UNIVERSITY OF LATVIA

Development of approach for exploration of autoantibody profiles in cancer patients and identification of autoantibody signature for diagnosis of gastric cancer

Doctoral thesis
Department of Molecular Biology
Faculty of biology

Author: Pāvels Zajakins
Supervisor: Dr. biol. Aija Linē

Riga 2011

The doctoral study was carried out in the
Latvian Biomedical Research and Study Centre during 2005–2011.

The doctoral thesis is a summary of publications.

This research was supported by:

ERDF project No 2010/0231/2DP/2.1.1.1.0/10/APIA/VIAA/044, grant No 09.1288;

European Social Fund project
"Support for Doctoral Studies at the University of Latvia"
Nr.2009/0138/1DP/1.1.2.1.2/09/IPIA/VIAA/004;

University of Latvia Promotion Council Biology Riga, Latvia

Chairman: **Prof., Dr. biol. Viesturs Baumanis**
(University of Latvia, Riga, Latvia)

Opponents:

ABSTRACT

The spontaneous production of autoantibodies against tumour-derived proteins has been observed in the most, if not all cancer patients and they seem to be very attractive biomarkers for the diagnosis, early detection and prognosis of cancer and prediction of response to immunotherapy. However, so far their clinical utility has been hampered by the low frequency of antibodies against each individual antigen, heterogeneity of the autoantibody repertoire and its overlap with that in the response to tissue damage by viral infections or autoimmune diseases. These limitations at least partially could be overcome by exploiting protein microarray technology that allows detecting the autoantibodies against thousands of antigens simultaneously using a microliter of serum and hence represent a valuable tool for exploring autoantibody profiles in human sera and defining autoantibody signatures with the diagnostic or prognostic relevance.

This thesis is focused on the development of technology for the production and processing of phage-displayed antigen microarrays and procedures for the data analysis, and their application for the exploration of autoantibody profiles in patients with gastric cancer (GC), gastric inflammatory diseases and healthy individuals, and assessment of humoral immune response against sperm-associated antigens.

At first, series of experiments were performed to select the most suitable slide surface chemistry, printing conditions and the optimal method for amplification of phages, and to optimise the serum and antibody dilutions, incubation times and method of pre-absorption of serum. After having established the reproducibility of the technique, we applied it for the characterization of the humoral immune response against a heterogeneous group of proteins called sperm associated antigens (SPAG) that are commonly expressed in male germ cells, are capable to elicit immune response underlying infertility, and several of them recently have been shown to be expressed in cancers and proposed to be implicated in the development of cancer. Cancer-associated spontaneous humoral immune response was detected against SPAG1, SPAG6, SPAG8 and a novel testis-restricted splice variant SPAG17-A1 that allowed to classify them as novel CT antigens with potential relevance as immunotherapeutic targets and serological biomarkers.

Next, 1150-feature antigen microarrays were produced by printing the recombinant phage particles from the previously established cancer antigen clone collection on FAST slides and applied for the exploration of autoantibody profiles in melanoma patients and healthy individuals. Initial data analysis demonstrated that the conventional methods, such as LOWESS or OLIN commonly used for the normalisation of DNA microarray data and the standard statistical methods such as t-tests and regression analysis as well as the artificial neural network-based approaches that are widely used for the analysis of gene expression data, are not suitable for the analysis of antigen microarray data. Next, we evaluated the performance of support vector machines (SVM)-based approach with an improved kernel for the generation of biomarker model for discriminating sera from melanoma patients and healthy controls. Although, the improvement of SVM with ranking-based topological kernel resulted in higher classification accuracy, detailed examination of the biomarker model showed that it is based on the detection of small but consistent differences in the signal intensities between the cases and controls. In our opinion, the biological significance of such differences is unclear and the development of clinically applicable biomarker assay based on such model would be technically extremely challenging. Therefore next, we endeavoured to

establish the procedures for the data normalisation, defining cutoffs that allow to discriminate between sero-positive and sero-negative cases, and ranking of antigens considering the signal intensity and the frequency of reactivity. This approach was applied for the survey of autoantibody profiles in patients with GC, gastric inflammatory diseases and healthy individuals and resulted in the identification of 45- autoantibody signature that could discriminate between GC and healthy control sera with 74.5% accuracy (AUC of 0.79, 58% sensitivity and 91% specificity), GC and peptic ulcer with 73.0% accuracy, and GC and gastritis with 63.5% accuracy. Moreover, it could detect early GC with equal sensitivity than advanced GC thus demonstrating its relevance for the early detection of GC. Interestingly, the autoantibody production did not correlate with histological type, *H. pylori* status, grade, localization and size of the primary tumor while it appeared to be associated with the metastatic disease.

Contents

LIST OF THE ORIGINAL PUBLICATIONS

The current dissertation is based on the following publications referred in the text by their Roman numerals:

Original paper I

Kalniņa Z, Siliņa K, Meistere I, Zayakin P, Rivosh A, Ābols A, Leja M, Minenkova O, Schadendorf D and Linē A. Evaluation of T7 and Lambda phage display systems for survey of autoantibody profiles in cancer patients. *J Immunol Methods*, 2008 May 20;334(1-2):37-50. Epub 2008 Feb 21 (IF 2.3)

Original paper II

V. Jumutc*, P. Zayakin*, and A. Borisov. Ranking-based Kernels in Applied Biomedical Diagnostics using Support Vector Machine. (accepted to *International Journal of Neural Systems,* IF 4.2)

* The first two authors contributed equally to this work

Original paper III

K. Silina, P. Zayakin, Z. Kalnina, L. Ivanova, I. Meistere, E. Endzelins, A. Abols, A. Stengrevics, M. Leja, K. Ducena, V. Kozirovskis, A. Linē. Sperm associated antigens as targets for cancer immunotherapy: expression pattern and humoral immune response in cancer patients. *J Immunother*, 2011, Jan;34(1):28-44. (IF 3.59)

Original paper IV

Zayakin P, Kalniņa Z, Siliņa K, Meistere I, Ivanova L, Endzeliņš E, Jumutcs V, Stengrēvics A, Leja M, Wex T and Linē A. Tumour-associated autoantibody signatures for the early detection of gastric cancer (submitted to *Cancer Prevention Research,* IF 4.98)

In addition, the results of this work have been presented in 17 international conferences, from these 2 have been presented by the author:

P. Zayakin, Z. Kalnina, K. Silina, I. Meistere, L. Ivanova, A. Stengrevics, M. Leja, T. Wex, P. Malfertheiner, A. Line. The application of antigen microarray technology for the biomarker identification in gastric cancer. 5th Baltic Congress of Oncology. 14.05.2010-15.05.2010 Latvia, Acta Chirurgica Latviensis, Supplement 2010 (10/1) p. 18

Zayakin P, Kalniņa Z, Gailīte I, Siliņa K, Rivosh A, Pilāns D, Stengrēvics A, Linē A. Development of antigen microarray for analysis of the autoantibody repertoires in cancer patients. HHMI course on Modern Technologies on Gene Expression Detection and Data Integration, Debrecen, Hungary, 2006, July 18-26.

ABBREVIATIONS

APC	antigen presenting cell
BCR	B cell receptor
CSC	cancer stem cell
CT	cancer-testis
CTL	cytotoxic T lymphocyte
DC	dendritic cell
EMT	Epithelial-mesenchimal transition
HLA	human leukocyte antigen
IFN	interferon
IL	interleukin
GC	gastric cancer
mAb	monoclonal antibody
MDSC	myeloid-derived suppressor cell
NK cells	natural killer cells
ORF	open reading frame
PhD-AM	Phage-displayed antigen microarray
PhD-SEREX	Phage display-based SEREX
pfu	plaque forming unit
PSA	prostate-specific antigen
RT-PCR	reverse transcriptase polymerase chain reaction
SEREX	serological identification of antigens by recombinant expression cloning
SPT	serum pepsinogen test
SVM	support vector machines
TAA	tumour-associated antigen
TCR	T cell receptor
TIL	tumour-infiltrating lymphocytes
Treg	regulatory T cell
UTR	untranslated region
WHO	World Health Organisation

INTRODUCTION

Despite the enormous progress in the understanding of pathogenesis and molecular mechanisms of cancer, it remains a major health burden and the second leading cause of death worldwide. It accounts for ~7.6 million deaths (~13% of all deaths) per year worldwide and these figures are projected to continue to rise over 11 million in 2030 (1). Early detection and improved management of cancer patients are of paramount importance for the reduction of cancer mortality. Advancements in the early detection can be achieved by implementing systematic screening programmes for an asymptomatic population or improving diagnostic approaches for specific target groups that would allow to start the treatment before the disease has spread beyond the original site and complete, curative resection is possible in the most cases of solid tumours. Hence, the identification of cancer biomarkers that could be detected in body fluids such as plasma, serum or urine and are suitable for the development of non-invasive or minimally invasive tests applicable for the screening programmes, diagnosis, prognosis, monitoring or prediction of response to therapy would represent a significant step towards the reduction of morbidity and mortality caused by cancer. The discovery of prostate specific antigen (PSA) more than 30 years ago and its application in the screening, diagnosis and monitoring has changed the way how prostate cancer is diagnosed and treated (2). However, despite the extensive interest it triggered, few other serum biomarkers have been introduced to the clinic since that and there are many difficult-to-diagnose cancers, such as gastric cancer, for which no reliable biomarker assays are currently available.

Humoral immune response against cancer-derived proteins has been detected by the classical or modified SEREX (serological identification of tumour antigens by recombinant expression cloning) approaches in all cancer types analyzed so far (3,4). Autoantibodies, due to their specificity and stability in sera, seem to be very attractive targets for the development of non-invasive serological tests for the diagnosis of cancer. In contrary to the currently know serum biomarkers such as PSA, CEA or CA19-9, they are qualitative not quantitative biomarkers, which implies higher specificity. Furthermore, there is some evidence that the autoantibodies can be detected as early as several years before the clinical diagnosis (5) that demonstrates their value for the early detection. Moreover, even if they by themselves may have a minor role in the anti-tumour immune response, the autoantibody profile likely reflects the repertoire of activated CD4+ T cells, presumably, including both, the helper cells and the regulatory T cells. Hence, they may turn out to be valuable biomarkers for monitoring patient's response to immunotherapy. However, so far the development of autoantibody-based biomarker assays has been hampered by several factors: the frequency of antibodies against any individual TAA is generally relatively low, typically ranging from 1 to ~15%; autoantibody repertoire is heterogeneous and to some extent resemble the response to tissue damage by viral infections or autoimmune diseases, and autoantibodies against a number of TAAs, such as CTAG1B, TP53, c-MYC etc. are found in patients with different types of cancer (6-9). At least partially these limitations could be overcome by applying high-throughput proteomic techniques to cancer serology that allows definition of a comprehensive set of antigens in each type of cancer and analysing the whole autoantibody repertoire in patients' sera. In a previous study we applied T7 phage display-based SEREX technique to define the repertoire of antigens eliciting the autoantibody production in melanoma, breast, gastric and prostate cancer patients. This resulted in the identification of over 1300 different serum-reactive phage clones encoding known cancer-testis (CT) antigens such as CTAG1B/NY-ESO1, MAGEA, SSX2, DDX53, HORMAD1 etc., a number of non-CT

antigens that have been previously identified by applying conventional SEREX to various tumour types (ANXA11, LIG1, HDLBP, SC65, KTN1, TPM3, ZNF282, EEF1A1 etc), several ribosomal proteins and heat-shock proteins known to elicit humoral response both in cancer patients and patients with autoimmune diseases but the rest of the natural ORF antigens have not been previously implicated in autoimmune responses. However, the majority of the serum-reactive clones contained cDNAs fused to the phage coat protein 10B in a different reading frame, 5' or 3' UTRs, ribosomal RNA genes or mitochondrial DNA, thus expressing 4 to 80 aa long peptides that likely represent mimotopes. The nature of the antigens they represent is not known as they may mimic protein as well as non-protein antigens of cancer or normal cells or various pathogens.

The main objectives of the current study were:

- To develop technology for the production and processing of phage displayed antigen microarrays.
- To evaluate the performance of support vector machine (SVM) with improved kernel for the generation of biomarker models for the diagnosis of cancer.
- To develop procedures for the normalisation of microarray data, defining cutoffs that allow to discriminate between sero-positive and sero-negative cases, and ranking of antigens considering the signal intensity and the frequency of reactivity.
- To apply the phage displayed antigen microarray technology and the data analysis approach for the exploration of autoantibody profiles in gastric cancer and identification of autoantibody signature with the diagnostic relevance, and assessment of humoral responses in patients with various cancers against a family of sperm-associated antigens (SPAG).

1. LITERATURE OVERVIEW

1.1 . Nature of Cancer

Cancer or malignant neoplasm is the class of diseases defined by uncontrolled growth of cells and formation of metastasis (10). Cancer development is a complex multistage process, associated with the aggregation of genetic and epigenetic changes in the cells, leading to the transformation of normal cell into the cancer (11).

In 2000, Douglas Hanahan and Robert A. Weinberg have proposed six major features ("hallmarks") acquired by cancer cells: self-sufficiency in growth signals, insensitivity to antigrowth signals , evading apoptosis , limitless replicative potential, sustained angiogenesis and tissue invasion and metastasis (12).

Self-sufficiency in growth signals.

The ability to unlimited proliferation, sustained by proliferative signal, is the fundamental characteristic of cancer cells. Cells of multicellular organisms require extracellular signal to move to proliferative state. This prohibition can be bypassed in four ways. The necessary growth factors can be produced by cells themselves, the signal can be received from normal cells of cancer-associated stroma, which supplies the cancer cells by various growth factors or the signal can be enhanced by amplification of cell surface receptors as well as by acquisition of independence from growth factors (13,14). The structural change of receptor molecule or persistent activation of components of signaling pathways downstream from receptor in the circuit can be as example of such independence.

Insensitivity to antigrowth signals .

Cellular homeostasis is maintained in the tissue through the correct responses to inter-, extra-, and intracellular antigrowth signals. In normal cells, TGF-β, acting through signaling network, stops the cell cycle, induce differentiation, or trigger apoptosis. Cancer cells bypass this replicative barrier, and acquire the ability to divide indefinitely by mutations or downregulation of TGF-β receptors, inactivation of SMAD4 or $p15^{INK4B}$ (15). TGF-β due these changes frequently convert from a supressor of tumor formation to promother of metastasis in late stage tumors (15). Contact inhibition is a process of arresting cell growth when two or more cells come into contact with each other. Cancer cells typically lose this property due deregulation of LKB1 or Merlin pathways (16,17).

Evading apoptosis .

The cells can induce the senescence or apoptosis as the response on the excessively elevated signaling. Consequently, the relative intensity of proliferative signaling in tumor is a compromise between stimulation of growth and escape of these antiproliferative defense. Cancer cell may try to avoid limitation by changes in pathways of senescence or apoptosis (18,19). The inactivation of tumour suppressors is common in cancers and can be detected in majority of malignant tumors. Because activation of tumour suppressors, such as TP53, can halt a cellular cycle or trigger the program of apoptosis, it is necessary for a cancer cell to inactivate them. The TP53 regulates the transcription of many different genes in response to a wide variety of stress signals or abnormality sensors (20). The cell can also loose the function of TP53 itself or increases the expression of antiapoptotic regulators (Bcl-2) and survival signals (21). The proliferative signaling can be enhanced by the defects of negative-feedback loops, normally involved in homeostatic regulation (22).

Limitless replicative potential.

The telomerase, specialized DNA polymerase responsible for lengthening of the ends of

telomeric DNA, was detected in overwhelming majority of cancer cells the opposite to normal differentiated cells where it is not expressed (23). The telomeres play a key role in the ability of unlimited proliferation as suspected in different observations. So it is extremely important that the activity of telomerase is enabled by cancer cells. Perhaps the induction of telomerase is delayed or increases gradually, and due this process the cancer cell can get necessary genetic instability by deletion of chromosomal segments (24). The enhancement of cell proliferation, involvement in DNA repair and resistance to apoptosis are another additional features of telomerase (25,26).

Sustained angiogenesis.

The ability of tumour to induce the angiogenesis (process of formation of blood vessels) is another important factor for tumour growth up to macroscopic size. Known inducers of angiogenesis as VEGF-A and TSP-1 upregulated in the tumor tissues chronically (27,28).

Tissue invasion and metastasis.

The process of invasion and metastasis includes the following steps: the local invasion, the intervasation into the surrounding blood and lymphatic vessels, the escape of cancer cells into the parenchyma of distant tissues, the foundation of small nodules and finally growth of micrometastases into the macroscopic colonies (29). The cancer cells get the necessary properties using the program of epithelial-mesenchymal transition. Normally this program is used by fetal cells at the earliest stage of formation of body and differentiation of multiple tissues, as well as for the wound healing and repair of adult tissues (30). The existence of interaction between the distant metastases, primary tumor and cells of tumor stroma as well as its importance in metastasis has been shown in many observations (31). The majority of disseminated cancer cells are not entrenched in the new tissues due to changed conditions, which they have not been adapted to. One of opinion is that in case of new environment these cells need to create a new set of features for adaptation. As the result the development of other types of cancer and following reverse metastasis into primary tumour occurs (32,33).

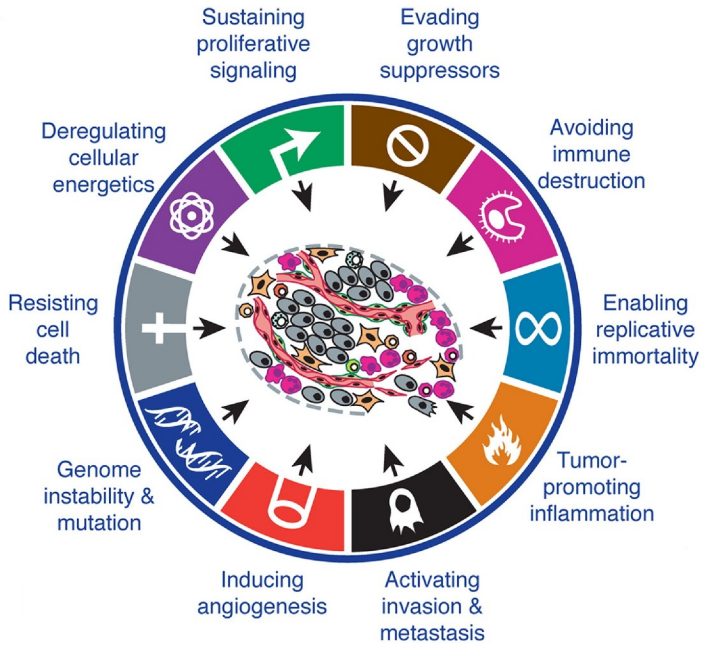

Figure 1. The hallmark features acquired by cancer cells. Adapted from Hanahan & Weinberg, 2011.

The modern theory extend the list of the features acquired by most of cancer cells by including two additional hallmarks: reprogrammed energy metabolism and avoiding immune destruction. Genome instability and tumour-promoting inflammation were established as enabling characteristic, that assist in obtaining of other hallmarks (34).

Reprogrammed energy metabolism.

In actively growing tissues, such as fetal tissue and tumours, metabolic regulation tends to differ from the normal adult tissues. They consume more glucose to produce lactate, even in the presence of ample oxygen. This phenomenon, known as the Warburg effect, enables rapidly dividing tumor cells to generate essential biosynthetic materials such as nucleic acids, amino acids and lipids from glycolytic intermediates (35).

Avoiding immune destruction.

Several lines of evidence suggest that at the beginning when genetic and epigenetic structural changes occurs in a single cell, the pre-neoplasm will be destroyed in case of significant immunogenicity and active immune system. The cancer with high amount of genetic changes contains new antigenic epitopes generated due to the expressed proteins. The immune system is able to recognize such tumour cells, classify them as foreign and activate a protective response, although the tumour cells are capable to develop characteristics that allow avoid the immune response (36). Tumors can escape from the immune response by process called immunoediting. This process consists of three phases: elimination, equilibrium and escape (37,38).

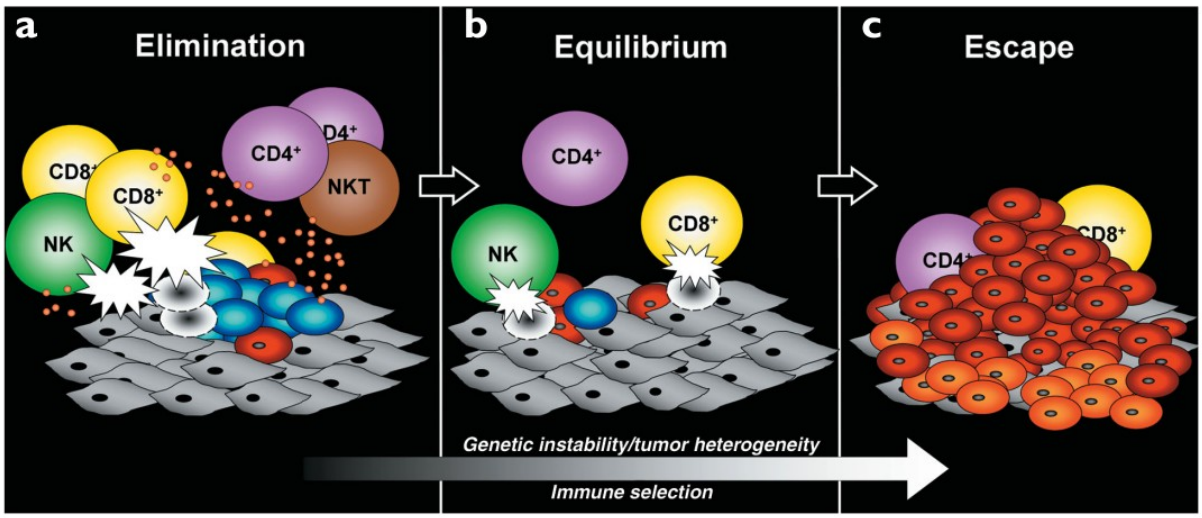

Figure 2. Cancer immunoediting include three process. (**a**) Elimination corresponds to immunesurveillance. (**b**) Equilibrium represents immunoselection of tumour cell variants with increasing capacities to survive immune attack. (**c**) Escape is the process wherein the immunologically sculpted tumour actively suppress the immune response and expands in the immunocompetent host. In panel **a** and **b**: developing tumour cells (blue), tumour cell variants (red), underlying stroma and nontransformed cells (grey), cytokines (small orange circles), white flashes represent cytotoxic activity of lymphocytes against tumour cells. In panel **c:** additional tumour variants (orange) that succeed to overcome immune pressure. Adapted from Dunn, Bruce, Ikeda, Old & Schreiber, 2002.

During the elimination phase, tumor-specific antigens trigger an immune response, which leads to the destruction of neoplastic cells, preventing development of cancer. The equilibrium phase of immunoediting describes a dynamic balance between the immune

system's containment of the tumor and acquisition of immune evasive strategies by subsets of cancer cells (39). Cancer cells evade immunesurveillance via two general ways: immunoselection and immunosubversion. Immunoselection is a selection of non-immunogenic tumour-cell variants. This process leads to down-regulation of MHC molecules or loss of expression of antigens recognized by the immune system and therefore reduced immunogenicity. Immunosubversion is a process by which tumor cells suppress the immune system through a complex network, leading to specific or generalized tolerance. The immune system contributes to tumor progression by selecting for the most aggressive clones or stimulating tumor cell proliferation (40). The tumor-induced tolerance is associated with expansion of non-functional T cells and suppression of other immune effector cells (41).

Genome instability.

The genes regulating cell growth and its differentiation must be altered within the transformation process of normal cell into the cancer cell. Genetic changes can be caused by carcinogenic factors – the radiation, the chemicals, infectious pathogens and specific endogenous reactions such as respiration and oxidative lipid peroxidation (10). The mutations can be accumulated by cells in case, when system of DNA repair fails (10). Epigenetic changes have been shown to play equally important role in the formation of malignancy (42,43). DNA methylation pattern in the cancer cell is altered leading to genome-wide alterations in the gene expression profiles as well as may contribute to the chromosomal instability (44). Epigenetic changes have been detected in the early stages of tumour formation and also in the normal tissues prior transformation to the cancer cells.

Tumour-promoting inflammation.

Paradoxically, the recent studies show that many immune system cells as macrophages, neutrophils and mast cells, as well as myeloid progenitors infiltrate tumour and help to maintain the angiogenesis as well as release the growth factors (27). The local tissue damage due to increased cell mass pressure on the surrounding tissues takes place at the time when the tumour is getting bigger and the tissue homeostasis is disturbed. As a result a variety of inflammatory factors, attracting immune cells have been produced. This leads to release of cytokines, chemokines and reactive oxygen compounds by macrophages and mast cells inducing the inflammation and infiltration of other immune cells, such as dendritic cell, into the tumor. The inflammatory process and formation of a stroma around the tumour can be induced by recruiting cells of innate immune system and normal cells of surrounding tissues. These cells, as recent studies show, are active participants in the process of formation of cancer, rather than passive bystanders (34). Due to inflammatory process these cells form the tumour associated stroma, providing the growth and survival factors, inducing the signals of activating of EMT and even acting as immunosuppressors (31,45). The modern theories of carcinogenesis suggest consider the cancerous tumour and surrounding supportive microenvironment joined (34).

It is well known that solid tumours are heterogeneous by histological structure and consist of a variety of tumour cells, the surrounding stroma, the vasculature, and inflammatory infiltrates, etc. Besides the populations of different types of cells present in the hematological malignancies as well as in the solid tumours, the subpopulation of cells with properties of stem, known as cancer stem cells (CSC), has been found. Subsequently, the cancer stem cell theory of carcinogenesis was created (46-48). This cell population is responsible for the following features of the tumour: the ability to self-regeneration, to metastasize, and recur of infiltrative growth, as well as resistance to chemotherapy and radiation treatment (49,50). The CSC can differentiate into a variety of cell types depending on the origin of tissues (51). Most of the studies show that CSCs is only a small minority of

the cells in a cancer (52), however in some cases of melanoma the CSC proportion can reach even 25% (53). The wide variation of proportion of CSC (from 20% to virtually all cells) has also been demonstrated in a colorectal cancer study (54). There is a hypothesis that these cells are formed by normal stem cells or progenitor cells as a result of transformation process influenced by mutations or epigenetic changes (55).

1.2 . Biomarkers

Biomarkers can be defined as a parameter that is objectively measured and evaluated as an indicator of normal biologic processes, pathogenic processes, or pharmacologic responses to a therapeutic intervention. The cancer biomarkers might be secreted by tumour itself or by the body as specific response to the cancer. For cancer diagnosis, prognosis and epidemiology the genetic, epigenetic, proteomic, glycomic and imaging biomarkers can be used.

The significant genetic, chromosomal and epigenetic changes as well as deregulation of the processes of splicing, translation and the activation of microRNA are ongoing in the cancer cells. This leads to appearance of protein atypical for normal tissues and wide oscillations of expression level of protein, found in normal tissues in small amount. These markers can be assessed in the biofluids collected non-invasively.

The tumour diagnostic biomarkers are used for general population screening, tumour volume estimation, differential diagnosis in symptomatic patients and clinical staging of cancer. Predictive biomarkers help to evaluate treatment response. Prognostic biomarkers indicate disease progression regardless of the type of treatment or lack of treatment (56). Unfortunately, there are not many reliable serum biomarkers currently in use at the clinic and tissue-based markers require an invasive procedure to obtain the samples for diagnostic purposes. However, majority of tumour biomarkers has low sensitivity, as its expression is increased in fewer than 50% of early-stage for most studied markers, as CA125 for ovarian cancer that has also been proposed as a possible screening test for this disease.(57,58).

One of the well – known examples of biomarkers widely used in diagnostic population screening is Prostate-Specific Antigen (PSA). It is a tumor marker discovered in 1980 and routinely tested in blood samples (59). The PSA is not prostate cancer specific and may be present in the normal prostate at the same levels as in the prostate cancer. Despite these limitations, the PSA screening has been implemented to the practice, although 17% of men with normal levels of PSA in blood have cancer cells within the prostate (60), and 78% of men, who has PSA level higher than 4 ng/ml, do not have prostate cancer (61). Many early stage patients with prostate cancer undergo radical treatment due to the lack of reliable biomarkers, which would allow to differentiate aggressive cancers from the relatively indolent and slowly progressing cancers (62).

Serum pepsinogen test based on the combination of the serum pepsinogen I level and pepsinogen I/II ratio and presence of *Helicobacter pylori* are the known biomarkers for diagnosis of gastric cancer. They both quantitative biomarkers and show relatively low specificity. Serum pepsinogen test show a sensitivity 77% and specificity 73% at pepsinogen I level < or = 70 ng/ml and pepsinogen I/II ratio < or = 3 (63). SPT has strong correlation with intestinal-type cancer (63).

A list of known serum gastric cancer biomarkers includes CEA, CA 19-9, CA50 and CA 72-4, all of which exhibit generally low sensitivity (<50%), especially for early stages (<20%) of GC (64-68). However, these biomarkers are shared between different types of cancer showing higher frequency in other, such as pancreatic, cancers (69). Not surprisingly, such biomarkers don't have a significant role in gastric cancer diagnosis and screening and are

more frequently used for the prognosis of recurrence after resection (64-68,70).
Several observations attribute to the CEA, CA19-9, CA50 and CA72-4 highest diagnostic accuracy 0.53, 0.64, 0.59 and 0.75, respectively (64-68,71-73).

1.3 . Autoantibodies

The tolerance against self-antigens is the critical function of immune system. The breakdown of self-tolerance leads to autoimmune disease involving appearance of autoantibodies, chronic inflammation and tissue destruction. The presence of autoantibodies is not exclusive for autoimmune disease. Most if not all patients' cancer sera contains the antibodies against tumor-derived proteins (3,74). The autoantibody repertoire is very heterogeneous and shared between cancer and autoimmune disease patients (6). The humoral immune response against self-antigens has the same features as the immune response referred to foreign antigens: high specificity to antigen, affinity maturation and immunoglobulin class switch from IgM to IgG (75).

If the immunosurveillance is a general phenomenon, the immune response is likely to occur before a clinically detectable cancer (37). In fact, several observations indicate that the antibodies to tumor-associated antigens occur long before the clinical symptoms (76-78). However, the mechanisms of production of autoantibodies and their biological significance are not completely understood. Many evidences show remarkable association of chronic inflammation and the development of cancer (6,9).

The only small part of autoantibodies react with antigens exclusively expressed in cancer cells, more frequently the autoantibodies recognize the self-antigens. The tumour specific antigens can be divided into the oncogenic viral proteins and altered proteins (neo-antigens), which appear in the cancer cell due to the genetic instability and deregulation of mRNA splicing (79). Tumor-associated antigens include overexpressed proteins that normally are expressed at very low levels.

The autoantibodies against many TAA, for example CTAG1B, TP53, c-MYC, cyclin B1 have been found in patients with different type of cancer (80). The cancer-testis (CT) antigens normally are expressed only in the immune-privileged tissues or during the fetal development. Autoantibodies against CT antigens have been associated with a poor prognosis. Antibodies to NY-ESO-1 (one of the CT antigens) has been shown to correlate with volume of tumour present in body (81), progression of disease in melanoma (82) and lower overall survival in prostate cancer (83).

The antigen p53 represents one of the best-studied tumor antigen. The TP53 tumor suppressor gene is a protein most frequently mutated in human cancers. Because mutated TP53 can induce an immune response and can occur early in the carcinogenic process for some tumors, p53 autoantibodies are the remarkable diagnostic biomarkers of cancer (84). Several studies indicate correlation between p53-specific autoantibodies and decreased overall and progression free survival in breast, lung, colon, and oral cancer (85,86), but in hepatocellular carcinoma, it has been suggested that the presence of p53-specific antibodies is associated with the increased overall survival, when compared with the antibody-negative patients (87).

Autoantibodies to CEA, CML66 and SOX1 show correlation with better outcomes in cancer patients (88-90).

1.4 . Microarray technology

Since its invention in 1995 (91), the microarray technology has become a principal tool for high-throughput detection of gene expression and analysis within scientific research and clinical studies. The microarray consist of an arrayed series of microscopic spots with a small amount of either DNA, protein, or carbohydrate target molecules immobilized on a solid medium (generally glass or plastic slides). Whichever the immobilized molecule in a probe is, all microarrays are based on the same principle as southern and dot blot is: the affinity of a sample to the probe allowing the evaluation of molecule/molecule interactions, but in comparison to a dot blot, microarray has a much larger scale. The possible pairs of interactions are DNA/DNA, DNA/protein, protein/protein, antigen/antibody and others. The samples such as serum specimen are labeled by the fluorescent tag molecule. The molecules immobilized in the microarray can then bind the target if present. The pattern of fluorescence is representative of the sample under test. This technology allows to test thousands of probes simultaneously. The commercial DNA microarrays can have up to 4.2 million features (Nimblegen 4.2M CGH arrays) allowing the evaluation of the whole genome in a single experiment.

The applications of protein microarrays include the expression profiling, serum-based diagnostics, protein / protein binding assays, and drug / target binding. Three different kinds of protein microarrays currently exist: the analytical protein microarrays, functional microarrays, and reverse-phase microarrays. The analytical protein microarrays are constructed by arraying of antibodies or non-folded antigens on a glass slide, which is then tested with a protein solution. This methodology is used to measure protein expression levels in a solution, similar to the sandwich ELISA technique. The functional protein microarrays consist of a full proteins or protein domains attached to a layer of a porous polymer (polyacrylamide, agarose, or gelatine) on solid medium. They are used for studying protein interactions with other proteins, macromolecules, or small molecules. The reverse phase protein microarrays (RPMA) correspond to the miniaturization of dot blot. In RPMA, proteins of interest such as recombinant proteins in purified form or in cell lysates or plasma are spotted onto a glass slide, and then the arrayed targets are tested by using antibodies against the protein of interest or sera. The complexity in comparison to DNA microarray is the main challenge of protein microarrays.

Although RPMA may be viewed by some experts as the microarray technology analogous to DNA microarrays, protein microarrays are technically in a distinct class due to the signal variability across the array (92). DNA microarrays are expected to have virtually identical mRNA levels for all samples, whereas protein microarrays can be markedly different total protein and target protein concentrations in different samples (92). The concentration of a protein is important for signal-to-noise ratio.

The technical variations occurred at manufacturing, processing and scanning microarray, such as probe labelling, incubation conditions, washing, signal and background detection, slide surface and batch effects, lead to a biased analysis of microarray data.

The normalization algorithms for microarray data include two different steps: within array and interarray normalizations. Within array step can include global and/or local types. The conventional methods for intra-slide normalisation are global, quantile and LOWESS normalizations.

