensitive markers, which enable clinicians to bring decisions with great accuracy and reliability. Detection of tumour-associated auto-antibodies from a few drops of blood may provide a possibility to screen patients with the suspicion of breast cancer or even before, through periodical examination. Tumour-associated antibodies can be identified through selective binding to special antigens, the so-called tumour-associated antigens. Tumour-associated antigens (TAAs) derived from aberrantly expressed proteins during the onset and progression of cancer development. These antigens display “non-self” epitopes, which trigger the immune system to remove them. The observed antigenicity have been attributed to multiple features of cancer growth, including accumulated mutations in cancer cells (*e.g.* point mutations, translocations), over-expression and translation of “differentiation genes”, which otherwise would not be present in adulthood, altered post-translational modifications or to gene products derived from aberrant splicing [3]. These molecules usually possess important functions in tumourigenesis, such as regulation of cell cycle, cell proliferation and apoptosis [4]. Previous studies have already elucidated several TAAs from the sera of breast cancer patients, such as MUC1, HSP90, HER2/neu, c-myc, NY-ESO1/LAGE1 and Lipophilin B [5-8]. Auto-antibodies against p53 tumour suppressor were also detected in the sera of 9-26% of women with breast cancer [9]. However, it has been shown that assaying of sera, reactivity for a single TAA is neither sensitive nor specific enough to discriminate between healthy individuals and cancer patients and thus rather a combination of multiple TAAs would be preferred to generate a classifier being employed for diagnostics.

Several methods have been developed to identify, screen and validate discriminative TAAs. SEREX (Serological Analysis of Recombinant Expressed cDNAs) and SERPA (Serological Proteomics Analysis) are such methods to identify *de novo* TAAs directly from tumour cells [10]. Although these methods have been used successfully to uncover new antigens [11-13], the drawback of these technologies is that they are labour intensive and small scale. Higher throughput methods like protein macro- and microarrays allow simultaneous quantification of serum reactivity of thousands of proteins. One of the major challenges of these applications is the requirement of a huge number of in frame cDNA clones and then expression and purification of the cognate proteins from these cDNA clones. The physicochemical properties (*i.e.* length vs. hydrophobic domains) of expressed proteins are usually highly variable and displaying of the reactive epitopes is cumbersome. Peptide microarray is another alternative solution since shorter peptide sequences may recapitulate the biological function (*i.e.* here, the antigenic epitope) of the corresponding protein [14,15]. Production of synthetic peptides is a well established technique and using peptide arrays as potential alternative to protein arrays would have several advantages. The concept of peptide array was first proposed by Southern in 1988 [16], then photolithographic peptide synthesis on a glass surface [17] and the SPOT-synthesis technology [18] accelerated their applications in microarray experiments [19].

In this chapter, the performance of a SPOT-synthesized peptide microarray was evaluated. This technology utilizes the traditional *fmoc* chemistry to synthesize peptides in single droplets *in situ* on the surface of slides. Based on a semi-empirical method developed by Kolaskar and Tongaonkar [20], antigenic peptides were deduced from a set of previously identified, protein microarray derived, antigenic proteins. These peptides were probed with sera of breast cancer patients, sera from individuals with benign breast nodules (benign and malignant) and compared them with samples from healthy donors. Further evaluation was done to identify the sero-reactive peptides using bioinformatics tools and panels of TAAs was defined, which are able to discriminate between samples of healthy control, malignant- and benign tumours.

Methods

Serum samples

Serum samples were obtained after the consent of the breast cancer patients and healthy women and were stored at -80°C. The study was approved by the Ethics Committee of the Medical University of Vienna and the General Hospital of Vienna (study number: 143/2007). For the protein and peptide microarray analysis of breast cancer serum biomarkers, 48 serum samples (malignant *n*=16; benign *n*=16; healthy *n*=16) were used. The clinical and the pathological cohorts of the serum samples are described in the table 1. All the 16 malignant samples were collected from patients diagnosed with invasive ductal carcinoma and were positive to the HER2/neu test. The benign samples were collected from the patients diagnosed with fibroadenoma. Healthy control individual's serum samples (n=16, mean age 76.9±7.15) were collected from healthy volunteers with no individual or familial history of breast or ovarian cancer.

Table 1. Clinical and pathological data of the patient-study cohort. Benign and maligant samples were collected from patients with fibroadenomas and invasive ductal carcinoma

	Benign (n=16)	**Malignant (n=16)**
Age (years)[a]	52.5±4.9	53.75±8
Grading[b]		
G1	-	1
G2	-	5
G3	-	10
Oestrogen receptor positive	-	9
pT stage[c]		
pT1; pT1b; pT1c; pT1mic; pT2	-	3; 3; 7; 1; 2
pN stage[d]		
pN0; pN1; pN1a; pN2; pN2a; pN3	-	7; 1; 1; 1; 3; 2
Metastasis stage[e]**: M0**	-	6
Menopause status[f]		
Pre-menopause	3	4
Post-menopause	8	11

[a]The age of the patients represented as mean age±standard deviation. [b]G1 (low-grade), G2 (intermediate grade) and G3 (high-grade). Low-grade tumours are usually slow growing and are less likely to spread. High-grade tumours are likely to grow more quickly and are more likely to spread. [c]pT1: Tumor 2.0 cm or less in greatest dimension; pT1b: Tumor more than 0.5 cm but not more than 1.0 cm in greatest dimension; pT1c: Tumor more than 1.0 cm but not more than 2.0 cm in greatest dimension; pT1mic: Microinvasion 0.1 cm or less in greatest dimension; pT2: Tumor more than 2.0 cm but not more than 5.0 cm in greatest dimension. [d]pN stage: information available for 15 patients. pN0: No regional lymph node metastasis; pN1: Metastasis to movable ipsilateral axillary lymph node(s); pN1a: Only micrometastasis (none larger than 0.2 cm); pN2: Metastasis to ipsilateral axillary lymph node(s) fixed to each other or to other structures; pN2a. Metastasis in 4-9 axillary lymph nodes, including at least one that is larger than 2 mm; pN3: Metastasis to ipsilateral internal mammary lymph node(s). [e]Metastasis stage: information from 6 patients. M0: No distant metastasis. [f]Information from 11 benign and 15 malignant samples.

Protein extraction and purification

In an earlier study 642 clones were identified, from a collection of 38,016 cDNA expression *E. coli* clones (hEx1 library [21]), which reacted positively to the sera from the breast cancer patients and the healthy control individuals. For the recombinant protein expression in *E. coli* and protein purification, the procedure developed by Stempfer *et al.* was followed [11].

In brief, the cDNA expression clones were cultured in 96 deep well plates and were induced by an autoinduction strategy for recombinant protein production. The expressed His-tagged proteins were then purified using Ni-NTA agarose and eluted in microarray spotting buffer (50 mM KH_2PO_4 and 50 mM K_2HPO_4, pH 8.0, 500 mM imidazole, 0.01% SDS and 0.01% NaN_3).

Design of Antigenic Peptides

Peptides corresponding to the 642 reactive proteins were designed as an alternative to the recombinant proteins found reactive in the initial membrane screening. To predict the antigenic peptides, the EMBOSS tool "Antigenic" (http://liv.bmc.uu.se/cgi-bin/emboss/antigenic) was used. The minimum length of the predicted peptide sequences is 6 amino acids (aa). The "Antigenic" tool employs a semi-empirical method developed by Kolaskar and Tongaonkar for the selection of antigenic peptide sequences. This method uses the physicochemical properties of amino acid residues and their frequencies of occurrence in experimentally known segmental epitopes to predict antigenic determinants on proteins [20].

The DNA sequence was available for 596 of the 642 clones; of those 581 clones were unique and were used for the antigenic peptide prediction. The default settings of the "Antigenic" tool were used and for each unique clone sequence, 2-3 different peptides were selected based on the antigenicity score and peptide-length. For technical reasons warranting uniform synthesis, peptides sized 8-10 aa were selected. Based on the maximum antigenicity score, those antigenic peptides which were longer than 10 aa were shortened but warranting to cover the maximum score aa-position. For antigenic motifs shorter than 8 aa peptides, N terminal aa corresponding to the template sequence were added. In addition tetanus specific antigenic peptides were designed for the NCBI reference sequence NP_783831; 56 tetanus specific peptides were selected from all potential antigenic peptides based on the max antigenicity score and peptides of 10 aa length selected for synthesis as described above.

Finding over-represented motifs in the peptide set, sequences were submitted to MEME motif search web-based tool (http://meme.nbcr.net). Motif was considered as enriched if it had at least 5 sequences (sites) with an E-value less than 0.001. Motif search was done also

with peptide sequences with high intensities (defined as median log2 intensities >13; min.: 6.21; max.: 15.84).

Microarray production

The procedure for the protein microarray production has been described in the chapter 1 which has been published in BMC Cancer [11]. In brief, the protein microarrays were generated using the purified recombinant proteins obtained from the cDNA expression *E. coli* clones. These purified proteins were spotted using an Omnigrid arrayer (GeneMachines) with SMP 3 pins (TeleChem International Inc.) under adjusted air humidity of between 55% and 60%. Spots were printed in duplicates on ARChip Epoxy slides [22]. Each microarray contained 4 identical subarrays. Crude clarified protein extract of the *E. coli* host was used for positive control spots and plain buffer spots were used as negative controls.

For the generation of peptide microarrays, 1212 clone-specific and 56 tetanus specific short peptides were synthesized using SPOT synthesis technology (JPT Peptide Technologies GmbH). Aminooxy-acetylated peptides were synthesized in parallel on cellulose membranes. Once the de-protection of the side chain was done, the solid phase-bound peptides were transferred to 96well microtitre filtration plates (Millipore). These peptides were cleaved from the cellulose membranes using 200 ml of aqueous triethylamine (0.5% v/v). The triethylamine-peptide solution was filtered and the evaporation under reduced pressure was conducted to remove the solvent. This was followed by re-dissolving of the resultant peptide derivatives (50 nmol) in 25 mL of spot buffer (70% DMSO, 25% 0.2 M sodium acetate pH 4.5, 5% v/v glycerol). The re-dissolved peptide solution was transferred into 384well microtitre plates and was used for the generation of the peptide arrays. Two droplets of 0.5 nL peptide solution (1 mM) were immobilized in triplicates on ARChip Epoxy slides [22], with 4 identical sub-arrays on each slide. For the immobilization of the peptide solution a non-contact printer Nanoplotter (GESIM) fitted with a piezoelectric NanoTip (GESIM) was used. Apart from the peptides derived from the cDNA clone-proteins human Igs (IgA, IgE, IgG and IgM) and 56 tetanus toxin (TT) specific peptides were also immobilized on the peptide microarrays. The human Igs and TT specific peptides were used as positive controls, while the empty buffer spots were used as negative spots.

Microarray processing

The microarrays were blocked with DIG easy Hyb (Roche Applied Science) for 30min and then washed twice in Phosphate Buffered Saline with 0.1% Tween 20 (PBST) for 5min. Breast cancer serum samples (benign; n=16 and malignant; n=16) and control sera (n=16) diluted in 1:10 with PBST were applied onto the microarrays and incubated for 2h. Then, the microarrays were washed twice in PBST for 5min. This was followed by incubation for 30min with goat anti human IgG detection antibody fluorescently labelled with Alexa647 dye (Invitrogen, Vienna, Austria) diluted 1:500 in PBST+3% non-fat dry milk powder. Later, the microarrays were washed twice in PBST for 5min. The array images of the processed slides were captured using an Axon Genepix 4000A microarray scanner (Molecular Devices).

Data analysis

Fluorescence intensity values (median after subtraction of the local background) were calculated from the scanned images using Genepix software (Molecular Devices). Statistical analysis of the microarray experiments was performed using the BRB-ArrayTools software 3.8.1 [http://linus.nci.nih.gov/BRB-ArrayTools.html] developed by Dr. R Simon and Amy Peng Lam [23]. The $\log_2$-transformed values of the signal intensities obtained from the scanned images of the processed microarrays were used for the analysis. The peptide microarray data were normalized using the "house-keeping gene" normalisation option within BRB-ArrayTools using the "Tetanus peptides" and "Igs" spots as normalisation features. For the data from protein microarrays, a global normalization was used to normalize each array using the relative median over all the log intensity values within one experiment. To identify the proteins/peptides that expressed differentially between classes, a random-variance t-test was applied to the data sets [24]. Significance of differentially expressed proteins/peptides was ranked using the p-value of the univariate test. In addition, the false discovery rate (FDR) was calculated using the method of Benjamini and Hochberg as provided within BRB-ArrayTools software. Further statistical data analysis was performed using R version 2.6.2 [25].

For defining a classifier set of antigenic proteins and peptides, the class prediction tools implemented in BRB-ArrayTools were used and leave one out cross validation was

conducted. Different classification algorithms (compound covariate, k nearest neighbour, nearest centroid, support vector machines, diagonal linear discriminant analyses and Bayesian compound covariate prediction) were run for model generation. The model incorporated the peptides and proteins that were differentially expressed among genes at the 0.01 and 0.05 significance level as assessed by the random variance t-test, respectively [24]. The estimation of the prediction error of each model was done using leave-one-out cross-validation (LOOCV) as described by Simon and colleagues [26]. For each LOOCV training set, the entire model building process was repeated, including the peptide and protein selection process. It was also evaluated whether the cross-validated error rate estimate for a model was significantly less than one would expect from random prediction. The class labels were randomly permuted and the entire LOOCV process was repeated. The significance level is the proportion of the random permutations that gave a cross-validated error rate no greater than the cross-validated error rate obtained with the real data. Cross-Validation ROC (receiver operating curve) analyses from the Bayesian Compound Covariate Predictor were conducted and area under curve (AUC) values were calculated as implemented in BRB-ArrayTools' class prediction tools.

Over-representation analysis

An over-representation analysis (ORA) of the classifiers derived from the microarray experiments was done using the gene set enrichment analysis tool "GeneTrail" [27]. The classifiers from the peptide microarray analysis were traced back to the proteins they were derived from and the ORA was performed using the corresponding gene Ids. Similarly, the classifiers from the protein array analysis were used for the ORA. The ORA was conducted using a "hypergeometric distribution test" statistical test method. As a reference set, all human genes were used to which the gene Ids corresponding to the classifiers were compared. The significance value of 0.05 was chosen and was adjusted by false discovery rate (FDR) adjustment.

Results

Antigenic motif search

Out of 642 clone-proteins which were used for the protein microarray production, sequences of 596 proteins were available. All 3 possible reading frames of DNA sequences coding for proteins were collected and checked for the longest uninterrupted sequence of ORF. After eliminating the duplicates 581 unique sequences were found which were used for antigenic motif search. Using the "Antigenic" tool yielded 4492 antigenic peptides for these 581 clone-sequences, resembling an average of 8.4 peptides per clone. When the length of the 4492 antigenic peptides were plotted against frequency of occurrence, a high frequency of occurrence was observed with peptides of length ranging from 6 to 20 amino acids (Fig.1) and also, a uniform distribution of antigenic motifs was found along the 581 clone-sequences subjected to peptide design (Fig.2).

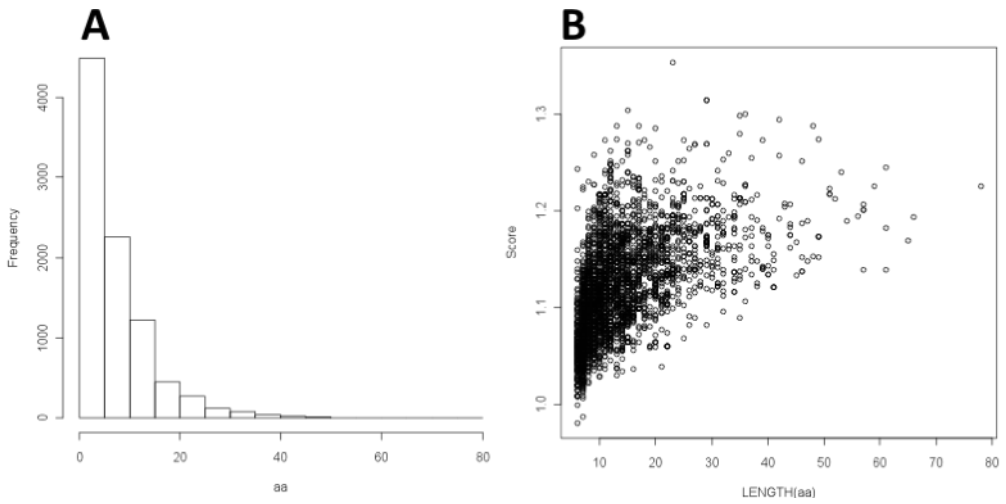

Figure 1: Length distribution of 4492 peptides. This figure A shows the frequency of occurrence (Y-axis) of the peptides with regards to the length of the antigenic motif (X-axis). A relatively high frequency of occurrence was observed for the short-length peptides. The figure B shows the distribution of 4492 antigenic motifs along the 581 clone sequences. The x-axis depicts the start amino acid position within the targeted clone-sequences; on the y-axis the length of the antigenic-peptides is depicted. A uniform distribution of the antigenic motifs was observed along the clone sequences. Density of plotted antigenic motifs is highest for short peptides (<20mers; y-axis).

Of the 4492 antigenic motifs 2866 were unique motifs. From of these 2866 antigenic motifs, 2-3 peptides per clone were selected which had maximum scores and thus identified 1212

peptides. Out of these 1212 peptides, 53%, 33.7% and 13.2% of the peptides were 8-10 aa, 7-14 aa and more than 14 aa long, respectively (Tab. 2). Peptides with lengths ranging 6 to14 were present at highest frequency compared to the longer peptides (Fig.2). These 1185 single peptides including the human Igs and the 56 TT specific peptides were used for the peptide array production.

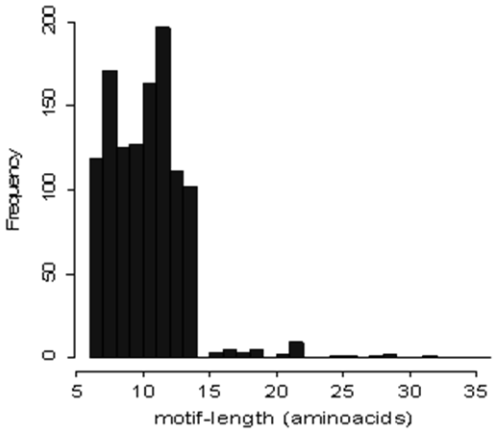

Figure 2: Length distribution of 1185 single peptides. This figure shows the frequency of occurrence (Y-axis) of the peptides with regards to the length of the antigenic motif (X-axis).

Table 2: Number of peptides with regards to the length of the antigenic motifs and the number of the corresponding clones.

Length (aa)	Number of peptides	Number of clones
8-10	643	329
7-14	409	79
>14	160	55

Serum reactivity of “antigenic” peptide arrays

Median intensities of each duplicate peptide spot from 48 microarray analyses were calculated and used for evaluation of any correlation of the serum-reactivity with the “antigenicity score” and the influence on addition / removal of aa from the antigenic motif to synthesized and spotted 8-10 aa peptides. It was not possible to find any correlation of microarray signal intensities with “length adjustment of peptides” and “antigenicity scores” (Fig. 3).

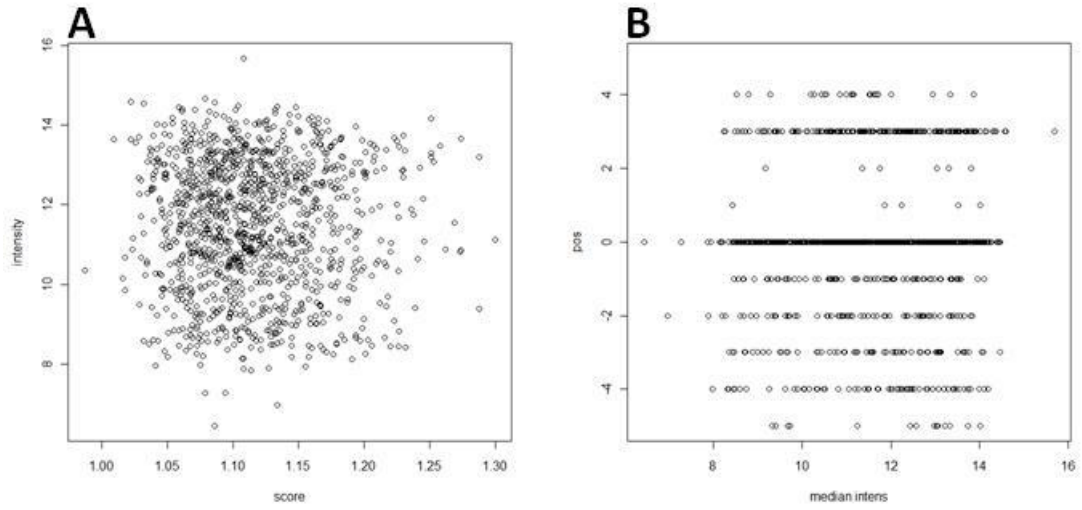

Figure 3: Antigenic reactivity derived from 48 samples. Median peptide array intensities were plotted versus the "antigenicity score" (A), and the "length adjustment" (denoted "pos") - positive values correspond to the number of aa's added, negative values to aa's removed from antigenic motifs for generation of 8-10aa peptides for array spotting (B).

Motif enrichment analysis

For screening of enriched peptide motifs in the microarray peptide-set, the MEME motif discovery tool [28] was used. The most significant and highly represented motif found was similar to Zinc-finger domains of Zn-H2C2-type. Motif logo consisted of 26 sequences and the diagram clearly depicts the highly weighted two central cysteines, separated by two other amino acids (Fig. 4A, see pfam13465: zf-H2C2_2). Seemingly the first two amino acids (proline and tyrosine) have also conserved role to constitute these domains (Fig. 4A). In a second screen, only those peptides that were considered highly antigenic were used to see if it is possible to identify back to these motifs. The analysis of highly active peptides elucidated similar results (Fig. 4B). Although these motifs were "narrower" (since less sequences), they clearly depicted the Zinc-finger domain characteristic for the superfamily. The finding that Zinc-finger domains are antigenic is concomitant with previous reports, which found several members of Zinc-finger proteins as tumour-associated antigens [29].

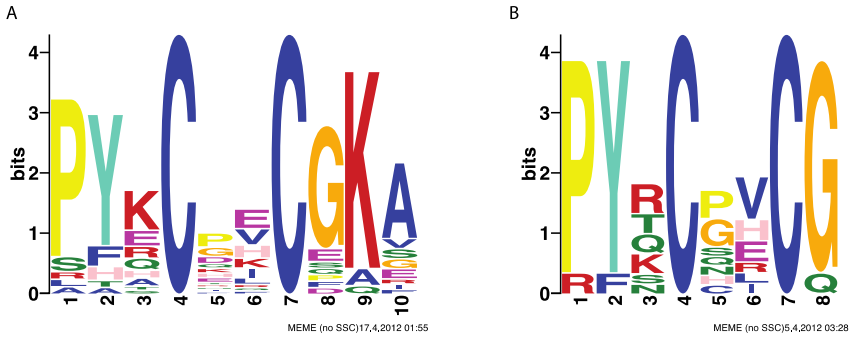

Fig 4: Sequence logos of enriched motifs. A: Sequence logo depicting the most significant motif (E-value: 1.0-e100, 26 sites). B: Analysis of peptides with high experimental signal intensity giving very similar result (E-value: 9.8e-23, 11 sites). MEME sequence logos represent probability matrices that specify the probability of each letter in all possible position.

Microarray analysis

The data obtained upon processing the protein and the peptide microarrays with the breast cancer and the healthy control sera was subjected to statistical evaluation. The class prediction elucidated a marker-set of 54 peptides which enabled the classification of malignant samples and healthy controls with 75% sensitivity and 81.2% specificity and 78% correct classification (compound covariate classifier) (Tab. 3). The list of the classifier peptides are shown in the Table 4. The ROC curve derived from this class prediction (Fig.5A) showed the AUC values of 0.758. For the prediction of the same classes on protein array, a marker-set of 57 proteins was deduced (Tab. 5). These proteins enabled the 69% correct classification of the malignant samples and healthy controls with 62.5% sensitivity, 75% specificity (Tab. 3) and AUC value of 0.68 (support vector machine classifier) (Fig. 5B).

Table 3: Class predication of benign, malignant and control samples using peptide and protein arrays.

Classes	Microarray	Classification method	Correctly classified (%)	Sensitivity (%)	Specificity (%)	PPV	NPV	AUC
Malignant vs. Control	Peptide	CCP	78	75	81.2	0.8	0.765	0.758
	Protein	SVM	69	62.5	75	0.714	0.667	0.68
Benign vs. Malignant	Peptide	3-NN	72	62.5	81.2	0.769	0.684	0.6
	Protein	1-NN	59	87.5	31.2	0.56	0.714	0.461
Benign vs. Control	Peptide	1-NN	94	93.8	93.8	0.938	0.938	0.852
	Protein	CCP	59	62.5	56.2	0.588	0.6	0.648

PPV: positve predictive value. NPV: negative predictive value. PPV and NPV are the proportions of the samples with positive test and negative test results, respectively, which are correctly diagnosed. Compound covariate predictor; Support vector machine; 3-Nearest neighbours; 1-Nearest neighbour

Table 4: Classifier peptides derived from class prediction which correctly classified malignant samples and healthy controls.

Peptide Id	p-value	t-value	Fold-change	Peptide Id	p-value	t-value	Fold-change
1127_429_MPMGp800P04517_2_32q0	0.0002	-4.213	0.31	394_147_MPMGp800F12540_3_131q0	0.00613	-2.912	0.39
628_236_MPMGp800I16557_3_196q-3	0.0004	-3.916	0.3	84_31_MPMGp800B03538_3_7q0	0.00627	-2.903	0.43
1041_396_MPMGp800N20541_2_70q0	0.0004	-3.905	0.44	106_40_MPMGp800B13557_2_132q0	0.00630	-2.901	0.43
1132_431_MPMGp800P05578_2_230q3	0.0006	-3.772	0.16	865_326_MPMGp800L08582_2_84q-39	0.00641	-2.894	0.34
723_274_MPMGp800J23565_2_201q0	0.0006	-3.735	0.39	1151_438_MPMGp800P09510_2_121q-68	0.00656	-2.886	0.46
651_245_MPMGp800J02584_2_54q0	0.0007	-3.729	0.34	1177_448_MPMGp800P17553_1_229q-3	0.00692	-2.865	0.56
900_340_MPMGp800L19532_3_33q-24	0.0009	-3.625	0.38	730_277_MPMGp800K03508_3_75q0	0.00692	-2.865	0.35
653_245_MPMGp800J02584_2_147q0	0.0010	-3.596	0.3	255_93_MPMGp800D09600_2_70q-1	0.00711	-2.854	0.47
367_136_MPMGp800F02569_2_50q0	0.0013	-3.491	0.36	1027_389_MPMGp800N14584_2_62q0	0.00717	-2.851	0.47
1037_393_MPMGp800N17552_1_121q-3	0.0020	-3.328	0.39	292_107_MPMGp800D20603_3_135q-2	0.00738	-2.84	0.37
543_204_MPMGp800H14523_1_58q0	0.0022	-3.299	0.55	599_226_MPMGp800I07544_2_217q0	0.00748	-2.834	0.54
104_39_MPMGp800B13514_2_12q-4	0.0022	-3.295	0.33	1103_420_MPMGp800O17527_1_139q-11	0.00769	-2.824	0.46
1124_428_MPMGp800P03549_2_137q-1	0.0027	-3.227	0.28	IgA-human	0.00790	-2.813	0.39
867_326_MPMGp800L08582_2_217q4	0.0027	-3.224	0.31	136_52_MPMGp800C02520_1_34q-17	0.00799	-2.808	0.44
4_1_MPMGp800A02550_2_41q0	0.0036	-3.11	0.5	21_6_MPMGp800A07545_2_216q-1	0.00811	-2.803	0.32
1118_426_MPMGp800P01541_1_211q-36	0.0037	-3.1	0.3	887_335_MPMGp800L14578_2_14q-1	0.00830	-2.793	0.52
955_360_MPMGp800M10592_2_164q0	0.0039	-3.082	0.33	766_290_MPMGp800K11529_2_81q0	0.00848	-2.785	0.47
792_300_MPMGp800K17510_2_129q-11	0.0041	-3.069	0.41	549_206_MPMGp800H15530_2_14q0	0.00883	-2.769	0.48
8_Tet_NP783831_174q3	0.0042	-3.06	0.46	478_178_MPMGp800G20576_1_138q-41	0.00909	-2.757	0.42
455_169_MPMGp800G13583_3_126q0	0.0043	-3.051	0.33	835_316_MPMGp800L02571_2_72q-2	0.00928	-2.749	0.46
282_103_MPMGp800D17584_2_187q-35	0.0043	-3.048	0.38	42_Tet_NP783831_783q-4	0.00939	-2.745	0.48
1149_437_MPMGp800P08580_1_211q-21	0.0045	-3.032	0.42	886_334_MPMGp800L14571_2_158q0	0.00974	-2.73	0.42
663_249_MPMGp800J05542_1_195q0	0.0046	-3.019	0.49	82_30_MPMGp800A24590_3_34q-6	0.00996	-2.721	0.47
825_312_MPMGp800K22574_2_189q-1	0.0048	-3.006	0.41	459_170_MPMGp800G14567_3_151q-5	0.01000	2.719	1.8
696_262_MPMGp800J16572_2_25q-41	0.0051	-2.986	0.45	381_142_MPMGp800F08524_2_4q4	0.00974	2.73	3.98
22_Tet_NP783831_511q-6	0.0059	-2.928	0.44	86_31_MPMGp800B03538_3_199q0	0.00746	2.836	1.62
935_353_MPMGp800M07508_1_43q-2	0.0061	-2.916	0.38	IgM-human	0.00201	3.331	2.28

Table 5: Classifier proteins derived from class prediction which correctly classified malignant samples and healthy controls.

Clone Id	p-value	t-value	Fold-change	Clone Id	p-value	t-value	Fold-change
MPMGp800L18590	0.01	-2.806	0.52	MPMGp800D07572	0.03	2.196	1.46
MPMGp800L10514	0.01	-2.726	0.54	MPMGp800H07544	0.03	2.208	1.64
MPMGp800J20511	0.01	-2.665	0.61	MPMGp800D02584	0.03	2.215	1.54
MPMGp800F17571	0.01	-2.584	0.64	MPMGp800D17584	0.03	2.274	1.59
MPMGp800G09586	0.02	-2.513	0.66	MPMGp800I01598	0.03	2.278	1.6
MPMGp800J11531	0.02	-2.505	0.62	MPMGp800M10592	0.03	2.294	1.58
MPMGp800J10529	0.02	-2.487	0.65	MPMGp800D17517	0.03	2.312	1.65
MPMGp800J09577	0.02	-2.486	0.52	MPMGp800C06579	0.02	2.338	1.64
MPMGp800O18529	0.02	-2.424	0.42	MPMGp800K20532	0.02	2.391	2
MPMGp800J16581	0.02	-2.392	0.53	MPMGp800D17529	0.02	2.393	1.64
MPMGp800G15509	0.03	-2.307	0.62	MPMGp800A03524	0.02	2.405	1.65
MPMGp800M10544	0.03	-2.291	0.65	MPMGp800D08553	0.02	2.501	1.78
MPMGp800J02584	0.03	-2.279	0.55	MPMGp800B10579	0.01	2.687	2.21
MPMGp800E24584	0.03	-2.201	0.66	MPMGp800D14551	0.01	2.704	1.68
MPMGp800N22556	0.04	-2.18	0.65	MPMGp800I06598	0.01	2.738	2.24
MPMGp800D22540	0.04	-2.162	0.5	MPMGp800K22533	0.01	2.828	1.61
MPMGp800O14536	0.04	-2.161	0.66	MPMGp800C02520	0.01	2.838	2.09
MPMGp800E21533	0.04	-2.15	0.53	MPMGp800P10579	0.01	2.885	2.17
MPMGp800M21592	0.04	-2.126	0.59	MPMGp800F10589	0.01	2.895	2.26
MPMGp800K15583	0.04	-2.124	0.69	MPMGp800D06601	0.01	2.918	1.82
MPMGp800P21572	0.04	-2.119	0.53	MPMGp800D14506	0.01	3.4001	1.95
MPMGp800L07522	0.04	-2.118	0.68	MPMGp800O10518	0.01	3.4366	2.06
MPMGp800E11583	0.05	-2.046	0.62	MPMGp800D13550	0.01	2.956	2.16
MPMGp800G07549	0.05	-2.034	0.68	MPMGp800H22541	0.00	3.028	1.87
MPMGp800F02519	0.05	-2.033	0.65	MPMGp800C20586	0.00	3.199	1.96
MPMGp800I14518	0.05	2.041	1.57	MPMGp800C22515	0.00	3.235	2.43
MPMGp800H01584	0.05	2.061	1.39	MPMGp800C06590	0.00	3.235	2.02
MPMGp800C13512	0.04	2.145	1.57	MPMGp800D09526	0.00	3.786	2.14
MPMGp800D17570	0.03	2.189	1.51				

The class prediction of the benign and malignant samples on peptide array elucidated 9 peptides (Tab. 6) which enabled correct classification of 72% (3-Nearest Neighbours classifier) with 62.5% sensitivity and 81.2% specificity (Tab. 3) An AUC value of 0.6 was observed for this class prediction (Fig. 5D). Similarly on protein array, 17 proteins (Tab. 7) enabled the class prediction of benign and malignant samples with 59% 62.5% correct classification (1-Nearest neighbour), 87.5% sensitivity, 31.2% specificity (Tab. 3) and AUC value of 0.461 (Fig. 5E) (Tab. 3).

Table 6: Classifier peptides derived from class prediction which correctly classified malignant samples and benign samples.

Peptide Id	p-value	t-value	Fold-change
88_33_MPMGp800B05530_2_13q-6	0.0044	-3.038	0.63
956_360_MPMGp800M10592_2_188q0	0.0049	-2.995	0.44
47_Tet_NP783831_847q0	0.0095	2.741	3.38
950_358_MPMGp800M08591_2_70q0	0.0082	2.796	1.56
651_245_MPMGp800J02584_2_54q0	0.0060	2.919	2.47
1132_431_MPMGp800P05578_2_230q3	0.0059	2.927	3.19
1041_396_MPMGp800N20541_2_70q0	0.0050	2.992	1.95
653_245_MPMGp800J02584_2_147q0	0.0048	3.003	2.32
1055_402_MPMGp800N24595_2_98q0	0.0018	3.367	2.97

Table 7: Classifier proteins derived from class prediction which correctly classified malignant samples and benign samples.