All normalization algorithms assume that less than 20% of probes vary between arrays. Traditionally for verification and visualization of quality of normalized data use MA plots where on the X axis: $A= (\log_2 R + \log_2 G)/2$ and on the Y axis: $M= \log_2 R - \log_2 G$.

The global normalization refers to method where intensity of each spot is scaled by an

array specific constant so that the mean/median of all arrays is the same.

The quantile normalization suggests that the intensities of each array have the same distribution. For the quantile normalization the highest value on all arrays becomes the mean of the highest values, the second highest value becomes the mean of the second highest values, and so on.

LOWESS stands for Locally Weighted Linear Regression. It is based on the MA plot. It takes in all the data points, log transforms data and fits them to localize linear regression line. Then, the cut-off is applied, dismissing spots whose signal and control are both below the cut-off (93).

The conventional methods of the analysis of microarray data is a fold change, t-test, analysis of variance (ANOVA), clustering, regression analysis, maximum likelihood and classification. The most widely used classifiers are Artificial Neural Networks, Gaussian Mixture Models, Naive Bayes, Decision Trees, RBF classifiers, K-Nearest Neighbours and Support Vector Machines.

The Support Vector Machines (SVM) dominates as the most popular technique for multicategory classification of microarray data (94), significantly out performing other classifiers. The Support Vector Machines is the “state-of-the-art” machine leaning approach that uses Statistical Learning Theory for successful implication of the best possible separation of two classes in the binary classification task. This approach could be easily extended for the multicategory classification purposes as well. The basic idea of SVMs is inferred as follows: find two parallel hyperplanes in the hypothesis space with the largest possible margin such that the majority of data points of the class “+1” is located in the first half-space (separated by the first hyperplane) and the majority of other data points (denoted by class “-1”) in the second half-space (NB: we should note that in between of these half-spaces exists a margin-area that is maximized by the SVM optimization problem). The extensions of SVM method suppose hard-margin, soft-margin cases as well as a nonlinear separation using so-called “kernel trick” that transfers initial hypothesis space to possibly infinite one with perfectly available separation. This trick enables SVM with direct embedding of similarity measures encoded into kernels via “kernel trick”. As close is this measure to the original classification problem and original distributions of similarity among samples as precise and accurate will be final generalization ability of SVM. For more detailed information about theory of SVMs or applications for microarray data analysis refer to (94,95).

2. MATERIALS AND METHODS

2.1 . Serum samples

Serum samples and information about the clinical status of patients with melanoma, gastric, lung, breast and prostate cancer as well as healthy individuals and patients with gastric inflammatory disease were received from Genome Database of Latvian population. Another set of sera of patients with gastric cancer were collected at Latvian Oncology Center and Clinic of Gastroenterology, Hepatology and Infectious Diseases, Otto-von-Guericke University Magdeburg, Germany.

The serum collection includes 39 breast, 24 lung, 339 gastric, 52 prostate cancer, 190 melanoma, 313 healthy controls, 150 patients with gastric inflammatory disease.

All samples were stored at -80ºC.

The tissue and serum specimens were collected after the patients' informed consent was obtained in accordance with the regulations of Committee of Medical Ethics of Latvia and the ethical committee of the Otto-von-Guericke University Magdeburg.

2.2 . Printing protocol

E.coli BLT5615 cells were grown from fresh culture until OD600 reached 0.8, the expression of the phage coat protein 10B was induced by IPTG for 30 minutes, aliquoted in 96 deep-well plates (Whatman, 500mkl per well) and infected with 5mkl of low-titre monoclonal phage stocks, grown at 37 ºC, 220 rpm until complete cell lysis. Then 6.4% glycerol was added and plates were centrifuged to remove cell debris and subjected to PCR-based quality control. The lysates were clarified by centrifugation and arrayed in 2 replicates onto nitrocellulose-coated FAST slides (Whatman) with a QArray Mini microarrayer (Genetix) using a protocol adapted for printing protein microarrays:

maintained air humidity 45-55%,

DNA microarray protocol with the following wash cycle of needles:

4x 1.5s 0.05% Tween

4x 1.5s double distilled water.

2.3 . Processing of antigen microarrays

Dry slides for 30 min at +37ºC, store at +4ºC;

Day 1:

- Make serum dilutions 1:200 in TBS, 0.5% Marvel:
  Prepare 5% Marvel in TBS, 0.05% Tween (0.5g/10ml)
  Serum-dilution buffer (7 ml): 5.7 ml TBS
  0.7 ml 5% Marvel
  0.3 ml *E. coli*-T7 phage lysate
  0.3 ml *E. coli*-T7-C3-Strep phage lysate
  Aliquot 200 µl in 2.2 ml deep-well plates +1 µl serum
  Incubate on shaker (200 rpm) at +4ºC.

Day 2:

- Blocking: 7% Marvel in TBS, 0.05% Tween (0.7g/10ml TBS, 0.05% Tween).

Put slides in incubation plates, pour over blocking buffer (10ml for 10x10cm plates), incubate 1h, RT, 38 rpm on shaker. Pour off the blocking solution and briefly wash in TBS-0.5% Tween.

- Put the slides in incubation chambers & Fast frame.
- Serum:
  - Cf the serum plate at 3000 rpm for 10 min
  - Add 80 µl of the diluted & preabsorbed serum per well, incubate for 2h, RT, 38 rpm on shaker
  - Pipette off the serum, rinse the chambers 3 × with 0.1 ml TBS-0.5% Tween using 8-channel pipette, take out the slides from the incubation chambers, put into the High Throughput washing chamber and wash 4×15 min in TBS-0.5% Tween on magnetic stirrer.
- T7 tail antibody:
  - Prepare T7 tail antibody dilution 1:10 000 in TBS, 0.5% Marvel
  - prepare 63 ml Ab dilution buffer: 58.5 ml TBS + 4.5 ml 7% Marvel
  - 3 µl T7 tail antibody + 30 ml Ab dilution buffer
- Put the slides back in the incubation plates
- Put 10ml Ab for 8x8cm plate, incubate for 45 min, RT, 38 rpm on shaker
- Pour off the Ab, wash 4 ×10 min in chambers.
- Secondary antibodies:
  - Cy5 conjugated Goat Anti-Human IgG (minimal cross-reaction with bovine, horse, mouse serum proteins) (Jackson ImmunoResearch, #109-175-098) 1:1500 diluted in Ab dilution buffer
  - Cy3 conjugated Goat Anti-Mouse IgG (minimal cross-reaction to human serum proteins) (Jackson ImmunoResearch, # 115-165-071) 1:3000 diluted in Ab dilution buffer
  - 30 ml Ab dilution buffer
  - 40 µl Cy5 conjugated Goat Anti-Human IgG 1:2 dil. glycerol stocks
  - 20 µl Cy3 conjugated Goat Anti-Mouse IgG 1:2 dil. glycerol stocks
- Put the slides back in the incubation plates
- Put 10ml Ab for 8x8cm plate and 4.5ml for black tray, incubate for 45 min, RT, 38 rpm on shaker
- Pour off the Secondary Ab, put the slides into the High Throughput washing chamber and wash 4×15 min in TBS-0.5% Tween on magnetic stirrer, rinse one in dist. H_2O and dry by centrifugation for 1 min at 1000 rpm. Wash the chambers with detergent, rinse in dist. H_2O and dry.

2.4 . Data processing and analysis

Slides scanned on Tecan Power Scanner with 532 and 635 nm lasers at 10µm resolution and the images were saved as TIFF files.

The spot data were extracted by GenePix Pro software using the proprietary spot recognition algorithm. R-language software (96) with additional packages: limma(97), marray, ROCR, OLIN(98), survival, verification, REvolution R were used for following data processing and computational analysis.

Median foreground and background intensities, as well as median, mean and standard deviation of Cy5/Cy3 ratios for pixels within a spot were obtained for each spot and imported into the script developed in-house. The dataset was filtered to remove unqualitative spots sets within print batch that for more then half of set had high morphological heterogeneity (for Cy5/Cy3 ratios of pixels within a spot: ratio standard deviation to mean bigger then 1.5 or difference mean and median more then 50%) and any elements that had been manually flagged as poor quality.

The signals Cy3 and Cy5 were separately normalized within a each slide by data centering on the basis of middle 80% of intensities (median of this set on each channel will be zero and the standard deviation will be 1).

Interslide normalization was performed for each antigen spots series within printing batch by data centering on the basis of middle 80% of intensities (median of this set on each channel will be zero and the standard deviation will be 1).

Cy5/Cy3 ratios were calculated and averaged between replicates.

The threshold value (T) for each antigen was calculated as follows:

$$T = mean\ (I_{HD}) + 3 \times SD(I_{HD})$$

, where I_{HD} is the signal intensities in healthy controls.

Then the rank for each antigen rank was calculated, on the base of intensities of positive signals within gastric cancer patients compared to healthy donors, using the following formula:

$$R_i = \left(\Sigma I_{GC_i} / N_{GC_i}\right) - 2\left(\Sigma I_{HD_i} / N_{HD_i}\right).$$

Finally, a score for each serum was calculated as follows: .

$$S = \sum_{i=1}^{n} \sqrt{R_i} \times I_i$$

Support Vector Machine

All measurements and experiments were performed using MATLAB framework (99) and some highly robust and optimized linear algebra packages: LAPACK (100), Metis, Ipopt. As the reference implementation of SVM algorithm another highly recognized calculus package was used: LibSVM (101). As the cornerstone of our implementation was SimpleMKL framework (102). We used it extensively for identification of optimal hyperparameters of RBF kernel and our proposed Ranking-Based kernel.

3. RESULTS

The current dissertation is based on the following original publications referred in the text by their Roman numerals. The author's contribution to the enclosed original publications:

Original paper I

Kalniņa Z, Siliņa K, Meistere I, Zayakin P, Rivosh A, Ābols A, Leja M, Minenkova O, Schadendorf D and Linē A. Evaluation of T7 and Lambda phage display systems for survey of autoantibody profiles in cancer patients. *J Immunol Methods*, 2008 May 20;334(1-2):37-50. Epub 2008 Feb 21

Contribution: development of the methodology for the production of phage-displayed antigen microarrays, development of the data normalization method, scanning, data acquisition, normalization and statistical analysis.

Original paper II

V. Jumutc* , P. Zayakin* , and A. Borisov. Ranking-based Kernels in Applied Biomedical Diagnostics using Support Vector Machine. *International Journal of Neural Systems,* (accepted).

* The first two authors contributed equally to this work

Contribution: development of normalization approach and idea of ranking-based kernel, preparation of data, the graphical information and partially manuscript.

Original paper III

K. Silina, P. Zayakin, Z. Kalnina, L. Ivanova, I. Meistere, E. Endzelins, A. Abols, A. Stengrevics, M. Leja, K. Ducena, V. Kozirovskis, A. Linē. Sperm associated antigens as targets for cancer immunotherapy: expression pattern and humoral immune response in cancer patients. *J Immunother*, 2011, Jan;34(1):28-44.

Contribution: production of phage-displayed antigen microarrays, development of approach for the data normalisation and determination of cutoffs, data acquisition, normalization and statistical analysis.

Original paper IV

Zayakin P, Kalniņa Z, Siliņa K, Meistere I, Ivanova L, Endzeliņš E, Jumutcs V, Stengrēvics A, Leja M, Wex T and Linē A. Tumour-associated autoantibody signatures for the early detection of gastric cancer (submitted to *Cancer Prevention Research*)

Contribution: development of normalization approach and ranking based procedure, production of phage-displayed antigen microarrays, data acquisition, normalization and statistical analysis, preparation of data and the graphical information.

3.1 . Evaluation of T7 and Lambda phage display systems for survey of autoantibody profiles in cancer patients

ELSEVIER

Journal of Immunological Methods 334 (2008) 37–50

JIM Journal of Immunological Methods

www.elsevier.com/locate/jim

Research paper

Evaluation of T7 and lambda phage display systems for survey of autoantibody profiles in cancer patients

Zane Kalniņa [a], Karīna Siliņa [a], Irēna Meistere [a], Pawel Zayakin [a], Alexander Rivosh [a], Artūrs Ābols [a], Mārcis Leja [b], Olga Minenkova [c], Dirk Schadendorf [d], Aija Linē [a,*]

[a] *Biomedical Research and Study Centre of Latvia, Riga, Latvia*
[b] *Faculty of Medicine, University of Latvia, Riga, Latvia*
[c] *Kenton Labs, c/o Sigma-Tau, Pomezia, Italy*
[d] *German Cancer Research Center, Skin Cancer Unit, Heidelberg, Germany*

Received 3 January 2008; received in revised form 28 January 2008; accepted 29 January 2008
Available online 21 February 2008

Abstract

In the current study we attempted to evaluate the suitability of T7 Select 10-3b and λKM8 phage display systems for the identification of antigens eliciting B cell responses in cancer patients and the production of phage-displayed antigen microarrays that could be exploited for the monitoring of autoantibody profiles. Members of 15 tumour-associated antigen (TAA) families were cloned into both phage display vectors and the TAA mini-libraries were immunoscreened with 22 melanoma patients' sera resulting in the detection of reactivity against members of 5 antigen families in both systems, yet with variable sensitivity. T7 phage display system showed greater sensitivity for the detection of antibodies against members of CTAG, MAGEA and GAGE families, both systems showed equal performance in detecting the reactivity against MAGEC and SSX2 while only λKM8 allowed the detection of anti-CTAGE5 antibodies. The biological properties of both phages turned out to be equally suitable for the production of antigen microarrays however in line with the plaque assay the sensitivity for the detection of various autoantibodies differed between the vectors. However, presumably due to the higher variability of the background signals in the microarray assay, it turned out to have comparable, in some cases even slightly lower sensitivity than the plaque assay.

Next, we explored the repertoire of antigens that could be identified by screening T7 phage-displayed testis cDNA library with sera from melanoma patients. From the 243 antigens identified, only 24 represented known genes translated in their natural reading frame and included known TAAs like Annexin XI-A and a novel potential CT antigen SPAG8. Another 12 were uncharacterised genes but the remaining clones contained DNA fragments in non-natural reading frames that most likely represent mimotopes, nevertheless, they may turn out to be valid biomarkers.

Keywords: Autoantibodies; Phage display; Antigen microarray; SEREX; Melanoma antigens

1. Introduction

Circulating autoantibodies against tumour-derived proteins have been observed in the most if not all cancer patients hence they may serve as biomarkers for the screening, diagnosis, prognosis or monitoring of cancer. In fact, autoantibodies against a number of CT antigens, including

Abbreviations: SEREX, serological identification of antigens by recombinant expression cloning; TAA, tumour-associated antigens; pfu, plaque forming units; CT, cancer-testis; UTR, untranslated region, ORF, open reading frame.

* Corresponding author. Biomedical Research and Study Centre of Latvia, Ratsupites Str 1, LV-1067, Riga, Latvia. Tel.: +371 7808208, fax: +371 7442407.
E-mail address: aija@biomed.lu.lv (A. Linē).

0022-1759/$ - see front matter © 2008 Elsevier B.V. All rights reserved.
doi:10.1016/j.jim.2008.01.022

NY-ESO-1, SSX2, MAGEA1 and 3 (Stockert et al., 1998) and several members of CTAGE family (Usener et al., 2003), and melanocyte differentiation antigens such as RAB38 (Zippelius et al., 2007) have been detected exclusively in the sera from cancer patients, thus suggesting they may serve as highly specific diagnostic markers. Autoantibodies against another set of antigens such as GLEA2, PHF3 and endostatin have been shown to correlate with the clinical outcome in glioblastoma and breast cancer patients, respectively (Pallasch et al., 2005; Bachelot et al., 2006). Moreover, appearance of anti-p53 autoantibodies may predict subsequent development of cancer (Li et al., 2005), while their prognostic significance still remains a subject of debate (Soussi, 2000). However, so far the clinical utility of tumour-associated autoantibodies has been hampered by the low frequencies to each particular antigen and the heterogeneity of antibody repertoires in cancer patients. This could be overcome by developing high-throughput techniques that would allow to identify a comprehensive set of immunogenic proteins in a given type of cancer and subsequently analyse the occurrence of antibodies against them in large sets of sera from patients with cancer, autoimmune disorders and healthy donors.

SEREX (serological identification of tumour antigens by recombinant expression cloning) is a widely used technique for the identification of immunogenic proteins in cancer patients that is based on the construction of cDNA expression libraries from tumour tissues, expression of the recombinant proteins in *E. coli* and immunoscreening of the libraries with sera from cancer patients (Sahin et al., 1995). The application of this technique to a variety of tumour entities has led to the identification of more than 2000 genes encoding potential tumour antigens, most of which are deposited in the Cancer Immunome database (http://www2.licr.org/CancerImmunomeDB). However, this approach is extremely time-consuming and labour-intensive and therefore is generally not applicable for the profiling autoantibody responses in multiple serum samples. Recently, several laboratories have exploited various phage display strategies, including M13 filamentous phage (Sioud and Hansen, 2001; Somers et al., 2002), pJuFo system (Fossa et al., 2004), T7 phage (Hansen et al., 2001; Zhong et al., 2004) and lambda phage (Mincnkova et al., 2003; Pavoni et al., 2004) for the construction of cDNA expression libraries and successfully applied them for the isolation of cDNAs encoding tumour-associated antigens. A feature that makes the phage display-based strategies particularly attractive for the searching of antigens is that tumour-derived cDNAs are expressed as fusion proteins with one of the phage coat proteins (or linked via a Jun–Fos interaction in case of pJuFo system) and exposed on the surface of the phage thus allowing the selection of serum-reactive phage clones by biopanning, which is much faster as well as more cost and labour-effective approach than the conventional immunoscreening. Moreover, the recombinant phage particles can directly be printed onto glass slides to produce antigen microarrays that allow monitoring the antibody responses against hundreds to thousands of antigens simultaneously using a microlitre of serum (Fernandez-Madrid et al., 1999; Cekaite et al., 2004; Zhong et al., 2005; Wang et al., 2005; Chatterjee et al., 2006). However, the main drawback of the phage display-based systems for the expression of cDNA libraries is the bias in the repertoire of proteins that can be displayed on the surface of the phage resulting from the biological peculiarities of a given phage species. For example, the assembly of M13 phage takes place in the periplasm of *E. coli*, therefore only those fusion proteins that can be exported through the bacterial inner membrane will be displayed on the phage capsid (Hufton et al., 1999). Capsids of lytic phages such as T7 and lambda phages are assembled in the cytoplasm therefore the repertoires of cDNA libraries do not have the biological constraints similar to M13 phage (Krumpe et al., 2006). The capability of a protein to be displayed on the surface of a lytic phage depends mostly on its impact on the assembly process, where the length, biochemical properties, hydrophilicity and hydrophobicity as well as folding characteristics play the major role. Moreover, the copy number of recombinant proteins per phage particle may vary in these systems allowing the assembly of chimerical capsids, thus affecting the sensitivity of the assay.

In the current study we assessed the capability of T7 Select 10-3b and λKM8 phages to display a range of clinically relevant tumour antigens by cloning the members of 15 TAA families into these display vectors, compared the sensitivity of the systems for the detection of autoantibodies by immunoscreening of the TAA mini-libraries with sera from 22 melanoma patients, and evaluated the suitability of these phages for the production of antigen microarrays. Moreover, in order to characterise the repertoire of melanoma associated antigens that can be identified using T7 Select 10-3b system, we constructed T7 phage-displayed testis cDNA library and screened it with sera from 9 melanoma patients resulting in the identification of 243 different sero-reactive phage clones.

2. Materials and methods

2.1. Serum samples

Serum samples from melanoma patients (stage II–IV) were collected at the Skin Cancer Unit, German Cancer Research Center and Latvian Oncology Center after the

patients' informed consent was obtained in accordance with the regulations of local ethics committees. The samples were aliquoted and stored at −70 °C.

2.2. Construction of phage-displayed cDNA expression libraries

2.2.1. T7 and lambda phage-displayed tumour-associated antigen (TAA) mini-libraries

Either the entire ORFs or the antigenic regions predicted using an algorithm developed by Welling et al. (1985) of 15 tumour antigens or antigen families were amplified by PCR using High Fidelity PCR enzyme mix (Fermentas) and cDNA from testis or melanoma tissue as a template. The antigens and the regions chosen for cloning are listed in Table 1, and the sequences of PCR primers are available on request. All reverse primers contained *NotI* site and forward primers contained either *SpeI* or *SalI* sites for cloning into λKM8 or T7 Select 10-3b vectors, respectively. For the amplification of genes from multigene families degenerated primers were designed to amplify as many as possible members of the same gene family. If more than one region of a gene was amplified, PCR products were combined prior to ligation in the vector. PCR products were digested with *NotI* and *SpeI* or *SalI* (Fermentas) and purified using GFX PCR DNA and Gel Band Purification kit (Amersham Biosciences). λKM8 (Pavoni et al., 2004) vector DNA was digested with *SpeI* and *NotI* but T7 Select 10-3b vector (Novagen) was digested with *SalI* and *NotI*, treated with shrimp alkaline phosphatase (Fermentas) followed by phenol/chloroform extraction and ethanol precipitation. Each digested PCR product (5 ng) was ligated into λKM8 or T7 Select 10-3b vector (100 ng), and 1/5 of each ligation mixture was subjected to *in vitro* packaging using 5 μl of Gigapack III Gold Packaging extract (Stratagene) or T7 Select Packaging extract (Novagen), respectively, resulting in mini-libraries of $1–5\times10^5$ pfu for lambda phage and $0.5–5\times10^4$ pfu for T7 phage. The obtained TAA mini-libraries were amplified once using *E. coli* BB4 cells or IPTG-induced BLT5615 cells, respectively. The lysates were centrifuged to remove cell debris and stored at +4 °C or as glycerol stocks at −80 °C.

2.2.2. T7 phage-displayed testis cDNA expression library

Testis cDNA library was constructed using T7 Select 10-3b vector and OrientExpress cDNA library construction system (Novagen) according to the manufacturer's instructions. Briefly, mRNA was isolated from 150 μg of testis total RNA (Ambion) using Dynabeads mRNA purification kit (Invitrogen) and converted to cDNA using HindIII Random primers (5′-TTNNNNNN-3′). Then cDNA was ligated to directional EcoRI and HindIII linkers, digested with the corresponding restriction enzymes and ligated into pre-digested T7 Select 10-3b vector followed by *in vitro* packaging resulting in a

Table 1
Antigens comprising TAA mini-libraries

Antigen family	Number of mRNAs[a]	Regions cloned (nt positions/NCBI RefSeq No)	Different genes found by sequencing 5 random clones	
			T7 Select 10-3B	λKM8
CTAG1B (NY-ESO)	2	86–400; 338–628/NM_001327	CTAG1B, CTAG2	CTAG1B, CTAG2
MAGEA	10	209–618; 590–1137/NM_005362	MAGEA1, 2, 3, 4, 9	MAGEA1, 3, 4, 6
SSX	10	58–401; 58–617/NM_005636	SSX2	SSX1, 2
BAGE	1	195–330/NM_001187	BAGE	BAGE
GAGE	8	84–410/NM_001472	GAGE2, 7, 8	GAGE1, 8
MAGEB	4	98–516/NM_002364	MAGEB2,4	MAGEB2
MAGEC	2	931–1453/NM_016249	MAGEC1, 2	MAGEC1, 2
SPANX	6	57–347/NM_022661	SPANXA2, SPANXE	SPANXD
LDHC	1	245–785/NM_017448	LDHC	LDHC
CT45	4	246–798/NM_001017417	CT45-1	CT45-1
THEG	1	778–1176/NM_016585	THEG	THEG
CTAGE	6	130–1070; 311–1070; 1052–1616; 1589–2354/203354; 11–235/NM_022663	CTAGE1, 3, 5, 5Δex7	CTAGE1, 5
MTA1	1	1394–1772/NM_004689	MTA1	MTA1
TYR	1	494–710/NM_000372	TYR	TYR
MLANA	1	54–408/NM_005511	MLANA	MLANA

[a] Number of known members of the antigen family that theoretically could be amplified with the selected primers.

library of 8×10^6 pfu. The library was amplified once in IPTG-induced BLT5615 cells.

2.3. Selection of serum-reactive phage clones

2.3.1. Immunoscreening of phage-displayed TAA mini-libraries

For screening of T7 phage-displayed TAA mini-libraries, ~10^3 pfu from each library was spotted on gridded LB/carbenicillin agar plates pre-coated with LB top agarose containing BLT5615 cells grown in LB supplemented with 1×M9 salts, 0.4% glucose, 1 mM $MgSO_4$ and carbenicillin (50 μg/ml) to OD_{600}=0.5 and induced with IPTG for 30 min. After ~2 h incubation at 37 °C when the plaques reached ~1 mm in diameter, plates were overlaid with Protan nitrocellulose (NC) filters (Schleicher & Schuell) and incubated for 1 h at 37 °C. For screening of lambda phage-displayed TAA mini-libraries, BB4 cells grown in LB supplemented with 0.2% maltose and 10 mM $MgSO_4$ were plated on NZY agar plates, infected with ~10^3 pfu from each mini-library and the plates were incubated at 37 °C until visible plaques appeared (6–8 h). The filters were blocked with 5% (w/vol) milk powder in TBS, 0.05% Tween 20 for 1 h, and then incubated overnight with 1/200 diluted patients' serum preabsorbed with *E.coli*-phage lysate. The serum-reactive clones were detected by incubating the filters with alkaline phosphatase conjugated anti-human IgG, Fcγ specific secondary antibody (Pierce) and NBT/BCIP (Fermentas). Ten serum-reactive clones from each mini-library were subjected to secondary screenings and subcloned to monoclonality.

2.3.2. Biopanning of T7 phage-displayed testis cDNA library

Approximately 5×10^{10} pfu from T7 phage-displayed testis cDNA library were incubated overnight with 2 μl of patient's serum that had been preabsorbed with BLT5615 and T7 phage lysate coupled with BrCN-activated sepharose. A hundred microlitre of Protein G coated magnetic beads (Pierce) were washed twice with blocking solution (5% milk powder in TBS, 0.05% Tween 20), added to the phage–serum mixture and incubated for 2 h at RT under agitation. The beads were washed 10 times with 1 ml TBS, 0.05% Tween and the bound phages were either amplified and subjected to the second round of biopanning or titrated and used for immunoscreening as described above.

2.3.3. Identification of cDNAs

The inserts of serum-reactive phages were amplified by 35-cycle PCR using primers flanking the insert (T7-Up2: 5′-CTTCGCCCAGAAGCTGCA-3′, T7-Down: 5′-AACCCCTCAAGACCCGTTTA-3′, λKM8-Up: 5′-CAATCTGTGTGGGCACTCG-3′, λKM8-Down: 5′-CGGCTGGTAATGGGTAAAGG-3′) and 1 μl of phage solution as a template. PCR products were purified and directly sequenced using ABI Prism BigDye Terminator v3.1 cycle sequencing kit and 3130 Genetic Analyser (Applied Biosystems). DNA sequences were analysed using BLAST tool at www.ncbi.nlm.nih.gov, Translate tool at www.expasy.org and compared against sequences available at Cancer Immunome Database (www2.licr.org/CancerImmunomeDB).

2.4. Production of GST fusion proteins and Western blot analysis

To confirm the recognition of the phage-displayed proteins, GST fusion proteins were produced and used for standard Western blot analysis. cDNA inserts of serum-reactive clones were amplified by PCR and cloned into a prokaryotic GST expression vector pGEX-4T-1 (Amersham Biosciences) to produce either the natural products of these genes or the out-of-frame peptides displayed by the serum-reactive clones.

Purified recombinant proteins (500 ng) were resuspended in Laemmli sample buffer, denatured for 5 min at 100 °C and separated by SDS/polyacrilamide gel electrophoresis (PAGE) on two gels simultaneously and transferred onto Protan NC membranes (Schleicher & Schuell). The filters were blocked for 1 h in 5% milk powder in TBS, 0.05% Tween 20 and incubated with 1/20 000 diluted HRP conjugated goat anti-GST antibody (Amersham Biosciences) or 1/200 diluted patient's sera followed by incubation with HRP conjugated goat anti-human IgG antibody 1/10 000 (Sigma) and detected using ECL Plus Western Blotting Detection Reagents (Amersham Biosciences).

To determine the copy number of fusion proteins per T7 phage particle, phages were amplified in *E. coli* BLT5615 cells, the phage particles were precipitated from the lysates with 3.5% PEG 8000, 0.4 M NaCl, titrated and ~10^9 pfu of each phage were used for Western blot analysis with 1/5000 diluted rabbit HRP conjugated anti-T7 tag antibody (ABcam) recognising the N-terminus of T7 coat protein 10B. To ascertain that unincorporated recombinant proteins are not co-precipitated with the phage particles, NaCl/PEG precipitated phage solution containing ~10^{10} pfu was filtrated through 100 000 NMWL filter units (Millipore) and the flow-through fraction was run in parallel with the phage samples as a negative control. For quantity calibration a standard curve was constructed from 5 serial 3-fold dilutions of one phage clone. The images were analysed using GelWorks software (Ultra-Violet Products).

2.5. Production of antigen microarrays

Nine corresponding T7 and lambda phage clones isolated by immunoscreening from TAA mini-libraries and four non-recombinant phage clones were grown in BLT5615 and BB4 cells, respectively, until complete lysis. The lysates were clarified by centrifugation and arrayed in 5 replicates onto nitrocellulose-coated 16-pad FAST slides (Whatman) using a QArray Mini microarrayer (Genetix) to generate T7 and lambda phage-displayed antigen microarrays. The microarrays were blocked in 5% (w/vol) milk powder in TBS, 0.05% Tween 20 for 1 h, and then incubated with 1/200 diluted patients' sera (preabsorbed with *E.coli*-phage lysates) for 2 h at room temperature. The slides were rinsed with TBS and washed 4 times in TBS, 0.5% Tween 20 for 15 min each and then incubated with anti-T7 tail fiber (Novagen) or anti-gpV (λ tail protein) monoclonal antibody (kindly provided by Dr. Mauruzio Cianfriglia) at a dilution of 1/10 000 or 1/1500, respectively to determine the amount of phages in each spot. After 3 washes in TBS, 0.5% Tween 20 for 10 min, the microarrays were incubated with Cy5 labelled goat anti-human IgG antibody (1/1500) and Cy3 labelled goat anti-mouse IgG antibody (1/3000) (Jackson ImmunoResearch) for 1 h, then washed thrice in TBS, 0.5% Tween 20, rinsed with distilled water and dried by centrifugation. The microarrays were read using AQuire scanner (Genetix) and the images were analysed using Genetix QScan software. For each spot the mean Cy5 and Cy3 signals were background subtracted, averaged between replicates, and the Cy5/Cy3 ratios were calculated for each antigen and normalised by that of non-recombinant phages. A cut-off value for defining serum-reactive antigens was set as >3 SDs above the mean ratio for all the spots.

3. Results

3.1. Construction of lambda and T7 phage-displayed TAA mini-libraries

Members of 12 CT antigen families and 3 other tumour antigens were cloned into lambda KM8 and T7 Select 10-3b phage display vectors to produce TAA mini-libraries. In λKM8 vector cDNAs are fused to the N-terminus of the coat protein gpD and are separated by a flexible GS linker while in T7 phage they are fused to the C-terminus of the coat protein 10B. The antigen families and individual genes as well as the protein regions chosen for cloning are listed in Table 1. Complete ORFs were amplified for shorter transcripts, while for genes with long ORFs the potential antigenic regions were predicted using an algorithm developed by Welling et al. (1985) and chosen for cloning. In the case of tyrosinase the immunodominant region that has been shown to react with sera from melanoma patients

Table 2
Antigens identified by immunoscreening lambda and T7 phage-displayed TAA mini-libraries and individual sera recognising the respective antigens

TAA mini-library	Antigen	Reactive sera[a]	
		T7 Select 10-3B	λKM8
CTAG	CTAG1B	MA001643, MA000703, MA000161, MA000445, MA00SK, MA00AM, MA00550, MA000513	MA001643, MA000703, MA000161, MA000445
	CTAG2	MA001643, MA000703, MA000161, MA000445, MA00SK, MA00AM, MA00550	MA001643, MA000703, MA000445
	CTAG1B-ORF2	UKRV-Mel31	–
MAGEA	MAGEA1	MA000513	MA000513
	MAGEA2	MA000513	MA000513
	MAGEA3	MA000513	MA000513
	MAGEA12	MA000513	–
	MAGEA2 antisense	MA00WF	–
MAGEC	MAGEC1	MA000703	MA000703
	MAGEC2	MA000703	–
SSX	SSX2	MA000951	MA000951
CTAGE	CTAGE5		MA000273, MA000525
	CTAGE pseudogene	MA000525, MA000273, MA001816, MA001111	–
	CTAGE1 antisense	MA001404	–
GAGE	GAGE3-7 subgroup	MA000445, MA001111, MA001404	MA000445
	GAGE1, 8	MA000445	MA000445
	GAGE2-ORF2	MA00GG	–
	GAGE7-ORF2	MA001643	–

[a] Sera are ranked by the signal intensities in immunoscreening.

but not with sera from healthy donors (Lucchese et al., 2005) was chosen. For the amplification of genes from multigene families, degenerated PCR primers capable of amplifying as many members of these families as possible were designed. The cloning of the respective PCR products into λKM8 and T7 Select 10-3b vectors resulted in TAA mini-libraries in the size of $1–5\times10^5$ pfu and $0.5–5\times10^4$ pfu, respectively. The complexity of the mini-libraries and the insertion of cDNAs in the correct reading frame relative to the phage coat proteins were determined by sequencing 5 random clones from each mini-library (Table 1).

3.2. Comparison of serum reactivity against TAA displayed on lambda and T7 phages

Approximately 2×10^3 pfu from each of the 15 lambdas and T7 phage-displayed TAA mini-libraries were immunoscreened with sera from 22 melanoma patients. Ten sero-reactive phage clones from each sero-reactive TAA mini-library were purified and identified by sequencing their cDNA inserts. Generally, relatively good concordance in the recognition of antigens displayed on lambda and T7 phages was observed — the reactivity against antigens representing CTAG (NY-ESO-1), MAGEA, MAGEC, SSX and GAGE families was detected in both expression systems (yet with different sensitivity) whereas no reactivity against BAGE, MAGEB, SPANX, LDHC, CT45, THEG, MTA1, TYR and MLANA was observed in any of them (Table 2).