Clone Id	p-value	t-value	Fold-change
MPMGp800H15573	0.0080	-2.795	0.55
MPMGp800C06590	0.0167	-2.5	0.6
MPMGp800K20532	0.0304	-2.246	0.53
MPMGp800F24553	0.0351	-2.183	0.61
MPMGp800A17595	0.0409	-2.114	0.64
MPMGp800H12591	0.0438	2.082	1.7
MPMGp800O18530	0.0381	2.146	1.52
MPMGp800A15533	0.0380	2.147	1.51
MPMGp800E21533	0.0375	2.153	1.84
MPMGp800L24584	0.0335	2.202	1.52
MPMGp800K16540	0.0335	2.203	1.89
MPMGp800M08528	0.0294	2.26	2.09
MPMGp800L18590	0.0283	2.277	1.72
MPMGp800O18529	0.0212	2.399	2.62
MPMGp800M08567	0.0197	2.432	2.36
MPMGp800P21572	0.0120	2.633	2.12
MPMGp800P03594	0.0067	2.862	2.52

The class prediction between the benign and the control samples yielded 17 peptides (Tab. 8) which gave 93.8% sensitivity and specificity with 94% correct classification (1-Nearest Neighbour classifier method) (Tab. 3). The observed AUC value of ROC curve for this class prediction was 0.852 (Fig. 5F). On the protein microarray data, a panel of 35 proteins (Tab. 9) enabled a 59% correct classification (compound covariate classifier) of benign and the control samples with 62.5%sensitivity, 56.2% specificity (Tab. 3) and AUC 0.648 (Fig. 5G). There were 9 proteins corresponding to the genes*, PCSK1, DGKK, ZNF598, TBC1D9, TMEM199, EPB41L3, SAMD6, PRPF38A* and *C1orf9* which were found in both the peptide and protein array classifiers' list derived from all the class predictions.

Table 8: Classifier peptides derived from class prediction which correctly classified control samples and benign samples.

Peptide Id	p-value	t-value	Fold-change
723_274_MPMGp800J23565_2_201q0	0.001	-3.676	0.44
1048_398_MPMGp800N22556_1_158q-7	0.001	-3.442	0.23
143_55_MPMGp800C04592_2_86q0	0.003	-3.214	0.21
367_136_MPMGp800F02569_2_50q0	0.003	-3.169	0.39
39_12_MPMGp800A14528_2_100q-1	0.006	-2.93	0.37
38_12_MPMGp800A14528_2_61q3	0.006	-2.897	0.5
15_Tet_NP783831_325q-2	0.007	-2.882	0.49
241_88_MPMGp800D06520_1_33q-1	0.007	-2.858	0.41
455_169_MPMGp800G13583_3_126q0	0.009	-2.763	0.49
916_346_MPMGp800M02514_3_69q-22	0.01	-2.721	0.5
543_204_MPMGp800H14523_1_58q0	0.01	-2.716	0.45
462_173_MPMGp800G17568_2_31q-1	0.01	2.725	1.69
IgM-human	0.006	2.912	1.89
1156_440_MPMGp800P10579_3_87q3	0.005	3.008	4.14
621_234_MPMGp800I14597_1_163q-6	0.004	3.107	4.42
943_356_MPMGp800M08567_3_9q0	0.003	3.123	2.55
358_132_MPMGp800E24584_2_201q3	0.003	3.185	1.74

Table 9. Classifier proteins derived from class prediction which correctly classified control samples and benign samples.

Clone Id	p-value	t-value	Fold-change
MPMGp800J10529	0.00026	-4.14	0.54
MPMGp800G09586	0.002	-3.482	0.56
MPMGp800J11531	0.003	-3.24	0.54
MPMGp800G09554	0.004	-3.119	0.56
MPMGp800F17571	0.006	-2.927	0.6
MPMGp800J09577	0.008	-2.864	0.65
MPMGp800G07584	0.009	-2.801	0.66
MPMGp800J22534	0.012	-2.69	0.61
MPMGp800K16550	0.015	-2.575	0.68
MPMGp800K12586	0.021	-2.437	0.65
MPMGp800J16581	0.022	-2.413	0.66
MPMGp800J02584	0.027	-2.331	0.68
MPMGp800K03508	0.028	-2.307	0.69
MPMGp800N22556	0.029	-2.295	0.68
MPMGp800J17590	0.031	-2.26	0.64
MPMGp800B14526	0.038	-2.167	0.66
MPMGp800B21550	0.039	-2.164	0.67
MPMGp800G16536	0.041	-2.138	0.65
MPMGp800G15509	0.043	-2.11	0.65
MPMGp800F14572	0.044	-2.106	0.64
MPMGp800H11522	0.048	-2.064	0.6
MPMGp800L07522	0.049	-2.05	0.74
MPMGp800H18598	0.046	2.083	1.87
MPMGp800H24517	0.037	2.179	1.46
MPMGp800B18580	0.024	2.37	1.8
MPMGp800D13550	0.022	2.412	1.89
MPMGp800P20598	0.019	2.476	1.42
MPMGp800C22515	0.019	2.491	1.72
MPMGp800C20586	0.018	2.502	1.51
MPMGp800H22541	0.013	2.642	1.89
MPMGp800I06598	0.012	2.67	2.17
MPMGp800J13577	0.010	2.742	1.51
MPMGp800B10579	0.008	2.833	1.95
MPMGp800D14506	0.006	2.961	1.77
MPMGp800P10579	0.001	3.571	2.46

Using the clone-proteins corresponding to the classifier peptides derived from the class prediction of malignant and control samples, a set of 4 proteins (Tab. 10) were deduced which enabled a correct classification of 66% (compound covariate classifier) with 56.2% sensitivity, 75% specificity and AUC value 0.688 (Fig. 5C). The class prediction of malignant and benign samples was not possible using the clone-proteins corresponding to the classifier

proteins derived from the class prediction on peptide array. The clone-proteins corresponding to the classifier from the class prediction of benign and the control samples enabled 72% correct classification (compound covariate classifier) of the same classes with 2 clone-proteins (Tab. 11). For this class prediction 68.8% sensitivity, 75% specificity and AUC value 0.793 (Fig. 5H) were observed.

Table 10. Class prediction of malignant samples from healthy controls on protein array using the proteins corresponding to the peptide in the Table 4.

Clone Id	p-value	t-value	Fold-change
MPMGp800J02584	0.03	-2.279	0.55
MPMGp800D17584	0.03	2.274	1.59
MPMGp800M10592	0.03	2.294	1.58
MPMGp800C02520	0.007	2.838	2.09

Table 11. Class prediction of benign samples from healthy controls on protein array using the proteins corresponding to the peptide in the Table 8.

Clone Id	p-value	t-value	Fold-change
MPMGp800N22556	0.03	-2.295	0.68
MPMGp800P10579	0.001	3.571	2.46

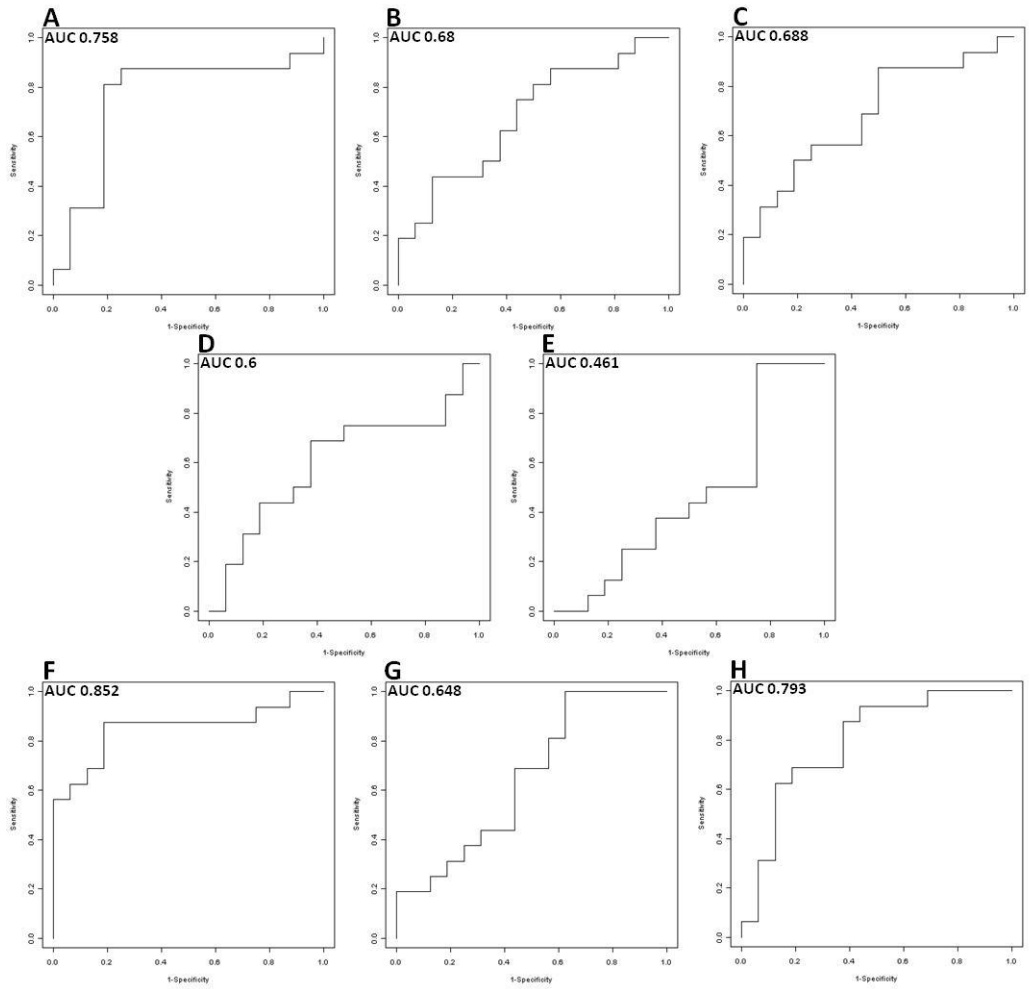

Figure 5: Cross-Validation ROC curves from the Bayesian Compound Covariate Predictor. The figures A, D and F represent the ROC curves obtained from the class predictions performed using the data from the peptide arrays. The figures B, E and G) represent the ROC curves from the class predictions obtained from the protein microarray analysis. The figures C and H represent the ROC curves from the class predictions obtained from the protein array analyzed using the proteins corresponding to the respective peptide array classifiers. The x-axes and y-axes represent the false positive rate (1-specificity) and true positive rate (sensitivity), respectively. The ROC curves A, B and C represent the class prediction of malignant and control samples. The class prediction of classes benign and malignant samples are represented by the ROC curves D and E. Similarly, the ROC curves F, G and H represent the class prediction of the benign and control samples.

The signal intensities from all the peptide array classifiers were compared with the signal intensities from the corresponding proteins on the protein arrays. Similarly, the signal intensities from all the protein array classifiers were compared to the signal intensities from the corresponding peptides. These comparisons failed to give any correlation between the peptide and protein array data.

Over-representation analysis

Peptide array classifiers

Over-representation analysis of the genes encoding the classifier peptides from all class predictions on peptide arrays was done using a gene set analysis tool "GeneTrail" [27]. Out of 57 genes representing the peptide classifiers, 3 genes namely *RPL7A, RPL24* and *RPL6* were involved in the KEGG ribosome pathway. *PGLS* and *ALDOA* were found to be involved in the pentose phosphate pathway, while *ISY1* and *PRPF38A* were involved in the Spliceosome pathway (Tab. 12).

Table 12: Over-represented genes from the peptide array classifier and the corresponding KEGG pathways

KEGG pathways	p-value	Expected number of genes	Observed number of genes	Genes
Pentose phosphate pathway	0.003	0.06	2	*PGLS, ALDOA*
Ribosome	0.003	0.2	3	*RPL7A, RPL24 RPL6*
Spliceosome	0.04	0.3	2	*ISY1, PRPF38A*

Protein array classifiers

Similarly, ORA was performed using the genes encoding the classifiers from all the class predictions on protein array. It was found that 2 genes (*RPS3A* and *RPS13*) out of 59 genes representing the protein classifiers were involved in the KEGG ribosome pathway. *GIT1, CHMP4C, EHD2* and *GRK1* were involved in the KEGG endocytosis pathway (Tab. 13). An enrichment of secondary structures was found among the genes represented by the classifiers. It was found that 28.07% and 31.58% of the genes represented by the classifier proteins contain sequence motifs like coiled coils and ELR motifs at p-values 0.0004 and 0.003, respectively. An enrichment of the protein family domains like UBA/TS-N domain and TBC domain was also observed.

Table 13: Over-represented genes from the protein array classifier and the corresponding KEGG pathways

KEGG pathways	p-value	Expected number of genes	Observed number of genes	Genes
Endocytosis	0.006	0.6	4	*GIT1*, *CHMP4C*, *EHD2*, *GRK1*
Ribosome	0.04	0.2	2	*RPS3A*, *RPS13*

Discussion

In this experiment, peptide microarrays were used to identify the auto-antibody signatures against the predicted antigenic peptides in the breast cancer patients' sera. The prediction of the antigenic peptides was done basing on the occurrence of the hydrophobic residues, cysteine, leucine and valine, on the surface of the proteins. For the identification and characterization of antibody epitope short peptide sequences can be exploited. The short peptides with lengths ranging from 4 to 15aa are effective enough to identify antibody epitopes [30]. In this study, 86.7% of the predicted antigenic peptides have varying length from 7 to 14 amino acids. These peptides sequences were used for deducing peptides and for generation of peptide arrays. Peptides with highest-seroreactivity of all 48 samples showed enrichment of motifs similar to Zinc-finger domain that can be explained because of their central cysteine's are highly hydrophobic, moreover many Zinc-finger protein contain several Zinc-finger domains [31]. These features allowed the EMBOSS tool to list many of these peptide to be antigenic.

Using the synthetic peptides, the peptide microarrays were generated. Simultaneously, protein microarrays were produced using the recombinantly expressed proteins from the cDNA expression *E. coli* clones. The peptide and protein microarrays were used for the evaluation of same set of serum samples. On peptide arrays classification success for distinguishing the 3 classes of malignant, benign and control serum samples outperformed protein arrays during class prediction analyses. Apart from the better sensitivities and specificities, ROC analyses on peptide array data provided higher AUC values (Tab. 4; Fig. 7) compared to that of protein microarrays. Apart from the better sensitivities and specificities, class prediction analyses on peptide array provided higher AUC values compared to that of protein microarrays. The binding ability of an antibody to a protein majorly depends on the conformation at the region of binding. The antibodies specific to the proteins have the same specificity as long as the binding site is located on the surface of the molecule [32]. Expression of recombinant protein in *E. coli* often leads to the production of misfolded proteins [33], as well as microarray immobilization of proteins will dramatically change the conformation and accessibility of proteins. These effects might in total lead to low reproducibility and controversial findings when array-platforms are changed. As done here, sero-reactive clones were identified by a macro-membrane based screening. On those

membranes *E. coli* clones are grown and protein-expression is induced and proteins are immobilized directly on the site of clone-growth. For elucidation of the diagnostic value of identified antigenic proteins microarrays provide today's best option for confirmation and validation of thousands of proteins in parallel. Biomarker-validation requires analyses of many patient samples would thus be best done on microarrays. However, when switching from macro-membranes used in the biomarker identification step to microarrays requires the isolation of proteins and subsequent spotting on the microarray surfaces. Thus conditions are dramatically changed. This might be also the reason why protein-microarrays in this study have shown up with moderate to low classification success of malignant breast cancer, benign breast nodules and controls. Since the prediction of the antigenic peptides was done independent of the protein conformation, there may be a better chance for the auto-antibodies to bind to the antigenic sites presented on the peptide arrays compared to the protein arrays. This may explain why the classifiers from peptide and protein arrays were so different and elucidated varying results. However, upon the comparing the classifiers from peptide and protein array class prediction analyses, 9 genes namely; *PCSK1, DGKK, ZNF598, TBC1D9, TMEM199, EPB41L3, SAMD6, PRPF38A* and *C1orf9* were found. Among these proteins, EPB41L3 (Dal1) is a tumour suppressor molecule and often lost in various cancers, including breast cancer [34]. Zinc-finger proteins (represented here as ZNF598) are also frequently found as antigenic [3] most probably due to their conserved Zinc-finger domains. These proteins usually localized in the nucleus and many of them expressed only during embryogenesis, thus over-expression in various cancers might be able to elicit immune responses. Another protein that might be relevant in tumour biology is Dyacilglycerol-kinase-kappa (DGKK). Diacylglycerol kinases catalyze the phosphorylation of diacylglycerol, which is a key intracellular signalling molecule that activates the protein kinase C pathways, one of the most important targets of oncotherapy [35].

Using the genes corresponding to all the classifiers obtained from peptide and protein arrays, an over-representation analysis (ORA) was performed. An over-representation of the genes associated with cell organelles like spliceosome was observed in the classifiers from both, peptide and protein, arrays. The plausible explanation of this over-representation can be deduced from the hypothesis put forward by Tan [36] and Hardin [37] that auto-antibodies often target cell organelles (*i.e.* protein complexes) rather than a single protein.

One conceivable explanation might be that cancer growth and invasion releases cell debris into the circulation and as a consequence evoke immune response. Spliceosomes which are involved in alternative spicing may have a role in tumourigenesis. Processes like cell cycle control, signal transduction, angiogenesis, metastasis and apoptosis might be affected as alternative splicing affects majority of the human genes. Two-thirds of all the human genes' transcripts are known to undergo alternative splicing. Although the function of the encoded protein does not alter in most of the cases, some may exhibit a malignant phenotype [38]. ORA of the classifiers from the peptide array revealed the over-representation of genes associated with ribosome and pentose phosphate pathway. Like spliceosome, ribosomes are frequently targeted by auto-antibodies [3]. Apart from playing a pivotal role in translational regulation, the ribosomal proteins are also associated with the processes like cellular transformation, tumour growth, aggressiveness and metastasis [39]. Similarly, the pentose phosphate pathway plays an important role in tumour proliferation by supplying reduced levels of Nicotinamide adenine dinucleotide phosphate (NADP) and carbons for intracellular anabolic processes in cancerous cells [40]. Over-representation of the genes associated with endocytosis was found among the genes corresponding to the protein array classifiers. Deregulated expression of the endocytosis proteins may play a role in human cancers by affecting the control of cell proliferation. The enhancement of cell replication may be promoted through impaired endocytosis as a result of prolonged signalling by growth-factor receptors [41]. The genes from the protein array classifier also showed enrichment of the sequence motifs like coiled coils and ELR motifs. The sequence motifs like coiled coil and ELR motifs may have autoantigenic potentiality [3,42]. Chemokines with the ELR motifs activate the leukocytes, thus triggering the immune response [43].

Although recombinant protein expression in *E. coli* has been a method of choice, the process is spiked with problems like amount of desired protein expression, length of the protein and biologically active forms [44]. Expression of recombinant protein in *E. coli* often leads to the formation of biologically inactive inclusion bodies [45]. Above all else, the process of high-throughput recombinant protein expression and purification is time consuming and cumbersome. Shorter peptide sequences of the protein can recapitulate the corresponding biological activities of the protein and hence can act as alternative to full-length recombinant proteins [46]. Synthetic peptides can mimic biological activities of the proteins and, also,

rather easy to synthesize and manipulate. They are inexpensive to synthesize and highly stable [14,15]. When using protein arrays for the evaluation of patient sera, one may encounter the problems associated with the *E. coli* specific reactivity. With the usage of short synthetic peptides, the problem of *E. coli* specific reactivity can be avoided. These salient features make them a good candidate to replace protein arrays.

Conclusion

Classification success of the serum samples was moderate using protein-microarrays based on 642 sero-reactive clones identified using an initial macro-membrane screen. The corresponding peptides were able to classify the serum samples with reasonable sensitivities and specificities. Through the usage of peptide arrays, the difficulties associated with the protein arrays can be circumvented and thus providing the possibility of building a robust platform for early diagnosis of cancer. However, to establish peptide arrays as a potential breast-cancer diagnostic tool, test-sensitivities and specificities should be increased by additional antigenic peptides which then have to be thoroughly validated on larger sets of serum samples. This study provides evidence that *in silico* designed peptides could improve classification success and thus be a good alternative to protein-arrays for auto-antibody based biomarker-development.

Acknowledgements

I want to thank Ronald Kulovics (AIT) for his help with the *E. coli* cultivation, protein isolation and printing microarrays.

Reference list

1. Pisani P, Bray F, Parkin DM: **Estimates of the world-wide prevalence of cancer for 25 sites in the adult population.** *Int J Cancer* 2002, **97:** 72-81.

2. Sturgeon CM, Duffy MJ, Stenman UH, Lilja H, Brunner N, Chan DW *et al*.: **National Academy of Clinical Biochemistry laboratory medicine practice guidelines for use of tumor markers in testicular, prostate, colorectal, breast, and ovarian cancers.** *Clin Chem* 2008, **54:** e11-e79.

3. Backes C, Ludwig N, Leidinger P, Harz C, Hoffmann J, Keller A *et al*.: **Immunogenicity of autoantigens.** *BMC Genomics* 2011, **12:** 340.

4. Ullah MF, Aatif M: **The footprints of cancer development: Cancer biomarkers.** *Cancer Treat Rev* 2009, **35:** 193-200.

5. Chapman C, Murray A, Chakrabarti J, Thorpe A, Woolston C, Sahin U *et al*.: **Autoantibodies in breast cancer: their use as an aid to early diagnosis.** *Ann Oncol* 2007, **18:** 868-873.

6. Disis ML, Calenoff E, McLaughlin G, Murphy AE, Chen W, Groner B *et al*.: **Existent T-cell and antibody immunity to HER-2/neu protein in patients with breast cancer.** *Cancer Res* 1994, **54:** 16-20.

7. Conroy SE, Gibson SL, Brunstrom G, Isenberg D, Luqmani Y, Latchman DS: **Autoantibodies to 90 kD heat-shock protein in sera of breast cancer patients.** *Lancet* 1995, **345:** 126.

8. Carter D, Dillon DC, Reynolds LD, Retter MW, Fanger G, Molesh DA *et al*.: **Serum antibodies to lipophilin B detected in late stage breast cancer patients.** *Clin Cancer Res* 2003, **9:** 749-754.

9. Montenarh M: **Chapter 20 - Humoral Immune Response Against the Growth Suppressor p53 in Human Malignancies.** In *Cancer and Autoimmunity*. Edited by Yehuda S, Eric MG. Amsterdam: Elsevier; 2000:193-203.

10. Lu H, Goodell V, Disis ML: **Humoral immunity directed against tumor-associated antigens as potential biomarkers for the early diagnosis of cancer.** *J Proteome Res* 2008, **7:** 1388-1394.

11. Stempfer R, Syed P, Vierlinger K, Pichler R, Meese E, Leidinger P *et al*.: **Tumour auto-antibody screening: performance of protein microarrays using SEREX derived antigens.** *BMC Cancer* 2010, **10:** 627.

12. Qian F, Odunsi K, Blatt LM, Scanlan MJ, Mannan M, Shah N *et al*.: **Tumor associated antigen recognition by autologous serum in patients with breast cancer.** *Int J Mol Med* 2005, **15:** 137-144.

13. Hamrita B, Chahed K, Kabbage M, Guillier CL, Trimeche M, Chaieb A *et al*.: **Identification of tumor antigens that elicit a humoral immune response in breast cancer patients' sera by serological proteome analysis (SERPA).** *Clin Chim Acta* 2008, **393:** 95-102.

14. Cretich M, Damin F, Pirri G, Chiari M: **Protein and peptide arrays: recent trends and new directions.** *Biomol Eng* 2006, **23:** 77-88.

15. Uttamchandani M, Yao SQ: **Peptide microarrays: next generation biochips for detection, diagnostics and high-throughput screening.** *Curr Pharm Des* 2008, **14:** 2428-2438.

16. Southern EM: **Analysing polynucleotide sequences.** 1988, Great Britain patent application GB 8810400.5.

17. Fodor SP, Read JL, Pirrung MC, Stryer L, Lu AT, Solas D: **Light-directed, spatially addressable parallel chemical synthesis.** *Science* 1991, **251:** 767-773.

18. Frank R: **The SPOT-synthesis technique. Synthetic peptide arrays on membrane supports--principles and applications.** *J Immunol Methods* 2002, **267:** 13-26.

19. Shin DS, Kim DH, Chung WJ, Lee YS: **Combinatorial solid phase peptide synthesis and bioassays.** *J Biochem Mol Biol* 2005, **38:** 517-525.

20. Kolaskar AS, Tongaonkar PC: **A semi-empirical method for prediction of antigenic determinants on protein antigens.** *FEBS Lett* 1990, **276:** 172-174.

21. Bussow K, Nordhoff E, Lubbert C, Lehrach H, Walter G: **A human cDNA library for high-throughput protein expression screening.** *Genomics* 2000, **65:** 1-8.

22. Preininger C, Bodrossy L, Sauer U, Pichler R, Weilharter A: **ARChip epoxy and ARChip UV for covalent on-chip immobilization of pmoA gene-specific oligonucleotides.** *Anal Biochem* 2004, **330:** 29-36.

23. Simon R, Lam A, Li MC, Ngan M, Menenzes S, Zhao Y: **Analysis of gene expression data using BRB-ArrayTools.** *Cancer Inform* 2007, **3:** 11-17.

24. Wright GW, Simon RM: **A random variance model for detection of differential gene expression in small microarray experiments.** *Bioinformatics* 2003, **19:** 2448-2455.

25. R Development Core Team: **R: A Language and Environment for Statistical Computing.** *R Foundation for Statistical Computing* 2005.

26. Simon R, Radmacher MD, Dobbin K, McShane LM: **Pitfalls in the use of DNA microarray data for diagnostic and prognostic classification.** *J Natl Cancer Inst* 2003, **95:** 14-18.

27. Keller A, Backes C, Al-Awadhi M, Gerasch A, Kuntzer J, Kohlbacher O *et al.*: **GeneTrailExpress: a web-based pipeline for the statistical evaluation of microarray experiments.** *BMC Bioinformatics* 2008, **9:** 552.

28. Bailey TL, Elkan C: **The value of prior knowledge in discovering motifs with MEME.** *Proc Int Conf Intell Syst Mol Biol* 1995, **3:** 21-29.

29. Ludwig N, Keller A, Leidinger P, Harz C, Backes C, Lenhof HP *et al.*: **Is there a general autoantibody signature for cancer?** *Eur J Cancer* 2012.

30. Reineke U, Sabat R: **Antibody epitope mapping using SPOT peptide arrays.** *Methods Mol Biol* 2009, **524:** 145-167.

31. Iuchi S: **Three classes of C2H2 zinc finger proteins.** *Cell Mol Life Sci* 2001, **58:** 625-635.

32. Geysen HM, Barteling SJ, Meloen RH: **Small peptides induce antibodies with a sequence and structural requirement for binding antigen comparable to antibodies raised against the native protein.** *Proc Natl Acad Sci U S A* 1985, **82:** 178-182.

33. Baneyx F, Mujacic M: **Recombinant protein folding and misfolding in Escherichia coli.** *Nat Biotechnol* 2004, **22:** 1399-1408.

34. Heller G, Geradts J, Ziegler B, Newsham I, Filipits M, Markis-Ritzinger EM *et al.*: **Downregulation of TSLC1 and DAL-1 expression occurs frequently in breast cancer.** *Breast Cancer Res Treat* 2007, **103:** 283-291.

35. Ron D, Kazanietz MG: **New insights into the regulation of protein kinase C and novel phorbol ester receptors.** *FASEB J* 1999, **13:** 1658-1676.

36. Tan EM: **Antinuclear antibodies: diagnostic markers for autoimmune diseases and probes for cell biology.** *Adv Immunol* 1989, **44:** 93-151.

37. Hardin JA: **The lupus autoantigens and the pathogenesis of systemic lupus erythematosus.** *Arthritis Rheum* 1986, **29:** 457-460.

38. van Alphen RJ, Wiemer EA, Burger H, Eskens FA: **The spliceosome as target for anticancer treatment.** *Br J Cancer* 2009, **100:** 228-232.

39. Zhu Y, Lin H, Li Z, Wang M, Luo J: **Modulation of expression of ribosomal protein L7a (rpL7a) by ethanol in human breast cancer cells.** *Breast Cancer Res Treat* 2001, **69:** 29-38.

40. Boros LG, Lee PW, Brandes JL, Cascante M, Muscarella P, Schirmer WJ *et al.*: **Nonoxidative pentose phosphate pathways and their direct role in ribose synthesis in tumors: is cancer a disease of cellular glucose metabolism?** *Med Hypotheses* 1998, **50:** 55-59.

41. Floyd S, De CP: **Endocytosis proteins and cancer: a potential link?** *Trends Cell Biol* 1998, **8:** 299-301.

42. Dohlman JG, Lupas A, Carson M: **Long charge-rich alpha-helices in systemic autoantigens.** *Biochem Biophys Res Commun* 1993, **195:** 686-696.

43. Strieter RM, Belperio JA, Phillips RJ, Keane MP: **CXC chemokines in angiogenesis of cancer.** *Semin Cancer Biol* 2004, **14:** 195-200.

44. Baneyx F: **Recombinant protein expression in Escherichia coli.** *Curr Opin Biotechnol* 1999, **10:** 411-421.

45. Singh SM, Panda AK: **Solubilization and refolding of bacterial inclusion body proteins.** *J Biosci Bioeng* 2005, **99:** 303-310.

46. Min DH, Mrksich M: **Peptide arrays: towards routine implementation.** *Curr Opin Chem Biol* 2004, **8:** 554-558.

Chapter 4: Identification of novel tumour-associated antigens in breast cancer

---

Keywords: SEREX, cDNA libraries, auto-antibodies, fusion genes, over-representation analysis

Abstract

There is an urgent need for the development of a diagnostic technique for the early identification of breast cancer. The aim of this study is to identify novel TAAs from breast cancer sera using SEREX which would enable an early and non invasive cancer diagnosis. cDNA expression libraries were made from 7 and 11 benign and malignant breast cancer tissue samples, respectively and these libraries were transfected in *E. coli*. These cDNA expression clones were screened with the autologous and allogeneic sera from the malignant and benign breast cancer patients. The sero-reactive clones were isolated and sequenced. The corresponding genes were compared with the SEREX database and over-representation analysis was performed. A mutation analysis was performed to see which of these genes are reported to be mutated across various cancer studies. The genes derived from these clones were used to analyze the microarray data downloaded from GEO.

Immunoscreening of the cDNA expression clones yielded 192 sero-reactive clones derived from malignant breast tumours which expressed 107 unique genes. Twenty-two clones expressed different genes when sequenced from 5' and 3' ends, which might represent fusion genes. Out of 107 genes, 22 genes were reported in the SEREX database from various cancer studies. The over-representation analyses of these 107 genes revealed enrichment of pathways like MAPK signalling pathway and spliceosome and secondary structures like coiled coils and GrB motifs. Fourteen of these 107 genes, when using microarray data from another study, were found to be differentially expressed in breast cancer samples and healthy controls. This study enabled the identification of 107 genes, out of which 22 genes have been reported in the SEREX database. This panel of antigens has been found enriched for MAPK signalling pathway proteins, which is unique when compared to the over-representation analysis of various other cancers. Twenty-two out of 192 clones identified in this study, express novel potential fusion genes in breast cancer. Furthermore 14 of 107 genes showed significant ($p<0.05$) differential expression in breast cancer patients and healthy controls.

Background

In 2008, an estimated 3.2 million new cases of cancer were reported from Europe. The most common cases of the cancers were colorectal cancer followed by breast, lung and prostate cancers which constitute 13.6%, 13.1%, 12.2% and 11.9% of the all the cancers, respectively [1]. Early diagnosis of breast cancer can improve the survival rate and hence, there is a great anticipation for the identification of biomarkers which can provide an early diagnosis. Protein biomarkers like tumour auto-antibodies can of use for such a need for early diagnosis. Auto-antibodies can be detected in the patient serum months or even years before the onset of the symptoms [2-4].

The discovery of MAGE-1 in melanoma by Thierry Boon's group using T-cell response has indicated the presence of proteins specifically expressed in tumour cells [5]. An immunological response can be evoked by a mutated or an aberrantly expressed protein, tumour-associated antigens (TAA), resulting in the production of auto-antibodies [6,7]. The corresponding auto-antibodies could be used as biomarkers for early diagnosis and prognosis of cancer [7,8]. These tumour associated antibodies may give an insight about the host-tumour interactions. The concerted action between the cellular and humoral immune systems results in the recognition of tumour antigens. SEREX (serological identification of antigens by recombinant expression cloning) employs immunoscreening of recombinant tumour cDNA libraries for specific interactions with the serum antibodies. SEREX plays a role in identifying the antigens recognized by B cells [9]. Besides identifying the antigens expressed on the cell surface, SEREX can detect the intracellular antigens [6,10]. This strategy was exploited by Sahin *et al.*[11] to come up with the concept of SEREX. Few of the typical features of SEREX are as follows: 1) the usage of fresh tumour specimens restricts the analysis to genes that are expressed by the tumour cells *in vivo* and thus, circumventing the experiments involving *in vitro* artefacts associated with short- and long-term tumour cell culture, 2) the usage of patients' serum which contains polyclonal antibodies which are used as probe for immunoscreening allowing the identification of multiple antigens with a single screening course [12]. More than 2700 tumour antigens have been identified using the SEREX method and enlisted in a database [13].

Identification of biomarkers for diagnosis and prognosis purpose has been quite a task for a long time now. As far as serum markers are concerned, few clinically beneficial antigenic markers have been identified. Jäger *et al*[14], identified *NY-ESO-1, SSX2* and ING1 along with 27 other genes in breast cancer. In another study, Jäger *et al.* [15] identified NY-BR-1 in breast cancer. The other TAAs associated with breast cancer include MAGE-3, MAGE-6, p53, ATK2, TMF1, TPD52, MAGE D, NY-BR-62, NY-BR-75, NY-BR-85, and NY-BR-96 [16].