In both systems, phage capsids are composed of wild-type capsid proteins and hybrid proteins, however the display density and its regulation are different. T7 Select 10-3b system has been shown to display 5–15 copies of recombinant protein per phage but the rest of the wild-type 10A capsid protein (415 in total) is provided by a plasmid in a complementing host (BLT5615) (Rosenberg et al., 1996). In the λKM8 system, wild-type copies of protein D are provided by a copy of the respective gene in the genome of the phage and the capsid was shown to be composed of ~50% recombinant proteins at least in the case of scFv antibody display (Vaccaro et al., 2006). Since it has been suggested that the display density on T7 phage is variable and depends on the size of the hybrid coat protein (Zucconi et al., 2001), we tried to determine the copy number of fusion proteins per phage particle in order to verify that the lack of reactivity against the mentioned antigens is not due to the failure to display the respective antigens. Individual T7 phage clones encoding 11 different antigens were amplified, the phage particles were precipitated from the bacterial lysates with PEG/NaCl and subjected to Western blot analysis using antibody against the N-terminus of T7 coat protein 10B that should recognise both, the fusion protein and the non-recombinant coat protein. Five 3-fold dilutions of the clone encoding MLANA were used to construct a standard curve and the quantity of each band was calculated using GelWorks software. The fusion proteins of the corresponding length were detected in all phage clones analysed, however the copy number per phage varied markedly being the highest in phages expressing N-terminus of CTAG1B (~17%) and lowest in phages expressing N-terminus of CTAGE5 (~1%) (Fig. 1).

Concerning the sensitivity, T7 Select phage display system showed a greater sensitivity in detecting the reactivity against members of CTAG, MAGEA and GAGE

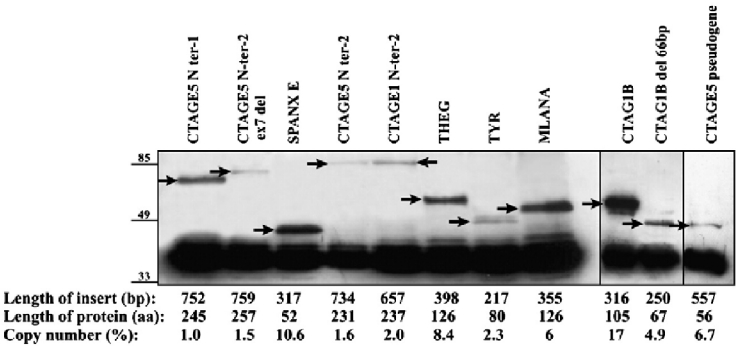

Fig. 1. Copy number of TAA fusion proteins per T7 phage particle determined by Western blotting using antibody against N-terminus of the coat protein 10A/10B. The lower band represents *wt* protein 10A, arrows indicate the fusion proteins of the respective sizes. Quantification of *wt*/fusion protein ratio was done by construction of a standard curve from a five 3-fold dilutions of one phage on separate gels (not shown) and the intensities of bands were calculated using GelWorks software.

antigen families both in terms of signal intensities and the number of positive sera. Both systems showed equal performance in detection of reactivity against MAGEC and SSX2 while only λKM8 allowed the detection of anti-CTAGE5 antibodies.

Eight sera reacted with CTAG1B (7 with CTAG2) when these antigens were displayed on T7 phage but only 4 of them reacted with CTAG1B (3 with CTAG2) when expressed by λKM8. Since the signal intensities for those four sera that did not recognise λ phage-displayed antigens were low, this inconsistency most likely resulted from more efficient display of this particular antigen on T7 phage (in fact, display density of CTAG1B was the highest among all the phages analysed), and not from the inability of lambda phage to display CTAG antigens.

An opposite situation was observed in the case of CTAGE5, where 2 sera recognised its N-terminal fragment when it was displayed on lambda phage but did not react with the corresponding T7 phage. At the same time, 4 sera reacted with T7 phage clones expressing 56 aa polypeptide derived from CTAGE pseudogene containing a stop codon. Western blot analysis showed that T7 phages display ~30 copies of the polypeptide encoded by the pseudogene but only 4 copies of CTAGE5 hence suggesting that very low display density did not allow to detect the presence of serum antibodies against CTAGE5. Alternatively, there could be conformational differences of this antigen when displayed on T7 resulting in poor accessibility of the epitope.

Moreover, 5 sera reacted with T7 phage clones expressing CTAG1B, GAGE2 and GAGE7 ORF 2 peptides and MAGEA2 and CTAGE1 antisense peptides. All these cDNAs contain stop codons when translated as fusion proteins with the coat protein 10B. Since cDNAs are expressed as N-terminal fusion proteins to gpD in λKM8 system, such peptides per definition cannot be displayed on lambda phage. UKRV-Mel-31 serum reacted with 3 different CTAG1B-encoding phage clones containing frame-shifting deletions (most likely resulting from PCR errors) resulting in the expression of CTAG1B-ORF2 peptide, which is well known CD4+ and CD8+ T cell antigen whose immunogenicity is supposed to be associated with the translation of CTAG1B alternative ORF in cancer cells (Slager et al., 2003). Whether or not MAGEA2, CTAGE1, GAGE2 and 7 are also translated in alternative ORFs in cancers and therefore could be considered as tumour specific antigens remains to be determined. The variable display density and the detection of reactivity against the clones expressing out-of-frame peptides could result in false negative and positive calls, respectively, that provides a disadvantage of T7 phage display system in studies aiming to determine the presence of antibodies against definite antigens, therefore λKM8 would be the expression system of choice for these kinds of studies. At the same time, the out-of-frame peptides might turn out to be novel tumour specific antigens or valid biomarkers.

3.3. Suitability of T7 and lambda phages for the production of antigen microarrays

In order to assess the suitability of both phage display systems for the production of antigen microarrays, nine

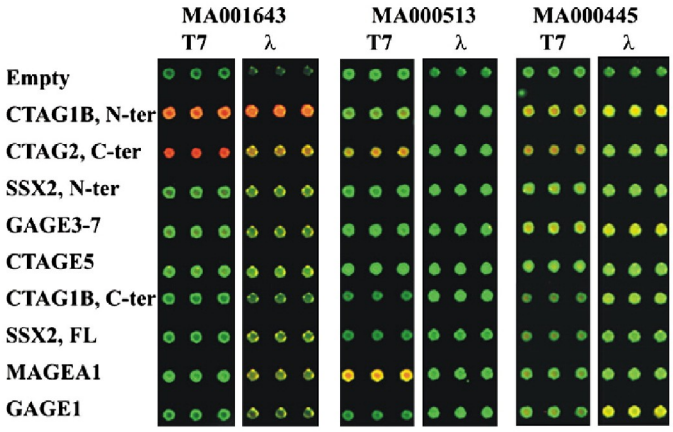

Fig. 2. T7 and lambda phage-displayed antigen microarrays. Nine recombinant T7 and lambda phage clones expressing corresponding TAAs and four non-recombinant phages were amplified, spotted on FAST slides in quintuplicate and tested with sera from melanoma patients (detected with Cy5 labelled secondary antibody) and with monoclonal antibody against tail protein of T7 or lambda phage (detected with Cy3 labelled secondary antibody) to quantify the amount of phages in each spot. Only partial array images are shown.

Table 3
Detection of autoantibodies against T7 and lambda phage-displayed TAAs using microarrays

	MA001643		MA000513		MA000445		MA000951		MA000161		MA000273		MA000525	
	T7	λ	T7	λ	T7	λ	T7	λ	T7	λ	T7	λ	T7	λ
CTAG1B, N-ter	**37.9**	**17.5**	**28.8**	**3.9**	**8.2**	**5.1**	**2.1**	1.4	**39.3**	**2.8**	0.9	1.5	1.9	0.9
CTAG2, C-ter	**58.8**	3.2	**64.4**	1.5	**9.8**	2.1	1.3	1.0	**43.2**	1.6	1.4	1.6	1.4	0.9
SSX2, N-ter	2.3	1.2	2.9	1.0	2.5	1.7	**3.5**	**1.9**	3.5	1.1	1.1	1.6	1.6	0.9
GAGE3-7	2.2	1.2	3.0	1.0	**5.2**	**4.8**	1.5	1.4	3.4	1.4	1.2	1.6	1.4	1.0
CTAGE5	2.0	1.2	2.5	2.1	1.7	1.8	1.2	1.1	3.3	1.4	0.9	**2.5**	1.3	1.2
CTAG1B, C-ter	4.6	2.7	3.0	0.9	2.3	2.8	1.6	1.2	1.1	1.6	0.9	1.6	1.5	0.9
SSX2, FL	2.3	1.0	2.8	0.6	2.4	1.2	**4.3**	**2.1**	1.8	0.8	1.0	1.4	1.2	0.9
MAGEA1	2.3	1.2	**119.7**	**2.7**	1.7	1.7	1.6	1.4	1.6	1.3	1.2	1.7	1.4	1.0
GAGE1	2.2	1.3	2.8	1.3	3.0	**5.7**	1.6	1.2	2.9	1.7	0.8	1.5	1.5	1.1

Cy5/Cy3 ratios for each antigen were normalised by that of non-recombinant phages and the cut-off value for serum-positive clones was set as >3 SDs above the average of all the spots (marked with bold).

T7 and lambda phage clones encoding corresponding TAAs and 4 non-recombinant phages were amplified and spotted on FAST slides, and the microarrays were tested with 7 serum samples used for the immunoscreening of the TAA mini-libraries (Fig. 2). The variation between replicates was less than 10% for both display systems. Similarly to plaque immunoscreening, T7 Select system showed a higher sensitivity for the detection of autoantibodies against CTAG1B and MAGEA1, both systems detected anti-SSX2 and GAGE antibodies with comparable signal-to-noise ratios but anti-CTAGE5 antibodies were detected only in λKM8 system (yet only in one out of two CTAGE-positive serum samples) (Table 3). In most of the cases, the antigens that were found to be serum-positive by the immunoscreening of TAA mini-libraries were defined as positive using the selected cut-off value (>3 SDs above the average of all spots in the array) in microarray screening. However, although T7 phage clones expressing CTAG1B C-terminus and GAGE1 were identified by plaque immunoscreening using MA001643 and MA000445 sera, respectively, the signal intensities for these clones using the respective sera did not reach the defined cut-off value. Similarly, lambda phage clones expressing CTAG1B and CTAG2 C-termini and CTAGE5 also were called sero-negative against the sera that were used to isolate these clones from TAA mini-libraries. Nonetheless, lowering the cut-off value most likely would result in false-positive calls due to the variability in the background signal intensities of serum non-reactive clones. At the same time MA000951 serum was defined as CTAG1B positive by microarray screening while no reactivity against CTAG1B was detected by plaque assay. Hence the microarray screening and plaque immunoscreening have comparable yet not identical sensitivity for the detection of autoantibodies.

3.4. Selection of sero-reactive clones from T7 phage-displayed testis cDNA library

As the ultimate goal of our study is to identify a comprehensive set of antigens that could be further used for profiling autoantibody repertoires in patients' sera, we next decided to explore the repertoire of antigens that could be identified using T7 Select phage display system. Since the germ cell transcriptome shares many characteristics with cancers, we used testis as a source of RNA for the construction of T7 phage-displayed cDNA expression library and screened it with sera from 9 melanoma patients. The selection of serum-reactive clones was based on the biopanning using protein G coated magnetic beads followed by the immunoscreening of the enriched library. Commonly, 4 to 5 rounds of biopanning are performed in order to achieve sufficient enrichment with the phage of interest (Somers et al., 2002; Willats, 2002). We reasoned that in an experiment where polyclonal antibodies with different titres and affinities are used to select potential antigens from a library of phages expressing fusion proteins of different sizes and biochemical properties, performing multiple rounds of biopanning may result in the enrichment of more viable phages recognised by high-titre antibodies and the under-representation of phages whose infectivity or viability is affected by the fusion protein they express. Therefore, initially we assessed the repertoire of antigens selected after the first and the second round of biopanning using two different serum samples. Approximately 5×10^{10} pfu were used for biopanning with MA002079 and LGP-Mel 150 sera that resulted in the recovery of $\sim5\times10^{5}$ pfu in both cases. Approx. 8×10^{3} pfu of the enriched libraries were subjected to immunoscreening that resulted in the detection of ~80 and 60 serum-reactive clones with MA002079 and LGP-Mel 150 sera,

respectively. The remaining phages were amplified and subjected to the second round of biopanning followed by the immunoscreening with the respective serum. Although a higher enrichment with serum-reactive clones (>130 positive clones per 8×10^3 pfu) was achieved, a considerable reduction in the diversity of the antigen repertoire was observed. The representation of different antigens among clones detected in biopan 1 and 2 fell from 73 to 10% and 92 to 50% when screened with MA002079 and LGP-Mel 150 sera, respectively. Therefore we chose to use a single round of biopanning throughout the study.

Next, serum-reactive clones were selected from T7 phage-displayed testis cDNA library using two more individual sera and a pool of 5 melanoma patients' sera. Approximately 10^4 pfu from the enriched libraries were immunoscreened and that resulted in the detection of 40–150 reactive clones per serum. In total 436 serum-reactive phage clones were isolated, purified via several rounds of immunoscreening and their cDNA inserts were amplified by PCR and sequenced.

3.5. Sequence analysis and characterisation of the identified antigens

The sequence analysis of the identified serum-reactive phage clones revealed that they represent 243 different antigens. However, only in 24 cases cDNAs of known genes were fused in-frame to the phage coat protein 10B thus ensuring that the natural products of these genes are exposed on the surface of the phage (Table 4). Six of these antigens have been previously detected by conventional SEREX, another 10 represent protein families whose other members have been detected by SEREX, and 3 are autoantigens known to

Table 4
In-frame antigens identified by screening T7 phage-displayed testis cDNA library with melanoma patients' sera

Gene symbol	Protein	Serum	Position (aa) on Ref Seq	SEREX DB ID [a]
CCDC84	Coiled-coil domain containing 84	MA002079	186–280, NP_940891.1	–
HDLBP	High density lipoprotein binding protein (vigilin)	MA002079	1161–1268, NP_005327.1	309, gastric Ca
KIF27	Kinesin family member 27	MA002079	715–940 and 715–1128, [b] NP_060046.1	–
MASK-BP3	MASK-4E-BP3 alternate reading frame gene	MA002079	1482–1540, NP_065741.3	1082, renal cell Ca
LIG1	DNA ligase I	Pool-5	1–141, NP_000225	–
LMOD1	Leiomodin 1	Pool-5, MA00GG, LGP-Mel 150 [c]	463–507, NP_036266.2	2374, fibrosarcoma
COPS4	COP9 signalosome subunit 4	Pool-5	15–136, NP_057213.2	–
R3HDM2	R3H domain-containing protein 2	Pool-5	285–513, NP_055740	1233, breast Ca
ANXA11	Annexin A11	Pool-5	1–71, NP_665876.1	81, lung Ca
DKFZP566E164	Hypothetical protein LOC25858	Pool-5	3–76, NP_001034585	–
SENP1	Sentrin/SUMO-specific protease 1	LGP-Mel 143	355–618, NP_055369.1	–
KIF1B	Kinesin family member 1B isoform b	LGP-Mel 143	316–458, NP_055889.2	–
RPS2	Ribosomal protein S2	LGP-Mel 143	263–278, NP_002943.2	182, CTCL, prostate Ca
ATP2C1/ANKRD24	Calcium-transporting ATPase 2C2/ankyrin repeat domain 24	LGP-Mel 143, LGP-Mel 150 [c]	775–785, NP_055676.2/614–650, NP_597732.1	–
SPAG8	Sperm associated antigen 8	LGP-Mel 143, LGP-Mel 150 [c]	9–99, NP_758516.1	–
AKAP12	A-kinase anchor protein 12	LGP-Mel 143, LGP-Mel 150	1465–1622 or 1382–1529, [b] NP_005091.2	–
CCDC92	Coiled-coil domain-containing protein 92 (Limkain beta-2)	LGP-Mel 150	75–200, NP_079416.1	–
ATP5G2	ATP synthase, H+ transporting, mitochondrial F0 complex	LGP-Mel 150	56–107, NP_005167.2	–
LRRC50	Leucine rich repeat containing 50	LGP-Mel 150	626-725, NP_848547.3	–
SLU7	Step II splicing factor	LGP-Mel 150	110–228, NP_006416.3	–
Similar to: LRRC37A	Similar to: leucine rich repeat containing 37A	MA00GG	803–870, NP_055649.3	–
TEF	Thyrotrophic embryonic factor (novel splice variant)	MA00GG	1–52 and 162–215, NP_003207.1	–

[a] Identification number of an antigen in the Cancer Immunome Database (www2.licr.org//CancerImmunomeDB).
[b] Two partially overlapping clones encoding the same antigen were isolated;
[c] The same phage clone was isolated with multiple sera.

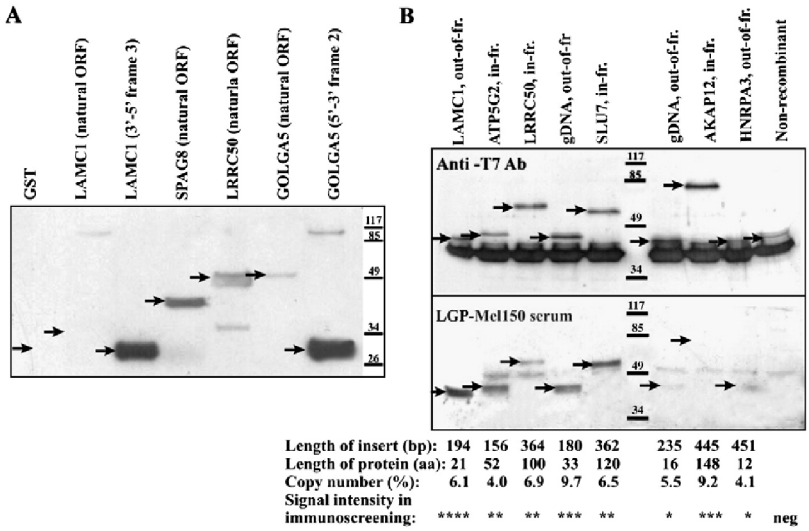

Fig. 3. Western blot analysis of serum reactivity against selected antigens. (A) In-frame and out-of-frame peptides encoded by the selected serum-reactive phages were expressed as GST C-terminal fusion proteins and tested for the reactivity with LGP-Mel 150 serum and anti-GST antibody to confirm the expression of the fusion protein and to determine its size (not shown). The arrows indicate the location of GST fusion proteins. (B) Lysates of ~10^9 pfu of the selected phage particles were separated by SDS/PAGE and tested with anti-10B antibody to determine the copy number of recombinant proteins per phage particle (quantification was done as described in Fig. 1) with patient's serum used for the immunoselection.

induce autoantibody production in autoimmune disorders (LMOD1, AKAP12) or infertility (SPAG8), while no immune responses against 4 antigens — SENP1, TEF, SLU7 and DKFZP566E164 have been reported before. One of the clones contained a hybrid cDNA generated by in-frame fusion of *ATP2C1* and *ANKRD24* genes; however no evidence for the expression of this fusion mRNAs in the testis tissue was found, suggesting that most likely the clone arose as a cloning artefact. Of the remaining 219 clones, 12 represent novel splice variants of known genes or uncharacterised genes for which the natural ORFs have not been determined, 143 represent 5′ or 3′ UTRs or cDNAs fused to 10B in a different reading frame, 52 clones represent intergenic DNA with no corresponding ESTs, 3 clones encode for ribosomal RNA genes and 9 clones contain mitochondrial DNA.

The antibody reactivity against two in-frame antigens (SPAG8 and LRRC50) and two out-of-frame peptides (encoded by *GOLGA5* and *LAMC1*) was confirmed by Western blot analysis. The proteins encoded by the corresponding serum-reactive clones were expressed as GST fusion proteins (for *GOLGA5* and *LAMC1* both — the natural products of these genes and 5′-3′ frame 2 peptide of *GOLGA5* and 3′-5′ frame 3 peptide of *LAMC1* were produced), purified and used for Western blot analysis with anti-GST antibody (not shown) and LGP-Mel 150 serum (Fig. 3A). This confirmed the reactivity against *SPAG8* and *LRRC50* encoded antigens and *GOLGA5* and *LAMC1* out-of-frame peptides.

We further hypothesised that the relatively high proportion of out-of-frame peptides among the serum-reactive clones identified by screening T7 phage-displayed library could be associated with the higher display density of these peptides than natural ORFs due to their shorter size (the size of the out-of-frame peptides ranged from 2 to 56 aa, average 21 aa, while in-frame proteins ranged from 10 to 414 aa, average 120 aa). To test this, a panel of 8 serum-reactive clones (4 in-frame, 4 out-of-frame antigens) and an empty T7 phage was assembled and subjected to Western blot analysis with anti-T7 10B antibody and LGP-Mel 150 serum (Fig. 3B) as described above. The copy number of fusion proteins per phage particle ranged from 17 to 42 (4–10%). No correlation between the copy number and (1) the size of the fusion protein, (2) the size of the insert and (3) signal intensity in Western blot and plaque immunoscreening with patient's serum was observed. This shows that the detection of serum reactivity against out-of-frame peptides is an intrinsic feature of T7 Select 10-3b phage display system that is not associated with the variations in the copy number of recombinant proteins per phage. The serum reactivity was detected against all fusion proteins, except for the AKAP12, hence suggesting that the anti-AKAP12

antibodies recognise a conformational epitope that is absent on denatured protein, while all the other antigens carry linear epitopes.

4. Discussion

In the current study we attempted to evaluate the suitability of T7 Select 10-3b and λKM8 phage display systems for monitoring autoantibody responses against a range of clinically relevant tumour antigens. This showed that both systems are capable of displaying members of all antigen families that were cloned and a relatively good concordance in the detection of autoantibodies by plaque immunoscreening was observed — reactivity against 5 antigen families was detected using both display systems while no reactivity against the remaining 9 antigen families was detected. As the display of the non-reactive antigens at least on the T7 phages was confirmed by Western blot analysis, these are likely to represent true negative calls. However, the sensitivity of the detection of autoantibodies against various antigens seems to differ between the display systems — T7 Select system exhibited a higher sensitivity (in terms of signal intensities and the number of reactive sera) for the detection of autoantibodies against the members of CTAG, MAGEA and GAGE antigen families. In line with the immunoscreening results, T7 phage display system turned out to be more sensitive for the detection of antibodies against CTAG1B/CTAG2 and MAGEA antigens in microarray screening also. This was an unexpected finding, since lambda phage has been shown to be able to assemble capsids consisting of up to ~50% N-terminal fusion protein D (Vaccaro et al., 2006), while T7 phage capsids can incorporate only 1.2–3.6% recombinant proteins according to the manufacturer's description (Novagen) or 1–17.0% according to our results. This suggests that the N-terminus of the recombinant protein D on lambda phage may be spatially less accessible for antibodies than C-terminus of 10B on T7 phage. In fact, the analysis of the crystal structure of gpD demonstrated that the N-termini up to Ser 15 are disordered and are located near the three-fold axis of gpD trimer on the side that binds to the capsid surface and hence at least partially may be hidden under the gpD trimer (Yang et al., 2000).

At the same time, the reactivity against CTAGE5 could be detected only when it was displayed on lambda phage but not on T7 phage. Since the copy number of fusion proteins on CTAGE5 encoding T7 phages was the lowest among all the phages analysed (~1%), most likely, it did not reach the detection limit thus preventing the serum reactivity it to be detected. Hence, the very variable display density of T7 Select system confers a risk of false negative calls due to an insufficient copy number and furthermore suggests that the signal intensity in plaque assay or microarrays may depend not only on the antibody titre but also on the copy number of recombinant proteins. However, this disadvantage could be overcome by constructing a novel vector that would allow monitoring the copy number of fusion proteins per phage particle.

Next, we applied T7 Select 10-3b phage display-based SEREX approach to search for antigens eliciting immune responses in melanoma patients that resulted in the identification of 436 serum-reactive clones representing 243 different antigens. However, only 24 of them represented known genes translated in their natural reading frames and another 12 were novel splice variants or uncharacterised genes (with at least two ESTs confirming that these sequences are transcribed) with unknown protein sequences. Six of the in-frame antigens have been previously detected by conventional SEREX and another 10 represent protein families whose other members have been detected by SEREX but no immune responses to 4 antigens have been reported before thus demonstrating that the repertoire of antigens identified by T7 phage display-based SEREX approach overlaps with conventional SEREX, at the same time may allow the detection of novel antigens.

Some of them have been previously shown to have a diagnostic value or may play a significant role in cancer progression. For example, the presence of autoantibodies against Annexin XI-A could significantly discriminate between breast cancer and non-cancer control sera and their frequency was higher in patients with ductal carcinoma *in situ* than in invasive ductal carcinoma (Fernandez-Madrid et al., 1999). Another of the identified antigens, sperm-associated antigen 8 (SPAG8), so far was implicated in a rare form of female infertility, where anti-SPAG8 antibodies have been shown to cause sperm agglutination (Zhang et al., 2000). This protein was shown to be predominantly expressed on the acrosome of sperm and functionally involved in the acrosome reaction and sperm binding to the zona pellucida (Cheng et al., 2007). Very recently it was found to be overexpressed in HPV18 infected cervical cancer cells (Vazquez-Ortiz et al., 2007). At the same time, a closely related protein, SPAG1, was shown to be expressed at high levels in a large proportion of pancreatic ductal adenocarcinomas and contribute to cancer cell motility. Hence, SPAG may represent a novel CT antigen family, however more detailed analysis of their expression in normal and cancerous tissues is required. Moreover, considering its localisation on cell surface at least on spermatozoa, it seems to be an attractive target for antibody-based therapeutical approaches.

The remaining 219 clones contained fragments of intergenic DNA (52), mtDNA (9), rRNA (3), 5′or 3′ UTRs or cDNAs (143) cloned out-of-frame relatively to the coat

protein 10B. It cannot be excluded that some of the 143 clones encoding UTRs or cDNAs in alternative ORFs indeed represent cancer-specific antigens generated by frame-shifting mutations, defects in pre-mRNA splicing or aberrations in translational controls in cancer cells, as evidenced by the detection of serum-reactive clones expressing CTAG1B-ORF2 peptide by screening TAA mini-libraries. Nevertheless, most likely the majority of these clones, particularly those 52 clones containing intergenic regions with no evidence of expression, display peptides that are not naturally expressed and therefore can be considered as mimotopes. A similar proportion of in-frame and out-of-frame antigens has been found by Chatterjee et al. (2006) in a study where serum-reactive clones were selected from T7 phage-displayed ovarian cancer cDNA library. Out-of-frame peptides, yet in smaller number, also have been detected by using pJuFo phage display system (Fossa et al., 2004) and lambda phage surface display (Minenkova et al., 2003). The nature of the antigens they mimic is not known and we believe — cannot be unambiguously determined by BLAST search through protein databases because the exact epitope sequences are not known and, moreover, they may mimic protein as well as non-protein antigens of various pathogens not only cancer cells. Consequently, the probability of finding cancer-associated biomarkers among them should be lower than among the in-frame antigens.

Experience of our and other groups (personal communication Dr. G. Li; Nottingham and Dr. S. Eichmuller, Heidelberg) has shown that approximately 1/3 of the serum-reactive clones identified by conventional SEREX also contain cDNAs fused to β-galactosidase gene in a non-natural reading frame. Although it is possible that the natural products of these genes could be produced by means of using alternative ribosomal binding sites, to our knowledge it has never been experimentally confirmed. Therefore we assume that a considerable fraction of these antigens also represents mimotopes, though their incidence is much lower than in T7 phage display-based SEREX approach. We reasoned, the identification of so high percentage of mimotopes using T7 phage could be caused by a higher display density of the out-of-frame peptides and subsequent detection of low affinity antibodies due to a higher valence of the epitope carrier. However, the analysis of the display density on the selected phage particles did not confirm this hypothesis, hence either the overrepresentation of the phages expressing shorter peptides in the amplified libraries or the folding properties and accessibility of the C-terminus of the coat protein 10B are likely to be responsible for this phenomenon.

The application of phage display technology to cancer serology offers time-, labour- and cost-effective alternative to conventional SEREX. Furthermore, it gives an opportunity to produce antigen chips by printing the recombinant phage particles on microarray slides thus allowing avoidance of production and purification of recombinant proteins that would not be feasible for all the antigens identified. This in turn enables the analysis of the whole autoantibody profile in patients' sera that would allow establishing the significance of the autoantibody profiles, not individual autoantibodies, as biomarkers for the early detection and prognosis of cancer and prediction of response to immunotherapy. So far, T7 phage, presumably due to its favourable biological properties (fast growing, chemically resistant and easy to obtain high-titre stocks) and availability of good commercial antibodies against the phage tail protein, has been the vector of choice for the production of antigen microarrays (Fernandez-Madrid et al., 1999; Zhong et al., 2005; Wang et al., 2005; Chatterjee et al., 2006). However, the variable copy number and the display of very high percentage of mimotopes constitute the two main drawbacks for exploiting T7 Select system for the analysis of autoantibody profiles. Here we demonstrated that the lambda phage is equally suitable for the production of phage-displayed antigen microarrays — it turned out to be possible to obtain high-titre phage stocks without any concentration or purification steps, the phage capsid appeared to be sufficiently stable and in the most cases the displayed proteins retained their capacity to be recognised by antibodies, however it was less sensitive than T7 phage for the detection of antibodies against several antigens. Nevertheless, as the lambda phage is capable to assemble capsids with N-terminal gpD fusion proteins that efficiently diminish display of out-of-frame peptides, it could be a preferential display system for the studies aiming to define novel potential therapeutic targets or to assess the presence of autoantibodies against known tumour antigens.

In conclusion, the exploitation of phage display-based approaches for the identification of tumour antigens provides a time- and labour-effective alternative to the conventional SEREX allowing the identification of a diverse antigen repertoire that partially overlaps with SEREX. Moreover, both T7 Select and λKM8 display systems are equally suitable for the production of antigen microarrays allowing the monitoring of autoantibody profiles, however they differ in the sensitivity of the detection of antibodies against various antigens.

Acknowledgments

We are thankful to Dr. Mauruzio Cianfriglia for providing anti-gpV monoclonal antibody and to Dr.

Aivars Stengrēvics for providing a part of the collection of sera from melanoma patients. We appreciate the help of Antje Sucker and Dr. Stefan Eichmüller for selecting appropriate serum samples.

This study was supported by EU 6th Framework Program ENACT (LSHC-CT-2004-503306) and a grant from Latvian State Research Program No. 07-VP-2, and fellowships from ESF.

References

Bachelot, T., Ratel, D., Menetrier-Caux, C., Wion, D., Blay, J.Y., Berger, F., 2006. Autoantibodies to endostatin in patients with breast cancer: correlation to endostatin levels and clinical outcome. Br. J. Cancer 94, 1066.

Cekaite, L., Haug, O., Myklebost, O., Aldrin, M., Ostenstad, B., Holden, M., Frigessi, A., Hovig, E., Sioud, M., 2004. Analysis of the humoral immune response to immunoselected phage-displayed peptides by a microarray-based method. Proteomics. 4, 2572.

Chatterjee, M., Mohapatra, S., Ionan, A., Bawa, G., li-Fehmi, R., Wang, X., Nowak, J., Ye, B., Nahhas, F.A., Lu, K., Witkin, S.S., Fishman, D., Munkarah, A., Morris, R., Levin, N.K., Shirley, N.N., Tromp, G., Abrams, J., Draghici, S., Tainsky, M.A., 2006. Diagnostic markers of ovarian cancer by high-throughput antigen cloning and detection on arrays. Cancer Res. 66, 1181.

Cheng, G.Y., Shi, J.L., Wang, M., Hu, Y.Q., Liu, C.M., Wang, Y.F., Xu, C., 2007. Inhibition of mouse acrosome reaction and sperm-zona pellucida binding by anti-human sperm membrane protein 1 antibody. Asian J. Androl. 9, 23.

Fernandez-Madrid, F., VandeVord, P.J., Yang, X., Karvonen, R.L., Simpson, P.M., Kraut, M.J., Granda, J.L., Tomkiel, J.E., 1999. Antinuclear antibodies as potential markers of lung cancer. Clin. Cancer Res. 5, 1393.

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

Hansen, M.H., Ostenstad, B., Sioud, M., 2001. Identification of immunogenic antigens using a phage-displayed cDNA library from an invasive ductal breast carcinoma tumour. Int. J. Oncol. 19, 1303.

Hufton, S.E., Moerkerk, P.T., Meulemans, E.V., de Bruine, A., Arends, J.W., Hoogenboom, H.R., 1999. Phage display of cDNA repertoires: the pVI display system and its applications for the selection of immunogenic ligands. J. Immunol. Methods 231, 39.

Krumpe, L.R., Atkinson, A.J., Smythers, G.W., Kandel, A., Schumacher, K.M., McMahon, J.B., Makowski, L., Mori, T., 2006. T7 lytic phage-displayed peptide libraries exhibit less sequence bias than M13 filamentous phage-displayed peptide libraries. Proteomics 6, 4210.

Li, Y., Karjalainen, A., Koskinen, H., Hemminki, K., Vainio, H., Shnaidman, M., Ying, Z., Pukkala, E., Brandt-Rauf, P.W., 2005. p53 autoantibodies predict subsequent development of cancer. Int. J. Cancer 114, 157.

Lucchese, A., Willers, J., Mittelman, A., Kanduc, D., Dummer, R., 2005. Proteomic scan for tyrosinase peptide antigenic pattern in vitiligo and melanoma: role of sequence similarity and HLA-DR1 affinity. J. Immunol. 175, 7009.

Minenkova, O., Pucci, A., Pavoni, E., De, T.A., Fortugno, P., Gargano, N., Cianfriglia, M., Barca, S., De, P.S., Martignetti, A., Felici, F., Cortese, R., Monaci, P., 2003. Identification of tumor-associated antigens by screening phage-displayed human cDNA libraries with sera from tumor patients. Int. J. Cancer 106, 534.

Pallasch, C.P., Struss, A.K., Munnia, A., Konig, J., Steudel, W.I., Fischer, U., Meese, E., 2005. Autoantibodies against GLEA2 and PHF3 in glioblastoma: tumor-associated autoantibodies correlated with prolonged survival. Int. J. Cancer 117, 456.

Pavoni, E., Vaccaro, P., Pucci, A., Monteriu, G., Beghetto, E., Barca, S., Dupuis, M.L., De Pasquale, C.A., Lugini, A., Cianfriglia, M., Cortesi, E., Felici, F., Minenkova, O., 2004. Identification of a panel of tumor-associated antigens from breast carcinoma cell lines, solid tumors and testis cDNA libraries displayed on lambda phage. BMC Cancer 4, 78.