In this study, 192 SEREX derived cDNA expression clones were identified which reacted to the sera from benign and malignant breast cancer patients and these clones express 107 genes. Out of 192 clones, 22 clones probably express fusion genes. A comparison was made between the lists of genes expressed by the SEREX derived clones and genes enlisted in the SEREX database. For systematically extracting the biological meaning and to find the over-represented genes from the list of genes expressed by the clones identified by SEREX, an over-representation analysis (ORA) was done using GeneTrail online tool [17]. ORA showed over-representation of genes associated with MAPK signalling pathway and spliceosome. Over-representation of genes associated with spliceosome was observed in colon and thyroid cancer studies obtained from the SEREX database. However, none of the 9 SEREX database derived cancer studies showed enrichment for MAPK signalling pathway. Using the Roche Cancer Genome Database [18], the information on different types of somatic mutations for each gene expressed by the cDNA expression clones was gathered. The comparison showed only 2 genes which are reported as mutated in the breast cancer samples analyzed. These 107 genes were analyzed using the data derived from a microarray experiment and found 14 genes which differentially expressed in breast cancer and healthy control samples.

Method

Tissue and serum samples

Tissue and serum samples were obtained after the approval from the patients and stored at -80°C until used. The study was approved by the Ethics Committee of the Medical University of Vienna and the General Hospital of Vienna (study number: 143/2007) and the research was carried out in compliance with the Helsinki Declaration. For the construction of two

cDNA libraries, 7 and 11 tissue samples from benign and malignant breast cancer patients were used, respectively. The subsequent immunoscreening of the cDNA expression clones was done using sera pools made from patient serum samples. The benign serum pool was made from 8 serum samples. Four of these 8 samples were obtained from the patients whose tissue was used for the cDNA library construction. The malignant serum pool was made from the collection of 13 malignant patients. The median ages of the benign breast cancer patients, whose tissue and serum samples were used for the cDNA library construction and the immunoscreening, were 44±13.4 and 47.5±10.5 years (median±standard deviation), respectively. The clinical cohorts and the pathological characteristics of the malignant breast cancer tissue and the serum samples can be found in the table 1.

Table 1. Clinical and pathological data of the patient-study cohort. Malignant samples were collected from patients invasive ductal carcinoma and lobular carcinoma

	Malignant (tissue)	Malignant (serum)
Age (years)[a]	49±13.4	64±12.4
Ductal carcinoma	10	7
Lobular carcinoma	1	6
Grading[b]	G1 [1]; G2 [3]; G3 [7]	G1 [3]; G2 [6]; G3 [4]
pT stage[c]	pT1b [1]; pT1c [2]; pT2 [5]; pT3 [1]; pTis [2]	pT1 [1]; pT1b [2]; pT1c [6]; pT2 [3]; pT4b [1]
pN stage[d]	pN0 [4]; pN1 [2]; pN1a [1]; pN2a [1]; pN3 [2]; pNX [1]	pN0[5]; pN1 [1]; pN1a [1]; pN1mi [2]; pN2a [2]; pNX [1]
Her2/neu positive	2	13
Oestrogen positive	5	9
Progesteron positive	4	2

[a]The age of the patients represented as median age±standard deviation
[b]G1 (low-grade), G2 (intermediate grade) and G3 (high-grade). Low-grade tumours are usually slow growing and are less likely to spread. High-grade tumours are likely to grow more quickly and are more likely to spread.
[c]pT1: Tumour 2.0 cm or less in greatest dimension; pT1b: Tumour more than 0.5 cm but not more than 1.0 cm in greatest dimension; pT1c: Tumour more than 1.0 cm but not more than 2.0 cm in greatest dimension; pT2: Tumour more than 2.0 cm but not more than 5.0 cm in greatest dimension.pT3: Tumour more than 5.0 cm; pT4b: oedema or ulceration of the breast; pTis: Carcinoma in situ
[d]pN stage: pN0: No regional lymph node metastasis; pN1: Metastasis to movable ipsilateral axillary lymph node(s); pN1a: Only micrometastasis (none larger than 0.2 cm); pN1mi: Micrometastases (greater than 0.2 mm, but none greater than 2.0 mm); pN2a. Metastasis in 4-9 axillary lymph nodes, including at least one that is larger than 2 mm; pN3: Metastasis to ipsilateral internal mammary lymph node(s); pNX: Regional lymph nodes cannot be assessed. Information of 12 serum samples was available. The numbers within [] represent the number of samples.

Total RNA extraction and cDNA library construction

Total RNA was isolated from 7 benign and 11 malignant breast cancer tissues using AllPrep DNA/RNA Mini Kit (Qiagen). Poly $(A)^+$ mRNA was isolated from the total RNA using Oligotex mRNA Spin-Column protocol (Qiagen) and pools of 250 µL each were made. The malignant

and the benign pools contained 1.72 µg contained 0.57 µg of mRNA, respectively. For the generation of cDNA library, a cDNA Synthesis Kit, ZAP express vector, oligo (dT) primer, *Eco*RI adaptors, ZAP-cDNA Gigapack III Gold Cloning Kit and the *E. coli* competent strains (XL1-Blue-MRF´ and XLOLR) were purchased from Stratagene and the prescribed protocol was followed. In brief, 5 µg of the Poly $(A)^+$ mRNA was used for the preparation of cDNA expression library. An oligo (dT) primer with an internal *XhoI* site and 5-methyl-CTP was used to synthesize the first-strand. Prior to the cloning into ZAP express vector, the cDNA fragments were ligated with *Eco*RI adaptors. This was followed by packaging into phage particles, using ZAP-cDNA Gigapack III Gold Cloning Kit, which were later on used to transfect *E. coli* (XL1-Blue-MRF´). Blue-white screening was conducted to select the transfectants. Libraries containing 6.1 X 10^4 plaque forming units (pfu)/mL recombinants were amplified and later used for immunoscreening.

Over-night bacterial culture

A single colony of *E. coli* XL1-Blue-MRF´or XLOLR was inoculated in 25 mL of LB medium (for 1 L: 10 g of bactotryptone, 5 g of yeast extract and 10 g of NaCl) with 0.0125 mg/ml tetracycline, 10 mM $MgSO_4$ and 10 mM 20% maltose. The inoculated media was incubated over-night at 37°C and 225 rpm. The over-night culture was then centrifuged at 2000 rpm for 20 min and the OD_{600} was adjusted to 0.5 or 1.0 with 10 mM $MgSO_4$.

Amplification of the library

To the 200 µL of the over-night culture of *E. coli* XL1-Blue-MRF´ (OD_{600} 0.5), +5 X 10^4 pfu of the library was added and incubated for 20 min at 37°C. To this mixture, 7.5 mL of pre-warmed (48°C) top agar (for 100 mL: 2.3 g of NZCYM and 0.75 g of agar) was added and was vortexed briefly. This reaction mixture was plated on NZCYM-agar (for 1 L: 23 g of NZCYM and 15 g of agar) plates with 0.0125 mg/mL tetracycline and incubated at 37°C until the plaques were visible. When the plaques appeared on the culture plate, 10 mL of SM buffer was added to the plate and incubated over-night at 4°C on a shaker. The following morning the SM buffer was removed and stored in a falcon tube. With a sterile spatula the top agar was scraped off and was transferred into the same falcon tube. The plate was rinsed with 2 mL of SM buffer and was collected into the falcon tube. To the falcon tube, chloroform (v/v 5%) was added and incubated for 30 min at room temperature. The tube was then

centrifuged at 350 rpm for 10 min and the supernatant was transferred into a new falcon tube. To this tube, 10% Polyethyleneglycol (PEG 6000) and 1 M NaCl were added and incubated over-night at 4°C. The overnight reaction mixture was centrifuged at 10,000 rpm for 15 min at 4°C and the supernatant was discarded. The pellet was resuspended using 2 mL of SM buffer and chloroform (v/v 10%) was added. After a brief vortex, the mixture was centrifuged at 10,000 rpm for 15 min and the supernatant containing the phages was transferred to a new falcon tube.

Titering

The packaged reaction mixtures and the amplified cDNA libraries were diluted 1:10, 1:100, 1:1000 and 1:10,000. To the aliquots of 200 µL of the over-night culture of *E. coli* XL1-Blue-MRF´ (OD_{600} 0.5), 2 µL of the aforementioned dilutions and the undiluted samples were added and incubated for 20 min at 37°C. To this mixture 2.5 mL of pre-warmed (48°C) top agar was added and briefly vortexed. This reaction mixture was plated on NZCYM-agar plates with 0.0125 mg/mL tetracycline and incubated over-night at 37°C. The following morning the number of clear plaques for each dilution was counted and the titre values were determined using the formula mentioned below.

$$\text{Titre} = \left[\frac{\text{pfu X Dilution factor}}{\text{Volume plated (µL)}}\right] \text{X } 1000 \text{ µL/mL}$$

Depletion of *E. coli* specific antibodies

To the 600 µL of over-night culture of *E. coli* XL1-Blue-MRF´ (OD_{600} 0.5) 6 µL of the non-transfected *E. coli* cells were added. 7.5 mL of pre-warmed (48°C) Top agar was added and the mixture was plated onto NZCYM agar plates (for 1 L: 23 g NZCYM medium, 15 g agar) and incubated for 4 h at 42°C. Nitrocellulose membrane (Whatman) soaked in 10 mM of Isopropyl β-D-thiogalactoside (IPTG) was put on the NZCYM agar plate with bacterial culture and was incubated for 4 h at 37°C. Then the membrane was peeled off the plate and washed thrice for 10 min each with Tris-buffered saline and 0.1% Tween20 (TBS-T). The membrane was blocked with 5% low-fat milk powder in TBS, followed by washing of the membrane thrice for 10 min each with TBS. In order to deplete the *E. coli* specific antibodies pre-absorption of pooled sera from 10 malignant breast cancer patients was done. 30 mL of sera

pool diluted 1:100 in TBS, 0.5% low-fat milk powder and 0.01% thimerasol was applied onto the membranes and incubated at 4°C over-night. After the collection of pooled sera, the membrane was washed thrice for 10 min each with (TBS-T) and blocked for 1 h with 5% low-fat milk powder in TBS. After washing the membrane for thrice for 10 min each with TBS, pooled sera was applied onto the membrane and incubated at 4°C over-night. The collected sera pool was used for the subsequent immunoscreening of the cDNA library.

Immunoscreening of the cDNA library

Immunoscreening was performed thrice to eliminate the false positives and to obtain the monoclonality of the positive clones. For the primary screening nearly 500,000 clones, obtained from the cDNA libraries constructed from malignant and benign breast cancer tissues each, were screened. The positive clones obtained from the primary screening were subjected to another round of screening. For the secondary screening each culture plate was divided into 5 parts, allowing culturing of 5 clones from the primary screening at the same time. At the end of the secondary screening duplicates of the positive signals from each part of the culture plates were collected. Similarly, this method was followed for the tertiary screening and two positive clones were picked.

Immunoscreening of the transfectants was done using the pre-absorbed pool of 10 malignant and 8 benign breast cancer sera. The 10,000 pfu were added to 600 µL of *E. coli* (XL1-Blue-MRF´) suspension and incubated for 20 min at 37°C. This was followed by addition of 7.5 mL of pre-warmed (48°C) Top agar to the reaction mixture. The culture was plated on pre-warmed (42°C) NZCYM agar plates and incubated for 4 h at 42°C. Nitrocellulose membrane soaked in 10 mM of IPTG was put on the NZCYM agar plate with bacterial culture and was incubated for 4 h at 37°C. The membrane was stripped and the culture plate was stored at 4°C. The stripped membrane was then washed three times for 10 min each with TBS-T and blocked with 5% low-fat milk powder in TBS for 1 h. Then the membrane was washed thrice for 10 min each with TBS and the pre-absorbed sera pool was applied onto the membrane and incubated at 4°C over-night. The following morning the sera pool was collected and the membrane was washed thrice for 10 min each with TBS. Then, the membrane was incubated with secondary antibody conjugated with alkaline phosphatase (Dianova) diluted 1:10,000 in TBS, 0.5% low-fat milk powder for 45 min at room

temperature. The membrane was then washed thrice for 10 min each with TBS and later incubated with detection buffer (colour detection solution, 0.025 mg/mL 5-bromo-4-chloro-3-indolyl phosphate (BCIP), 0.05 mg/mL nitroblue tetrazolium (NBT)) for 20 min in dark. The positive spots are represented by dark stained spots and the corresponding lytic plaques on the culture plates were picked and stored in 500 µL SM buffer (for 1 L: 5.8 g NaCl, 2.46 g $MgSO_4.7H_2O$, 6.06 g Tris, pH 7.5 and add 5 mL of 2% gelatin) at 4°C.

In vivo excision

The over-night cultures of *E. coli* XL1-Blue MRF´ and XLOLR were then centrifuged at 2000 rpm for 20 min and the OD_{600} was adjusted to 1.0 with 10 mM $MgSO_4$. 250 µL of phage solution and 1 µL of ExAssist helper phage were added to 200 µL of *E. coli* XL1-Blue MRF´ and incubated for 15 min at 37°C. To this mixture 3 mL of NZCYM medium was added and incubated for 3 h at 37°C and 225 rpm. The cells were then lysed at 60°C for 20 min and centrifuged for 20 min at 3000 rpm. pBKCMV phagemid is now in the supernatant 10 µL of the supernatant was added to 200 µL *E. coli* XLOLR and incubated for 15 min at 37°C. To this mixture 300 µL of NZCYM medium was added and incubated at 37°C, 225 rpm for 45 min. Later the culture was spread onto LB agar (for 1 L: 10 g of bactotryptone, 5 g of yeast extract, 10 g of NaCl and 15 g of agar) plate with 0.05 mg/mL kanamycin and incubated over-night at 37°C. The following morning a single colony was picked and added to 5 mL LB medium with 0.05 mg/mL kanamycin and incubated over-night at 37°C and 225 rpm. An aliquot of 500 µL was mixed with 500 µL autoclaved glycerol and stored at -80°C as a back-up. The rest of the culture was used for the isolation of plasmid.

Sequence analysis of identified antigens

The isolated plasmid DNA was subjected to restriction digestion using *EcoR I* and *XhoI* and checked on gel for the cDNA insert. The digested product was then sequenced using M13 forward and reverse primers. The resulting genes were compared to the SEREX database. To identify over-representation of the genes, expressed by the SEREX derived clones, with specific function or characteristics ORA was performed using GeneTrail [17]. Simultaneously, genes reported from various cancers in the SEREX database were used for the ORA. These sets of genes were compared to a reference set “all human genes” by using the

hypergeometric distribution test implemented in GeneTrail. The significance threshold value was set to 0.05 and only the KEGG pathways with p-values less than 0.05 were considered significant.

Mutation analysis

For elucidation of genes known mutated in various cancers, information was gathered from the Roche Cancer Genome Database [18]. This database utilizes various sources of human mutation databases like the *Catalogue of Somatic Mutations in Cancer* (COSMIC), the *Cancer Genome Atlas* project, *Online Mendelian Inheritance in Man* (OMIM), the *IARC TP53 database*, *KinMutBase* and the *L1CAM mutation database* [19].

Comparison with the GEO dataset

The gene expression dataset GSE20437 which used for the differentiation of normal and the breast cancer sample was downloaded from Gene Expression Omnibus (GEO). This dataset was generated by Graham *et al* [20]., wherein they analyzed 36 breast cancer samples and 6 healthy control samples. Expression-data of genes corresponding to the identified SEREX expression clones were selected based on the Gene-Symbol from this GEO-dataset. A class comparison was conducted to identify the differentially expressed genes using BRB-ArrayTools software 3.8.1 [http://linus.nci.nih.gov/BRB-ArrayTools.html] developed by Simon *et al.*[21]. The genes which were expressed differentially at a p-value ≤0.05 were considered as significant.

Results

cDNA library construction and immunoscreening

The average concentrations of total RNA pools for cDNA libraries construction made from the malignant and benign tissue samples were 522 ng/µL and 305 ng/µL, respectively. After packaging the synthesized cDNA libraries into the phages, the titre values of the libraries were determined and amplified. The titre values of the amplified malignant and benign cDNA libraries were found to be 0.94 X 10^8 pfu/mL and 0.46 X 10^8 pfu/mL, respectively.

Three rounds of immunoscreening of the cDNA expression clones were performed and the positive clones were picked in duplicates. From the primary screening 434 and 61 clones were identified using malignant and benign sera pools, respectively. From the secondary screening, 165 clones were found reacting positively to the malignant serum pool, while only 2 clones reacted to the benign serum pool. From the tertiary screening of the malignant cDNA expression library a total of 189 cDNA expression clones were identified.

Sequence analysis

The plasmid DNA from the positive clones, identified by the immunoscreening, was sequenced using M13 forward and reverse primers. BLAST alignment elucidated 105 and 2 unique genes expressed by the clones identified by malignant and benign breast cancer sera, respectively (Tab. 2). It was observed that 22 of the 192 clones encoded 2 entirely different gene products when sequenced from 5’ and 3’ ends using M13-forward primer and M13-reverse primer (Tab. 3). This observation suggests the possibility that these cDNA expression clones may well be expressing fused genes, which is a common phenomenon observed in cancers [22].