Rosenberg, A., Griffin, K., Studier, W.F., McCormick, M., Berg, J., Novy, R., Mierendorf, R., 1996. T7 Select Phage Display System: a powerful new protein display system based on bacteriophage T7. Innovations 1–6.

Sahin, U., Tureci, O., Schmitt, H., Cochlovius, B., Johannes, T., Schmits, R., Stenner, F., Luo, G., Schobert, I., Pfreundschuh, M., 1995. Human neoplasms elicit multiple specific immune responses in the autologous host. Proc. Natl. Acad. Sci. U. S. A 92, 11810.

Sioud, M., Hansen, M.H., 2001. Profiling the immune response in patients with breast cancer by phage-displayed cDNA libraries. Eur. J. Immunol. 31, 716.

Slager, E.H., Borghi, M., van der Minne, C.E., Aarnoudse, C.A., Havenga, M.J., Schrier, P.I., Osanto, S., Griffioen, M., 2003. CD4+ Th2 cell recognition of HLA-DR-restricted epitopes derived from CAMEL: a tumor antigen translated in an alternative open reading frame. J. Immunol. 170, 1490.

Somers, V.A., Brandwijk, R.J., Joosten, B., Moerkerk, P.T., Arends, J.W., Menheere, P., Pieterse, W.O., Claessen, A., Scheper, R.J., Hoogenboom, H.R., Hufton, S.E., 2002. A panel of candidate tumor antigens in colorectal cancer revealed by the serological selection of a phage displayed cDNA expression library. J. Immunol. 169, 2772.

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

Stockert, E., Jager, E., Chen, Y.T., Scanlan, M.J., Gout, I., Karbach, J., Arand, M., Knuth, A., Old, L.J., 1998. A survey of the humoral immune response of cancer patients to a panel of human tumor antigens. J. Exp. Med. 187, 1349.

Usener, D., Schadendorf, D., Koch, J., Dubel, S., Eichmuller, S., 2003. cTAGE: a cutaneous T cell lymphoma associated antigen family with tumor-specific splicing. J. Invest. Dermatol. 121, 198.

Vaccaro, P., Pavoni, E., Monteriu, G., Andrea, P., Felici, F., Minenkova, O., 2006. Efficient display of scFv antibodies on bacteriophage lambda. J. Immunol. Methods 310, 149.

Vazquez-Ortiz, G., Garcia, J.A., Ciudad, C.J., Noe, V., Penuelas, S., Lopez-Romero, R., Mendoza-Lorenzo, P., Pina-Sanchez, P., Salcedo, M., 2007. Differentially expressed genes between high-risk human papillomavirus types in human cervical cancer cells. Int. J. Gynecol. Cancer 17, 484.

Wang, X., Yu, J., Sreekumar, A., Varambally, S., Shen, R., Giacherio, D., Mehra, R., Montie, J.E., Pienta, K.J., Sanda, M.G., Kantoff, P.W., Rubin, M.A., Wei, J.T., Ghosh, D., Chinnaiyan, A.M., 2005. Autoantibody signatures in prostate cancer. N. Engl. J. Med. 353, 1224.

Welling, G.W., Weijer, W.J., van der Zee, R., Welling-Wester, S., 1985. Prediction of sequential antigenic regions in proteins. FEBS Lett. 188, 215.

Willats, W.G., 2002. Phage display: practicalities and prospects. Plant Mol. Biol. 50, 837.

Yang, F., Forrer, P., Dauter, Z., Conway, J.F., Cheng, N., Cerritelli, M.E., Steven, A.C., Pluckthun, A., Wlodawer, A., 2000. Novel fold and

capsid-binding properties of the lambda-phage display platform protein gpD. Nat. Struct. Biol. 7, 230.

Zhang, X.D., Miao, S.Y., Wang, L.F., Li, Y., Zong, S.D., Yan, Y.C., Koide, S.S., 2000. Human sperm membrane protein (hSMP-1): a developmental testis-specific component during germ cell differentiation. Arch. Androl. 45, 239.

Zhong, L., Hidalgo, G.E., Stromberg, A.J., Khattar, N.H., Jett, J.R., Hirschowitz, E.A., 2005. Using protein microarray as a diagnostic assay for non-small cell lung cancer. Am. J. Respir. Crit Care Med. 172, 1308.

Zhong, L., Peng, X., Hidalgo, G.E., Doherty, D.E., Stromberg, A.J., Hirschowitz, E.A., 2004. Identification of circulating antibodies to tumor-associated proteins for combined use as markers of non-small cell lung cancer. Proteomics 4, 1216.

Zippelius, A., Gati, A., Bartnick, T., Walton, S., Odermatt, B., Jaeger, E., Dummer, R., Urosevic, M., Filonenko, V., Osanai, K., Moch, H., Chen, Y.T., Old, L.J., Knuth, A., Jaeger, D., 2007. Melanocyte differentiation antigen RAB38/NY-MEL-1 induces frequent antibody responses exclusively in melanoma patients. Cancer Immunol. Immunother. 56, 249.

Zucconi, A., Dente, L., Santonico, E., Castagnoli, L., Cesareni, G., 2001. Selection of ligands by panning of domain libraries displayed on phage lambda reveals new potential partners of synaptojanin 1. J. Mol. Biol. 307, 1329.

3.2 . Development of phage-displayed antigen microarray

In order to establish and optimize the technique for the production of PhD-AM, the following issues were addressed:

Choice of slide surface chemistry. A test set of 380 phage clones comprising 335 sero-reactive clones and 45 non-recombinant T7 phage clones was assembled, amplified and printed on slides coated with different surface: epoxy, aldehyde and nitrocellulose coated slides produced by two different companies. The slides were tested with serum and anti-T7 antibodies and spots morphology and signal-to-noise ratios were compared. For this study we selected nitrocellulose coated FAST slides produced by Whathman, that showed the best spot morphology and highest signal to noise ratio.

Choice of method for the amplification of phages. Plate lysates, 3 ml liquid culture lysates in individual tubes with or without subsequent precipitation with PEG/NaCl, and 0.5 ml liquid culture lysates in deepwell plates were prepared, spotted on FAST slides and tested with anti-T7 antibody. The phages that were grown in deepwell plates showed the less variation in spot size and intensity across the array therefore this amplification method was used for the rest of the study.

Antibody dilutions. Series of experiments were performed in order to determine the optimal dilutions of serum, anti-T7 antibody and secondary antibodies. The best signal-to-noise ratio was observed when 1:200 diluted serum, 1:1000 diluted anti-T7 antibody, 1:1500 anti-human IgG and 1:3000 anti-mouse IgG secondary antibodies were used.

Serum preabsorbtion. Pre-incubation of serum with *E. coli* and T7 phage lysates for 1 hour before putting it on a slide significantly reduced the non-specific Cy3 signal (serum reactivity against phage and *E. coli* proteins) and increased the Cy3/Cy5 signal ratio for positive clones, therefore we concluded that serum pre-absorption is an essential step in microarray processing protocol. To standardise this, large amounts of *E. coli* and phage lysates were prepared and the same batch of lysates was used all the way through the study.

Printing conditions. The conventional procedure for printing DNA was amended due to the differences in the chemical properties of DNA molecules and proteins. The results of our tests have shown the need to use lower levels of humidity (45-55%) and the catastrophic deterioration of spots in case of use of ethanol in a cycle of needle washing. The *E. coli* lysates tend to form deposits on the surface of the needles. This sediment was fixed by ethanol during washing cycle and could not be removed by the following wash cycles. Therefore 0.05% Tween has been entered to the wash cycle and the ethanol solution has been replaced by distilled water.

Data acquisition. The slides were scanned on AQuire scanner (Genetix) with 532 and 635 nm lasers at 10μm resolution with as the highest possible PMT gain that did not shown the saturation. The spots data has been extracted by GenePix Pro software using the proprietary spot recognition algorithm. The analysis of data, passed the normalization as described below, shows the variation of not more than 3% for arrays scanned in a fairly wide range PMT gains in the absence of saturated spots. The following data processing and analysis was made in ad hoc composed scripts using R-language software (96).

By addressing the above mentioned issues, the protocol for production and processing of antigen microarrays was developed. In order to validate the assay, the set of 380 phages was printed on FAST slides and tested for reactivity with the same set of 22 sera that was used for TA mini-library and macroarray screening and sera from 8 healthy donors. Median of Cy3/ Cy5 ratios was calculated for each spot and averaged between replicates. For local normalization LOWESS were used. In order to normalize the signal intensities among

different arrays, an average Cy3/Cy5 ratio for all wild-type phage spots in an array was set to be equal to 1 and the values for the rest of the spots were re-calculated accordingly. A positive cut-off value was set as 3 -4 SDs above the average for wild type phage spots in each array. Using these settings, 70-90% of phages that were positive in plaque assay screening were defined as positive by microarrays. Lowering the cut-off value resulted in better sensitivity but increased a false positive background.

The intra-assay variability of the phage-displayed antigen microarray technology was 7%, while the average CV was ~13% in the inter-assay comparisons, which is generally acceptable variability for immunoassays (103).

Plaque assay and PhD-AM turned out to have comparable, presumably due to high difference in conditions of adsorbing phage particles onto nitrocellulose. In some cases PhD-AM demonstrate even slightly lower sensitivity than the plaque assay.

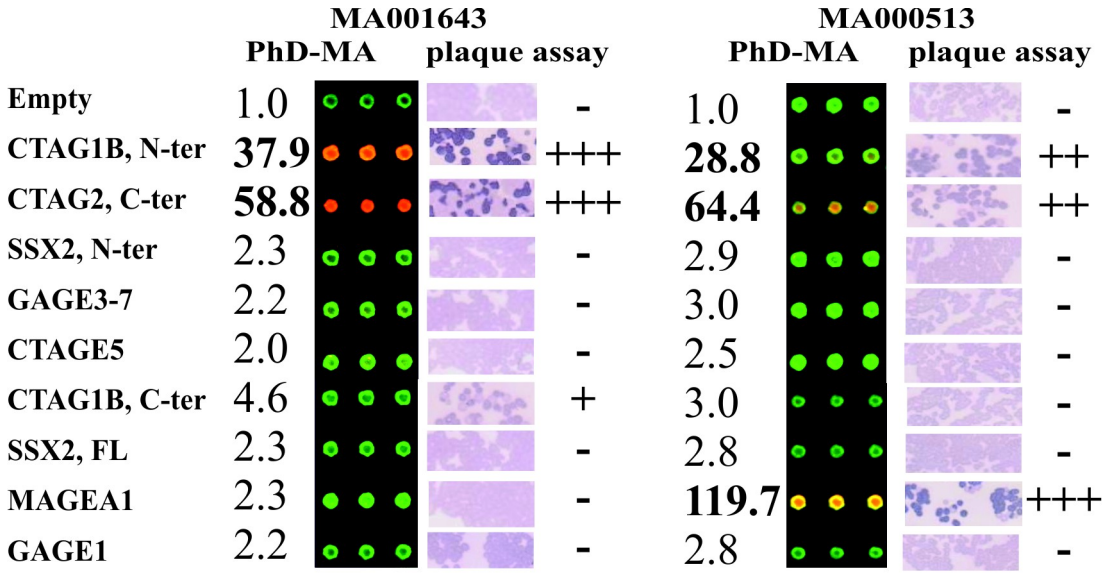

Figure 3. Comparison of sensitivity of PhD-AM technology and plaque assay.

3.3 . Ranking-based Kernels in Applied Biomedical Diagnostics using Support Vector Machine

International Journal of Neural Systems

RANKING-BASED KERNELS IN APPLIED BIOMEDICAL DIAGNOSTICS USING A SUPPORT VECTOR MACHINE

VILEN JUMUTC*
Riga Technical University, Meža 1/4
Riga, LV-1658, Latvia
E-mail: jumutc@gmail.com

PAWEL ZAYAKIN*
Latvian Biomedical Research & Study Center, Rātsupītes 1
Riga, LV-1067, Latvia
E-mail: pawel@biomed.lu.lv
http://biomed.lu.lv

ARKADY BORISOV
Riga Technical University, Meža 1/4
Riga, LV-1658, Latvia
E-mail: aborisov@cs.rtu.lv
http://www.cs.rtu.lv/dssg/dotnetnuke

This paper presents some essential findings and results on using ranking-based kernels for the analysis and utilization of high dimensional and noisy biomedical data in applied clinical diagnostics. We claim that presented kernels combined with a state-of-the-art classification technique — a Support Vector Machine (SVM) — could significantly improve the classification rate and predictive power of the wrapper method, e.g. SVM. Moreover, the advantage of such kernels could be potentially exploited for other kernel methods and essential computer-aided tasks such as novelty detection and clustering. Our experimental results and theoretical generalization bounds imply that ranking-based kernels outperform other traditionally employed SVM kernels on high dimensional biomedical and microarray data.

Keywords: ranking, SVM, kernel methods, z-score, diagnostics.

1. Introduction

Optimal kernel selection is one of the most crucial conditions for classification success and generalization bounds. In theory, good generalization depends on the right similarity measure encoded into kernels via appropriate expansion of the inner product. But in practice, the SVM method cannot find an optimal separating hyperplane if we employ some unreliable or even biased similarity measure in the original input space. In addition, any non-linear mapping and RBF expansion cannot help us to select an optimal decision surface if two classes collide into one point due to the lack of a precise similarity estimation.

Another shortcoming of a standard SVM kernel is a "curse of dimensionality" when the distribution of inner products or distances for the high dimensional spaces is characterized by a very small variance and the fourth central moment. This generally implies a higher model capacity and a higher fraction of Support Vectors that maintain the optimal decision boundary. Such distributions could be very unstable due to the peculiarities of the data (the number of meaningful dimensions, the number of samples, non-balanced target classes etc.). The latter conditions are very sensitive in applied biomedical analysis and potentially could bias even a very qualitatively and accurately handled experimental setup. For instance, during post-processing and analysis of the collected protein microarray data, we noticed that there was no significant difference between signal means for cancer specific and healthy samples, which resulted in a very limited applicability of

*The first two authors contributed equally to this work.

statistical inference for the related classification problem (see Figure 1). The latter fact indicates that for the classical L1- or L2-norm SVM with the standard RBF or another potentially highly discriminative kernel[1,15] the absence of major discriminating attributes could potentially lead to a highly non-sparse solution with as many Support Vectors as training samples. In this paper we confirm the previous conjecture for every scanned and evaluated microarray dataset.

To overcome the above-mentioned shortcomings we propose a topological ranking-based kernel[2] that measures similarity by means of rank information available for each attribute in the sample. This property helps to avoid expensive and time-consuming cross-sample normalization and provides classification with a robust and even more accurate estimation of similarity. In the following paper we extend a concept of a ranking-based kernel, adduce and analyze several generalization bounds for it and finally discuss possible applications of such kernels in applied clinical diagnostics.

The remainder of this paper is structured as follows. Section 2 reviews the problem outline and our primary data source. Section 3 gives a very brief overview of SVM, its generalization bounds, and the Multiple Kernel Learning method used for performing experiments. In Section 4 we recall previously proposed ranking-based kernels with some notable new extensions. In Section 5 we present generalization bounds for ranking-based kernels. In Section 6 we discuss the experimental setup followed by numerical results in Section 7 and statistical tests in Section 8 applied to our primary microarray dataset as well as to some public UCI datasets. Finally, we conclude with some discussion and findings about the proposed method and possible directions of future work.

2. Background

This section is dedicated to the problem outline, our primary data source and the description of the biomedical challenges in developing and implementing robust cancer diagnostic systems.

2.1. *Motivation*

We know that preventive measures, such as screening and awareness, could have a great impact on reducing cancer proliferation. Thus, the identification and validation of novel biomarkers for the early detection of cancer would contribute significantly to the decrease of cancer-related morbidity and mortality. The best known biomarkers still have low specificity or sensitivity. Microarrays are a convenient approach for detecting and observing tumor-derived proteins and involved genes. Thereafter they are popular in experiments for biomarker-related screening. Unfortunately, classic statistical inference methods cannot be applied to this data analysis due to the lack of discrimination power between averaged expected values for observed classes (healthy donors vs. cancer patients). And there is a great need for novel, more robust classification methods and diagnostic systems that could effectively overcome the above-mentioned problems.

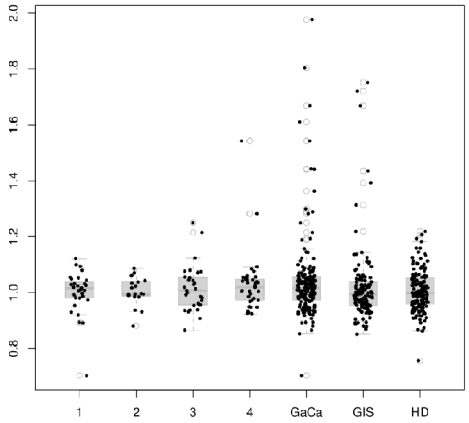

Fig. 1. Distribution of a "signal" intensities (*y* axis) within different classes (one spot – one person) for one of the promising biomarkers (1, 2, 3, 4 and GaCa – gastric cancer patients within different stages and overall picture, GIS – gastric intestinal disease, HD – healthy donors)

2.2. *Primary data source*

During the biomedical experimental setup we applied the T7 phage display-based SEREX technique to identify a representative set of antigens eliciting humoral responses in melanoma, prostate, gastric cancer and gastritis patients, as described in Refs.4–5. All identified antigens were used for the production of a phage displayed-antigen microarray (see Figure 2) that

was applied to survey autoantibody profiles in patients with melanoma (n=167), prostate (n=52), gastric cancer (n=176), various gastrointestinal inflammatory diseases (n=125) and healthy individuals (n=148). The microarray data were analyzed as the ratio of serum antibody signal intensities in cancer patients' sera and healthy donors (HD) to signal intensities of anti-T7 phage antibodies. Serum autoantibody profiling of 1322 element phage-displayed antigen microarrays comprising all immuno-selected antigens resulted in the identification of a high-dimensional biomarkers' panel to find antigens with potential high diagnostic and prognostic value.

The results of this study clearly show that the serum autoantibody signatures have a potential to detect the presence of cancer with higher accuracy than currently known tests. The latter signatures exploit a very limited number of biologically important outlying "signals" to the expected mean value for this autoantibody (biomarker) as a whole. Thus it was very important to evaluate a kernel that potentially could effectively incorporate this inner property of the data source.

3. Methods

In this section we briefly review the SVM method, a MKL extension for learning with multiple kernels and basic generalization bounds implied by Statistical Learning Theory[6]. This section gives only an insight into original SVM formalism and corresponding optimization problems.

3.1. *Support Vector Machine*

A Support Vector Machine is based on the concept of separating hyperplanes that define decision boundaries using Statistical Learning Theory. Using a kernel function, SVM is an alternative classification approach in which the optimal solution or decision surface is found by solving the quadratic programming problem with linear constraints, rather than by solving a non-convex, unconstrained minimization problem as stated in typical back-propagation neural network.

For more detailed and formal description of the SVM method we refer an interested reader to an extensive research and methodology papers dedicated to the practical usage of SVM in bioinformatics and engineering (see Ref.1 or Refs.15-18 for more detailed explanation), hence we do not provide any SVM-related theory in this subsection.

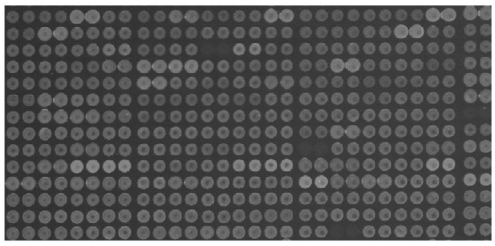

Fig. 2. Typical microarray resulted from immunoscreening of one individual (lighter spots stand for the higher "signal" intensity of evaluated antigens).

3.2. *Multiple Kernel Learning*

Multiple Kernel Learning aims at simultaneously learning the kernel and the associated predictor in the general SVM context. For the SVM, an efficient and general MKL algorithm, based on semi-infinite linear programming (SILP), has been proposed recently and could be easily applied to learning optimal kernel parameters without facing the cross-validation for large biomedical problems[7].

The latest applications of MKL have clearly proven that using multiple kernels instead of a single one can enhance the interpretability of the decision function and improve performances[8,9]. In such cases, a convenient approach is to consider that the kernel $K(x,x')$ is actually a convex combination of basis kernels:

$$K(x,x')=\sum_{m=1}^{M} d_m K_m(x,x'), \quad d_m \geq 0, \quad \sum_{m=1}^{M} d_m = 1 \qquad (3)$$

where M is the total number of kernels. Each basis kernel K_m may either use the full set of variables describing x or subsets of variables extracted from different data sources[8]. Alternatively, the kernels K_m can simply be any standard kernels (such as Gaussian or RBF kernels) with different parameterization. Within this framework, the problem of data representation through the kernel is then transferred to the choice of optimal weights d_m that minimizes the MKL objective function[9].

In our work we have used the SimpleMKL framework[3] to perform the general classification task and estimate the optimal bandwidth parameter (γ) of the basis RBF kernel being employed in our ranking-based

kernel as well.

3.3. *SVM generalization bounds*

The theory of Structural Risk Minimization relies on a VC-dimension for large margin classifiers and the following bound can be proven from VC-theory for the generalization error ε using hyperplanes in feature space[10]:

$$\varepsilon \le O\left(\frac{R^2}{m\gamma^2}\right), \quad (4)$$

where R is the radius of the smallest unit ball containing the all training set, m is the number of training samples and γ is the margin between separating hyperplanes.

Another useful bound on generalization error can be directly defined by means of a VC-dimension and an empirical error:

$$P\{Y \neq f(X)\} \le \hat{P}_n\{Y \neq f(X)\} + \sqrt{\left(\frac{h(\log(2n/h)+1)-\log(\eta/4)}{n}\right)}, \quad (5)$$

where $\hat{P}$ is the probability of making an empirical error, h is the VC dimension of the hypothesis's space, n is the number of training samples, and $\eta = 1-\delta$ is the corresponding confidence level (hereby and after by δ we indicate a significance level).

We search for a generalized bound that could be extended for any arbitrary kernel regardless of the commonly undefined VC-dimension. This bound (6) was recently yielded by Bartlett and Mendelson[11]; basically it incorporates a Rademacher complexity of the restricted function class F of $\{\pm 1\}$-valued functions defined on some domain $X \times \{\pm 1\}$ with respect to probability P. Additionally this bound considers an empirical distribution P^n and training samples $(X_i, Y_i)_{i=1}^n$ drawn according to this distribution (see Ref.11 and corresponding Theorem 5 of Bartlett and Mendelson).

$$P\{Y \neq f(X)\} \le \hat{P}_n\{Y \neq f(X)\} + \frac{R_n(F)}{2} + \sqrt{\frac{\ln(1/\delta)}{2n}}, \quad (6)$$

where

$$F = \left\{x \mapsto \sum_{i=1}^{n} \lambda_i k(x, x_i) : n \in \mathrm{N}, x_i \in \mathrm{X}, \sum_{i,j} \lambda_i \lambda_j k(x_i, x_j) \le \gamma^{-2}\right\}.$$

and the empirical Rademacher complexity of F on a sample $S = (x_1, x_2, \ldots, x_n) \in Z^n$ is defined as

$$\hat{R}_S(F) = \frac{2}{n} \mathrm{E}\left[\sup_{f \in F}\left|\sum_{i=1}^{n} \sigma_i f(x_i)\right| S\right],$$

where $\sigma_1, \sigma_2, \ldots, \sigma_n$ are independent random Rademacher variables.

This bound implies one stronger corollary that incorporates the kernel function into the Rademacher complexity and using McDiarmid's and Jensen's inequalities makes the stated above bound tighter:

$$R_n(F) = \mathrm{E}\hat{R}_n(F) \le \frac{2}{\gamma}\sqrt{\frac{\mathrm{E}k(X,X)}{n}}, \quad (7)$$

(see Ref.11 and corresponding Lemma 22 of Bartlett and Mendelson for proof and further details).

The latest bound is of the greatest interest because of the direct impact of the kernel on generalization bound and lower misclassification rates. Further we will show that ranking-based kernels proposed in this paper imply more accurate and robust generalization bounds in comparison with simpler kernels.

4. Ranking-based kernels

In this section we propose a new ranking-based kernel and present some loose and tight generalization bounds obtained for this kernel in terms of the Rademacher complexity.

4.1. *Simple kernel*

Our proposed kernel uses rank information available for each attribute in the sample. This information is acquired by introducing a so-called topological measure of every attribute that has its own relative disposition in the sample that does not depend on other samples and can be completely regarded as an ordinal ranking of this attribute within each sample. The formal description of proposed "upper-ranking" and "lower-ranking"-based topological measures for *i-th* attribute of each sample one can find in our previous paper (see Ref.2 for details) and for the sake of completeness they are given by:

$$\Omega_{up}(x^{(i)}) = \frac{1}{|\tau|}\sum_{j=1}^{m} I(\tau_j > x^{(i)}), \quad (8)$$

$$\Omega_{low}(x^{(i)}) = \frac{1}{|\tau|}\sum_{j=1}^{m} I(\tau_j \le x^{(i)}), \quad (9)$$

Hereby I is an indicator function (defined on $I(A,x): X \to \{0,1\}$, where A is a condition and x is a

related variable), $x^{(i)}$ is the value under *i-th* attribute, m is the total number of attributes and τ is a vector of all possible unique attribute "values" within sample x.

Finally, the first-order representation of the ranking-based kernel can be obtained as follows:

$$\hat{K}(x_i, x_j) = \langle \Omega_{up}(x_i), \Omega_{up}(x_j) \rangle + \langle \Omega_{low}(x_i), \Omega_{low}(x_j) \rangle, \quad (10)$$

where $\langle x,y \rangle$ represents inner product between vectors x and y, while $\Omega_{up}(x)$ and $\Omega_{low}(x)$ stand for "upper-ranking" and "lower-ranking"-based vectors of topological measures of sample x.

Analogically, the RBF kernel can be formalized by:

$$K_{RBF}(x_i, x_j) = \exp(-\gamma \cdot C(x_i, x_j)), \quad (11)$$
$$s.t. \quad C(x_i, x_j) = \|\Omega_{up}(x_i) - \Omega_{up}(x_j)\|^2 + \|\Omega_{low}(x_i) - \Omega_{low}(x_j)\|^2.$$

Hereby γ is a bandwidth parameter of RBF kernel while corresponding proof can be found in Appendix D.

Finally, we present a slightly corrected ranking-based kernel in order to imply further stronger bounds on the generalization error. The major difference is in the applied square-root transformation to initial topological similarity measures. Since the kernel matrix is positive semi-definite, this transformation has no restrictions and is given as follows:

$$\hat{\Omega}_{up}(x^{(i)}) = \sqrt{\frac{1}{|\tau|} \sum_{j=1}^{m} I(\tau_j > x^{(i)})}, \quad (12)$$

$$\hat{\Omega}_{low}(x^{(i)}) = \sqrt{\frac{1}{|\tau|} \sum_{j=1}^{m} I(\tau_j \le x^{(i)})}. \quad (13)$$

And finally we get the first-order representation of the ranking-based kernel upper-bounded by the following term (see Appendix A for details):

$$\hat{K}(x, x') \le m. \quad (14)$$

4.2. *Z-score based kernel*

To extend our work and develop some novel extension to the original ranking-based kernel, we propose a z-score related modification that incorporates the same ranking-based topological measures (8)-(10), but performs additional transformation of these measures by the means of a standard score:

$$z = \frac{x - \mu}{\sigma}. \quad (15)$$

Instead of calculations in the original input space, the z-score is evaluated by a set of topological measures that are within the discrete ranking-based space. This directly implies necessary replacement of the mean and standard deviation in equation (15) by the means of rank information calculated over sample x:

$$\mu_{rank}(x) = \Omega_{low}(\hat{\mu})$$
$$s.t. \quad \hat{\mu}(x) = median(x), \quad (16)$$

and

$$\sigma_{rank}(x) = |\tau| \cdot |\Omega_{low}(\hat{\mu}) - \Omega_{low}(\hat{\mu} - \sigma)|, \quad (17)$$

or alternatively balanced ranking-based standard deviation is given by:

$$\sigma_{rank}(x) = \frac{1}{2}|\tau| \cdot [\tilde{\sigma}_{rank}(x, \sigma) + \tilde{\sigma}_{rank}(x, -\sigma)]$$
$$s.t. \quad \tilde{\sigma}_{rank}(x, \sigma) = |\Omega_{low}(\hat{\mu}) - \Omega_{low}(\hat{\mu} - \sigma)|, \quad (18)$$

Where the incorporated "lower-ranking" based topological measure could be effectively replaced by the "upper-ranking" one due to the inner symmetry of both topological measures:

$$\mu_{rank}(x) = \Omega_{up}(\hat{\mu})$$
$$\sigma_{rank}(x) = |\tau| \cdot |\Omega_{up}(\hat{\mu}) - \Omega_{up}(\hat{\mu} + \sigma)|$$
$$s.t. \quad \Omega_{up}(\hat{\mu}) = 1 - \Omega_{low}(\hat{\mu}) \quad (19)$$

In the formal definition of the z-score based kernel we employ statistical measures (σ_{rank}, μ_{rank}) inferred according to Ω_{low} topological similarity measure. This basic assumption is inferred from the similarity of resulted transformations in discrete ranking-based space. Therefore, applying equation (15) to the previously given definition of the ranking-based kernel[a], we can derive necessary formalization of a z-score based kernel according to above stated in (16) and (18) statistical measures:

$$\hat{K}(x_i, x_j) = \sum_{k=1}^{m} \frac{(\Omega_{up}(x_i^{(k)}) - \mu_{rank}(x_i)) \cdot (\Omega_{up}(x_j^{(k)}) - \mu_{rank}(x_j))}{\sigma_{rank}(x_i) \cdot \sigma_{rank}(x_j)} \quad (20)$$

Similar to (11) when we apply the extension to some higher dimensional Hilbert space, we derive the corresponding z-score based RBF kernel and perform additional tests for it.

[a] See equation (10) for a linear representation of kernel

Finally, to prove experimental results on the discussed kernel and make some fair comparison with the simpler z-score based transformation (15) in original input space, we define the inner product or basic linear kernel as follows:

$$\langle x,x' \rangle_z = \hat{K}_z(x,x') = \langle z,z' \rangle. \quad (21)$$

5. Generalization bounds

In this section we present loose and tight generalization bounds on the ranking-based kernels proposed in Section 4 in terms of the Rademacher complexity and previously presented upper bounds on the kernel itself.

5.1. *Loose bounds*

The basic idea provided by Bartlett and Mendelson[11] of the generalization bound in (6) is to restrict the Rademacher complexity to some kernel-derived function class and infer an appropriate upper bound for this complexity in terms of the expected value of the kernel function (see Lemma 22 of Bartlett and Mendelson in Ref.11). But for the loose bound, we could do a somewhat different inference and use the initial Rademacher complexity term unaffected. Instead, we bound the norm of the decision function f in the Rademacher complexity:

$$R_n(F) = \frac{2}{n} \sup_{f \in F} \left| \sum_{i=1}^{n} \sigma_i f(X_i) \right| \quad (22)$$

We state our loose generalization bound for the simple ranking-based kernel defined by topological similarity measures (12), (13) similarly to inequality (6) with the corresponding Rademacher complexity bounded in equation (23):

$$R_n(F) = E\hat{R}_n(F) \leq \frac{2}{n\gamma} E\left[\sum_{i=1}^{n} \sigma_i \cdot \sqrt{m} \right], \quad (23)$$

where σ is a vector of Rademacher independent random variables, m is the number of training samples and γ the margin between discrimination hyperplanes and F is given as in (6). Proof of the presented upper bound on the generalization error can be found in Appendix B.

For the initial ranking-based topological measures (8), (9) we infer the following upper bound on the Rademacher complexity:

$$R_n(F) = E\hat{R}_n(F) \leq \frac{2}{n\gamma} E\left[\sum_{i=1}^{n} \sigma_i \cdot \sqrt{2m-1} \right], \quad (24)$$

where m is an input dimensionality of the classification problem. The additive term of inequality is inferred according to the following basic upper bounds on ranking-based similarity measures:

$$\Omega_{low}(x^{(i)}) = \mathrm{E}(I_{x^{(i)} \geq x}) = \Pr(x^{(i)} \geq x) \leq 1, \quad (25)$$

and the reciprocal probability on the "upper-ranking" based similarity measure for the *i-th* attribute is surely upper-bounded by:

$$\Omega_{up}(x^{(i)}) = \mathrm{E}(I_{x^{(i)} < x}) = \Pr(x^{(i)} < x) \leq \frac{m-1}{m}. \quad (26)$$

From the bounds stated above we can clearly see that square-root transformation of similarity measures (12), (13) implies stronger bounds on the Rademacher complexity for every $m > 1$. Proof of the presented upper bound can be found in Appendix C.

In the same manner we infer the generalization bounds for the z-score ranking-based kernel that incorporates newly proposed statistical measures (σ_{rank}, μ_{rank}).

5.2. *Stronger bounds*

For the stronger upper bounds we have directly applied the Lemma 22 of Bartlett and Mendelson[11] and the resulting upper bound on the Rademacher complexity in (7).

In Ref.11 the authors yielded this bound with respect to the empirical Rademacher complexity that is defined as an expectation of the kernel function for a symmetric case of equal samples. Further we could get an even stronger bound using inequality (14) and transformed similarity measures (12), (13):

$$R_n(F) = \mathrm{E}\hat{R}_n(F) \leq \frac{2}{\gamma} \sqrt{\frac{\mathrm{E}(\hat{K}(x,x))}{n}} = \frac{2}{\gamma} \sqrt{\frac{m}{n}}. \quad (27)$$

For the initial ranking-based topological measures (8), (9) we infer the corresponding upper bound on the Rademacher complexity as follows:

$$R_n(F) = \mathrm{E}\hat{R}_n(F) \leq \frac{2}{\gamma} \sqrt{\frac{2m-1}{n}}. \quad (28)$$

Final generalization error bounds are given by (6) with Rademacher complexities provided by previously

yielded inequalities (27), (28).

5.3. *Remarks*

By analyzing the above given bounds and supported by ranking-based kernels generalization we see that error bounds seem to be more stable and predictable. This is obvious if we compare the bounded Rademacher complexities in (7) and (27). The first equation contains the expectation of the kernel function on the evaluated dataset. The latter can be very different from dataset to dataset and exhibits greater stability in the case of applied scaling. As opposed to the bound given by inequality (7), ranking-based kernels imply very simple and stable upper bounds that rely only on dimensionality of any classification problem and could be computed without explicit construction of kernel matrices.

6. Experimental setup

In our experiments we have tested the proposed model under the predefined C=10 (error trade-off) value of the soft-margin SVM. The experiments show the most comprehensible performance for high dimensional sparsely sampled data set and varying γ value of RBF kernel that trade-offs kernel smoothness and can be effectively estimated via the SimpleMKL framework[3]. To exhaustively test our dataset(s), we decided to use two different versions of it:

(i) a non-normalized dataset directly obtained by a scanning microarray chip,

(ii) a dataset normalized via OLIN[13] and other ad-hoc techniques.

To verify and test the topological and RBF kernels under selected performance measures we decided to conduct the following experimental setup that consisted of some prefixed number of iterations (I=100 in our experimental setup) where in every iteration a verification set was composed of i.i.d. selected N samples ($N=50$ for all datasets) and all remaining samples were used as the training set. After fixing the number of iterations for each independent trial we have employed the MKL approach to estimate optimal "tuning" parameters for the SVM classifier. Additionally, we have conducted two separate sets of experiments for scaled and non-scaled invariants of evaluated datasets to estimate the importance of scaling for the proposed ranking-based and RBF kernels.

Finally, each set of experiments was divided according to the different origins of data sources (melanoma, gastric, prostate etc.) and every separate data source was evaluated jointly in all available invariants (normalized vs. non-normalized) on the same verification set.

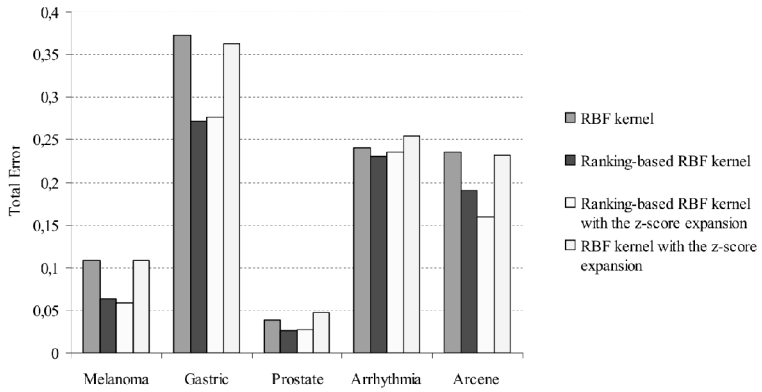

Fig. 3. The lowest attained error rates for every evaluated dataset

The aforementioned scheme was not employed for UCI datasets[14]. Because of the initial separation of the Arcene training and validation datasets where both sets were reasonably normalized and contributed to SVM in an already refined manner, we tested the Arcene dataset only under scaled/non-scaled invariants. Note that the experimental setup and similar reasons correspondingly apply for the Arrhythmia dataset.

6.1. *Additional scaling*

Verifying the performance of the proposed kernel, we have found that pseudo-normality of a signal distribution within each sample plays an essential role in classification success for all datasets, while scaling of the input space sometimes leads to performance degradation due to the "corrupted" normality.

Therefore we have decided to fix the normality of

Table 1. Averaged performance measures for simple ranking-based kernel.

Dataset	Number of selected kernels	Total Error	BER
Melanoma A (RBF kernel)	1.980±0.140	0.108±0.038	0.112±0.056
Melanoma A (Ranking-based kernel)	2.190±2.312	0.101±0.040	0.103±0.060
Melanoma B (RBF kernel)	14.45±3.851	0.167±0.057	0.169±0.081
Melanoma B (Ranking-based kernel)	5.540±8.407	**0.071±0.036**	**0.074±0.052**
Melanoma C (RBF kernel)	1.950±0.219	0.113±0.046	0.119±0.062
Melanoma C (Ranking-based kernel)	2.000±0.000	**0.088±0.043**	**0.091±0.060**
Melanoma D (RBF kernel)	16.74±12.14	0.163±0.050	0.164±0.078
Melanoma D (Ranking-based kernel)	5.760±9.342	**0.064±0.031**	**0.068±0.045**
Gastric A (RBF kernel)	15.98±5.596	0.401±0.069	0.425±0.084
Gastric A (Ranking-based kernel)	19.36±2.772	0.396±0.062	0.421±0.078
Gastric B (RBF kernel)	5.810±4.948	0.381±0.070	0.385±0.126
Gastric B (Ranking-based kernel)	1.960±0.197	**0.314±0.063**	**0.322±0.099**
Gastric C (RBF kernel)	8.910±9.151	0.416±0.076	0.431±0.100
Gastric C (Ranking-based kernel)	14.97±9.066	0.389±0.081	0.399±0.120
Gastric D (RBF kernel)	12.39±12.31	0.376±0.068	0.374±0.122
Gastric D (Ranking-based kernel)	1.810±0.394	**0.338±0.077**	**0.350±0.114**
Gastric E (RBF kernel)	11.78±7.819	0.419±0.066	0.438±0.078
Gastric E (Ranking-based kernel)	13.58±4.522	0.411±0.066	0.431±0.076
Gastric F (RBF kernel)	5.870±5.438	0.380±0.075	0.382±0.121
Gastric F (Ranking-based kernel)	12.31±6.799	**0.272±0.062**	**0.271±0.103**
Prostate A (RBF kernel)	1.140±0.349	0.042±0.032	0.073±0.060
Prostate A (Ranking-based kernel)	1.000±0.000	0.039±0.032	0.062±0.064
Prostate B (RBF kernel)	10.36±5.410	0.131±0.040	0.206±0.081
Prostate B (Ranking-based kernel)	1.310±0.465	**0.028±0.020**	**0.048±0.043**
Prostate C (RBF kernel)	1.060±0.239	0.088±0.043	0.101±0.083
Prostate C (Ranking-based kernel)	1.080±0.273	0.102±0.048	0.196±0.073
Prostate D (RBF kernel)	9.380±10.02	0.118±0.047	0.189±0.088
Prostate D (Ranking-based kernel)	1.000±0.000	**0.026±0.023**	**0.050±0.047**
Arrhythmia A (RBF kernel)	3.590±1.634	0.240±0.052	0.244±0.081
Arrhythmia A (Ranking-based kernel)	3.050±3.448	**0.231±0.052**	**0.235±0.082**
Arrhythmia C (RBF kernel)	3.280±0.473	0.260±0.061	0.266±0.088
Arrhythmia C (Ranking-based kernel)	4.530±4.834	0.251±0.062	0.257±0.089
Arcene A (RBF kernel)	1.000±0.000	0.242±0.043	0.252±0.057
Arcene A (Ranking-based kernel)	2.000±0.000	0.252±0.043	0.266±0.055
Arcene C (RBF kernel)	1.000±0.000	0.439±0.058	0.500±0.500
Arcene C (Ranking-based kernel)	2.000±0.000	**0.191±0.036**	**0.205±0.052**

Table 2. Averaged performance measures for z-score related expansion of ranking-based kernel.

Dataset	Number of selected kernels	Total Error	BER
Melanoma A (RBF kernel)	1.990±0.100	0.111±0.045	0.116±0.063
Melanoma A (Ranking-based kernel)	2.090±0.321	0.107±0.044	0.111±0.063
Melanoma B (RBF kernel)	16.61±16.00	0.168±0.047	0.167±0.069
Melanoma B (Ranking-based kernel)	12.97±12.09	**0.059±0.032**	**0.062±0.047**
Melanoma C (RBF kernel)	2.820±0.411	0.119±0.038	0.124±0.052
Melanoma C (Ranking-based kernel)	1.470±0.502	0.123±0.046	0.123±0.068
Melanoma D (RBF kernel)	30.23±14.07	0.170±0.046	0.171±0.075
Melanoma D (Ranking-based kernel)	3.680±6.583	**0.058±0.031**	**0.059±0.044**
Gastric A (RBF kernel)	17.35±8.070	0.412±0.061	0.440±0.077
Gastric A (Ranking-based kernel)	22.70±9.029	0.407±0.059	0.436±0.073
Gastric B (RBF kernel)	6.800±10.13	0.382±0.063	0.385±0.124
Gastric B (Ranking-based kernel)	2.140±0.349	**0.309±0.055**	**0.320±0.089**
Gastric C (RBF kernel)	13.92±12.34	0.420±0.069	0.435±0.112
Gastric C (Ranking-based kernel)	2.380±2.260	0.409±0.077	0.424±0.115
Gastric D (RBF kernel)	19.55±20.07	0.372±0.066	0.374±0.126
Gastric D (Ranking-based kernel)	2.000±0.449	**0.327±0.063**	**0.334±0.107**
Gastric E (RBF kernel)	12.24±11.04	0.408±0.066	0.439±0.083
Gastric E (Ranking-based kernel)	17.17±10.33	0.397±0.066	0.427±0.084
Gastric F (RBF kernel)	11.18±13.73	0.373±0.068	0.379±0.111
Gastric F (Ranking-based kernel)	16.81±9.586	**0.277±0.066**	**0.277±0.108**
Prostate A (RBF kernel)	1.150±0.359	0.044±0.029	0.069±0.064
Prostate A (Ranking-based kernel)	1.930±0.355	**0.038±0.032**	**0.057±0.063**
Prostate B (RBF kernel)	15.01±11.74	0.129±0.042	0.198±0.081
Prostate B (Ranking-based kernel)	2.920±0.273	0.055±0.042	0.094±0.066
Prostate C (RBF kernel)	1.120±0.327	0.090±0.051	0.109±0.101
Prostate C (Ranking-based kernel)	1.570±0.498	0.153±0.062	0.279±0.087
Prostate D (RBF kernel)	3.810±0.929	0.115±0.040	0.181±0.077
Prostate D (Ranking-based kernel)	1.830±0.378	**0.028±0.022**	**0.051±0.043**
Arrhythmia A (RBF kernel)	3.590±1.832	0.246±0.060	0.251±0.088
Arrhythmia A (Ranking-based kernel)	4.140±2.331	0.246±0.058	0.249±0.084
Arrhythmia C (RBF kernel)	3.160±0.487	0.256±0.052	0.263±0.085
Arrhythmia C (Ranking-based kernel)	4.230±4.292	**0.236±0.057**	**0.242±0.090**
Arcene A (RBF kernel)	1.000±0.000	0.236±0.039	0.247±0.058
Arcene A (Ranking-based kernel)	2.000±0.000	0.255±0.044	0.271±0.060
Arcene C (RBF kernel)	1.000±0.000	0.438±0.048	0.500±0.000
Arcene C (Ranking-based kernel)	1.000±0.000	**0.160±0.036**	**0.176±0.052**

this dataset after scaling by the following δ_i term:

$$\delta_i = \mu\left(X^{(i)}\right) - \mu(X), \tag{29}$$

where μ is an averaging operator and δ_i applies to all samples with the fixed *i-th* attribute:

$$X^{(i)} = X^{(i)} - \delta. \tag{30}$$

7. Numerical results

In the following section we summarize the experimental results for all datasets under the fixed *C* parameter and enclosed subspace for the γ parameter with an initial guess of its corresponding scaling factor[b]. All experiments and kernel functions were implemented in MATLAB. In Figure 3 we briefly summarize all evaluated experiments by presenting only lowest attained error rates for every dataset. Further in Table 1 and Table 2 we present detailed performance measures

[b] We have defined range $b_\gamma \cdot 10^{[-10...10]}$ with the step 0.25 resulting in a total of 81 kernels where b_γ - a corresponding scaling factor of γ stated as follows is: $b_\gamma = 1/2\sqrt{median(X)}$ where X is given dataset.

(number of kernels, total error and BER – balanced error rate) obtained by the MKL approach for SVM with the standard RBF kernel and for SVM with the proposed ranking-based kernel (11) supported by the following linear expansions in (10) and (20). All the results are summarized according to the separately handled experimental evaluations when the RBF kernel was experimented more than once and in each experiment it was compared with another kernel to obtain necessary P-values. Best values are presented in bold.

In Tables 1 and 2, letters A through F abbreviate tested invariants of initial datasets: A stands for the normalized scaled dataset, B stands for the non-normalized scaled dataset, C stands for the normalized non-scaled dataset, D stands for the non-normalized non-scaled dataset, E stands for the normalized rescaled by δ_i dataset and finally F stands for the non-normalized rescaled by δ_i dataset.

To thoroughly test our newly proposed z-score based modification of the kernel, in Table 3 we present some additional results on the RBF kernel with linear expansion defined by (21) to prove that basic transformation to z-space gives lower generalization error rates but ranking-based expansion improves it even more.

Additionally we analyze and describe results under scaled and non-scaled invariants of evaluated datasets with the provided number of selected by SimpleMKL kernels. All results are averaged across 100 independent trials and provided with standard deviations.

Table 3. Averaged performance measures for simple z-score based expansion of RBF kernel.

Dataset	Number of selected kernels	Total Error	BER
Melanoma A (RBF kernel)	1.980±0.141	0.108±0.044	0.113±0.064
Melanoma A (Z-score based kernel)	2.950±0.219	0.104±0.043	0.109±0.061
Melanoma B (RBF kernel)	16.57±15.98	0.163±0.048	0.163±0.069
Melanoma B (Z-score based kernel)	2.000±0.000	**0.080±0.037**	**0.086±0.052**
Melanoma C (RBF kernel)	2.710±0.518	0.121±0.043	0.125±0.055
Melanoma C (Z-score based kernel)	3.000±0.000	**0.099±0.033**	**0.097±0.057**
Melanoma D (RBF kernel)	30.87±14.29	0.170±0.050	0.168±0.081
Melanoma D (Z-score based kernel)	3.040±0.281	**0.091±0.036**	**0.093±0.047**
Gastric A (RBF kernel)	16.43±7.276	0.404±0.073	0.431±0.076
Gastric A (Z-score based kernel)	18.76±7.168	0.407±0.072	0.436±0.071
Gastric B (RBF kernel)	8.220±11.79	0.370±0.059	0.372±0.117
Gastric B (Z-score based kernel)	3.250±0.609	0.386±0.065	0.396±0.122
Gastric C (RBF kernel)	15.37±13.59	0.426±0.060	0.447±0.103
Gastric C (Ranking-based kernel)	3.980±5.958	0.416±0.065	0.443±0.110
Gastric D (RBF kernel)	16.82±18.10	0.363±0.062	0.371±0.118
Gastric D (Z-score based kernel)	3.000±0.000	0.371±0.060	0.388±0.105
Gastric E (RBF kernel)	13.13±10.56	0.414±0.063	0.442±0.087
Gastric E (Z-score based kernel)	18.91±7.641	0.416±0.061	0.444±0.086
Gastric F (RBF kernel)	10.04±13.02	0.377±0.061	0.381±0.115
Gastric F (Z-score based kernel)	6.600±8.513	**0.329±0.066**	**0.329±0.115**
Prostate A (RBF kernel)	1.180±0.386	0.040±0.027	0.065±0.060
Prostate A (Z-score based kernel)	1.920±0.307	**0.032±0.028**	**0.050±0.058**
Prostate B (RBF kernel)	17.74±13.05	0.128±0.045	0.198±0.081
Prostate B (Z-score based kernel)	2.950±0.219	0.047±0.040	0.084±0.071
Prostate C (RBF kernel)	1.190±0.394	0.089±0.043	0.106±0.080
Prostate C (Z-score based kernel)	2.990±0.100	0.154±0.054	0.272±0.076
Prostate D (RBF kernel)	3.740±0.848	0.122±0.036	0.184±0.074
Prostate D (Z-score based kernel)	1.920±0.273	**0.029±0.028**	**0.048±0.051**
Arrhythmia A (RBF kernel)	3.790±2.591	0.254±0.060	0.261±0.083
Arrhythmia A (Z-score based kernel)	3.410±4.144	0.251±0.061	0.257±0.084
Arrhythmia C (RBF kernel)	3.310±0.545	0.264±0.059	0.272±0.089
Arrhythmia C (Z-score based kernel)	11.80±9.994	**0.243±0.062**	**0.252±0.084**
Arcene A (RBF kernel)	1.000±0.000	0.232±0.048	0.242±0.063
Arcene A (Z-score based kernel)	2.000±0.000	0.232±0.048	0.242±0.063
Arcene C (RBF kernel)	1.000±0.000	0.440±0.052	0.500±0.000
Arcene C (Z-score based kernel)	2.000±0.000	**0.193±0.037**	**0.211±0.054**

8. Statistical tests

In this section we present results of a two-tailed student's t-test applied to classification errors derived from 100 independent trials of the MKL method for each invariant of the evaluated dataset. The corresponding P-values that indicate the level of confidence of the paired comparisons under null-hypothesis of equal error means are given in Tables 4 – 9.

The tables are organized as follows: Table 4 and Table 5 summarize results on our original ranking-based kernel, Tables 6 – 7 present additional P-values on the z-score based expansion of our original kernel and finally Tables 8 – 9 summarize results on the simple z-score based expansion of the RBF kernel.

Table 4. P-values that indicate the confidence level of significance when comparing the proposed ranking-based kernel to the RBF one in terms of classification error

Dataset	Ranking-based kernel vs. RBF kernel
Melanoma A	0.17626
Melanoma B	0
Melanoma C	0.00014896
Melanoma D	0
Gastric A	0.59833
Gastric B	1.8182e-11
Gastric C	0.017626
Gastric D	0.00023403
Gastric E	0.42226
Gastric F	0
Prostate A	0.48182
Prostate B	0
Prostate C	0.026427
Prostate D	0
Arrhythmia A	0.23271
Arrhythmia C	0.2902
Arcene A	0.086388
Arcene C	0

In the Tables 5, 7, 9 letters AL through CL abbreviate the tested invariants of our initial datasets: AL stands for the scaled dataset, BL stands for for the non-scaled and CL stands for the rescaled by δ_i dataset.

Note that we have conducted additionally two joint comparisons for normalized and non-normalized invariants of microarray datasets (assuming identical validation sets) in order to show that the ranking-based kernel performs significantly (in terms of achieved P-values) better than the RBF kernel. This is especially true if we consider a non-normalized dataset.

Table 5. P-values that indicate the confidence level of significance when comparing normalized to non-normalized datasets (assuming strictly ranking-based vs. RBF kernel) in terms of classification error

Dataset	Normalized vs. non-normalized	Non-normalized vs. normalized
Melanoma AL	1.8181e-11	0
Melanoma BL	6.6613e-16	0
Gastric AL	0	0.11797
Gastric BL	1.0951e-11	0.22854
Gastric CL	0	0.0019808
Prostate AL	0.00016227	0
Prostate BL	0	0.020899

Evaluated two-tailed student's t-test provides necessary evidence of an achieved performance boost and significantly differed distribution means of generalization errors for 5% significance level and assumed equivalence of unknown variances (t-test was provided by MATLAB function *ttest2*).

Table 6. P-values that indicate the confidence level of significance when comparing the z-score based expansion of ranking-based kernel to the RBF one in terms of classification error

Dataset	Ranking-based kernel vs. RBF kernel
Melanoma A	0.54656
Melanoma B	0
Melanoma C	0.48558
Melanoma D	0
Gastric A	0.57038
Gastric B	8.8818e-16
Gastric C	0.30833
Gastric D	1.671e-06
Gastric E	0.22403
Gastric F	0
Prostate A	0.14991
Prostate B	0
Prostate C	4.3976e-13
Prostate D	0
Arrhythmia A	0.94243
Arrhythmia C	0.010252
Arcene A	0.001545
Arcene C	0

Table 7. P-values originated from comparing normalized to non-normalized datasets in terms of classification error (assuming strictly z-score based vs. RBF kernel)

Dataset	Normalized vs. non-normalized	Non-normalized vs. normalized
Melanoma AL	0	0
Melanoma BL	0	7.8958e-12
Gastric AL	0	0.0042497
Gastric BL	0	0.0002886
Gastric CL	0	0.0127130
Prostate AL	0.039284	0
Prostate BL	0	7.7827e-07

Table 8. P-values that indicate the confidence level of significance when comparing results on the simple z-score based expansion of the RBF kernel to RBF kernel as itself in terms of classification error

Dataset	Ranking-based kernel vs. RBF kernel
Melanoma A	0.57815
Melanoma B	0
Melanoma C	6.3343e-05
Melanoma D	0
Gastric A	0.77034
Gastric B	0.082055
Gastric C	0.26064
Gastric D	0.40635
Gastric E	0.85514
Gastric F	1.7908e-07
Prostate A	0.042597
Prostate B	0
Prostate C	0
Prostate D	0
Arrhythmia A	0.74462
Arrhythmia C	0.013176
Arcene A	1
Arcene C	0

Table 9. P-values originated from comparing normalized to non-normalized datasets in terms of classification error (assuming strictly simple z-score based expansion of RBF kernel vs. RBF kernel)

Dataset	Normalized vs. non-normalized	Non-normalized vs. normalized
Melanoma AL	0	0
Melanoma BL	2.609e-06	0
Gastric AL	0.059196	0.00010068
Gastric BL	4.8963e-10	1.9666e-08
Gastric CL	0	1.3545e-05
Prostate AL	0.14135	0
Prostate BL	0	2.1546e-06

The results discussed in this section give us a clear view of the statistical power behind the evaluated classification rates and the significance of the achieved performance boost.

9. Discussion

In this section we try to explain the presented results and the observed connection between generalization error bounds and real experimental evaluation of proposed ranking-based kernels. Finally, we summarize all provided data and conclude with a quick overview of the finished and ongoing work.

We can observe that performance measures on different data sources (see Table 1) almost everywhere attain the same or even better results for ranking-based kernels, bringing some useful discrimination capabilities to the SVM classifier. However for the normalized invariant of microarray datasets[e] the proposed kernels achieve just a slight improvement in a total generalization error and balanced error rate but for non-normalized one it drastically outperforms the RBF kernel, attaining even better results than for the normalized and scaled invariant. The significance of this result can be clearly proven by P-values obtained from the comparison of classification errors for different invariants of evaluated datasets. Poor results obtained for the normalized versions of some microarray datasets lead us to the assumption that even a very slight (and possibly improper) normalization can severely reduce discrimination power and generalization accuracy. The latter does not apply for UCI datasets where ranking-based kernels achieve good results too, whereas all the data were already preprocessed and properly normalized.

It is interesting that z-score transformation may improve accuracy regardless of the underlying kernel and its generalization power. This is one of the most intriguing aspects because of the upper bound on generalization error (6-7) that tends to be lower under rescaled by such transformation kernel-space. This is what we could clearly observe from the results in Table 2 and Table 3. The ranking-based kernel with applied transformation to "z-space" achieves even better

[e] Meaning global cross-sample normalization

accuracy rates for some microarray and UCI datasets than the initial kernel. The same result is clearly seen for the z-score based transformation of the linear kernel (21) that gives necessary extension for the RBF kernel as well.

To prove the more robust nature of the proposed ranking-based kernels, we presented some upper-bounds on generalization error in (23), (24) and (27), (28). These bounds give us a necessary clue and connection to the success of proposed kernels in generalization on unseen data. The key point of every bound is an input dimensionality of evaluated data source that does not depend on distribution or subset of samples being employed in the training set. The latter is a clear advantage to the bound of Bartlett and Mendelson in (7) where we deal with an expectation of the kernel matrix. This expectation could significantly vary from one training set to another, modifying every time the effective upper-bound on generalization error.

The above-mentioned fact totally explains the observed superiority of ranking-based kernels in generalization on very unstable and limited biomedical data.

10. Conclusion

In this paper we proposed the ranking-based kernels that perfectly fit the classification purposes on highly dimensional biomedical datasets and show very comprehensible results in comparison with standard RBF kernels. The major improvement and higher accuracy is observed for non-normalized and even non-scaled microarray datasets. This fact rejects the necessity for additional preprocessing and normalization of data from microarray chips and shortens the elaboration phase of every diagnostic system. In future we are considering further research in the field of new kernels and kernel methods for biomedical purposes and CAD systems.

Acknowledgements

We are very grateful to the anonymous reviewers for their valuable comments that helped us to improve this paper significantly. We would also like to thank Dr. Sharif Guseynov for helpful technical comments and Dr. Roman Erenshteyn and Dr. Fazel Famili for the careful and valuable proofreading. This study was supported by the fellowships from European Social Fund projects "Support for Doctoral Studies at the University of Latvia" and "Support for Doctoral Studies at Riga Technical University".

Appendix A. Proof of the kernel bound

The linear representation of the ranking-based kernel is upper-bounded (14) by the use of two similar arithmetic progressions that represent aligned values of similarity measures (12), (13) for the sample x. The resulting upper bound is given in terms of the following arithmetic progressions:

$$S_m(\Omega_{low}) = \sum_{i=1}^{m} \Omega_{low}\left(x^{(k)}\right) = \sum_{i=1}^{m} \left(\sqrt{\frac{1}{|\tau|} \sum_{j=1}^{m} I\left(\tau_j \le x^{(i)}\right)} \right)^2 = \frac{m+1}{2} \quad \text{(A.1)}$$

$$S_m(\Omega_{up}) = \sum_{i=1}^{m} \Omega_{up}\left(x^{(k)}\right) = \sum_{i=1}^{m} \left(\sqrt{\frac{1}{|\tau|} \sum_{j=1}^{m} I\left(\tau_j > x^{(i)}\right)} \right)^2 = \frac{m-1}{2} \quad \text{(A.2)}$$

In this proof we assume that the cardinality of the vector τ in (8), (9) is equal to the dimension m of the classification problem. For the beginning we infer an exact value of the kernel function for two equal samples x:

$$\begin{aligned} \hat{K}(x,x) &= \sum_{k=1}^{m} \left[\Omega_{low}\left(x^{(k)}\right) + \Omega_{up}\left(x^{(k)}\right) \right] \\ &= S_m(\Omega_{low}) + S_m(\Omega_{up}) \\ &= \frac{m}{2} \cdot \left(\frac{1}{m} + 1 \right) + \frac{m}{2} \cdot \left(0 + \frac{m-1}{m} \right) = m. \end{aligned} \quad \text{(A.3)}$$

To get out of this equation inequality and final upper bound we should consider a random permutation π of two values in (A.1) or (A.2):

$$\begin{aligned} &\pi(S_m) = \{ S_m \mapsto a_1, \ldots, a_{m-t}, \ldots, a_t, \ldots, a_m \}, \\ &\text{s.t. } a \in S_m,\ m \in \mathbb{N},\ t \in \mathbb{N},\ m \ge 2t \end{aligned} \quad \text{(A.4)}$$

After applying π and taking the inner product of the initial arithmetic progression and permuted arithmetic progression, it is obvious that resulted inner products differ only by:

$$a_{m-t} + a_t - 2\sqrt{a_{m-t} \cdot a_t} \ge 0, \quad \text{(A.5)}$$

and any further permutations will only strengthen this inequality. Equation (A.5) completely satisfies the necessary condition for inequality (14) and thus implies bound on the ranking-based kernel. □

Appendix B. Proof of the loose generalization bound

In this appendix we prove the generalization error bound for self-normalized version of the proposed ranking-based kernel presented in Section 5.

We infer the Rademacher complexity with respect to general kernel mapping and similarity measures provided in (12) and (13) as follows:

$$\begin{aligned} R_n(F) &= \frac{2}{n}\sup_{f\in F}\left|\sum_{i=1}^{n}\sigma_i f(X_i)\right| \\ &= \frac{2}{n}\sup_{\|w\|\le 1/\gamma}\mathrm{E}\left[\sum_{i=1}^{n}\sigma_i\langle w,\Phi(X_i)\rangle\right] \\ &\le \frac{2}{n}\sup_{\|w\|\le 1/\gamma}\|w\|\cdot\mathrm{E}\left[\sum_{i=1}^{n}\sigma_i\|\Phi(X_i)\|\right] \\ &\le \frac{2}{n\gamma}\mathrm{E}\left[\sum_{i=1}^{n}\sigma_i\|\Phi(X_i)\|\right] \\ &= \frac{2}{n\gamma}\mathrm{E}\left[\sum_{i=1}^{n}\sigma_i\sqrt{m}\right]. \end{aligned} \tag{B.1}$$

By inserting the last term of the equation into (6) as an effective upper bound on the Rademacher complexity we yield the following generalization bound in (B.2). □

$$P\{Y\neq f(X)\}\le \hat{P}_n\{Y\neq f(X)\}+\frac{1}{n\gamma}\mathrm{E}\left[\sum_{i=1}^{n}\sigma_i\sqrt{m}\right]+\sqrt{\frac{\ln(1/\delta)}{2n}}, \tag{B.2}$$

Appendix C. Proof of the Rademacher complexity bound for original similarity measures

For yielding the necessary upper bound on the Rademacher complexity for original ranking-based similarity measures we follow the basic inference of (B.1) but in the end we formulate the term $\|\Phi(X_i)\|$ by means of proposed upper bounds on topological similarity measures (25), (26) and additivity of these measures in generalized kernel formalization (10):

$$\begin{aligned} \|\Phi(X_i)\| &= \sqrt{\langle\Phi(X_i),\Phi(X_i)\rangle} \\ &= \sqrt{\sum_{j=1}^{m}\left[\Omega_{low}\left(x_i^{(j)}\right)^2+\Omega_{up}\left(x_i^{(j)}\right)^2\right]} \\ &\le \sqrt{\sum_{j=1}^{m}\left[1+\left(\frac{m-1}{m}\right)^2\right]} \\ &= \sqrt{\frac{2m^2-2m+1}{m}}\le\sqrt{2m\text{-}1}. \end{aligned} \tag{C.1}$$

By inserting the $\|\Phi(X_i)\|$ term into the Rademacher complexity bound of (B.1) we get the desired result. □

Appendix D. Proof of the RBF extension for the ranking-based kernel

A standard RBF kernel is given as follows:

$$K_{RBF}(x,y)=\exp\left(-\gamma\cdot\|x-y\|^2\right) \tag{D.1}$$

In the stated above kernel the term $\|x-y\|$ is the actual norm of the distance between two samples x and y while in (11) we calculate the distance between vectors representing ranking-based topological measures.

In (11) the term $C(x,y)$ represents the complete squared distance between samples x and y over two distinct feature spaces presented by "upper-ranking" and "lower-ranking"-based topological measures that should be considered separately in the similar way to (10).

Because of the supplemental nature of topological similarity measures the term $C(x,y)$ completes the definition of the RBF extension for the ranking-based kernel. □

References

1. Y. Yang and B.L. Lu, Protein subcellular multi-localization prediction using a min-max modular support vector machine, *International Journal of Neural Systems*, 20:1 (2010), pp. 13-28.
2. V. Jumutc, P. Zayakin, On the Power of Topological kernel in Microarray-Based Detection of Cancer, *in Proceedings of the 11th International Conference on Intelligent Data Engineering and Automated Learning – IDEAL,* Lecture Notes in Computer Science, Vol.6283, (Paisley, UK, September 1-3, 2010) pp. 70 – 77,
3. A. Rakotomamonjy *et al.*, SimpleMKL. *Journal of Machine Learning Research*, Vol. 9 (2008), pp. 2491-2521.
4. Z. Kalnina *et al.*, Autoantibody profiles as biomarkers for response to therapy and early detection of cancer. *Current Cancer Therapy Reviews*, Vol. 4(2) (2008), pp. 149-156.
5. Z. Kalnina *et al.*, Evaluation of T7 and Lambda phage display systems for survey of autoantibody profiles in cancer patients. *Journal of Immunological Methods*, Vol. 334(1-2) (2008), pp. 37-50.
6. V. Vapnik, *The Nature of Statistical Learning Theory*. (Springer-Verlag, New-York, 1995).
7. S. Sonnenburg *et al.*, Large scale multiple kernel learning. *Journal of Machine Learning Research*, Vol. 7(1) (2006), pp. 1531–1565.
8. G. Lanckriet *et al.*, Learning the Kernel Matrix with Semidefnite Programming. *Journal of Machine Learning*

Research, Vol. 5 (2004), pp. 27-72.
9. F. Bach *et al.*, Multiple kernel learning, conic duality, and the SMO algorithm. *In Proceedings of the 21st International Conference on Machine Learning*. (Montreal, Canada, 2004), pp. 41–48.
10. V. N. Vapnik and A. Y. Chervonenkis. On the uniform convergence of relative frequencies of events to their probabilities. *Theory of Probability and its Applications*, 16 (2) (1971), pp. 264-280.
11. P. L. Bartlett and S. Mendelson, Rademacher and Gaussian complexities: Risk bounds and structural results. Journal of Machine Learning Research, Vol. 3 (2002), pp. 463-482.
12. O. Chapelle *et al.*, Choosing Multiple Parameters for Support Vector Machines. *Machine Learning*, Vol. 46 (2002), pp. 131–159.
13. M. Futschik, Introduction to OLIN package, [http://www.bioconductor.org/packages/2.3/bioc/vignettes/OLIN/inst/doc/OLIN.pdf] (2010).
14. A. Frank, and A. Asuncion. UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science (2010).
15. F. Chu and L. Wang, Applications of Support Vector Machines to cancer classification with microarray data, *International Journal of Neural Systems*, 15:6 (2005), pp. 475-484.
16. D. Glotsos *et al.*, Automated diagnosis of brain tumours astrocytomas using probabilistic neural network clustering and support vector machines, *International Journal of Neural Systems*, 15:1-2 (2005), pp. 1-11.
17. Lin, D.T. and D.C. Pan, D.C., Integrating a Mixed-Feature Model and Multiclass Support Vector Machine for Facial Expression Recognition, *Integrated Computer-Aided Engineering*, 16:1 (2009), pp. 61-74.
18. E.D. Wandekokem, E. Mendel, F. Fabris, M. Valentim, R.J. Batista, F. M. Varejao, and T.W. Rauber, Diagnosing multiple faults in oil rig motor pumps using support vector machine classifier ensembles, *Integrated Computer-Aided Engineering*, 18:1 (2011), pp. 61-74.

3.4 . Sperm associated antigens as targets for cancer immunotherapy: expression pattern and humoral immune response in cancer patients

BASIC STUDY

Sperm-associated Antigens as Targets for Cancer Immunotherapy: Expression Pattern and Humoral Immune Response in Cancer Patients

Karīna Siliņa, Pavel Zayakin,* Zane Kalniņa,* Lāsma Ivanova,* Irēna Meistere,* Edgars Endzeliņš,* Artūrs Ābols,* Aivars Stengrēvics,† Mārcis Leja,† Kristīne Ducena,‡ Viktors Kozirovskis,§ and Aija Linē**

Summary: The identification of novel cancer-related and immunogenic proteins is still a challenge to be faced to improve antigen-specific tumor immunotherapy. The category of so-called cancer-testis (CT) antigens is one of the most perspective groups of proteins for anticancer immune response activation as normally they are expressed in immunoprivileged tissues and are immunogenic if aberrantly generated in tumors. The heterogeneous group of proteins called sperm-associated antigens (SPAG) might encompass novel CT antigens owing to their common expression in male germ cells, their ability to elicit immune response underlying infertility, and lately proposed oncogenic properties. We carried out a comprehensive analysis of the expression pattern in various normal and cancerous tissues and assessed the frequency of spontaneous humoral immune response against members of the SPAG group in cancer patients using phage-displayed antigen microarrays. Our results show that out of 15 analyzed SPAG genes only SPAG1, SPAG6, SPAG8, SPAG15, and SPAG17 are predominantly expressed in testis, whereas the others are ubiquitously expressed with only a testis-associated alternative splice variant of SPAG16. mRNA expression of SPAG1, SPAG6, and alternative splice variants of SPAG8, SPAG16, and SPAG17 was elevated in various tumors with frequencies ranging from approximately 10% to 70%. The upregulation of SPAG6 in lung and breast cancer was confirmed by immunohistochemical analysis of tumor and normal tissue microarrays. Cancer-associated spontaneous humoral immune response was detected against SPAG1, SPAG6, SPAG8, and a novel testis-specific splice variant of SPAG17 and ascribe these as novel CT antigens that potentially are applicable as immunotherapeutic targets and serologic biomarkers.