Table 2. Genes expressed by the SEREX-derived cDNA expression clones.

Gene Symbol	Entrez Gene Id	Gene Symbol	Entrez Gene Id	Gene Symbol	Entrez Gene Id	Gene Symbol	Entrez Gene Id
SRRT	51593	*IL33*	90865	*CLINT1*	9685	*MAPKAPK2*	9261
DDX3Y	8653	*SPHK1*	8877	*YIPF3*	25844	*HSP90AB3P*†	3327
SLC12A1	6557	*NCAPG*	64151	*EEF1A1**	1915	*CLUAP1**	23059
*GAPDH**	2597	*M11S1**	4076	*RNF12*	51132	*IGFBP3*	3486
B3GALNT2	148789	*SF3A1*	10291	*FAM149B1*	317662	*HSP90AB4P*†	664618
LOC253264†	253264	*C1S*	716	*MRPS16*	51021	*MAP3K3*	4215
UBE2K	3093	*BAG3**	9531	*RPL3*	6122	*DSP**	1832
CAPNS1	826	*MTRNR2L1*	100462977	-	100293090†	*SF1*	7536
LOC729082†	729082	*SCCPDH*	51097	*NLN*	57486	*VASH2*	79805
*P4HB**	5034	*GPKOW*	27238	*PGAP3*	93210	*SDPR*	8436
RPL36A	6173	*FOXK1*	221937	*SSR1*	6745	*NUDCD3*	23386
MTIF2	4528	*HSPA6*	3310	*FAM102A*	399665	*HNRNPA3*	220988
*RCD-8**	23644	*GPRC5C*	55890	-	100289576	*PQBP1*	10084
*D21S2056E**	8568	*ZC3H12A*	80149	-	389293†	*WDR73*	84942
TCF7	6932	*CAPN6*	827	*RNF8**	9025	*DDI2*	84301
PIK3C2B	5287	*DBNL**	28988	-	100293563†	*RNF157*	114804
LOC100129060†	100129060	*ELOVL5*	60481	*RRBP1**	6238	*CENPF**	1063
WDR6	11180	*ISCA1*	81689	*SPARCL1**	8404	*TRIM56*	81844
AKR7A2	8574	*GBA*	2629	*AMY2B**	280	*PDLIM1*	9124
COL4A2	1284	-	387933†	*LOC100130331*	100130331†	*LOC441155*†	441155
LCORL	254251	*WDR45L**	56270	*CCNB2*	9133	-	728534
-	100129211†	*SAFB2*	9667	*EGF*	1950	*SYNGR2*	9144
*PAIP1**	10605	*HSPA1A*	3303	*WDR74*	54663	*ZC3H11A**	9877
TRAK1	22906	*HSPA1L*	3305	*MFN1*	55669	*RPS9*	6203
TMEM236	653567	*SDCCAG33**	10194	*F13A1*	2162	*LOC642311*†	642311
ZNF197	10168	*AGAP11*	119385	*HOXC11*	3227	*LOC100288974*†	100288974
ATXNL2#	11273	**IGHV1-69***#†	28461	***PIP4K2A***#	5305		

#genes encoded by the clones derived from benign breast cancer tissue. *genes enlisted in the SEREX database. †genes not used for the over-representation analysis.

Table 3: Clones expressing possible fusion proteins.

Clone id	Primer	Gene Symbol	Entrez Gene Id	Chromosome	Clone id	Primer	Gene Symbol	Entrez Gene Id	Chromosome
Ben2	Forward	*ATXN2L*	28461	14	**Mal30A**	Forward	*TRAK1*	22906	3
	Reverse	*IGHV1-69**	11273	16		Reverse	*TMEM236*	653567	10
Mal12A	Forward	*SLC12A1*	6557	15	**Mal31A**	Forward	*CLINT1*	9685	5
	Reverse	-	100293563	-		Reverse	*YIPF3*	25844	6
Mal13A	Reverse	*GAPDH**	2597	12	**Mal345A1**	Forward	*FAM149B1*	317662	10
	Forward	*ZNF197*	10168	3		Reverse	*MRPS16*	51021	10
Mal140A	Forward	*LOC642311*	642311	15	**Mal37A**	Forward	-	100293090	5
	Reverse	*LOC100288974*	100288974	10		Reverse	*MTRNR2L1*	100462977	17
Mal15B	Forward	*AMY2B**	280	1	**Mal437B1**	Forward	*HSPA1A*	3303	6
	Reverse	*LOC100130331*	100130331	1		Reverse	*HSPA1L*	3305	6
Mal18A	Forward	*UBE2K*	3093	4	**Mal438A1**	Reverse	*HNRNPA3P5*	387933	13
	Reverse	*CAPNS1*	826	19		Forward	*HNRNPA3*	220988	2
Mal21A	Forward	*RRP1*	8568	21	**Mal454A1**	Reverse	*IGFBP3*	3486	7
	Reverse	*TCF7*	6932	5		Forward	*PIP4K2A*	5305	10
Mal21B	Forward	*PIK3C2B*	5287	1	**Mal460-2**	Forward	*WDR73*	84942	15
	Reverse	*MFN1*	55669	3		Reverse	*DDI2*	84301	1
Mal23A	Forward	*HSP90AB3P*	3327	4	**Mal4B**	Forward	*SSR1*	6745	6
	Reverse	*HSP90AB4P*	664618	15		Reverse	*FAM102A*	399665	9
Mal26A	Forward	*AKR7A2*	8574	1	**Mal5A**	Reverse	-	100289576	
	Reverse	*COL4A2*	1284	13		Forward	*GPRC5C*	55890	17
Mal27A	Forward	*LCORL*	254251	4	**Mal8B**	Reverse	*ISCA1P1*	389293	5
	Reverse	*NCAPG*	64151	4		Forward	*ISCA1*	81689	9

*genes enlisted in the SEREX database

These 107 genes were compared with the SEREX database and sorted out 22 genes which were already enlisted in the SEREX database from various cancer studies including breast cancer. Two genes, *GAPDH* and *RCD-8* were reported exclusively from breast cancer studies. While, genes like *IGHV1-69, P4HB*, *DBNL* and *CENPF* were reported in the studies from hepatocellular carcinoma, colon, lung, prostate and oesophageal cancers along with breast cancer study (Tab. 4).

Table 4. Genes expressed by the clones which are enlisted in the SEREX database. Gene-names are underlined which have been identified in previous Breast cancer studies enlisted in the SEREX database.

Gene Symbol	Full name	Cancer study in SEREX DB
AMY2B	Amylase, alpha 2B	prostate cancer
BAG3	BCL2-associated athanogene 3	stomach cancer
CENPF	Centromere protein F	hepatocellular carcinoma, colon cancer, breast cancer, lung cancer, prostate cancer, oesophageal cancer
CLUAP1	Clusterin associated protein 1	colon cancer, ovarian cancer
D21S2056E	Ribosomal RNA processing 1 homolog (*S. Cerevisiae*)	stomach cancer
DBNL	Drebrin-like	prostate cancer, breast cancer
DSP	Desmoplakin	lung cancer, prostate cancer
EEF1A1	Eukaryotic translation elongation factor 1 alpha 1	colon cancer
GAPDH	Glyceraldehyde-3-phosphate dehydrogenase	breast cancer
IGHV1-69	Immunoglobulin heavy variable 1-69	lung cancer, breast cancer
M11S1	Cell cycle associated protein 1	lung cancer
P4HB	Prolyl 4-hydroxylase, beta polypeptide	hepatocellular carcinoma, colon cancer, breast cancer
PAIP1	Poly(A) binding protein interacting protein 1	colon cancer
RCD-8	Enhancer of mRNA decapping 4	breast cancer
RNF12	Ring finger protein, LIM domain interacting	renal cancer
RNF8	Ring finger protein 8	pancreas cancer
RRBP1	Ribosome binding protein 1 homolog 180kda (dog)	melanoma
SDCCAG33	Teashirt zinc finger homeobox 1	renal cancer, colon cancer
SPARCL1	Sparc-like 1 (hevin)	normal testis
WDR45L	WDR45-like	melanoma
ZC3H11A	Zinc finger CCCH-type containing 11A	lung cancer, prostate cancer, stomach cancer

Over-representation of antigenic pathways and functional categories

To investigate the potential functions and the pathways in which the identified 107 unique genes from this study may possibly be involved in, ORA was performed using GeneTrail [17], with “all human genes” as reference set. Out of 107 sequences, 92 sequences were annotated with significant biological information. The sequences without annotation or limited biological information were eliminated from the analysis; these are indicated in Tab. 2. ORA of the 92 genes suggested that breast cancer sera were enriched with antibodies against *signalling proteins*, *heat shock proteins* and *various intracellular proteins*. Out of 92 genes, 9 genes were found to be involved in the KEGG *spliceosome* and *MAPK signalling* pathways. Three genes, *SF3A1, HNRNPA3* and *PQBP1,* were involved exclusively in the KEGG spliceosome pathway. Genes like *EGF, MAPKAPK2* and *MAP3K3* involved exclusively in the

MAPK signalling pathway. There were 3 genes, *HSPA6, HSPA1A* and *HSPA1L*, which were over-represented and were involved in both the aforementioned KEGG pathways.

The genes reported from breast, colon, colorectal, kidney, lung, ovary, prostate, stomach and thyroid cancers in the SEREX database were subjected to ORA, as well. The enriched KEGG pathways in the cDNA expression clones from this study (represented as "BrCa_SEREX Exp") were compared to the enriched KEGG pathways in the above-mentioned cancers. The spliceosome pathway was enriched in colon and thyroid cancers along with "BrCa_SEREX Exp". There was very limited overlap observed among the cancer data sets as shown in the figure 1.

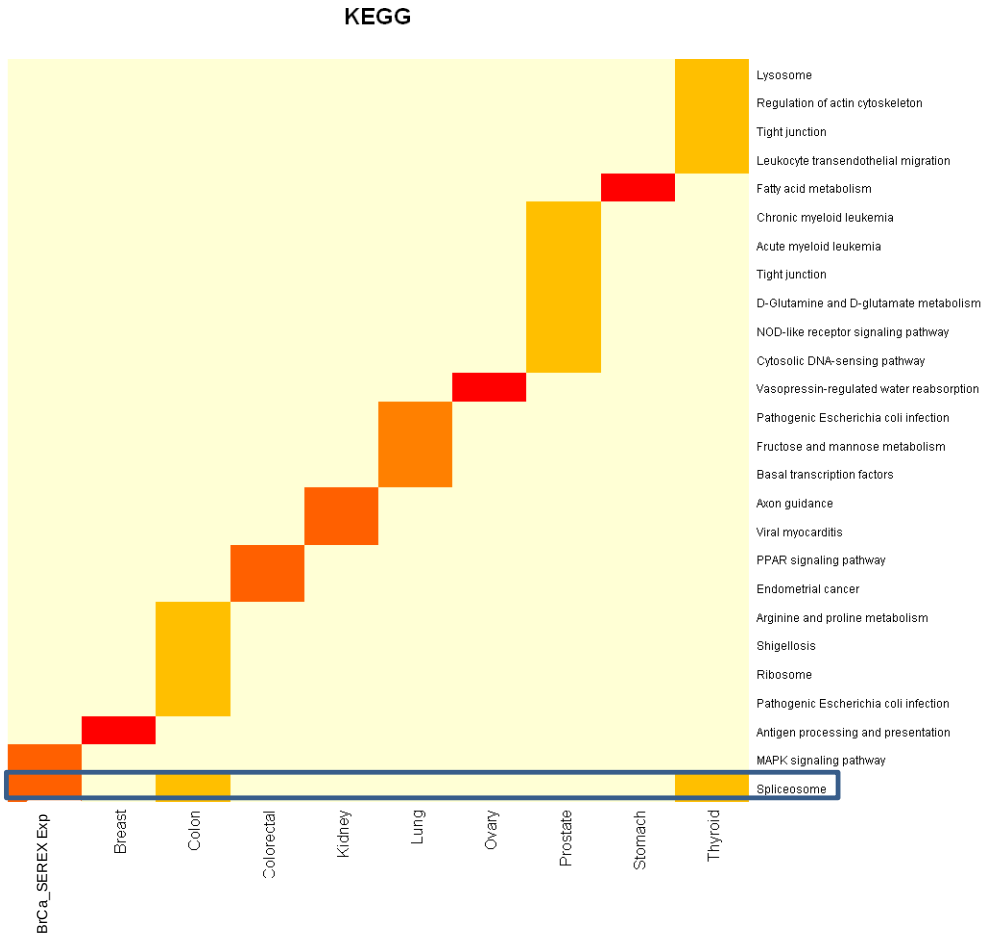

Figure 1: Heat map showing the enriched KEGG pathways in various cancer data sets. The heat map shows the p-values of the significantly enriched pathways in various cancers. The darkest colour represents the most significant (least p-value = 1.90913E-06) and the lightest colour represents the least significant (highest p-value = 0.05) entity. The p-values≤0.05 were considered significant. Of the significantly enriched pathways of "BrCa_SEREX Exp", and the SEREX-clones known for colon and thyroid cancers, only the spliceosome pathway overlapped.

The genes corresponding to the "BrCa_SEREX Exp" and the 9 cancers showed enrichment of secondary structures like Granzyme B (GrB), coiled coils, ELR and RGD (Tab. 6). Coiled coils and GrB motifs were enriched in all the cancers. The ELR motifs were enriched in all the cancers except "BrCa_SEREX Exp" and colorectal cancer; the RGD motif was enriched only in colon and prostate cancers (Tab. 6).

Table 6: Enriched secondary structures in various cancers. This table shows the p-values of the enriched secondary structures in various cancers, using the corresponding genes enlisted in the SEREX database

Secondary structure	BrCa_SEREX Exp	Breast	Colon	Colorectal	Kidney	Lung	Ovary	Prostate	Stomach
GrB	0.000577	1.07E-07	6.64E-14	0.006957	2.82E-05	4.56E-06	7.25E-07	2.13E-14	1.22E-08
Coiled coils	0.006896	7.58E-21	1.14E-20	0.008346	3.38E-14	1.49E-08	1.42E-14	6.21E-25	1.58E-22
ELR	-	0.002581	8.68E-08	-	0.001428	5.85E-05	6.48E-06	8.74E-09	0.001077
RGD	-	-	0.006211	-	-	-	-	0.004176	-

The role of mutations for antigenicity

Out of the 107 genes corresponding to the 192 cDNA expression clones, 23 genes were found to have somatic mutations across various cancers when compared with the Roche Cancer Genome Database (Tab. 7). The cancers from which these mutations were reported include glioma, colon cancer, astrocytoma, lung carcinoma, renal cell carcinoma, pancreatic cancer, breast cancer (infiltrating duct carcinoma, NOS (c50._)), melanoma, pleural mesothelioma, peritoneal adenocarcinoma, ovarian carcinoma and breast carcinoma. Out of these 23 genes, the Roche Cancer Genome Database enlisted 3 genes, namely *MAP3K3, SF1* and *B3GALNT2*, known mutated in breast cancer. The remaining 20 genes (SEREX clones) with a matched entry in the database of mutated genes have been found in single samples of various cancers (Table 7).

Table 7: Number of mutated samples for the given genes in the Roche Cancer Genome Database.

Gene symbol	Glioma	Colon cancer	Astrocytoma	Lung carcinoma	Renal cell carcinoma	Pancreatic cancer	Breast cancer (infiltrating duct carcinoma, NOS (c50._))	Breast carcinoma	Melanoma	Pleural mesothelioma	Peritoneal ductal adenocarcinoma	Ovarian carcinoma
F13A1	2	-	-	-	-	-	-	-	-	-	1	-
SDPR	-	-	-	-	-	1	-	-	-	-	-	-
HSPA1L	-	-	-	-	1	-	-	-	-	1	-	-
MFN1	-	1	-	-	-	-	-	-	-	-	-	-
WDR6	1	-	-	-	-	-	-	-	-	-	-	-
NCAPG	-	1	-	-	-	-	-	-	-	-	-	-
PQBP1	-	1	-	-	-	-	-	-	-	-	-	-
EGF	1	-	-	-	-	-	-	-	-	-	-	-
SPHK1	-	-	-	-	1	-	-	-	-	-	-	1
TRIM56	1	-	-	-	1	-	-	-	1	-	-	-
PIK3C2B	21	-	1	1	-	-	-	-	-	-	-	5
CENPF	33	-	-	-	-	-	-	-	-	-	-	-
SF1	-	-	-	-	-	-	-	1	1	-	-	-
COL4A2	1	-	-	-	-	-	-	-	-	-	-	-
EDC4	2	-	-	-	-	-	-	-	-	-	-	-
MAP3K3	-	1	-	-	-	-	1	-	-	-	-	-
IGFBP3	2	2	-	-	-	-	-	-	-	-	-	-
SF3A1	-	1	-	-	-	1	-	-	-	-	-	-
B3GALNT2	-	-	-	-	-	-	-	1	-	-	-	-
AKR7A2	1	-	-	-	-	-	-	-	-	-	-	-
CLUAP1	-	-	-	-	-	-	-	-	-	-	1	-
DSPP	1	-	-	-	-	-	-	-	-	-	-	-
DSP	36	-	-	-	-	-	-	-	-	-	-	-

Over-expression of SEREX derived clones in breast cancer

The microarray data reporting a gene expression study (GSE20437) of 36 breast cancer patients (18 samples from patients undergone reduction mammoplasty, 9 oestrogen receptor positive samples and 9 oestrogen receptor negative samples) and 6 healthy controls was downloaded from Gene Expression Omnibus. The gene-expression data of the 107 genes corresponding to the SEREX-cDNA expression clones were selected form the entire microarray data. Class comparison elucidated 14 genes (Tab. 8) which were differentially expressed in breast cancer patients versus healthy controls ($p<0.05$). Seven out of the 14 genes, namely *RRP1, RPS9, SAFB2, FAM149B1, AKR7A2, SRRT* and *PQBP1*, were found being over-expressed in breast cancer samples (Fig. 2). The other 7 genes, namely *MTIF2, ISCA1, DSP, PIP4K2A, ZC3H12A, TRAK1* and *SF3A1*, were over-expressed in the control samples (Fig. 3).