Received for publication July 9, 2010; accepted August 23, 2010.

From the *Latvian Biomedical Research and Study Centre; †Riga Eastern Clinical University Hospital, Oncology Centre; ‡Pauls Stradins Clinical University Hospital, Department of Endocrinology; and §Pauls Stradins Clinical University Hospital, Clinic of Oncology, Riga, Latvia.

Supported in parts by funds from Latvian Council of Science, grant numbers 09.1288 and 09.1310, Latvian State Research Program "Development of novel drugs: construction, transport forms and mechanisms of action", European 6th FP project ENACT, contract number LSHC-CT-2004-503306, and European Social Fund project number 1DP/1.1.2.1.2/09/IPIA/VIAA/004. The publishing costs are covered by ERDF project number 2DP/2.1.1.2.0/10/APIA/VIAA/004.

All authors have declared that there are no financial conflicts of interest in regards to this work.

Reprints: Karīna Siliņa, Latvian Biomedical Research and Study Centre, Ratsupites 1, Riga, LV-1067, Latvia (e-mail: karina@biomed.lu.lv).

Supplemental Digital Content is available for this article. Direct URL citations appear in the printed text and are provided in the HTML and PDF versions of this article on the Journal's website, www.immunotherapy-journal.com.

Key Words: sperm-associated antigen, CT antigen, CT-spliced antigen, antigen microarray, cancer immunotherapy

(*J Immunother* 2011;34:28–44)

Tumor antigen-specific immunotherapy holds a great potential to eradicate cancer in a highly specific and nontoxic manner, and allows direct monitoring of immune response, which is crucial for understanding immunologic mechanisms that underlie tumor regression in vivo.[1] However, the considerably low efficacy in clinical trials has hampered the translation of such therapy approaches into clinical practice and is caused by the selection of tumor escape variants through inefficient antigen presentation, the loss of antigen expression and the immunosuppressive activities of tumor, its stroma and/or innate, and adaptive tolerogenic immune cells.[2,3] Attempts have been made to overcome the above hurdles by combining antigen targeting with blocking of tolerance, adding adjuvants, altering ways of antigen delivery, TCR-based gene therapy etc., but there is still a risk of tumor escape through down-regulation of target antigen expression in case the antigen is not necessary for tumor survival and if polyclonal T-cell activation by epitope spreading was not achieved.[4] Hence, the need to develop polyvalent antigen targeting approaches as one of the means to avoid the selection of tumor escape variants, and to cover the heterogeneity of tumors, and the need to determine the protective antigens in tumors that do not express the currently exploited targets like WT1, MUC1, Her-2/neu, NY-ESO-1, and others[5] puts the discovery of novel tumor antigens in the frontline of successful antigen-specific immunotherapy development.

Cancer-testis (CT) antigens are perspective candidates for cancer immunotherapy, as naturally they are expressed in immunoprivileged tissues, but are detected in various neoplastic lesions, and they can induce spontaneous immune responses when aberrantly produced in tumors as there is no or weak central tolerance against proteins restricted to immunoprivileged sites. Emerging evidence suggests that CT antigens may possess functions related to stemness owing to their expression during germ cell and embryonic development, hence, providing space for crucial oncogenic properties in cancer cells.[6] One of the most successful antigen-specific immunotherapy targets to date is the CT antigen NY-ESO-1[5]; its targeting has shown to underlie tumor regression and induce integrated immune system activation involving both, cellular and humoral responses.[7–10]

During the last 2 decades, a group of proteins called sperm-associated antigens (SPAG) has grown to reach 15 members—SPAG1,[11] SPAG2/UAP1,[12] SPAG4,[13] SPAG5,[14–16] SPAG6,[17] SPAG7,[18] SPAG8,[19] SPAG9,[20] SPAG10/MFGE8,[21] SPAG11B,[22] SPAG12/NHP2L1,[23] SPAG13/SSFA2,[24] SPAG15/SPAM1,[25] SPAG16,[26] and SPAG17.[27] For further simplicity in this report, we use the names of the SPAG group even if they differ from the official gene name. These are functionally and evolutionary distinct proteins with a common expression in testis or sperm, and humoral immune response against some have been shown to underlie infertility. Several of them have been candidates for treating infertility[23] and for the development of immunocontraception.[28] Lately also, a role in tumorigenesis has been ascribed to several SPAG proteins, for example, SPAG1 has been shown to be a progression marker of pancreatic cancer and promote cell motility.[29] Decreased expression of SPAG5, a mitotic spindle protein, also known as astrin, is associated with good prognosis for estrogen receptor positive breast cancer patients as determined by cDNA microarray data analysis.[30] In our earlier study of large-scale identification of humoral tumor antigens, we found SPAG8 to react with sera from melanoma patients,[31] and it has been shown to be overexpressed in cervical carcinoma.[32] SPAG9 or JLP has been reported to facilitate migration and invasiveness of renal cell carcinoma[33] and proposed as an early marker for breast[34] and cervical[35] cancers. Besides, it was shown to be an AML[36] and epithelial ovarian cancer-specific antigen and was suggested for cancer immunotherapy applications,[37] and together with SPAG4 are designated as CT genes by the CT Database.[38] The vastly studied SPAG10 also known as lactadherin or the milk fat globule protein MFGE8 was identified as a breast cancer-specific antigen and was considered as a perspective serologic diagnostic marker and a radioimunotherapy target.[39,40] Elevated level of SPAG15 or testis-specific hyaluronidase PH-20 might contribute to the metastatic potential of breast and laryngeal cancer.[41–45] Besides, several SPAG proteins have been related in one way or another to the primary cilium, which is a crucial regulator of Hedgehog, PDGFaa, and WNT signaling pathways that can underlie various aspects of tumorigenesis.[46,47] Noteworthy, several of the SPAG proteins have been shown to be located on the sperm surface, which mediates infertility in patients with corresponding sperm agglutinating antibodies.[11,19] If SPAG proteins were also present on the tumor cell surface, novel-specific antibody therapeutics could be developed. Hence, SPAGs might represent a novel group of CT antigens with functional implications in cancer formation and might have a potential to be novel targets for tumor immunotherapy. However, many SPAG proteins have been analyzed in a narrow focused way concerning infertility. The aim of this study was to gain view on the expression profile and the frequency of spontaneous humoral immune responses against SPAG proteins by using phage-displayed antigen microarrays to determine if any member of the SPAG group might fulfill the requirements for a novel tumor immunotherapy target.

MATERIALS AND METHODS

Patient Material, Cell Lines

Tumor and adjacent normal tissue specimens were obtained from operation material of gastric, colon, melanoma, and breast cancer patients undergoing surgery in Latvian Oncology Centre and stored in RNALater (Applied Biosystems). Lung tumor specimens were obtained during standard diagnostic bronchoscopy procedure at Pauls Stradins Clinical University Hospital and stored in RNALater. All patients have signed an informed consent, and the study has been approved by the Central Commission of Medical Ethics of Latvia.

Sera were collected from the same patients whose tissue specimens were collected. Additional serum samples of melanoma, lymphocytic leukemia, gastric, lung, colon, breast, and thyroid cancer patients and healthy individuals with no known history of cancer, infertility, and autoimmune disorders were provided by the Genome Database of the Latvian population and sera from melanoma, gastric, and prostate cancer patients were kindly provided by the Skin Cancer Unit in German Cancer Research Center, the Clinic of Gastroenterology, Hepatology and Infectious Diseases, Otto-von-Guericke University Magdeburg, Norwegian Radium Hospital, and Onyvax Vaccine Therapies Ltd, UK.

RNA Extraction, cDNA Synthesis

Bead-based tissue homogenization was done by using the FastPrep-24 instrument and Lysing Matrix A (MP Biomedicals) in 1 mL of TRIsolution (Applied Biosystems) followed by the extraction of total RNA according to manufacturer's protocol. RNA from melanoma cell lines was kindly provided by Skin Cancer Unit in German Cancer Research Center. RNA of various normal tissues was purchased from Applied Bisystems, USA and Biocat, Germany. Total RNA was treated with DNase I using DNA-*free* DNase treatment and removal reagents (Applied Biosystems). cDNA was synthesized by random hexamer priming from 2 μg of total RNA by using RevertAidTM First StrandcDNA Synthesis Kit (Fermentas, Lithuania) according to manufacturer's instructions.

mRNA Expression Analysis

Qualitative RT-PCR reaction mixtures contained 1 × reaction buffer, 2.5 mM $MgCl_2$, 0.1 μM primers (Table 1), 0.75 U Taq DNA polymerase (Fermentas, Lithuania), and 1/60th of the prepared cDNA. Amplification of all target and reference genes was done at the same cycling conditions (45 s at 94°C, 30 s at 58°C, 45 s at 72°C), except for the number of cycles that was determined individually according to mRNA abundance (Table 1).

Quantitative RT-PCR (qPCR) reactions were done using 1/60th of cDNA reaction mixture, ABSolute Blue SYBR green Low ROX (Thermo Scientific) and ABI7500 sequence detection system (Applied Biosystems). Appropriate primer concentrations were established by serial dilution curves to ensure amplification efficiency over 95% (Table 1). To normalize the expression data a normalization factor was calculated for each cDNA from the expression values of the 3 most stable reference genes (ACTB, POLR2A, TUB3A) determined among 7 most often used housekeeping genes (GAPDH, ACTB, POLR2A, TUB3A, TBP, YWHAZ, PGK1) by using geNorm software.[48] All reactions were carried out in duplicates.

The statistical analysis of the expression data from tumor and normal samples was done using the nonparametric Mann-Whitney *U* test.

TABLE 1. Primers and Cycling Conditions Used in mRNA Expression Analyses

Gene	Forward Primer, 5′-3′	Reverse Primer, 5′-3′	RT-PCR Cycles	qPCR Ratio F/R (nM)
SPAG1	GAAAAGCATCTTCAAGCCTTGG	GGAGGTCAAGCACCAAGTTTG	35	100/100
SPAG4	TGGGTCTCCAGTAGTCTCTGA	TCCTCTGCACGACCAGTCG	35	—
SPAG5	CGCAGAGCAGGTTCAAACAC	GGAGAGGCACTTGAATGGGA	38	—
SPAG6	AGCAATGGCAGTCATCATTTC	GGATGAATGGTCGGGAACTT	35	50/100
SPAG8	CAAGCATGCAGGATGGCTCT	ATGGCTTCACGCTTCCCTCG	35	100/100
SPAG8-e2L	CTACAACTGGGAGGAAGAGAG	GTGGCTGGTACGAGTCTTTC	35	100/100
SPAG9	AGACCCGAGTGGAATCTTTAG	GTTGATCACTCCCTGAGAGC	35	—
SPAG9-C	CTCATACCAGCCTGAAGGTC	CCATCGGGTCCTTTGATCTT	35	100/50
SPAG11B	CACCCAGCCTCACTCCATC	CACTTTGCCCTGGAGAATGG	35	—
SPAG12	CAAGAAGCTACTGGACCTCG	GATGCACAAACACGCGAGG	35	—
SPAG13	GCACCACTGACAATACCATCC	CACGACTATCAACACTGTCACT	35	—
SPAG15	GTTGCTCTGGGTGCTTCTG	GGTCCTCGTTCCTCACACA	35	100/50
SPAG16	CGGAAAACAGTTCTTCCTTC	AGACTGAAAGCAATCAAGAG	35	100/100
SPAG16-L	CTGTCTATATGGGATGCAAGAAC	GACCGTTACTCCACTTAAAACTA	35	50/100
SPAG17	AACAGAAATCCTCAAGTGTGC	TGTGTTCACTTTTCCTCCAAC	35	100/100
SPAG17-A	GGAACATTGCTCTCCACTCC	GCTAATCGTCTTCTCCTCGC	35	50/100
SPAG17-A1	CAACATGAGTCTCTGGGTAA	GCTAATCGTCTTCTCCTCGC	38	100/100
SSX2	GCTCAAATACCAGAGAAGATCC	GTGGCCTTGAAACCTAGTTTAG	35	100/50
ACTB	AATCTCATCTTGTTTTCTGCGC	AGTGTGACGTGGACATCCG	25	100/100
PolR2A	GGGTCATCTTCCCAACTGGAG	CACCAGCTTCTTGCTCAATTCC	—	100/100
TUBA3	TATGGCAAGAAGTCCAAGCTG	TACCATGAAGGCACAATCAGAG	—	100/100

Immunohistochemical Analysis of Tissue Microarrays

Paraffin-embedded tissue microarrays (TMA) of various normal tissues (duplicated 45 tissues, A700 (III), AccuMax Array triplicated 33 tissues, FDA994, US Biomax, Inc.) and melanoma, gastric (ME481t and ST805t, US Biomax, Inc.), lung, and breast [A716 (III) and A712 (13), Accu Max Array] cancers were used for standard immunohistochemistry analysis including melanin bleaching procedure (Melanin Bleach Kit, Polysciences, Inc.) for normal tissue and melanoma TMAs according to manufacturers protocol. In brief, after standard deparaffination and hydration, TMAs were incubated in melanin bleaching solutions, rinsed, and continued with peroxidase quenching. Antigen retrieval was done by heating the TMA in 0.01-M citric acid, pH 6.0 (SIGMA-Aldrich, Germany) for 15 minutes. Commercial primary antibodies for SPAG6 (mouse monoclonal antibody, SantaCruz Inc), and for SPAG8 (rabbit polyclonal antibody, Proteintech Group, Inc.) were used at the dilution 1:50 and incubated overnight at +4°C. Secondary antibodies (HRP conjugated antimouse IgG antibody and antirabbit IgG antibody, SIGMA-Aldrich, Germany) were used at the dilution 1:100 for 1 hour at +37°C. DAB (SIGMA-Aldrich, Germany) was used for color development and counterstained with hematoxylin (SIGMA-Aldrich, Germany).

Analysis of Autoantibody Responses

The antigenic regions of SPAG proteins were predicted using algorithms developed by Welling et al[49] for SPAG1, SPAG6, SPAG8, SPAG9, SPAG16, and SPAG17, which, together with full CDS regions of well-known CT antigens NY-ESO-1 (CTAG1B), MAGEA1, HORMAD1, and SSX2 were amplified using high-fidelity PCR enzyme mix (Fermentas, Lithuania) from testis cDNA. Several different transcripts were obtained for SPAG6 (designated as SPAG6-A, B, and A1), SPAG9 (designated as SPAG9-A and C), and SPAG17 (designated as SPAG17-A and A1). All fragments were cloned into T7Select 10-3b Phage Display vector (Novagen) (Table, Supplemental Digital Content 1, http://links.lww.com/JIT/A84 for cloned antigens with corresponding amino acid positions). StrepII tag (Trp-Ser-His-Pro-Gln-Phe-Glu-Lys) was inserted into HindIII and NotI sites located downstream the cDNA cloning site of the vector DNA to monitor the copy number of the recombinant proteins on each phage. Obtained recombinant phages together with 9 nonrecombinant control phages were amplified, purified, and printed in triplicate on nitrocellulose slides (Whatman, GE Healthcare) to create an antigen microarray, which was screened with sera from 39 breast, 24 lung, 33 colon, 28 thyroid, 172 gastric (stage I—33, II—20, III—31, IV—37, not determined—53), 52 prostate cancer, 163 melanoma (stage I—22, II—23, III—22, IV—22, not determined—74), and 28 lymphocytic leukemia patients and 127 patients of gastrointestinal inflammatory diseases (gastric ulcer 17, duodenum ulcer—20, gastritis and duodenitis—10, acute hemorrhagic gastritis—13, chronic atrophic gastritis—48, dyspepsia—11, Crohn disease—8) and 147 healthy donors. The production and processing of antigen microarrays were done as described earlier.[31] In brief, the microarray slides were incubated with patients' sera and mouse anti-T7 phage tail antibody (Novagen). The serum reactivity was detected by Cy5 conjugated goat anti-human IgG antibody (Jackson ImmunoResearch) and the ratio against the total amount of printed phage, which was detected by Cy3 conjugated goat anti-mouse IgG antibody (Jackson ImmunoResearch) was calculated. A 2-step normalization strategy was used for the fluorescent signal ratios to eliminate variations introduced by custom production of microarrays and variable background intensities of different sera. At first, the values in each slide (each serum) were normalized by the median of all printed spots for each fluorescent channel separately. Next, the Cy5 and Cy3 signal intensities for each spot were divided by the median value of that spot within the print lot. The threshold value of a specific antibody response for each antigen was defined as 4 standard deviations above the

TABLE 2. Summary of Published Literature and Open Source Expression Data

SPAG Gene	Official Name	Evidence of Functions	Immunogencity	Expression Evidence	
				Normal Tissues*	Tumors†
SPAG1	SPAG1	Signaling from G protein-coupled receptors during spermatogenesis and fertilization[59,60] through PKC dependent EKR activation[61]; mediates maternal mtDNA inheritance,[62,63] promotes cell motility[29]	Infertility related[11,59]	Testis-selective (also in gastrointestinal tract, pancreas, tonsils, lung, skin, liver, kidney)[11,29,59,64]; colon, appendix[57]; tracheal, bronchial and nasal epithelial cells, colon[55]	Upregulated in pancreatic cancer,[29] seminomas,[65] renal, breast and other cancers,[53] colon cancer[55]
SPAG2	UAP1	Structural element of sperm axoneme in outer dense fiber[66]; UDP-N-acetylglucose-aminepyrophosphorylase activity[67]	Infertility related[12]	Testis-selective (also in muscle and liver)[12]; smooth muscle[57]; ubiquitous[51]	Upregulated in prostate and renal cancers,[51] prostate cancer, lymphoma and other cancers[53]
SPAG4	SPAG4	Role in protein localization to sperm axoneme and outer dense fibers[13,68]	—	Testis-specific[68]; testis and pancreas (little in, pituary, ovary, duodenum, lymph node)[69]; testis, and pancreatic islets[57]	Upregulated in liver, prostate, breast etc. cancer cell lines,[69] renal, lung, and other cancers[53]
SPAG5	SPAG5	Interacts with sperm axoneme and outer dense fiber structural proteins[14]; mitotic progression, centrosome integrity[70–74] and embryonic development of testis[75]	—	Testis-selective (also in thymus, pancreas, liver)[15]; testis-specific[14]; ubiquitous[51,70]	Upregulated in lung, bladder and other cancers[53]
SPAG6	SPAG6	Sperm axoneme integrity, element of flagellum central apparatus[17]; sperm motility, component of ependymal cilia,[76] ciliogenesis in bronchial epithelium cells[77]	Infertility related[17]	Testis-selective (also in lung)[17,57]; testis, oviduct, bronchial, tracheal, nasal epithelial cells[55]	Upregulated in bone marrow of AML patients,[78] prostate, colon and other cancers,[53] spinal cord neoplasm[55]
SPAG7	SPAG7	Possible structural element of sperm acrosome[18]	—	Ubiquitous[51]	Upregulated in synovial sarcoma,[79] brain, prostate and other cancers[53]
SPAG8	SPAG8	Element of sperm acrosome[19,80,81]; germ cell differentiation[82]; cell division regulation during spermatogenesis[83]	Infertility related[19,81,84]; cancer related[31]	Testis-specific[19,85]; fallopian tube, bronchial, tracheal epithelial cells[55]	Upregulated in cervical carcinoma,[32] lobular breast carcinoma, lung, and other cancers[53]
SPAG9	SPAG9	Positive regulator of JNK p38 MAPK signaling module, shuttling of preorganized signaling complexes[37,86–88]; male germ cell development, fertility, element of sperm acrosome[37,89,90]; cell migration[91]; membrane trafficking[92]; endodermal differentiation of stem cells[93]	Infertility related[20]; autoimmune disorder-related[86,94]; cancer related[33–37]	Testis-specific[20]; testis-selective alternative splice variant (also in fibroblasts, stomach, brain), total SPAG9—ubiquitous[86]	Upregulated in renal cell carcinoma[33], breast cancer,[34] epithelial ovarian cancer,[37] cervical carcinoma,[35] renal, brain and other cancers,[53] thyroid cancer[55]
SPAG10‡	MFGE8	Apoptotic cell clearance, sperm-egg binding, maintenance of epididymal and intestinal epithelia, mammary gland development, promotion of vascularization, exocytosis, facilitation of antigen presenation[95]	Cancer related[21,96,97]	Unbiquitous[98,99]	Upregulated in breast cancer,[21] breast, colon, and other cancers[53]
SPAG11	SPAG11B	Sperm binding and maturation[100–102]; antibacterial activity[103,104]	—	Testis-selective (also in epididymis, prostate, ovary)[22,104,105]	Upregulated in breast cancer, testicular seminoma, and other cancers[53]
SPAG12§	NHP2L1	Fertilization, element of sperm tail, midpiece, and postacrosome[106]; splicesome component[107]	Infertility related[108]	Testis-specific[109]; ubiquitous[107]	Upregulated in renal, bladder and other cancers[53]

TABLE 2. (*continued*)

SPAG Gene	Official Name	Evidence of Functions	Immunogencity	Expression Evidence: Normal Tissues*	Expression Evidence: Tumors†
SPAG13‖	SSFA2	Early cleavage of the fertilized oocyte[24]; actin regulation[110]; signal transduction[111]; energy homeostasis[112]	Infertility related[113]	Ubiquitous[110,111]	Upregulated in colon cancer,[110] brain, esophagus, and other cancers[53]
SPAG15	SPAM1	Sperm-egg adhesion[114,115]; cumulus penetration, hyaluronidase activity[116]; sperm maturation, element of sperm axoneme[117,118]; fluid reabsorption in kidney[119]; epididymosome, and uterosome component[120,121]; intracellular signaling[122]; angiogenesis[123]; increased metastatic potential of breast cancer[42]	Infertility related[124–126]	Testis-specific[25,127]; kidney, macrophages[119,128]; epididymis[129]; vagina, uterus, oviduct,[130] breast[42]; endothelium[131]; testis-selective (also in duodenum, small intestine, colon, bronchus)[51]	Upregulated in breast cancer,[43] laryngeal carcinoma,[44,45] colon cancer, melanoma, and glioblastoma cell lines,[123] lung cancer,[51] brain cancers[53]
SPAG16	SPAG16	Postmeiotic germ cell viability, sperm motility[132,133]; central apparatus element of cilia, and sperm flagella[26,134]	—	Testis-selective (also in prostate, spleen, ovary, thymus)[117]; testis-specific transcript SPAG16-L and testis-selective transcript SPAG16-S (also in trachea, brain, liver, kidney)[134]; fallopian tube, bronchial, tracheal epithelial cells, kidney[55]	Upregulated in ALL and breast and other cancers,[53] breast and urothelial cancer[55]
SPAG17	SPAG17	Sperm motility, central apparatus element of cilia and sperm flagella[135]	—	Testis-selective (also in brain, uterus, oviduct, lung)[135]; bronchial, tracheal epithelial cells[55]	Upregulated in pancreatic, brain, and other cancers[53]

*The major anatomical sites of expression reported by published literature as well as Human Protein Atlas,[51] BioGSP,[57] and Genevestigator[55] tools are named.

†Only the upregulation in cancer is summarized and, besides the literature data, the top two tumor types sorted by overexpression gene rank in the Oncomine database,[53] or by highest relative overexpression in Genevestigator[55] tool are named.

‡Owing to the vast amount of publications on SPAG10 only the review paper is cited.

§SPAG12 together with FA-1 are alternative names of the official gene symbol NHP2L1 in the current version of NCBI Gene database. The initial publications on FA-1 protein indicate its testis-specificity[108] and there is no precise cloned sequence of FA-1 as the previously published one[23] has been discontinued due to failure to prove existing as stated in the NCBI OMIM description of SSFA1—another alternative name for FA-1 (Scott, A. F. Personal Communication. Baltimore, MD, 2.8.2007). Hence it is not clear whether the initially described FA-1 protein is indeed the same as the ubiquitously expressed NHP2L1.

‖SPAG13 together with CS1 are alternative names of SSFA2 in the current version of Entrez Gene database, however, CS1 is also an alternative name of a completely distinct protein—the natural killer cell receptor SLAMF7 and should not be confused here.

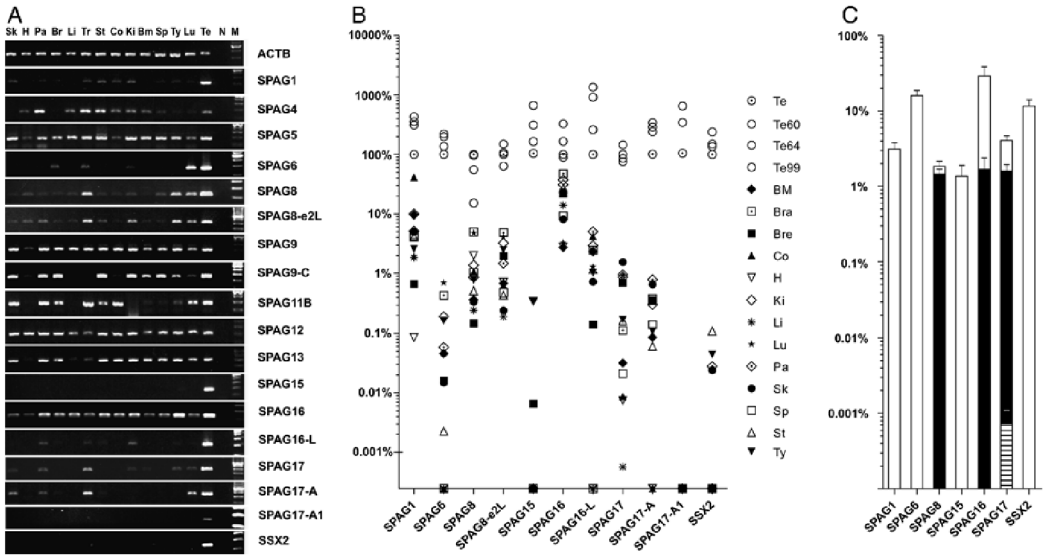

FIGURE 1. The mRNA expression of sperm-associated antigens (SPAG) in various normal tissues. A, RT-PCR was carried out on a set of 14 various normal tissues and the amplification products were visualized in a 1.2% agarose gel. B, qPCR was done on a set of 13 various normal tissues and testis samples from 4 different individuals. Mean quantities are displayed. Y axis represents percentage of the testis expression level in logarithmical scale. C, The average quantity of total SPAG (white bars) and corresponding splice variants obtained from 4 different testis samples is displayed. Black bars—SPAG8-e2L, SPAG16-L, SPAG17-A, striped bar—SPAG17-A1. Y axis represents percentage of the ACTB expression level in logarithmic scale. Error bars represent standard error of the mean. BM indicates bone marrow; Bra, brain; Bre, breast; Co, colon; H, heart; Ki, kidney; Li, liver; Lu, lung; M, size marker; N, no template control; Pa, pancreas; Sk, skin; Sp, spleen; St, stomach; Te, testis (60, 64, 99 designate 3 additional testis samples); Tr, trachea; Ty, thymus.

average of 70 empty phage controls. Statistical significance was calculated using χ^2 test. To validate SPAG17-A1 splice variant-specific antibodies, an antigen array comprising phage particles expressing SPAG17-A, SPAG17-A1, an unrelated-antigen HORMAD1, and empty phages was tested with serial 3-fold dilutions of 2 SPAG17-A1 positive gastric cancer sera. The normalized values were further normalized against StrepII signal detected by anti-Sterp II tag antibody (StrepMAB-Immo, IBA, Germany) to correct for copy number variations of recombinant proteins per phage particle.

RESULTS

Selection of Candidate CT Genes

Some of the SPAG genes have been studied extensively whereas for others only the coding sequence is known, hence, we carried out data mining in Human Protein Atlas,[50,51] Oncomine,[52,53] Genevestigator,[54,55] BioGSP,[56,57] and Entrez Gene[58] databases and the published literature to select CT gene candidates for experimental validation. The criteria for selecting a SPAG gene for further expression analyses were CT-associated expression profile, possible functional implications in oncogenesis and the cell surface localization. The decision was based on the obtained information summarized in Table 2 and included SPAG1, SPAG4, SPAG5, SPAG6, SPAG8, SPAG9, SPAG11B, SPAG15, SPAG16, and SPAG17. SPAG12 and SPAG13 were included to affirm the ubiquitous expression represented by the above mentioned online resources.

All SPAG genes are alternatively spliced as evidenced by the NCBI AceView database.[136] For the expression analysis, primers were designed to amplify a common region of all known transcripts. Additional sets of primers were designed to amplify earlier reported testis-associated splice variants of SPAG9[86] (designated as SPAG9-C), SPAG16[134] (designated as SPAG16-L), SPAG8 isoform earlier found to elicit antibody responses in melanoma patients[31] (designated as SPAG8-e2L), and SPAG17 splice variant containing the predicted antigenic region (designated as SPAG17-A). In addition, we identified novel transcripts during the cloning of the antigenic regions of SPAG6 and SPAG17 and designated them as variations of corresponding transcripts in the AceView nomenclature namely SPAG6-A1 and SPAG17-A1 (Figure, Supplemental Digital Content 2, http://links.lww.com/JIT/A85 for all alternative transcripts, amplicon positions, and cloned antigenic regions).

mRNA Expression Pattern in Normal Tissues

To describe the expression pattern of SPAG genes, we use the terms defined by Hofmann et al[137] where CT genes that are expressed mostly in testis but are detectable in low levels in a few other nongerm line tissues are designated as testis-selective, and the ones present only in germ line tissues and placenta are called testis-restricted. Expression of the selected SPAG genes together with the well-known CT antigen SSX2 as a control of testis-restricted expression (according to the CTDatabase[38]) was first analyzed by qualitative RT-PCR in a panel of 14 different normal tissues (Fig. 1A). SPAG1, SPAG6, SPAG8, SPAG8-e2L,

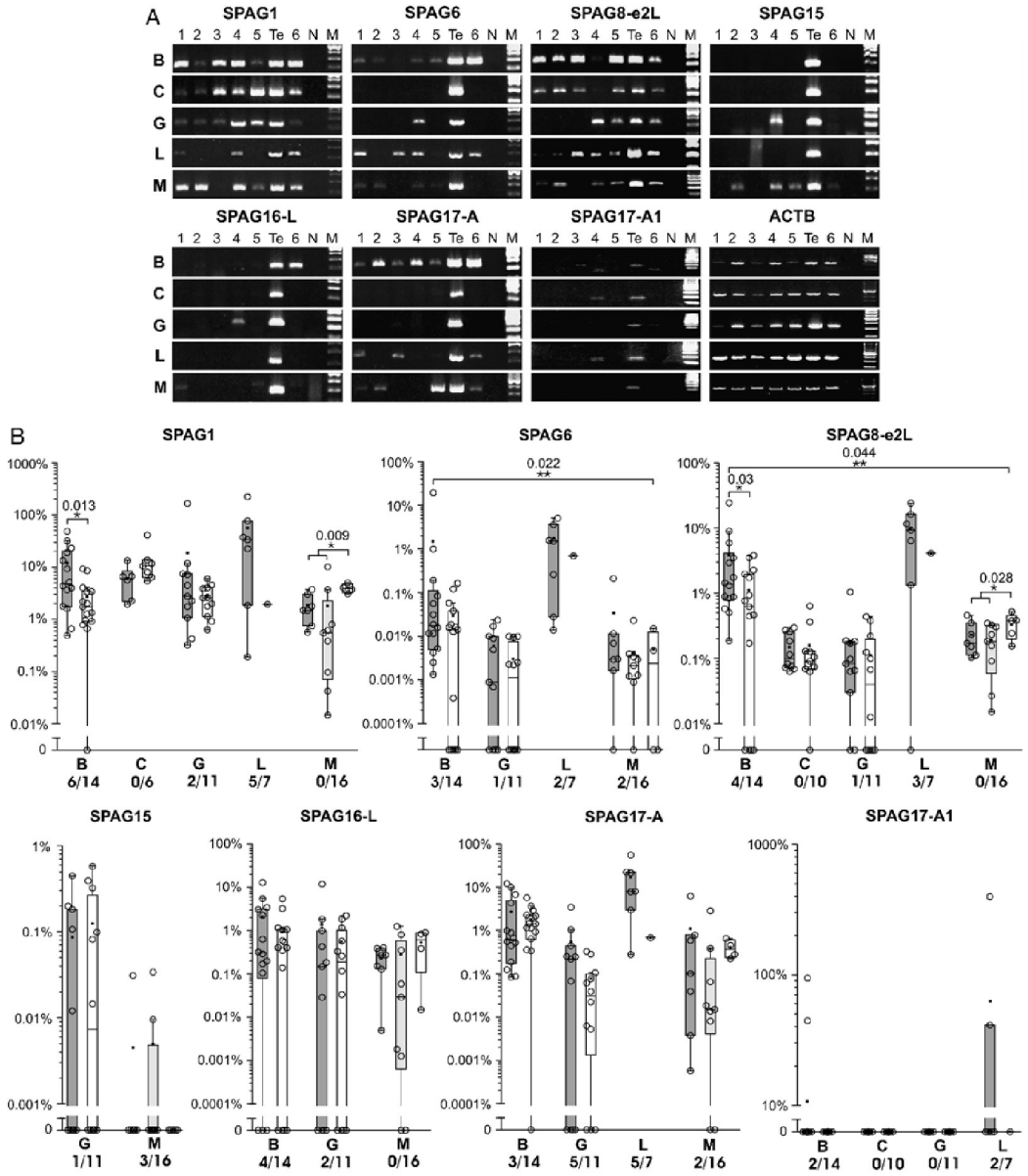

FIGURE 2. The mRNA expression of sperm-associated antigens (SPAG) in various tumors. A, RT-PCR was carried out on sets of 6 breast (B), colon (C), gastric (G), lung (L) cancer, and melanoma (M) tumor samples. Amplification products were visualized in a 1.2% agarose gel. Numbers 1-6 designate tumor tissues of cancer patients, TE-commercial normal testis, N-no template control, M-size marker. B, qPCR was carried out on sets of breast, colon, gastric, lung cancer, and melanoma samples (dark gray boxes), 9 melanoma cell lines (light gray boxes), and adjacent normal tissues and corresponding commercial normal samples (white boxes). Boxes show median with interquartile ranges for each sample set with Tukey whiskers, black dots represent the average value of the sample set, open circles designate expression values of individual tissue samples. Y axis represents percentage of the testis expression level in logarithmic scale. Single asterisk designates statistical significance reached within 1 tumor type, 2 asterisks—across all tumor types. One-tailed normal approximated *P* values, calculated using the nonparametric Mann-Whitney *U* test, are indicated if significant. Numbers below the tumor type designation indicate the number of overexpressed tumor samples out of all analyzed.