Table 8: Class comparison analyses elucidated 14 differentially expressed genes (p<0.05). Out of 107 SEREX-derived genes/ clones in breast cancer (n=36) and healthy (n=6) controls. Unique Id is the annotation used for the genes used in the microarray experiment. Fold change is the ratio of the median of the signal intensities derived from the cancer and control sample for a given gene.

Gene symbol	Unique id (Microarray)	p-value	Fold-change
SAFB2	32099_at	0.003	1.52
SF3A1	201356_at	0.005	0.7
ZC3H12A	218810_at	0.001	0.72
TRAK1	202080_s_at	0.001	0.7
PQBP1	210499_s_at	0.01	1.27
FAM149B1	213896_x_at	0.01	1.43
AKR7A2	202139_at	0.01	1.4
DSP	200606_at	0.02	0.74
MTIF2	203095_at	0.02	0.84
RRP1	218758_s_at	0.02	2.05
SRRT	222047_s_at	0.02	1.35
RPS9	214317_x_at	0.03	1.54
PIP4K2A	212829_at	0.04	0.72
ISCA1	209274_s_at	0.04	0.77

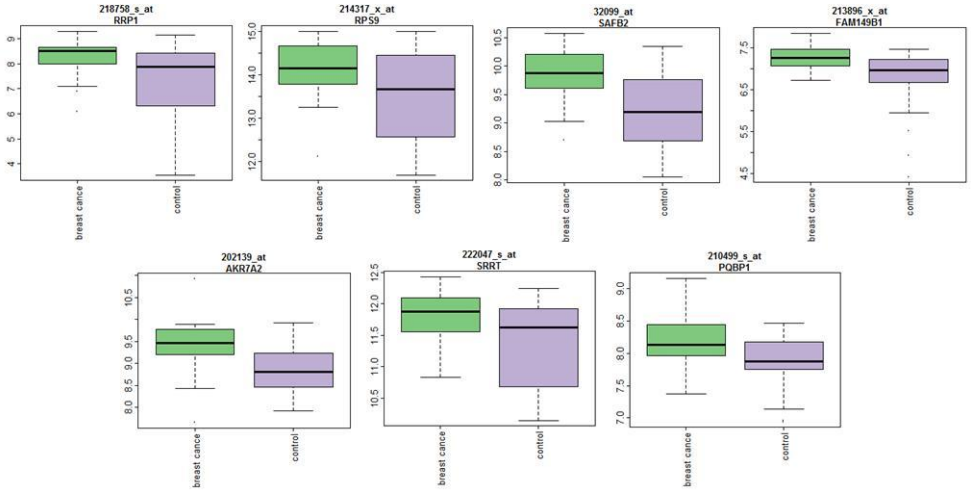

Figure 2: Box-plots of the log2 transformed signal intensities. The box-plots in this figure show the signal intensities of the genes that are over-expressed in the breast cancer samples.

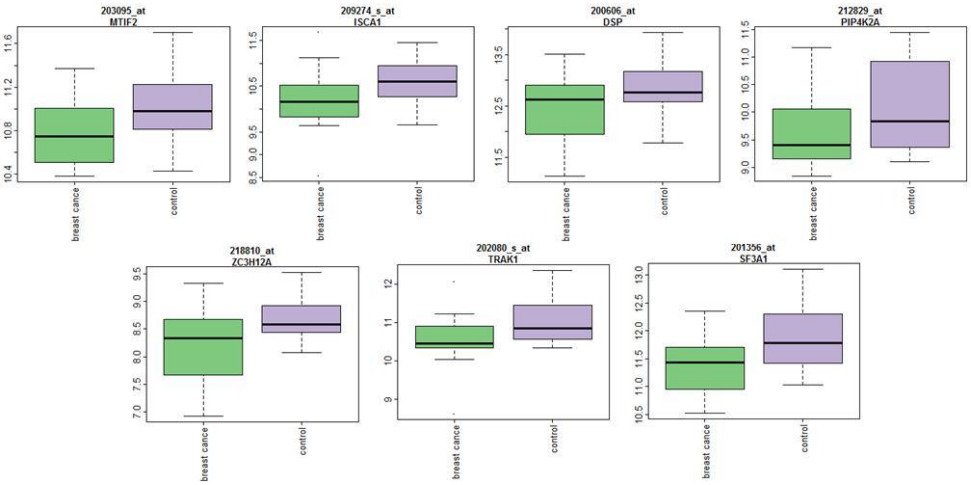

Figure 3: Box-plots of the log2 transformed signal intensities. The box-plots in this figure show the signal intensities of the genes which are over-expressed in the controls.

Discussion

Auto-antibodies against the TAAs have been reported as potential biomarkers for early diagnosis of cancers and SEREX has been used for a while for the identification of such auto-antibodies. Other prominent methods for identifying auto-antibodies are 2D gel electrophoresis and mass spectrometry [7,23].

This study confirmed the potential antigenicity by comparison to SEREX database, gene-mutations / fusions, gene-expression changes and overrepresentation analyses. Using

SEREX, 192 cDNA expression clones were identified that elucidated immunogenic response in breast cancer patients. These 192 clones expressed 107 unique genes/proteins, out of which 22 have been reported in the SEREX database from various cancer studies. Out of these 22 genes, *GAPDH* and *RCD-8* were reported from breast cancer studies while *P4HB*, *DBNL, IGHV1-69* and *CENPF* were reported from various cancer studies along with the breast cancer studies in the SEREX database. *CENPF, GAPDH, P4HB* and *DBNL* reported to be over-expressed in other breast cancer studies [24-27]. Although, reports of *IGHV1-69* being over-expressed in breast cancer could not be found, *IGHV1-69* over-expression is observed in chronic lymphocytic leukaemia [28]. *RCD-8* has been previously reported as autoantigen associated with autoimmune diseases like primary biliary cirrhosis and Sjögren's syndrome [29,30].

Although the rest of the 85 genes were not found to be reported in the SEREX database, one could expect their association with breast cancer as these genes were expressed by the cDNA expression clones selected by immunoscreening using autologous patient sera. It may be speculated that the humoral immune response against cancer related antigens arise from over-expression [31,32]. However, for establishing these antigens to as biomarkers, these antigens have to be validated using more robust technique like ELISA, MALDI, or protein microarrays.

It was found that 22 clones encoded hybrids of 2 different genes when sequenced from each of the ends. This observation gives a chance to speculate that these clones might be derived from tumours expressing fusion-genes. Although somatic rearrangements are often found in the cancer genomes and are intrachromosomal, their role in the development of cancer is poorly characterized [33]. Out of the 22 clones, 4 clones expressed sets of genes which are found to be located on the same chromosome. For example, *AMY2B* and *LOC100130331*, *LCORL* and *NCAPG*, *FAM149B1* and *MRPS16*, and *HSPA1A* and *HSPA1L* are located on the chromosome 1, 4, 10 and 6, respectively. Out of the 44 genes presented by these 22 clones, there were 3 genes, *SLC12A1, GAPDH* and *CAPNS1,* which reported to have fused with other genes. *SLC12A1* is known to be fused with *LEO1* in gliomas [34], *GAPDH* is fused with *BCL6* in B-cell lymphomas [35] and *CAPNS1* is fused with *WDR62* in ovarian cancer [36]. All these findings hint the possibility of the SEREX clones expressing fusion genes, hence suggesting

the possible reason for their immunogenicity. However, the real nature of the genes expressed by these 22 clones has yet to be validated.

Apart from the mutations, over-expressed proteins contribute to the humoral response against tumour antigens [37]. To find out the nature of the genes expressed by the cDNA expression clones and the sets of genes which are over-represented and clustering into certain KEGG pathways in breast cancer, ORA was performed using GeneTrail. This analysis identified enriched pathways like MAPK signalling pathway and spliceosome. MAPK signalling pathway is known to be associated with control of growth signals, cell survival, and invasion in cancer. In colorectal cancer, MAPK signalling pathway plays a role of regulating apoptosis [38]. Over-representation of the genes associated with the spliceosome was also observed. Majority of the human gene transcripts undergo alternative splicing. Although the functions of the encoded proteins do get affected by the alternative splicing, some may exhibit a malignant phenotype [41]. The ORA of the genes from other 9 cancer studies showed enrichment of various KEGG pathways (Fig. 1). However, there were no overlapping pathways found between the genes identified from this study and the genes representing the breast cancer study from the SEREX database. Although, the enrichment of spliceosome was also found in the colon and thyroid cancers, MAPK signalling pathway enrichment was unique to the set genes expressed by the clones identified from this study. Since MAPK signalling pathway is known to be associated with other cancers, it could be hypothesized that it may play a role in promoting immune response in breast cancer.

In addition to the enriched pathways, enrichment of sequence motifs like GrB cleavage sites and coiled coils was observed. GrB is found in the cytoplasmic granules of CTLs and NK cells and plays a vital role in inducing apoptosis. GrB cleavage sites are identified in autoimmune diseases and as well as meningioma [42,43]. From a study on autoantigens, conducted by Dohlman *et al.*, it was suggested that coiled-coils may have autoantigenic potential [44]. This indicates the immunogenicity of antigens with such sequence motifs against breast cancer. These motifs were enriched in all the other cancer studies from the SEREX database used for ORA.

A higher number of mutated proteins in cancer leads to the increase in the levels of auto-antibodies and hence are easily recognized [45]. Comparison of the genes expressed by the

cDNA expression clones with the Roche Cancer Genome Database lead to the identification of 23 genes which are known to have some sort of mutation and association with cancers (Tab. 7). Out of 23 genes, 12 genes were found to be mutated in glioma samples and only 2 genes were mutated in the breast cancer samples. This may be because most of the samples reported were from the glioma patients.

Irrespective of the nature of the gene/protein, the clones identified in this study showed immune-reactivity to the sera from breast cancer patients. To determine their usefulness as a possible panel for diagnosis, these proteins are needed to have some sort of confirmation and validation.

One way of validating the antigenic nature of these genes was to see if they are being differentially expressed in breast cancer patients and healthy control individuals. This was done using the microarray data, Graham *et al.*[20], derived from a study which was used for differentiating breast cancer samples from healthy control samples. Out of the 107 genes, from this study, 14 genes were found to be significant and differentially expressed in breast cancer samples and healthy control samples. Using this microarray, Graham and colleagues found the enrichment of MAPK pathway in their classifier. This finding along with the findings of the ORA suggests the role of MAPK signalling pathway in breast cancer.

Conclusion

From this study, 192 sero-reactive clones were identified which expressed 107 genes. There were 22 clones which gave different gene products when sequenced from 5’ and 3’ ends, suggesting the possibility of having gene fusions. The 107 genes were over-represented for the KEGG pathways like MAPK signalling pathway and spliceosome. Over-representation of the genes associated with spliceosome was also observed in chapter 3, where entirely different set of serum samples were used. Alternative splicing is one of the reasons for the immunogenicity of the self-proteins (autoantigens) [46]. The limited overlap between the enriched pathways of these over-represented genes and the genes from the other SEREX studies hints the novelty of the genes from this study. The genes, from this study, also showed enriched of sequence motifs which are characteristics of autoantigens [46]. Similar enrichment of sequence motifs in different sets of genes from chapter 2 and 3 were

observed. The mutation analysis showed that most of the genes reported from this study are known to be mutated in glioma and not breast cancer. These genes were also found to be differently expressed in breast cancer patients and the healthy individuals, which could be confirmed by analyzing a microarray data from another study. Having showed the qualities of the typical TAAs, these genes/proteins may prove to be important in differentiating breast cancer samples from healthy controls.

Reference list

1. Ferlay J, Parkin DM, Steliarova-Foucher E: **Estimates of cancer incidence and mortality in Europe in 2008.** *Eur J Cancer* 2010, **46:** 765-781.

2. Lubin R, Zalcman G, Bouchet L, Tredanel J, Legros Y, Cazals D *et al*.: **Serum p53 antibodies as early markers of lung cancer.** *Nat Med* 1995, **1:** 701-702.

3. Trivers GE, De B, V, Cawley HL, Caron G, Harrington AM, Bennett WP *et al*.: **Anti-p53 antibodies in sera from patients with chronic obstructive pulmonary disease can predate a diagnosis of cancer.** *Clin Cancer Res* 1996, **2:** 1767-1775.

4. Abendstein B, Marth C, Muller-Holzner E, Widschwendter M, Daxenbichler G, Zeimet AG: **Clinical significance of serum and ascitic p53 autoantibodies in epithelial ovarian carcinoma.** *Cancer* 2000, **88:** 1432-1437.

5. van der BP, Traversari C, Chomez P, Lurquin C, De PE, Van den Eynde BJ *et al*.: **A gene encoding an antigen recognized by cytolytic T lymphocytes on a human melanoma.** *J Immunol* 2007, **178:** 2617-2621.

6. Anderson KS, LaBaer J: **The sentinel within: exploiting the immune system for cancer biomarkers.** *J Proteome Res* 2005, **4:** 1123-1133.

7. Casiano CA, Mediavilla-Varela M, Tan EM: **Tumor-associated antigen arrays for the serological diagnosis of cancer.** *Mol Cell Proteomics* 2006, **5:** 1745-1759.

8. Sanchez-Carbayo M: **Antibody arrays: technical considerations and clinical applications in cancer.** *Clin Chem* 2006, **52:** 1651-1659.

9. Wang Y, Gu Q, Liu B, Zhu Z: **Perspectives of SEREX-defined antigens in diagnosis and immunotherapy for gastric cancer.** *Cancer Biol Ther* 2004, **3:** 806-811.

10. Chen YT, Scanlan MJ, Sahin U, Tureci O, Gure AO, Tsang S *et al*.: **A testicular antigen aberrantly expressed in human cancers detected by autologous antibody screening.** *Proc Natl Acad Sci U S A* 1997, **94:** 1914-1918.

11. Sahin U, Tureci O, Schmitt H, Cochlovius B, Johannes T, Schmits R *et al*.: **Human neoplasms elicit multiple specific immune responses in the autologous host.** *Proc Natl Acad Sci U S A* 1995, **92:** 11810-11813.

12. Li G, Miles A, Line A, Rees RC: **Identification of tumour antigens by serological analysis of cDNA expression cloning.** *Cancer Immunol Immunother* 2004, **53:** 139-143.

13. SEREX database. http://ludwig-sun5.unil.ch/CancerImmunomeDB/

14. Jager D, Stockert E, Scanlan MJ, Gure AO, Jager E, Knuth A *et al*.: **Cancer-testis antigens and ING1 tumor suppressor gene product are breast cancer antigens: characterization of tissue-specific ING1 transcripts and a homologue gene.** *Cancer Res* 1999, **59:** 6197-6204.

15. Jager D, Stockert E, Gure AO, Scanlan MJ, Karbach J, Jager E *et al*.: **Identification of a tissue-specific putative transcription factor in breast tissue by serological screening of a breast cancer library.** *Cancer Res* 2001, **61:** 2055-2061.

16. Scanlan MJ, Gout I, Gordon CM, Williamson B, Stockert E, Gure AO *et al*.: **Humoral immunity to human breast cancer: antigen definition and quantitative analysis of mRNA expression.** *Cancer Immun* 2001, **1:** 4.

17. Backes C, Keller A, Kuentzer J, Kneissl B, Comtesse N, Elnakady YA *et al*.: **GeneTrail--advanced gene set enrichment analysis.** *Nucleic Acids Res* 2007, **35:** W186-W192.

18. Küntzer J, Eggle D, Lenhof HP, Burtscher H, Klostermann S: **The Roche Cancer Genome Database (RCGDB).** *Hum Mutat* 2010, **31:** 407-413.

19. Küntzer J, Maisel D, Lenhof HP, Klostermann S, Burtscher H: **The Roche Cancer Genome Database 2.0.** *BMC Med Genomics* 2011, **4:** 43.

20. Graham K, de las MA, Tripathi A, King C, Kavanah M, Mendez J *et al*.: **Gene expression in histologically normal epithelium from breast cancer patients and from cancer-free prophylactic mastectomy patients shares a similar profile.** *Br J Cancer* 2010, **102:** 1284-1293.

21. Simon R, Lam A, Li MC, Ngan M, Menenzes S, Zhao Y: **Analysis of gene expression data using BRB-ArrayTools.** *Cancer Inform* 2007, **3:** 11-17.