SPAG16-L, SPAG17, and SPAG17-A showed testis-selective expression, and SPAG15 and SPAG17-A1 seemed to be testis-restricted. These were further validated by qPCR in a cDNA set composed of 13 different normal tissues and testis samples from 4 different individuals. Testis-selectivity was confirmed for all analyzed SPAG transcripts (Fig. 1B), however, SPAG1 showed considerably high expression in normal colon. A minute amount of SPAG15 was observed in thymus, heart, and breast and renders it testis-selective. SPAG17-A1 was approved as testis-restricted. The remaining SPAG4, SPAG5, SPAG9, SPAG9-C, SPAG11B, SPAG12, SPAG13, and the common region of SPAG16 were present in many normal tissues in comparable levels with that of the testis sample (Fig. 1A) and this observation was confirmed by qPCR for SPAG9-C (data not shown) and SPAG16 (Fig. 1B). The objective expression level of testis-selective SPAG gene transcripts was determined in 4 different testis samples (Fig. 1C) by comparing it with the level of ACTB. It is shown that the most prominently expressed SPAG gene in testis is SPAG16 reaching about 30% of the ACTB level, whereas the testis-selective transcript SPAG16-L comprises around 6% of the total SPAG16. The transcripts containing the predicted antigenic regions of SPAG8 (SPAG8-e2L) and SPAG17 (SPAG17-A) represent approximately 80% and approximately 40% of the corresponding total mRNA, respectively, whereas the testis-restricted transcript SPAG17-A1 represents only 0.02% of the total SPAG17.

mRNA Expression in Tumor Tissues

Next, we carried out a qualitative RT-PCR screening experiment in melanoma, gastric, colon, lung, and breast cancer specimens from 6 patients in each tumor type (Fig. 2A) to determine whether any of the testis-associated SPAG transcripts are present in tumors. It is evident that the mRNA pattern varies in different cancer types among SPAGs, SPAG1, and SPAG8-e2L are the most frequent, detected in about 80% of samples. SPAG6 and SPAG17-A can be detected in about 50%, and SPAG15, SPAG16-L, and the testis-restricted transcript SPAG17-A1 are the least frequent transcripts detected in 10% to 20% of analyzed tumor samples.

To confirm upregulation of the testis-associated SPAGs in the respective tumor types, their expression was examined by quantitative RT-PCR in larger panels of tumor and adjacent normal tissue pairs. The expression level in tumor samples was compared with the adjacent normal tissue and to the average signal detected in all corresponding normal samples including the commercial normal sample to account for the interindividual expression variations and for cases when the adjacent normal tissue was not available (such as for lung cancer biopsies and melanoma cell lines). Expression of a SPAG gene was considered upregulated in a tumor sample if it exceeded that of the adjacent normal sample by at least 2-fold and the average amount in all normal samples by at least 3-fold. We could detect overexpression of all SPAG genes in tumor samples of various tumor types, however the level of SPAG15 in overexpressed tumor samples is below 1% of the testis level (Fig. 2B). Overexpression in 12.5% melanoma samples was found only for SPAG6 and SPAG17-A, whereas no SPAG gene expression was upregulated in colon cancer sample set (Fig. 2B). All SPAG genes analyzed in breast cancer showed elevated expression level with frequencies varying from 14% (SPAG17-A1) to 43% (SPAG1), and overexpression in gastric cancer was noted for all except SPAG17-A1 from 9% (SPAG6, SPAG8-e2L) to 45% (SPAG17-A) of samples (Fig. 2B). The most prominent upregulation of SPAG genes is seen in lung cancer ranging from approximately 30% (SPAG6, SPAG17-A1) to 70% (SPAG1, SPAG17-A) (Fig. 2B).

Protein Expression Analysis

To evaluate the expression of testis-selective SPAG genes at the protein level, we used immunohistochemistry on tissue microarrays comprising various normal tissues and paired breast and lung tumor-normal and unpaired gastric tumors and normal stomach tissues. The choice of analyzed proteins was limited to the availability of commercial antibodies. SPAG6 was not detected in any of the 45 normal tissues represented on TMAs (from 2 different commercial providers in 2 repeated experiments), including lung and spermatozoa (Fig. 3A). However, its expression was observed in 7 out of 12 breast cancers and 1 adjacent normal breast specimen (Fig. 3B, normal tissue panel). All positive tumor samples showed distinct perinucleolar staining (Fig. 3B, arrows in the tumor images) and mostly weak nuclear staining with strong signals in 2 out of 7 tumors (Fig. 3B). A more prominent signal was observed in 11 out of 12 lung cancer specimens strongly staining either nucleus alone or also cytoplasm and little amount in all adjacent normal lung samples with exclusively nuclear localization (Fig. 3C) just as in the adjacent normal breast sample. The positive tumors could be divided by the proportion of the positive cells, 8 out of 11 samples showing around or less than 50% and 3 out of 11 staining close to 100% of cells (Fig. 3C), however, no correlation of tumor stage or metastatic status was noted.

SPAG8 was observed only in discrete cells in the parabasal layer of ectocervix (Fig. 3D), acinar, and ductal epithelium of the breast (Fig. 3E, normal tissue panel) and stomach fundus glands (Fig. 3F), whereas all other normal tissues including testis, skin, and other stratified squamous cell epithelia such as that of esophagus or larynx were negative (data not shown). The subcellular localization is mostly cytoplasmic, but in the cells of fundus glands also the nucleus is stained and in ectocervix also possibly the plasma membrane. Its expression was also detected in 60% of breast tumors (7 out of 12) with a strong staining characteristic to nonmetastatic lesions (3 out of 3), whereas metastatic breast carcinomas (assessed by the presence of dissemination to lymph nodes) were less prominently stained (4 out of 9) and showed a possible surface localization in some cells (2 out of 4 positive cases) (Fig. 3D, arrows in tumor panel pictures). SPAG8 in gastric cancer (Fig. 3E) was only rarely detected by a strong signal and a decreased frequency of positive cases was noted with advancing stage: from 55% of stage II (11 out of 20) to 44% of stage III (8 out of 18). All inflammatory, metaplastic, and dysplastic samples and a single stage I case were positive and the single stage IV case and 10 lymph node metastases were negative. Surface localization was suspected in 1 case (gastric adenocarcinoma, stage II) (Fig. 3E, arrow in tumor image). We observed no SPAG8 protein expression neither in 40 malignant melanoma tissues nor in 8 unpaired normal skin samples (data not shown).

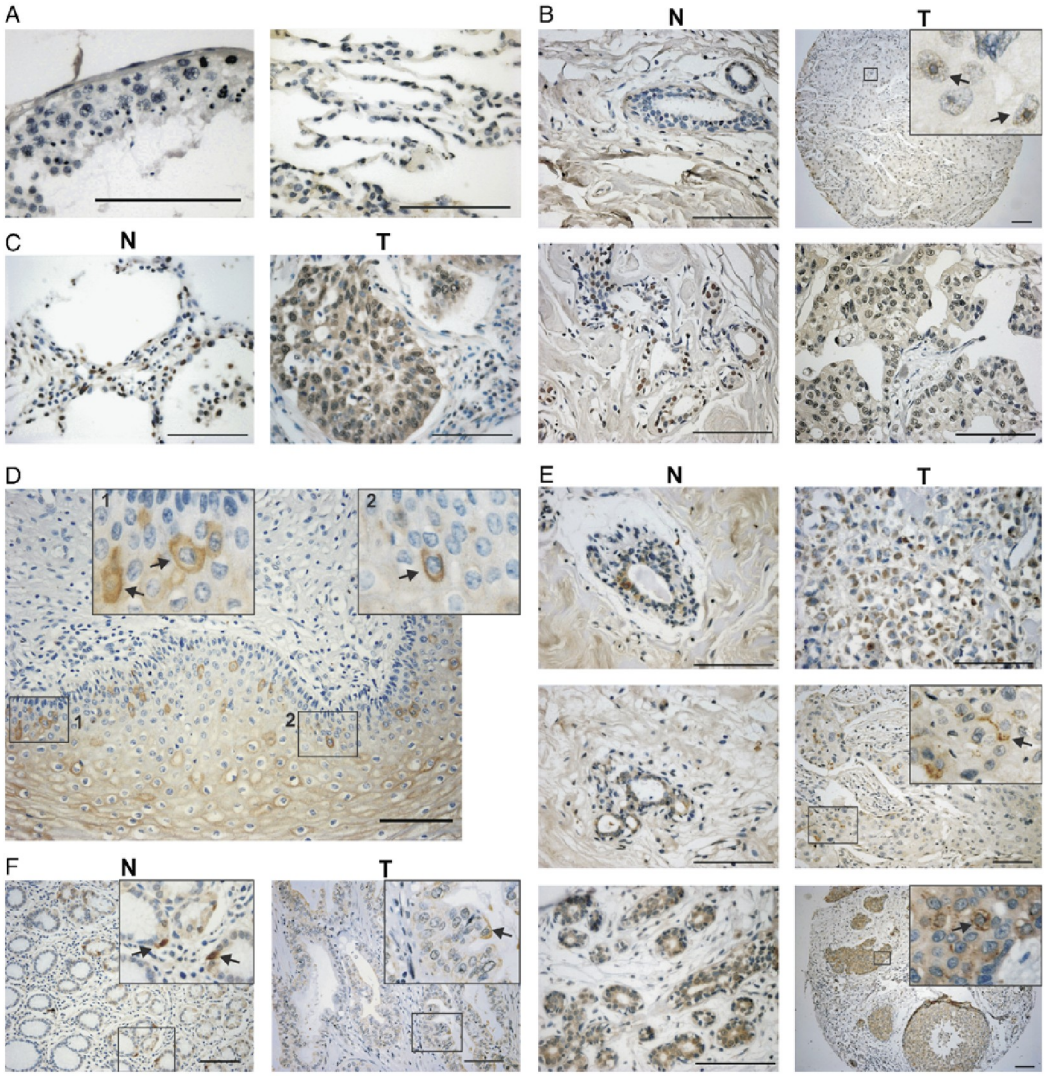

FIGURE 3. Protein expression of sperm-associated antigens (SPAG) 6 and 8. Immunohistochemical (IHC) analysis was done on paraffin-embedded tissue microarrays using anti-SPAG6 and anti-SPAG8 antibodies, HRP-conjugated secondary antibodies, DAB color development, and hematoxylin counterstaining. Black lines designate scale bars of 10 μm. A, SPAG6 IHC analysis in normal testis (left) and normal lung (right). B, SPAG6 IHC staining in normal breast (N) and paired tumor (T) tissues; arrows indicate cells with perinucleolar staining. C, SPAG6 IHC staining in normal lung (N) and paired tumor (T) tissues. D, SPAG8 IHC staining in normal stratified squamous cell epithelium of ectocervix; arrows indicate discrete strongly stained cells with possible surface staining. E, SPAG8 IHC staining in normal breast (N), and paired tumor (T) tissues, arrows indicate tumor cells with possible surface staining. F, SPAG8 staining in normal stomach (N), arrows indicate discrete strongly stained cells, and unpaired gastric tumor (T) tissues, arrows indicate tumor cells with possible surface staining.

Frequency of Autoantibodies

To analyze the immunogenicity of the cancer-associated SPAG proteins and the earlier described cancer serum marker SPAG9-C[33–37] in cancer patients, we determined the frequency of IgG class autoantibodies in sera from 539 cancer patients, 127 patients with inflammatory gastrointestinal disorders and 147 cancer-free individuals by using custom phage-displayed antigen microarray. The frequency of cancer autoantibodies against SPAG1, SPAG6-A and A1, SPAG9-C, and the novel testis-restricted splice variant SPAG17-A1 are comparable with the CT antigen HORMAD1 and higher than against MAGEA1 and SSX2, whereas no sera reacted with SPAG6-B and SPAG17-A transcripts (Table 3, Fig. 4A).

TABLE 3. The Frequency of Reactive Sera Against SPAG Proteins in Cancer Patients, Gastritis Patients, and Healthy Individuals*

Antigen	B (39)	C (33)	G (172)	L (24)	LE (28)	M (163)	P (52)	T (28)	All Ca (539)†	GI (127)	HD (147)
SPAG1	0	3	0	4.2	0	0	0	0	0.4	0	0
SPAG6-A	0	0	0	0	0	0	2	0	0.2	0	0
SPAG6-A1	0	0	0.6	0	0	0.6	0	0	0.4	0	0
SPAG6-B	0	0	0	0	0	0	0	0	0	0	0
SPAG8-e2L	0	0	1.7	0	3.6	3.1	2	3.6	2	3.1	0.7
SPAG9-A	0	0	1.2	0	0	0	0	0	0.4	0	0.7
SPAG9-C	0	0	0	0	0	0.6	0	0	0.2	0	0
SPAG16-L	0	0	0.6	0	3.6	0.6	5.8	0	1.1	3.1	2
SPAG17-A	0	0	0	0	0	0	0	0	0	0	0
SPAG17-A1	0	0	1.7	0	0	0.6	0	0	0.7	0	0
CTAG1B	10	6	7	12.5	7	10.4	9.6	10.7	8.9	0	0.7
HORMAD1	0	0	3.5	0	3.6	0.6	0	0	1.5	1.6	0
MAGEA1	0	0	0	0	0	0.6	0	0	0.2	0	0
SSX2	0	0	0	0	0	0.6	2	0	0.4	0	0

*The number of tested sera is indicated in brackets under the tested tumor type, and the frequency of reacting sera is indicated in percents.
†All Ca—frequency of antibody responses across all tumor types.
B indicates breast cancer; C, colon cancer; Ca, cancer; G, gastric cancer; GI, gastrointestinal inflammatory diseases; HD, healthy donors; L, lung cancer; LE, lymphocytic leukemia; M, melanoma; P, prostate cancer; T, thyroid cancer.

SPAG9-A and SPAG16-L were detected also by sera from healthy donors and gastritis patients in comparable frequencies with the tumor sera suggesting an inflammation rather than cancer-related response (Table 3). The most often recognized was SPAG8-e2L with high titer antibodies (antibody signal over 3) in 1 out of 167 melanoma sera (stage I) and low titer in 4 out of 167 melanoma, 3 out of 172 gastric, 1 out of 28 lymphocytic leukemia, 1 out of 52 prostate cancer, and 1 out of 28 thyroid cancer sera (stages I-III), and 4 out of 127 gastrointestinal disease patients (3 atrophic gastritis and 1 duodenum ulcer) and 1 out of 147 healthy donors (Fig. 4A, Table 3). Cancer-specific autoantibodies against SPAG17-A1 were detected in 3 out of 172 gastric cancer patients (1 stage III and 2 stage IV patients), whereas the major isoform SPAG17-A was very weakly recognized with the corresponding sera antibody signals just below the serum positivity threshold (Fig. 4A). SPAG17-A1 isoform lacks 171 amino acids compared with SPAG17-A (Table and Figure, Supplemental Digital Contents 1 and 2, http://links.lww.com/JIT/A84 and http://links.lww.com/JIT/A85) and hence, might be represented on the phage surface in a more efficient way leading to increased antibody signals. To exclude the possibility that differential serum reactivity against these 2 splice isoforms is owing to variable amount of recombinant protein on the phage surface and to determine the titer of the anti-SPAG17-A1 autoantibodies, we carried out additional screening with serial 3-fold dilutions of 2 positive gastric cancer sera on a separate array containing 5 replicates of SPAG17-A, SPAG17-A1, and irrelevant antigen HORMAD1 and the antibody signal intensities were normalized not only against the total amount of printed phage by anti-T7 phage tail antibody, but also against the copy number of recombinant protein per phage particle by anti-Strep Tag antibody (Materials and Methods). The presence of the anti-SPAG17-A1-specific antibodies was approved and the antibody signals are detected at the sera dilution higher than 1:2700 (Fig. 4B).

DISCUSSION

The specificity and oncogenicity of a potential cancer antigen are the dominant criteria for choosing it for further evaluation of immunogenicity and therapeutic functions as defined in the recent guidelines of prioritizing antigens for immunotherapy.[5] We analyzed online expression databases and published literature to define those SPAG genes that might match the above criteria and to select the candidate CT genes for experimental validation. Five of the SPAG genes (SPAG2, SPAG7, and SPAG10, SPAG12, SPAG13) were presented as ubiquitous by literature data and data deposited in online gene array databases, and this expression pattern was experimentally confirmed for SPAG12 and SPAG13. Qualitative mRNA expression analysis showed that SPAG4 is most abundant in pancreas fitting the earlier published results[69] and several other normal tissues, hence, while being a possible marker of various neoplasias,[69] it cannot be included in the CT gene category. SPAG5 was ubiquitously expressed differing from the earlier reported testis-selective distribution,[15,53] but in concordance with its functional involvement in centrosome integrity[70–74] and the protein expression pattern presented in HPA.[51] Among the most studied is SPAG9 or JLP, a scaffolding protein that is involved in various signaling events along the MAPK[87] and TNF-α or NF-κB[138] pathways and is important for sperm development[89] and retinoic acid-mediated endodermal differentiation.[93] In addition, the overexpression of a distinct testis-associated splice variant of SPAG9 (designated by AceView database[136] as SPAG9-C) is suggested to be important for tumorigenesis and proposed as a tumor immunotherapy target.[33 35,37] We show, however, that, despite the dominant expression in testis, SPAG9-C is present in various normal tissues in comparable levels and it cannot be considered as a CT antigen with implications in immunotherapy.

We continued the expression analyses of testis-selective SPAG genes in various tumor types and showed upregulation in cancer tissues of all except SPAG15. Earlier, SPAG15 has been shown as overexpressed in around 60% of breast cancer cases,[43] as well upregulation is noted in lung cancer by Oncomine,[53] however, no expression was detected in these tumors in our sample set. A larger sample size and/or other tumor types should be analyzed to see whether SPAG15 can be classified as a CT gene. According to the results of our expression analysis

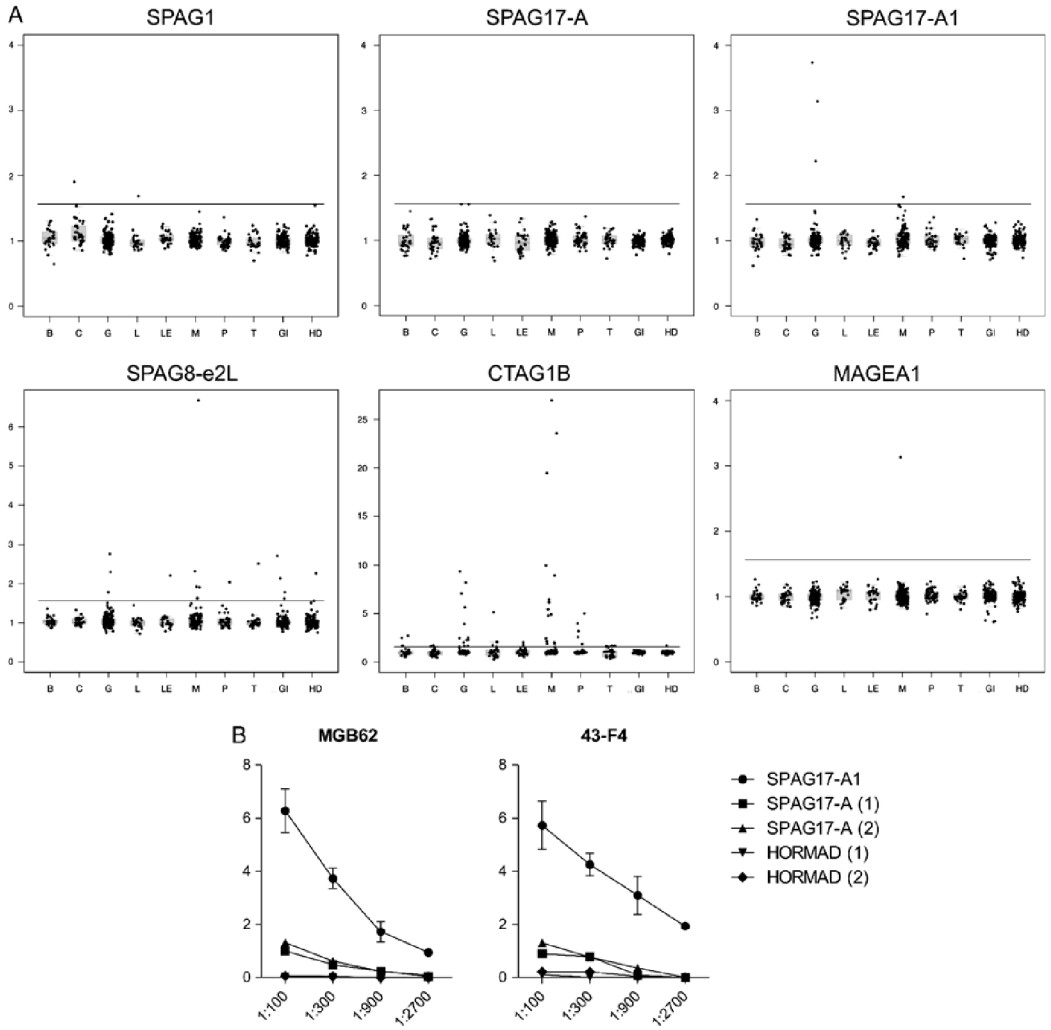

FIGURE 4. Autoantibody responses in sera of various cancer patients, gastritis patients, and healthy donors. A, Autoantibody reactivity against SPAG proteins and the well-known CT antigens was determined by screening phage-displayed antigen microarray with sera from 39 breast (B), 33 colon (C), 172 gastric (G), 24 lung cancer (L), 163 melanoma (M), 28 lymphocytic leukemia (LE), 28 thyroid (T), 52 prostate cancer (P) patients and 127 patients of gastrointestinal inflammatory diseases (GI), and 147 healthy donors (HD). Normalized average values of triplicate recombinant phages are displayed. The serum positivity threshold is defined as 4 standard deviations above the average signal of 70 nonrecombinant phage controls and is designated as a line across the graphs. B, Two gastric cancer sera MGB62 and 43-F4 showing strong reactivity against the novel splice variant SPAG17-A1 and not the dominant isoform SPAG17-A were used in serial 3-fold dilutions for screening of antigen microarray containing these phages and an irrelevant antigen HORMAD1. The obtained signals were normalized against the total amount of printed phage and the amount of recombinant surface protein per each phage. Y axis represents autoantibody signal values, X axis represents serum dilution, error bars designate standard deviations of 5 recombinant phage replicates. Average antibody signals for clone of SPAG17-A1, 2 clones of SPAG17-A, and 2 clones of irrelevant antigen HORMAD1 are displayed.

SPAG1, SPAG6, SPAG8, and SPAG17 genes are the new members of the CT gene category, whereas only distinct isoforms of SPAG16 (SPAG16-L) are testis-selective.

Tumor antigen-targeted immunotherapy trials have shown that the expansion of antigen-specific CTLs does not necessarily correlate with tumor regression,[139] whereas the proper activation of antigen-specific T helper cells has a great potential to underlie tumor control and eradication.[140] The presence of specific class switched autoantibodies in patients' sera against tumor antigens is indicative of spontaneous cell activation against the particular antigen and the detection of such autoantibodies can be used to

monitor the frequency of such responses in cancer patients. To do this, we used the phage-displayed antigen microarray system developed in our lab[31] and created a custom SPAG antigen microarray for serum screening of various cancer patients. Overall, the frequency of autoantibodies against SPAGs is low, yet higher than that of the well-known CT antigens such as MAGEA1 and SSX2. Humoral response against SPAG1, SPAG6, and the novel testis-restricted splice variant SPAG17-A1 was specific to cancer patients, and cancer-associated in the case of SPAG8 ascribing these genes as novel members of the CT antigen group, while anti-SPAG16-L antibodies are equally present in healthy individuals and gastritis patients leaving its CT antigen status in question.

SPAG1, a protein involved in G protein coupled receptor signaling during spermatogenesis and fertilization[59,60] and sperm mtDNA degradation in zygote,[62,63] has been earlier shown to be a progression marker of pancreatic cancer and a cell motility factor.[29] We, for the first time show here that SPAG1 can be immunogenic and is up-regulated prominently in lung and breast cancers, however, relatively high expression in normal colon was also detected, which correlates with gene array results presented by the Genevestigator tool,[55] noting that its application for immunotherapy might not be straightforward.

The fresh appreciation of an organelle present on most mammalian cells—the nonmotile primary cilium—as an important signal transduction hook-up has lead to the notion that centrosome and basal body (the nucleation center of ciliary microtubules) are interexchanging structures that respond to cell cycle regulation, and that cilium is necessary for the proper function of such crucial signaling pathways as the Hegdehog, Wnt and PDGFalfa, disruption of which results in various developmental disorders and cancer.[141] The proteome of primary cilium has now extended to more than 2500 proteins, many of which are shared between basal body and centrosome and are involved in cell cycle checkpoints.[142,143] It is interesting to note that 8 of the 15 SPAG genes: SPAG2, SPAG4, SPAG5, SPAG6, SPAG8, SPAG15, SPAG16, and SPAG17 have been shown to participate in centrosome and/or cilium-related events.

Initially identified as a sperm acrosome protein recognized by sperm agglutinating antibodies and necessary for sperm-egg binding,[19,80] SPAG8 has also been shown to participate in G2/M phase regulation delaying the exit form mitosis when overexpressed in CHO-K1 cells and colocalizes with microtubule organizing center (MTOC) in prophase and spindle microtubules during metaphase.[83] MTOC nucleates microtubules in both, cilium (basal body) and mitotic spindle (centrosome), and it would be of interest to determine whether the disbalanced expression of SPAG8 might participate in ciliary or mitotic defects. A relation of SPAG8 to oncogenesis comes from a study, in which cDNA microarrays showed a 5-fold overexpression in the more aggressive HPV18-type cervical carcinoma when compared with normal cervical epithelium.[32] It is interesting to note that our IHC results showed strong cytoplasmic and possibly membranal staining of distinct cells in the parabasal layer of ectocervix and, considering the frequency of the stained cells, they might correspond to the stem cells of this squamous-stratified epithelium. The basal and parabasal level of ectocervix is thought to be the location of the cervical squamous epithelium stem cells[144] and the HPV-induced cervical carcinogenesis is thought to arise from the deregulation of these stem cells.[145] It would be interesting to see if SPAG8 colocalizes to the same cells as the currently suggested cervical stem cell markers CK17 and p63[145] and whether its MTOC-related activities could be involved in the development of HPV-induced cervical cancer. In addition, we saw the staining of distinct cells in the glandular epithelia of the breast and stomach and the overexpression in the corresponding tumors was seen with a tendency to decrease with tumor stage and was not detected in any of the metastasis samples. In accordance with this observation, the autoantibodies were found mostly in sera of chronic atrophic gastritis that is an early preneoplastic condition, and sera of gastric cancer and melanoma patients with stages I to III. It is tempting to speculate that overexpression of SPAG8 could be advantageous for tumor evolvement in certain contexts but it is lost during tumor progression possibly owing to immune selection. With this in mind, autoantibodies against SPAG8 could be used as an early cancer biomarker; however, targeting of SPAG8 in tumor immunotherapy might be jeopardized. Nevertheless, the possible relation to stem cell functions and surface localization might provide another axis of treatment approaches using this antigen.

First described in *Chlamydomonas flagellum* SPAG6, SPAG16-L, and SPAG17 have been shown to mutually interact at the central apparatus of sperm axoneme—a central duplet of microtubules that is characteristic to the motile cilia, but is absent from nonmotile primary cilia, and are necessary for flagellar motility.[117,135] The expression of SPAG6 has been described in tracheal and bronchial epithelium[77] and detected in other motile cilia-covered epithelia of the respiratory tract as presented by gene array data,[55] in addition, it has been shown as a dynamically exchanging basal body component.[146] A double knockout mouse model of SPAG6 and SPAG16-L showed early mortality of litters owing to severe phenotypes of hydrocephalus and pneumonia indicating to their importance in proper functioning of mucus and fluid-propelling motile cilia, but not polycystic kidneys or left-right axis defects characteristic to nonmotile and nodal ciliopathies.[147] It is not known how these proteins contribute to the motility of mammalian cilia, as the ultrastructure of axonemes in the deficient animals is normal.[147]

Only a few studies have related these proteins to cancer. SPAG6 was reported to be overexpressed in the bone marrow of AML patients and suggested as a marker for minimal residual disease and relapse,[78] whereas SPAG17 has been suggested as a potential candidate gene of thyroid cancer susceptibility.[148] We have shown the overexpression of SPAG6 mRNA in breast and lung cancer specimens and verified this by IHC analysis. We also detected SPAG6 protein in 1 out of 12 adjacent normal breast samples and all adjacent normal lung alveoli samples, while none of these tissues from cancer-free individuals were positive. Whether the presence of SPAG6 in these adjacent normal samples represents a factor of early premalignant transformation or is a particular tumor microenvironment-induced phenomenon, remains to be determined. Taking into account the functional involvement of SPAG6 in the motile cilia of the brain and the respiratory tract, it's targeting in cancer immunotherapy raises caution. Our IHC results showed, however, that the overexpressed SPAG6 protein in breast and lung cancer had perinucleolar, nuclear, and cytoplasmic subcellular localization patterns, indicating to other possible functions

of SPAG6 besides participation in the motile cilium. The observed humoral immune response in cancer patients against transcripts of SPAG6 might also suggest that its ectopic overexpression can be immunogenic. Further studies are warranted to reveal the functional significance of SPAG6 overexpression and to define the molecular alterations underlying its ectopic expression and possibly motile cilium-unrelated activities in cancers, and might result in the identification of epitopes distinct from the normal ciliary SPAG6 providing ground for tumor-associated SPAG6 targeting.

SPAG17 showed testis-selective expression pattern and was upregulated in lung and gastric tumors, whereas its minor splice variant SPAG17-A1 was testis-restricted and present in a portion of lung and breast tumors, but it is not yet clear whether the low-level mRNA of SPAG17-A1 could result in significant quantity of protein; nevertheless, exactly this isoform is recognized by high titer autoantibodies in late-stage gastric cancer sera. The low-antibody signal against SPAG17-A in SPAG17-A1 reactive sera might represent a weak cross-reactivity between these splice variants. Testis is an organ with one of the most diverse transcriptomes owing to vastly rich alternative splicing. We and other researchers have earlier suggested that the deregulation of alternative splicing in cancer can result in the recognition of such testis-restricted splice sites, leading to production of immunogenic isoforms of otherwise tolerated proteins.[149 151] We show here for the first time a testis-restricted splice variant-specific immune response and suggest that such antigens are designated to a separate category called CT-spliced antigens. Further studies of SPAG6 and SPAG17 alternative isoform expression in various tumors and the capability to elicit protective T helper and/or CTL responses in patients bearing SPAG-positive tumors are warranted and ongoing. Furthermore, it would be of interest to find out whether these SPAG proteins might also participate in primary cilium regulated signaling pathways related to oncogenesis if their natural organelle—the motile cilium—is absent.

In conclusion, we have determined the expression pattern of SPAG genes in various normal tissues and showed that only 5 of the 15 genes in the SPAG group are actually testis-selective, although the earlier considered CT genes SPAG4 and SPAG9 are expressed in several normal tissues in comparable levels with the testis. Expression analysis in various tumor types revealed upregulation of SPAG1, SPAG6, SPAG8, the splice variants SPAG16-L, SPAG17-A, and novel testis-restricted alternative splice variant SPAG17-A1. Cancer-related humoral immune response was found against SPAG1, SPAG6, SPAG8, and SPAG17-A1, thus showing these as novel CT antigens that might be useful as cancer serum biomarkers. In addition, transcripts of SPAG6 and SPAG17-A1 are potential candidates for cancer immunotherapy.

ACKNOWLEDGMENTS

The authors thank Dr L. Ņikitina-Zaķe from the Genome Database of the Latvian population, Dr D. Schadendorf from Skin Cancer Unit in German Cancer Research Center, Dr T. Wex from the Clinic of Gastroenterology, Hepatology and Infectious Diseases, Otto-von-Guericke University Magdeburg, Dr G. Gaudernack from Norwegian Radium Hospital and Onyvax Vaccine Therapies Ltd, UK for the serum sample supply; Dr E. Liepiņš and Dr M. Dambrova from the Latvian Institute of Organic Synthesis, and Dr D. Pjanova and Dr R. Brūvere from Latvian Biomedical Research and Study Centre for excellent help in IHC and microscopy.