22. Mitelman F, Johansson B, Mertens F: **The impact of translocations and gene fusions on cancer causation.** *Nat Rev Cancer* 2007, **7:** 233-245.

23. Fernandez MF: **Autoantibodies in breast cancer sera: candidate biomarkers and reporters of tumorigenesis.** *Cancer Lett* 2005, **230:** 187-198.

24. Higashimura Y, Nakajima Y, Yamaji R, Harada N, Shibasaki F, Nakano Y *et al*.: **Up-regulation of glyceraldehyde-3-phosphate dehydrogenase gene expression by HIF-1 activity depending on Sp1 in hypoxic breast cancer cells.** *Arch Biochem Biophys* 2011, **509:** 1-8.

25. Zhang D, Tai LK, Wong LL, Chiu LL, Sethi SK, Koay ES: **Proteomic study reveals that proteins involved in metabolic and detoxification pathways are highly expressed in HER-2/neu-positive breast cancer.** *Mol Cell Proteomics* 2005, **4:** 1686-1696.

26. Kessels MM, Engqvist-Goldstein AE, Drubin DG: **Association of mouse actin-binding protein 1 (mAbp1/SH3P7), an Src kinase target, with dynamic regions of the cortical actin cytoskeleton in response to Rac1 activation.** *Mol Biol Cell* 2000, **11:** 393-412.

27. Brendle A, Brandt A, Johansson R, Enquist K, Hallmans G, Hemminki K *et al*.: **Single nucleotide polymorphisms in chromosomal instability genes and risk and clinical outcome of breast cancer: a Swedish prospective case-control study.** *Eur J Cancer* 2009, **45:** 435-442.

28. Murray F, Darzentas N, Hadzidimitriou A, Tobin G, Boudjogra M, Scielzo C *et al*.: **Stereotyped patterns of somatic hypermutation in subsets of patients with chronic lymphocytic leukemia: implications for the role of antigen selection in leukemogenesis.** *Blood* 2008, **111:** 1524-1533.

29. Yu JH, Yang WH, Gulick T, Bloch KD, Bloch DB: **Ge-1 is a central component of the mammalian cytoplasmic mRNA processing body.** *RNA* 2005, **11:** 1795-1802.

30. Bloch DB, Rabkina D, Quertermous T, Bloch KD: **The immunoreactive region in a novel autoantigen contains a nuclear localization sequence.** *Clin Immunol Immunopathol* 1994, **72:** 380-389.

31. Bauer C, Brass N, Diesinger I, Kayser K, Grasser FA, Meese E: **Overexpression of the eukaryotic translation initiation factor 4G (eIF4G-1) in squamous cell lung carcinoma.** *Int J Cancer* 2002, **98:** 181-185.

32. Comtesse N, Keller A, Diesinger I, Bauer C, Kayser K, Huwer H *et al*.: **Frequent overexpression of the genes FXR1, CLAPM1 and EIF4G located on amplicon 3q26-27 in squamous cell carcinoma of the lung.** *Int J Cancer* 2007, **120:** 2538-2544.

33. Stephens PJ, McBride DJ, Lin ML, Varela I, Pleasance ED, Simpson JT *et al*.: **Complex landscapes of somatic rearrangement in human breast cancer genomes.** *Nature* 2009, **462:** 1005-1010.

34. Bralten LB, Kloosterhof NK, Gravendeel LA, Sacchetti A, Duijm EJ, Kros JM *et al*.: **Integrated genomic profiling identifies candidate genes implicated in glioma-genesis and a novel LEO1-SLC12A1 fusion gene.** *Genes Chromosomes Cancer* 2010, **49:** 509-517.

35. Montesinos-Rongen M, Akasaka T, Zuhlke-Jenisch R, Schaller C, Van RD, Wiestler OD *et al*.: **Molecular characterization of BCL6 breakpoints in primary diffuse large B-cell lymphomas of the central nervous system identifies GAPD as novel translocation partner.** *Brain Pathol* 2003, **13:** 534-538.

36. McPherson A, Hormozdiari F, Zayed A, Giuliany R, Ha G, Sun MG *et al*.: **deFuse: an algorithm for gene fusion discovery in tumor RNA-Seq data.** *PLoS Comput Biol* 2011, **7:** e1001138.

37. Reuschenbach M, von Knebel DM, Wentzensen N: **A systematic review of humoral immune responses against tumor antigens.** *Cancer Immunol Immunother* 2009, **58:** 1535-1544.

38. Fang JY, Richardson BC: **The MAPK signalling pathways and colorectal cancer.** *Lancet Oncol* 2005, **6:** 322-327.

39. Tan EM: **Antinuclear antibodies: diagnostic markers for autoimmune diseases and probes for cell biology.** *Adv Immunol* 1989, **44:** 93-151.

40. Hardin JA: **The lupus autoantigens and the pathogenesis of systemic lupus erythematosus.** *Arthritis Rheum* 1986, **29:** 457-460.

41. van Alphen RJ, Wiemer EA, Burger H, Eskens FA: **The spliceosome as target for anticancer treatment.** *Br J Cancer* 2009, **100:** 228-232.

42. Casciola-Rosen L, Andrade F, Ulanet D, Wong WB, Rosen A: **Cleavage by granzyme B is strongly predictive of autoantigen status: implications for initiation of autoimmunity.** *J Exp Med* 1999, **190:** 815-826.

43. Comtesse N, Zippel A, Walle S, Monz D, Backes C, Fischer U *et al*.: **Complex humoral immune response against a benign tumor: frequent antibody response against specific antigens as diagnostic targets.** *Proc Natl Acad Sci U S A* 2005, **102:** 9601-9606.

44. Dohlman JG, Lupas A, Carson M: **Long charge-rich alpha-helices in systemic autoantigens.** *Biochem Biophys Res Commun* 1993, **195:** 686-696.

45. Babel I, Barderas R, az-Uriarte R, Martinez-Torrecuadrada JL, Sanchez-Carbayo M, Casal JI: **Identification of tumor-associated autoantigens for the diagnosis of colorectal cancer in serum using high density protein microarrays.** *Mol Cell Proteomics* 2009, **8:** 2382-2395.

46. Backes C, Ludwig N, Leidinger P, Harz C, Hoffmann J, Keller A *et al*.: **Immunogenicity of autoantigens.** *BMC Genomics* 2011, **12:** 340.

Conclusions and outlook

Conclusions

The aim of this thesis was to identify novel auto-antibody signatures which enable early diagnosis of breast cancer. SEREX technique has been exploited the identification of tumour associated autoantigens for a while now. Such autoantigens identified using sera from brain and lung cancer patients were used for the generation of protein macro- and microarrays. A comparison of the performances of both the arrays led to the observation that protein microarrays display better correlation between the replicates of the same sample compared to the macroarrays [1]. Using protein microarrays the brain and lung cancer samples could be correctly classified. This study reported in the chapter 1 helped in establishing a protein microarray platform for the identification of auto-antibody signatures against tumour autoantigens using only few microliters of sera.

Although, the comparison between the microarrays and macroarrays yielded results in favour of microarray (Chapter 1), macroarrays are advantageous when a large collection of cDNA expression clones is used for evaluation. Since the macroarrays harbour the bacterial colonies, the expression of the recombinant proteins was done on the macroarray itself. In the study reported in the chapter 2, these macroarrays were used to short list a sizable panel of clones which were reactive to the serum samples. These sero-reactive clones were identified and were cultured for the recombinant proteins. The purified His-tag proteins were used for the production of protein microarrays and were used for the evaluation of sera from the breast cancer patients and the healthy individuals. Although, the differentiation of the benign and the malignant samples could not be achieved, the breast cancer samples could be differentiated from the healthy controls. One reason which could be hypothesized, for not being able to differentiate the benign and the malignant samples, is that the expression of recombinant proteins in *E. coli* leads to the formation of misfolded proteins [2]. These misfolded proteins may not be exposing all the immunogenic epitopes and hence the mediocre result. Over-expression of autoantigens is often observed in cancers [3] and the genes identified in chapter 2 showed over-representation of genes associated with various pathways and protein families associated with cancers.

Considering the possibility of having misfolded proteins which may interfere with the identification of all the possible auto-antibody signatures in breast cancer, a decision was made to use the synthetic peptides. These potential antigenic peptides sequences were deduced from the proteins expressed by the cDNA expression clones from the chapter 2. The usage of the synthetic peptides could solve the problem associated with the *E. coli*-expression. In chapter 3, basing on the maximum antigenic scores the peptide sequences were deduced and were synthesized. These peptides sequences showed enrichment of Zinc-finger domain motif which is a characteristic of tumour-associated antigens [4]. The synthetic peptides were used for the generation of the peptide microarrays. A comparison between the protein and peptide microarrays was made upon evaluation of same set of sera from benign and malignant breast cancer patients and healthy controls. A higher degree of discrimination of samples was obtained using peptide microarrays when compared to protein microarrays. The over-representation analysis of the classifiers showed the enrichment of MAPK signalling pathway, which is known to be associated with breast cancer [5]. Secondary structural motifs, which are the features for autoantigens in autoimmune diseases, were enriched as well [3].

Simultaneously, SEREX strategy was used for the identification of tumour auto-antibodies against autoantigens in the sera from breast cancer patients (Chapter 4). cDNA expression clones were made using the cDNA obtained from the benign and malignant breast cancer patients and the sero-reactive clones were picked. The corresponding genes from these clones were found to be mutated in various cancers. However, only couple of them were reported being mutated in breast cancer. Mutated- or aberrantly expressed proteins act as antigens evoking immune response which results in the production of auto-antibodies [6]. Basing on the information obtained from sequencing of the plasmid DNA obtained from these clones, it could be speculated that some of these clones may be are expressing fusion genes. When these genes were used for evaluation on a microarray data, some of these genes showed differential expression. When all the genes expressed by the SEREX derived clones were subjected to over-representation analysis, enrichment of pathways which are associated with cancers was observed. However, very little overlap between the enriched pathways of this data and the data obtained from 9 other cancer studies was observed. This

suggests the novelty of the genes identified in this experiment. In this study, potential fusion genes, a common phenomenon observed in cancers [7].

Outlook

The conventional diagnostic methods for breast cancer like mammography, ultra sound imaging and magnetic sound resonance imaging, lack sensitivity and specificity and even may necessitate biopsy [8]. Since the auto-antibodies can be detected in the patient sera prior to the onset of the disease symptoms, they serve as perfect alternative to the existing diagnostic methods. The microarrays can be employed for the identification of such auto-antibodies using only few microliters of patient sera.

The microarray studies conducted during the course of my thesis have led to the identification of panels of potential tumour auto-antibodies. However, for these potential auto-antibody biomarkers to be dubbed as clinically validated biomarkers, validation has to be performed using much larger sample size. The usage of *E. coli* expressed recombinant proteins may restrict the identification of auto-antibodies up to some extent as such recombinant proteins are usually misfolded. The usage of synthetic peptides, on the other hand, could solve the problem associated with the *E. coli*-expression. Although, peptide sequences with maximum score were used in this thesis, one could use the tiling peptide arrays for a better diagnostic approach. Tiling peptide array strategy would ensure that all the possible antigenic motifs within the given protein sequence are presented on the array surface and thus can provide a robust platform for serum auto-antibody evaluation. The usage of peptide array would cut short the money required for the recombinant protein expression and saves valuable time.

The SEREX study identified the cDNA expression clones which express potential fusion genes and express genes which are over-represented for pathways relevant to cancer. The recombinant proteins from these clones can be expressed and used for the production of protein microarrays. These protein microarrays can be used for validating the antigenic nature of the proteins expressed by these cDNA expression clones. Alternatively, the corresponding protein sequences can be used for the deduction of potential antigenic peptide sequences and can be used for the generation of peptide microarrays.

The features like the enriched pathways, the possibility of having fusion genes, over-expression and mutation analysis are the typical features of tumour associated antigens. Thus, I conclude that the SEREX derived clones and the classifiers from protein and peptide microarray approaches can be used to construct a minimal invasive platform for early diagnosis of breast cancer.

Reference list

1. Stempfer R, Syed P, Vierlinger K, Pichler R, Meese E, Leidinger P *et al*.: **Tumour auto-antibody screening: performance of protein microarrays using SEREX derived antigens.** *BMC Cancer* 2010, **10:** 627.

2. Baneyx F, Mujacic M: **Recombinant protein folding and misfolding in Escherichia coli.** *Nat Biotechnol* 2004, **22:** 1399-1408.

3. Backes C, Ludwig N, Leidinger P, Harz C, Hoffmann J, Keller A *et al*.: **Immunogenicity of autoantigens.** *BMC Genomics* 2011, **12:** 340.

4. Ludwig N, Keller A, Leidinger P, Harz C, Backes C, Lenhof HP *et al*.: **Is there a general autoantibody signature for cancer?** *Eur J Cancer* 2012.

5. Santen RJ, Song RX, McPherson R, Kumar R, Adam L, Jeng MH *et al*.: **The role of mitogen-activated protein (MAP) kinase in breast cancer.** *J Steroid Biochem Mol Biol.* 2002, **80**: 239-56.

6. Anderson KS, LaBaer J: **The sentinel within: exploiting the immune system for cancer biomarkers.** *J Proteome Res* 2005, **4:** 1123-1133.

7. Mitelman F, Johansson B, Mertens F: **The impact of translocations and gene fusions on cancer causation.** *Nat Rev Cancer* 2007, **7:** 233-245.

8. Gundry KR: **The application of breast MRI in staging and screening for breast cancer.** *Oncology (Williston Park)* 2005, **19**: 159-69.

Curriculum vitae

Personal Details

Name: **Parvez Syed**
Date of birth: 12th of August, 1982
Nationality: Indian
Marital status: Married
Languages: English, Urdu, Telugu, Hindi

Address:
Goldschlagstraße 21/2
1150, Vienna, Austria

Current work address:
Department of Health & Environment,
Molecular Diagnostics,
AIT Austrian Institute of Technology,
Muthgasse 11, 1190 Vienna

Telephone: +43 6641822123
e-mail: parvezsyedsyed@gmail.com
parvez.syed@ait.ac.at

Education

2008-Present: **Doctoral studies;** "Serum-autoantibody testing for early diagnosis of breast cancer", AIT Austrian Institute of Technology GmbH, Vienna, Austria

2006-2008: **Master of Science** in Molecular Biology with specialization in Biotechnology, University of Skövde, Skövde, Sweden

2008: **Master degree thesis;** "Evaluation of Autoantibody-Serum-Biomarkers for Breast Cancer Screening". Austrian Research Centers, Seibersdorf, Austria

2001-2004: **Bachelor of Science** (Genetics, Microbiology and Chemistry), Osmania University, Hyderabad, India

Expertise

Recombinant protein expression in E. coli, purification (His-Tag)

Tumour auto-antibody screening using protein microarrays

Serological expression of cDNA expression libraries (SEREX)

Publications

Stempfer R*, Syed P*, Vierlinger K, Pichler R, Meese E, Leidinger P, Ludwig N, Kriegner A, Nöhammer C, Weinhäusel A: Tumour auto-antibody screening: performance of protein microarrays using SEREX derived antigens. *BMC Cancer* 2010, 10: 627.

Luna-Coronell J*, Syed P*, Sergelen K, Gyurján I, Weinhäusel A: The current status of cancer biomarker research using tumor-associated antigens for minimal invasive and early cancer diagnostics. *J Proteomics*. Under review.

Syed P, Vierlinger K, Kriegner A, Sergelen K, Luna-Coronell J, Rappaport C, Nöhammer C, Singer C, Weinhäusel A: Evaluation of auto-antibody serum biomarkers for breast cancer screening *and in silico* analysis of sero-reactive proteins. *J Mol Biochem.* Under review.

Syed P, Gyurján I, Kriegner A, Vierlinger K, Singer CF, Rappaport-Fürhauser C, Zerweck J, Söllner J, Weinhäusel A: *In silico* design and performance of peptide microarrays for breast cancer tumour-auto-antibody testing. *J Mol Biochem.* Under review.

*equal contribution

Posters

Syed P, Kriegner A, Vierlinger K, Nöhammer C, Weinhäusel A: “Performance of Peptide array versus Protein array”; The EMBO Meeting 2011, Vienna; 10.09.2011 – 13.09.2011.

Syed P, Stempfer R, Vierlinger K, Pichler R, Meese E, Leidinger P, Ludwig N, Kriegner A, Nöhammer C, Weinhäusel A: "Performance of protein-microarrays using SEREX-derived antigens for auto-antibody Screening"; ÖGMBT Annual Meeting 2010, Vienna; 27.09.2010 - 29.09.2010.

Syed P, Fürhauser C, Winkler S, Schönthaler S, Stempfer R, Nöhammer C, Muhr D, Singer C, Weinhäusel A: "Evaluation of Autoantibody-Serum-Biomarkers for Breast Cancer Screening"; European Human Genetics Conference 2009, Vienna; 23.05.2009 - 26.05.2009.

Oral presentation

Syed P, Vierlinger K, Fürhauser C, Kriegner A, Schönthaler S, Nöhammer C, Singer C, Weinhäusel A: "Breast cancer serum testing using protein-micorarrays and *in silico* analysis of reactive antigens"; 12th Status Seminar Chip Technologies: Sequencing and Functional Genomics, Frankfurt, Germany: 04.02.2010 - 05.02.2010.