REFERENCES

1. Jager E, Jager D, Knuth A. Antigen-specific immunotherapy and cancer vaccines. *Int J Cancer*. 2003;106:817–820.
2. Finn OJ. Cancer immunology. *N Engl J Med*. 2008;358:2704–2715.
3. Rosenberg SA. Development of effective immunotherapy for the treatment of patients with cancer. *J Am Coll Surg*. 2004;198:685–696.
4. Reiman JM, Kmieciak M, Manjili MH, et al. Tumor immunoediting and immunosculpting pathways to cancer progression. *Semin Cancer Biol*. 2007;17:275–287.
5. Cheever MA, Allison JP, Ferris AS, et al. The prioritization of cancer antigens: a national cancer institute pilot project for the acceleration of translational research. *Clin Cancer Res*. 2009;15:5323–5337.
6. Simpson AJ, Caballero OL, Jungbluth A, et al. Cancer/testis antigens, gametogenesis and cancer. *Nat Rev Cancer*. 2005;5:615–625.
7. Jager E, Nagata Y, Gnjatic S, et al. Monitoring CD8 T-cell responses to NY-ESO-1: correlation of humoral and cellular immune responses. *Proc Natl Acad Sci U S A*. 2000;97:4760–4765.
8. Jager E, Gnjatic S, Nagata Y, et al. Induction of primary NY-ESO-1 immunity: $CD8^+$ T lymphocyte and antibody responses in peptide-vaccinated patients with NY-ESO-1^+ cancers. *Proc Natl Acad Sci U S A*. 2000;97:12198–12203.
9. Jager E, Chen YT, Drijfhout JW, et al. Simultaneous humoral and cellular immune response against cancer-testis antigen NY-ESO-1: definition of human histocompatibility leukocyte antigen (HLA)-A2-binding peptide epitopes. *J Exp Med*. 1998;187:265–270.
10. Gnjatic S, Atanackovic D, Jager E, et al. Survey of naturally occurring $CD4^+$ T-cell responses against NY-ESO-1 in cancer patients: correlation with antibody responses. *Proc Natl Acad Sci U S A*. 2003;100:8862–8867.
11. Zhang ML, Wang LF, Miao SY, et al. Isolation and sequencing of the cDNA encoding the 75-kD human sperm protein related to infertility. *Chin Med J (Engl)*. 1992;105:998–1003.
12. Diekman AB, Goldberg E. Characterization of a human antigen with sera from infertile patients. *Biol Reprod*. 1994;50:1087–1093.
13. Tarnasky H, Gill D, Murthy S, et al. A novel testis-specific gene, SPAG4, whose product interacts specifically with outer dense fiber protein ODF27, maps to human chromosome 20q11.2. *Cytogenet Cell Genet*. 1998;81:65–67.
14. Shao X, Xue J, van der Hoorn FA. Testicular protein Spag5 has similarity to mitotic spindle protein deepest and binds outer dense fiber protein Odf1. *Mol Reprod Dev*. 2001;59:410–416.
15. Chang MS, Huang CJ, Chen ML, et al. Cloning and characterization of hMAP126, a new member of mitotic spindle-associated proteins. *Biochem Biophys Res Commun*. 2001;287:116–121.
16. Mack GJ, Compton DA. Analysis of mitotic microtubule-associated proteins using mass spectrometry identifies astrin, a spindle-associated protein. *Proc Natl Acad Sci U S A*. 2001;98:14434–14439.
17. Neilson LI, Schneider PA, Van Deerlin PG, et al. cDNA cloning and characterization of a human sperm antigen (SPAG6) with homology to the product of the Chlamydomonas PF16 locus. *Genomics*. 1999;60:272–280.

18. Beaton S, Cleary A, ten Have J, et al. Cloning and characterization of a fox sperm protein FSA-1. *Reprod Fertil Dev*. 1994;6:761–770.
19. Liu QY, Wang LF, Miao SY, et al. Expression and characterization of a novel human sperm membrane protein. *Biol Reprod*. 1996;54:323–330.
20. Shankar S, Mohapatra B, Suri A. Cloning of a novel human testis mRNA specifically expressed in testicular haploid germ cells, having unique palindromic sequences and encoding a leucine zipper dimerization motif. *Biochem Biophys Res Commun*. 1998;243:561–565.
21. Larocca D, Peterson JA, Urrea R, et al. A Mr 46,000 human milk fat globule protein that is highly expressed in human breast tumors contains factor VIII-like domains. *Cancer Res*. 1991;51:4994–4998.
22. Kirchhoff C, Osterhoff C, Habben I, et al. Cloning and analysis of mRNAs expressed specifically in the human epididymis. *Int J Androl*. 1990;13:155–167.
23. Naz RK, Zhu X. Molecular cloning and sequencing of cDNA encoding for human FA-1 antigen. *Mol Reprod Dev*. 2002; 63:256–268.
24. Javed AA, Naz RK. Human cleavage signal-1 protein; cDNA cloning, transcription and immunological analysis. *Gene*. 1992;112:205–211.
25. Lin Y, Kimmel LH, Myles DG, et al. Molecular cloning of the human and monkey sperm surface protein PH-20. *Proc Natl Acad Sci U S A*. 1993;90:10071–10075.
26. Smith EF, Lefebvre PA. PF20 gene product contains WD repeats and localizes to the intermicrotubule bridges in Chlamydomonas flagella. *Mol Biol Cell*. 1997;8:455–467.
27. Rupp G, O'Toole E, Porter ME. The Chlamydomonas PF6 locus encodes a large alanine/proline-rich polypeptide that is required for assembly of a central pair projection and regulates flagellar motility. *Mol Biol Cell*. 2001;12:739–751.
28. Suri A. Sperm specific proteins-potential candidate molecules for fertility control. *Reprod Biol Endocrinol*. 2004;2:10.
29. Neesse A, Gangeswaran R, Luettges J, et al. Sperm-associated antigen 1 is expressed early in pancreatic tumorigenesis and promotes motility of cancer cells. *Oncogene*. 2007;26:1533–1545.
30. Buechler S. Low expression of a few genes indicates good prognosis in estrogen receptor positive breast cancer. *BMC Cancer*. 2009;9:243.
31. Kalnina Z, Silina K, Meistere I, et al. Evaluation of T7 and lambda phage display systems for survey of autoantibody profiles in cancer patients. *J Immunol Methods*. 2008;334:37–50.
32. Vazquez-Ortiz G, Garcia JA, Ciudad CJ, et al. Differentially expressed genes between high-risk human papillomavirus types in human cervical cancer cells. *Int J Gynecol Cancer*. 2007;17:484–491.
33. Garg M, Kanojia D, Khosla A, et al. Sperm-associated antigen 9 is associated with tumor growth, migration, and invasion in renal cell carcinoma. *Cancer Res*. 2008;68: 8240–8248.
34. Kanojia D, Garg M, Gupta S, et al. Sperm-associated antigen 9, a novel biomarker for early detection of breast cancer. *Cancer Epidemiol Biomarkers Prev*. 2009;18:630–639.
35. Garg M, Kanojia D, Salhan S, et al. Sperm-associated antigen 9 is a biomarker for early cervical carcinoma. *Cancer*. 2009; 115:2671–2683.
36. Guinn BA, Bland EA, Lodi U, et al. Humoral detection of leukaemia-associated antigens in presentation acute myeloid leukaemia. *Biochem Biophys Res Commun*. 2005;335: 1293–1304.
37. Garg M, Chaurasiya D, Rana R, et al. Sperm-associated antigen 9, a novel cancer testis antigen, is a potential target for immunotherapy in epithelial ovarian cancer. *Clin Cancer Res*. 2007;13:1421–1428.
38. CTDatabase, [database online]. New York, NY, São Paulo, Petrópolis, RJ: Ludwig Institute for Cancer Research, Weill-Cornell Medical Center, Laboratório Nacional de Computação Científica; 2005. Updated 2009.
39. Ceriani RL, Sasaki M, Sussman H, et al. Circulating human mammary epithelial antigens in breast cancer. *Proc Natl Acad Sci U S A*. 1982;79:5420–5424.
40. Ceriani RL, Blank EW. Experimental therapy of human breast tumors with 131I-labeled monoclonal antibodies prepared against the human milk fat globule. *Cancer Res*. 1988;48:4664–4672.
41. Madan AK, Yu K, Dhurandhar N, et al. Association of hyaluronidase and breast adenocarcinoma invasiveness. *Oncol Rep*. 1999;6:607–609.
42. Beech DJ, Madan AK, Deng N. Expression of PH-20 in normal and neoplastic breast tissue. *J Surg Res*. 2002;103: 203–207.
43. Wang LP, Xu XM, Ning HY, et al. Expression of PH20 in primary and metastatic breast cancer and its pathological significance. *Zhonghua Bing Li Xue Za Zhi*. 2004;33: 320–323.
44. Godin DA, Fitzpatrick PC, Scandurro AB, et al. PH20: a novel tumor marker for laryngeal cancer. *Arch Otolaryngol Head Neck Surg*. 2000;126:402–404.
45. Christopoulos TA, Papageorgakopoulou N, Theocharis DA, et al. Hyaluronidase and CD44 hyaluronan receptor expression in squamous cell laryngeal carcinoma. *Biochim Biophys Acta*. 2006;1760:1039–1045.
46. Michaud EJ, Yoder BK. The primary cilium in cell signaling and cancer. *Cancer Res*. 2006;66:6463–6467.
47. Barakat MT, Scott MP. Tail wags dog: primary cilia and tumorigenesis. *Cancer Cell*. 2009;16:276–277.
48. Vandesompele J, De Preter K, Pattyn F, et al. Accurate normalization of real-time quantitative RT-PCR data by geometric averaging of multiple internal control genes. *Genome Biol*. 2002;3:research0034.1-0034.12
49. Welling GW, Weijer WJ, van der Zee R, et al. Prediction of sequential antigenic regions in proteins. *FEBS Lett*. 1985; 188:215–218.
50. Berglund L, Bjorling E, Oksvold P, et al. A genecentric Human Protein Atlas for expression profiles based on antibodies. *Mol Cell Proteomics*. 2008;7:2019–2027.
51. Human Protein Atlas, Version 6.0 [database online]. Stockholm, Uppsala, Mumbai: Royal Institute of Technology, Rudbeck Laboratory, Lab Surgpath; 2005. Updated March 26, 2010.
52. Rhodes DR, Yu J, Shanker K, et al. ONCOMINE: a cancer microarray database and integrated data-mining platform. *Neoplasia*. 2004;6:1–6.
53. Oncomine Research Edition, Version 4.3 [database online]. Ann Arbor, MI: Compendia Bioscience, Inc; 2008. Updated May 11, 2010.
54. Hruz T, Laule O, Szabo G, et al. Genevestigator v3: a reference expression database for the meta-analysis of transcriptomes. *Adv Bioinformatics*. 2008;2008:420747.
55. Genevestigator, Version 3 Zurich: Gruissem Laboratory, Widmayer Laboratory, ETH Zurich; 2008. Updated 2010.
56. Wu C, Orozco C, Boyer J, et al. BioGPS: an extensible and customizable portal for querying and organizing gene annotation resources. *Genome Biol*. 2009;10:R130.
57. BioGSP, [database online]. San Diego, CA: Genomics Institute of the Novartis Research Foundation; 2009. Updated 2010.
58. Entrez Gene, [database online]. Bethesda, MD: National Center for Biotechnology Information; 1988. Updated 2010.
59. Lin W, Zhou X, Zhang M, et al. Expression and function of the HSD-3.8 gene encoding a testis-specific protein. *Mol Hum Reprod*. 2001;7:811–818.
60. Lin W, Miao SY, Zhang L, et al. Study on the function of HSD-3.8 gene encoding a testis-specific protein with yeast two-hybrid system. *Zhongguo Yi Xue Ke Xue Yuan Xue Bao*. 2002;24:582–587.
61. Liu N, Qiao Y, Cai C, et al. A sperm component, HSD-3.8 (SPAG1), interacts with G-protein beta 1 subunit and activates extracellular signal-regulated kinases (ERK). *Front Biosci*. 2006;11:1679–1689.

62. Hayashida K, Omagari K, Masuda J, et al. The sperm mitochondria-specific translocator has a key role in maternal mitochondrial inheritance. *Cell Biol Int.* 2005;29:472–481.
63. Hayashida K, Omagari K, Masuda J, et al. An integrase of endogenous retrovirus is involved in maternal mitochondrial DNA inheritance of the mouse. *Biochem Biophys Res Commun.* 2008;366:206–211.
64. Takaishi M, Huh N. A tetratricopeptide repeat-containing protein gene, tpis, whose expression is induced with differentiation of spermatogenic cells. *Biochem Biophys Res Commun.* 1999;264:81–85.
65. Biermann K, Heukamp LC, Steger K, et al. Genome-wide expression profiling reveals new insights into pathogenesis and progression of testicular germ cell tumors. *Cancer Genomics Proteomics.* 2007;4:359–367.
66. Diekman AB, Olson G, Goldberg E. Expression of the human antigen SPAG2 in the testis and localization to the outer dense fibers in spermatozoa. *Mol Reprod Dev.* 1998;50:284–293.
67. Mio T, Yabe T, Arisawa M, et al. The eukaryotic UDP-N-acetylglucosamine pyrophosphorylases. Gene cloning, protein expression, and catalytic mechanism. *J Biol Chem.* 1998;273:14392–14397.
68. Shao X, Tarnasky HA, Lee JP, et al. Spag4, a novel sperm protein, binds outer dense-fiber protein Odf1 and localizes to microtubules of manchette and axoneme. *Dev Biol.* 1999;211:109–123.
69. Kennedy C, Sebire K, de Kretser DM, et al. Human sperm associated antigen 4 (SPAG4) is a potential cancer marker. *Cell Tissue Res.* 2004;315:279–283.
70. Gruber J, Harborth J, Schnabel J, et al. The mitotic-spindle-associated protein astrin is essential for progression through mitosis. *J Cell Sci.* 2002;115(Pt 21):4053–4059.
71. Thein KH, Kleylein-Sohn J, Nigg EA, et al. Astrin is required for the maintenance of sister chromatid cohesion and centrosome integrity. *J Cell Biol.* 2007;178:345–354.
72. Liu L, Akhter S, Bae JB, et al. SNM1B/Apollo interacts with astrin and is required for the prophase cell cycle checkpoint. *Cell Cycle.* 2009;8:628–638.
73. Du J, Jablonski S, Yen TJ, et al. Astrin regulates Aurora-A localization. *Biochem Biophys Res Commun.* 2008;370:213–219.
74. Cheng TS, Hsiao YL, Lin CC, et al. hNinein is required for targeting spindle-associated protein Astrin to the centrosome during the S and G2 phases. *Exp Cell Res.* 2007;313:1710–1721.
75. Suzuki H, Yagi M, Suzuki K. Duplicated insertion mutation in the microtubule-associated protein Spag5 (astrin/MAP126) and defective proliferation of immature Sertoli cells in rat hypogonadic (hgn/hgn) testes. *Reproduction.* 2006;132:79–93.
76. Sapiro R, Kostetskii I, Olds-Clarke P, et al. Male infertility, impaired sperm motility, and hydrocephalus in mice deficient in sperm-associated antigen 6. *Mol Cell Biol.* 2002;22:6298–6305.
77. Lonergan KM, Chari R, Deleeuw RJ, et al. Identification of novel lung genes in bronchial epithelium by serial analysis of gene expression. *Am J Respir Cell Mol Biol.* 2006;35:651–661.
78. Steinbach D, Schramm A, Eggert A, et al. Identification of a set of seven genes for the monitoring of minimal residual disease in pediatric acute myeloid leukemia. *Clin Cancer Res.* 2006;12:2434–2441.
79. Fernebro J, Francis P, Eden P, et al. Gene expression profiles relate to SS18/SSX fusion type in synovial sarcoma. *Int J Cancer.* 2006;118:1165–1172.
80. Cheng GY, Shi JL, Wang M, et al. Inhibition of mouse acrosome reaction and sperm-zona pellucida binding by anti-human sperm membrane protein 1 antibody. *Asian J Androl.* 2007;9:23–29.
81. Kuang Y, Yan YC, Gao AW, et al. Immune responses in rats following oral immunization with attenuated Salmonella typhimurium expressing human sperm antigen. *Arch Androl.* 2000;45:169–180.
82. Tang X, Zhang J, Cai Y, et al. Sperm membrane protein (hSMP-1) and RanBPM complex in the microtubule-organizing centre. *J Mol Med.* 2004;82:383–388.
83. Li R, Tang XL, Miao SY, et al. Regulation of the G2/M phase of the cell cycle by sperm associated antigen 8 (SPAG8) protein. *Cell Biochem Funct.* 2009;27:264–268.
84. Wu YW, Chen DH, Miao SY, et al. Eliciting an immune response by plasmid DNA encoding a human sperm protein (HSD-1). *Arch Androl.* 1999;42:127–136.
85. Zhang XD, Miao SY, Wang LF, et al. Human sperm membrane protein (hSMP-1): a developmental testis-specific component during germ cell differentiation. *Arch Androl.* 2000;45:239–246.
86. Yasuoka H, Ihn H, Medsger TA Jr, et al. A novel protein highly expressed in testis is overexpressed in systemic sclerosis fibroblasts and targeted by autoantibodies. *J Immunol.* 2003;171:6883–6890.
87. Lee CM, Onesime D, Reddy CD, et al. JLP: a scaffolding protein that tethers JNK/p38 MAPK signaling modules and transcription factors. *Proc Natl Acad Sci U S A.* 2002;99:14189–14194.
88. Nguyen Q, Lee CM, Le A, et al. JLP associates with kinesin light chain 1 through a novel leucine zipper-like domain. *J Biol Chem.* 2005;280:30185–30191.
89. Iwanaga A, Wang G, Gantulga D, et al. Ablation of the scaffold protein JLP causes reduced fertility in male mice. *Transgenic Res.* 2008;17:1045–1058.
90. Jagadish N, Rana R, Selvi R, et al. Characterization of a novel human sperm-associated antigen 9 (SPAG9) having structural homology with c-Jun N-terminal kinase-interacting protein. *Biochem J.* 2005;389(Pt 1):73–82.
91. Gantulga D, Tuvshintugs B, Endo Y, et al. The scaffold protein c-Jun NH2-terminal kinase-associated leucine zipper protein regulates cell migration through interaction with the G protein G(alpha 13). *J Biochem.* 2008;144:693–700.
92. Ikonomov OC, Fligger J, Sbrissa D, et al. Kinesin adapter JLP links PIKfyve to microtubule-based endosome-to-trans-Golgi network traffic of furin. *J Biol Chem.* 2009;284:3750–3761.
93. Kashef K, Xu H, Reddy EP, et al. Endodermal differentiation of murine embryonic carcinoma cells by retinoic acid requires JLP, a JNK-scaffolding protein. *J Cell Biochem.* 2006;98:715–722.
94. Yasuoka H, Kuwana M. Autoantibody response against a novel testicular antigen protein highly expressed in testis (PHET) in SSc patients. *Autoimmun Rev.* 2007;6:228–231.
95. Raymond A, Ensslin MA, Shur BD. SED1/MFG-E8: a bi-motif protein that orchestrates diverse cellular interactions. *J Cell Biochem.* 2009;106:957–966.
96. Liu Y, Chiriva-Internati M, You C, et al. Use and specificity of breast cancer antigen/milk protein BA46 for generating anti-self-cytotoxic T lymphocytes by recombinant adeno-associated virus-based gene loading of dendritic cells. *Cancer Gene Ther.* 2005;12:304–312.
97. Zeelenberg IS, Ostrowski M, Krumeich S, et al. Targeting tumor antigens to secreted membrane vesicles in vivo induces efficient antitumor immune responses. *Cancer Res.* 2008;68:1228–1235.
98. Ensslin M, Vogel T, Calvete JJ, et al. Molecular cloning and characterization of P47, a novel boar sperm-associated zona pellucida-binding protein homologous to a family of mammalian secretory proteins. *Biol Reprod.* 1998;58:1057–1064.
99. Oshima K, Aoki N, Negi M, et al. Lactation-dependent expression of an mRNA splice variant with an exon for a multiply O-glycosylated domain of mouse milk fat globule glycoprotein MFG-E8. *Biochem Biophys Res Commun.* 1999;254:522–528.
100. Hamil KG, Sivashanmugam P, Richardson RT, et al. HE2beta and HE2gamma, new members of an epididymis-specific family of androgen-regulated proteins in the human. *Endocrinology.* 2000;141:1245–1253.

101. Osterhoff C, Kirchhoff C, Krull N, et al. Molecular cloning and characterization of a novel human sperm antigen (HE2) specifically expressed in the proximal epididymis. *Biol Reprod.* 1994;50:516–525.
102. Frohlich O, Po C, Young LG. Organization of the human gene encoding the epididymis-specific EP2 protein variants and its relationship to defensin genes. *Biol Reprod.* 2001; 64:1072–1079.
103. Yenugu S, Hamil KG, Birse CE, et al. Antibacterial properties of the sperm-binding proteins and peptides of human epididymis 2 (HE2) family; salt sensitivity, structural dependence and their interaction with outer and cytoplasmic membranes of *Escherichia coli*. *Biochem J.* 2003;372(Pt 2): 473–483.
104. Yenugu S, Hamil KG, Grossman G, et al. Identification, cloning and functional characterization of novel sperm associated antigen 11 (SPAG11) isoforms in the rat. *Reprod Biol Endocrinol.* 2006;4:23.
105. Avellar MC, Honda L, Hamil KG, et al. Novel aspects of the sperm-associated antigen 11 (SPAG11) gene organization and expression in cattle (Bos taurus). *Biol Reprod.* 2007;76: 1103–1116.
106. Naz RK. Effects of sperm-reactive antibodies present in human infertile sera on fertility of female rabbits. *J Reprod Immunol.* 1990;18:161–177.
107. Nottrott S, Hartmuth K, Fabrizio P, et al. Functional interaction of a novel 15.5 kD (U4/U6.U5) tri-snRNP protein with the 5′ stem-loop of U4 snRNA. *EMBO J.* 1999;18: 6119–6133.
108. Naz RK, Alexander NJ, Isahakia M, et al. Monoclonal antibody to a human germ cell membrane glycoprotein that inhibits fertilization. *Science.* 1984;225:342–344.
109. Zhu X, Naz RK. Fertilization antigen-1: cDNA cloning, testis-specific expression, and immunocontraceptive effects. *Proc Natl Acad Sci U S A.* 1997;94:4704–4709.
110. Inokuchi J, Komiya M, Baba I, et al. Deregulated expression of KRAP, a novel gene encoding actin-interacting protein, in human colon cancer cells. *J Hum Genet.* 2004; 49:46–52.
111. Fujimoto T, Koyanagi M, Baba I, et al. Analysis of KRAP expression and localization, and genes regulated by KRAP in a human colon cancer cell line. *J Hum Genet.* 2007;52: 978–984.
112. Fujimoto T, Miyasaka K, Koyanagi M, et al. Altered energy homeostasis and resistance to diet-induced obesity in KRAP-deficient mice. *PLoS One.* 2009;4:e4240.
113. Naz RK. Effects of antisperm antibodies on early cleavage of fertilized ova. *Biol Reprod.* 1992;46:130–139.
114. Myles DG, Primakoff P. Localized surface antigens of guinea pig sperm migrate to new regions prior to fertilization. *J Cell Biol.* 1984;99:1634–1641.
115. Griffiths GS, Miller KA, Galileo DS, et al. Murine SPAM1 is secreted by the estrous uterus and oviduct in a form that can bind to sperm during capacitation: acquisition enhances hyaluronic acid-binding ability and cumulus dispersal efficiency. *Reproduction.* 2008;135:293–301.
116. Gmachl M, Sagan S, Ketter S, et al. The human sperm protein PH-20 has hyaluronidase activity. *FEBS Lett.* 1993; 336:545–548.
117. Zhang Z, Sapiro R, Kapfhamer D, et al. A sperm-associated WD repeat protein orthologous to Chlamydomonas PF20 associates with Spag6, the mammalian orthologue of Chlamydomonas PF16. *Mol Cell Biol.* 2002;22: 7993–8004.
118. Zhang H, Martin-Deleon PA. Mouse epididymal Spam1 (pH-20) is released in the luminal fluid with its lipid anchor. *J Androl.* 2003;24:51–58.
119. Grigorieva A, Griffiths GS, Zhang H, et al. Expression of SPAM1 (PH-20) in the murine kidney is not accompanied by hyaluronidase activity: evidence for potential roles in fluid and water reabsorption. *Kidney Blood Press Res.* 2007;30: 145–155.
120. Sullivan R, Frenette G, Girouard J. Epididymosomes are involved in the acquisition of new sperm proteins during epididymal transit. *Asian J Androl.* 2007;9:483–491.
121. Griffiths GS, Galileo DS, Reese K, et al. Investigating the role of murine epididymosomes and uterosomes in GPI-linked protein transfer to sperm using SPAM1 as a model. *Mol Reprod Dev.* 2008;75:1627–1636.
122. Cherr GN, Yudin AI, Overstreet JW. The dual functions of GPI-anchored PH-20: hyaluronidase and intracellular signaling. *Matrix Biol.* 2001;20:515–525.
123. Liu D, Pearlman E, Diaconu E, et al. Expression of hyaluronidase by tumor cells induces angiogenesis in vivo. *Proc Natl Acad Sci U S A.* 1996;93:7832–7837.
124. Primakoff P, Lathrop W, Woolman L, et al. Fully effective contraception in male and female guinea pigs immunized with the sperm protein PH-20. *Nature.* 1988;335:543–546.
125. Tung KS, Primakoff P, Woolman-Gamer L, et al. Mechanism of infertility in male guinea pigs immunized with sperm PH-20. *Biol Reprod.* 1997;56:1133–1141.
126. Deng X, Meyers SA, Tollner TL, et al. Immunological response of female macaques to the PH-20 sperm protein following injection of recombinant proteins or synthesized peptides. *J Reprod Immunol.* 2002;54:93–115.
127. Jones MH, Davey PM, Aplin H, et al. Expression analysis, genomic structure, and mapping to 7q31 of the human sperm adhesion molecule gene SPAM1. *Genomics.* 1995;29:796–800.
128. Sun L, Feusi E, Sibalic A, et al. Expression profile of hyaluronidase mRNA transcripts in the kidney and in renal cells. *Kidney Blood Press Res.* 1998;21:413–418.
129. Deng X, He Y, Martin-Deleon PA. Mouse Spam1 (PH-20): evidence for its expression in the epididymis and for a new category of spermatogenic-expressed genes. *J Androl.* 2000; 21:822–832.
130. Zhang H, Martin-DeLeon PA. Mouse Spam1 (PH-20) is a multifunctional protein: evidence for its expression in the female reproductive tract. *Biol Reprod.* 2003;69:446–454.
131. Mohamadzadeh M, DeGrendele H, Arizpe H, et al. Proinflammatory stimuli regulate endothelial hyaluronan expression and CD44/HA-dependent primary adhesion. *J Clin Invest.* 1998;101:97–108.
132. Zhang Z, Kostetskii I, Tang W, et al. Deficiency of SPAG16L causes male infertility associated with impaired sperm motility. *Biol Reprod.* 2006;74:751–759.
133. Zhang Z, Kostetskii I, Moss SB, et al. Haploinsufficiency for the murine orthologue of Chlamydomonas PF20 disrupts spermatogenesis. *Proc Natl Acad Sci U S A.* 2004;101: 12946–12951.
134. Pennarun G, Bridoux AM, Escudier E, et al. Isolation and expression of the human hPF20 gene orthologous to Chlamydomonas PF20: evaluation as a candidate for axonemal defects of respiratory cilia and sperm flagella. *Am J Respir Cell Mol Biol.* 2002;26:362–370.
135. Zhang Z, Jones BH, Tang W, et al. Dissecting the axoneme interactome: the mammalian orthologue of Chlamydomonas PF6 interacts with sperm-associated antigen 6, the mammalian orthologue of Chlamydomonas PF16. *Mol Cell Proteomics.* 2005;4:914–923.
136. AceView, [database online]. Bethesda, MD: National Center for Biotechnology Information; 2000. Updated December 12, 2009.
137. Hofmann O, Caballero OL, Stevenson BJ, et al. Genome-wide analysis of cancer/testis gene expression. *Proc Natl Acad Sci U S A.* 2008;105:20422–20427.
138. Bouwmeester T, Bauch A, Ruffner H, et al. A physical and functional map of the human TNF-alpha/NF-kappa B signal transduction pathway. *Nat Cell Biol.* 2004;6:97–105.
139. Rosenberg SA, Sherry RM, Morton KE, et al. Tumor progression can occur despite the induction of very high levels of self/tumor antigen-specific $CD8^+$ T cells in patients with melanoma. *J Immunol.* 2005;175:6169–6176.
140. Muranski P, Restifo NP. Adoptive immunotherapy of cancer using CD4(+) T cells. *Curr Opin Immunol.* 2009;21:200–208.

141. Veland IR, Awan A, Pedersen LB, et al. Primary cilia and signaling pathways in mammalian development, health and disease. *Nephron Physiol.* 2009;111:p39–p53.
142. Plotnikova OV, Golemis EA, Pugacheva EN. Cell cycle-dependent ciliogenesis and cancer. *Cancer Res.* 2008;68: 2058–2061.
143. Ciliary proteome database, Version 3 [database online]. Baltimore, MD: Johns Hopkins University, McKusick-Nathans Institute of Genetic Medicine; 2006. Updated 2008.
144. Zwillenberg LO. At 40 years of the "Golden Chain". Which are the stem cells in ectocervical epithelium? *Gynecol Obstet Invest.* 1998;46:247–251.
145. Martens JE, Arends J, Van der Linden PJ, et al. Cytokeratin 17 and p63 are markers of the HPV target cell, the cervical stem cell. *Anticancer Res.* 2004;24:771–775.
146. Pearson CG, Giddings TH Jr, Winey M. Basal body components exhibit differential protein dynamics during nascent basal body assembly. *Mol Biol Cell.* 2009;20:904–914.
147. Zhang Z, Tang W, Zhou R, et al. Accelerated mortality from hydrocephalus and pneumonia in mice with a combined deficiency of SPAG6 and SPAG16L reveals a functional interrelationship between the two central apparatus proteins. *Cell Motil Cytoskeleton.* 2007;64:360–376.
148. Baida A, Akdi M, Gonzalez-Flores E, et al. Strong association of chromosome 1p12 loci with thyroid cancer susceptibility. *Cancer Epidemiol Biomarkers Prev.* 2008;17:1499–1504.
149. Kalnina Z, Zayakin P, Silina K, et al. Alterations of pre-mRNA splicing in cancer. *Genes Chromosomes Cancer.* 2005; 42:342–357.
150. Kalnina Z, Silina K, Line A. Autoantibody profiles as biomarkers for response to therapy and early detection of cancer. *Curr Cancer Ther Rev.* 2008;4:149–156.
151. Yang F, Chen III, Xiong Z, et al. Model of stimulation-responsive splicing and strategies in identification of immunogenic isoforms of tumor antigens and autoantigens. *Clin Immunol.* 2006;121:121–133.

3.5 . Tumour-associated autoantibody signatures for the early detection of gastric cancer

Tumor-associated autoantibody signature for the early detection of gastric cancer

Pawel Zayakin[1], Guntis Ancāns[2,3], Karīna Siliņa[1], Irēna Meistere[1], Zane Kalniņa[1], Diāna Andrejeva[1], Edgars Endzeliņš[1], Lāsma Ivanova[1], Angelina Pismennaja[1], Agnese Sudraba[3], Simona Doniņa[3], Thomas Wex[4], Peter Malfertheiner[4], Mārcis Leja[2,3] and Aija Linē[1]

[1] Latvian Biomedical Research and Study Centre, Riga, Latvia
[2] Latvian Oncology Center, Riga Eastern Clinical University hospital, Riga, Latvia
[3] Faculty of Medicine, University of Latvia, Riga, Latvia
[4] Clinic of Gastroenterology, Hepatology and Infectious Diseases, Otto-von-Guericke University Magdeburg, Magdeburg, Germany

Running title: Autoantibody signature in gastric cancer

Key words: autoantibodies, phage-displayed antigen microarray, gastric cancer, biomarker, early detection

Grant support: This study was supported in parts by ERDF project No 2010/0231/2DP/2.1.1.1.0/10/APIA/VIAA/044, ESF project No. 009/0220/1DP/1.1.1.2.0/09/APIA/VIAA/016, grant No 09.1288, Latvian State Research Program and individual fellowships from ESF No. 2009/0138/1DP/1.1.2.1.2/09/IPIA/VIAA/004.

Corresponding Author: Aija Linē, Latvian Biomedical Research and Study Centre, Ratsupites Str 1, LV-1067, Riga, Latvia. Phone: 371 67808208, Fax: 371 67442407, E-mail: aija@biomed.lu.lv

Conflicts of interest: The authors declare that they have no competing interests.

Abstract

Autoantibodies against tumor-associated antigens due to their specificity and stability in the sera are very attractive biomarkers for the development of non-invasive serological tests for the early detection of cancer. In the current study we applied T7 phage display-based SEREX technique to identify a representative set of antigens eliciting humoral responses in gastric cancer (GC) patients, produced phage-antigen microarrays and exploited them for the survey of autoantibody repertoire in patients with GC and inflammatory diseases. We developed procedures for data normalization and cutoff determination in order to define sero-positive signals and rank them by the signal intensity and frequency of reactivity. To identify autoantibodies with the highest diagnostic value, a 1150-feature microarray was tested with sera from 100 patients with GC and 100 cancer-free controls and then the top-ranked 86 antigens were used for the production of focused array that was tested with an independent validation set comprising serum samples from 239 GC patients, 150 peptic ulcer and gastritis patients and 213 healthy controls. ROC curve analysis showed that 45-autoantibody signature could discriminate GC and healthy controls with AUC of 0.79 (58% sensitivity and 91% specificity), GC and peptic ulcer with AUC of 0.76, and GC and gastritis with AUC of 0.64. Moreover, it could detect early GC with equal sensitivity than advanced GC. Interestingly, the autoantibody production did not correlate with histological type, *H. pylori* status, grade, localization and size of the primary tumor while it appeared to be associated with the metastatic disease.

Introduction

Despite the overall global decrease in incidence, gastric cancer (GC) with the estimated ~989 600 new cases and ~ 738 000 deaths per year remains the fourth most common type of cancer and the second most common cause of cancer-related death worldwide (1). The high mortality rate in GC is mostly due to its detection at advanced stage (IIIA-IV), when the estimated 5-year survival rate ranges from 4 to 20% and the median overall survival is around 8-12 months. Only less than 20% of cases are detected at an early localized stage when the complete, curative resection is possible and the 5-year survival rate reaches 75% (2, 3). The early detection of GC is hampered by the lack of specific symptoms before it has spread beyond the original site and the lack of reliable non-invasive screening tests. Currently, the diagnosis is based on endoscopic examination followed by the histological analysis of gastric biopsy that is an invasive technique not applicable for the screening of asymptomatic population. Hence the identification and validation of GC biomarkers that could be detected in body fluids such as plasma, serum or urine and are suitable for the development of non-invasive or minimally invasive tests applicable for screening high-risk groups would represent a significant step towards the reduction of morbidity and mortality caused by GC.

Autoantibodies against tumor-associated antigens (TAA) have been detected by the classical or modified SEREX approaches in all cancer types analyzed so far (4, 5) and due to their specificity and stability in the sera they seem to be very attractive targets for the development of non-invasive serological tests for the diagnosis or early detection of cancer. Furthermore, in contrary to the currently known serum biomarkers such as PSA, CEA or CA19-9, they are qualitative not quantitative biomarkers. However, so far the exploitation of tumor-associated autoantibodies for cancer diagnosis has been hampered by several factors: the frequency of antibodies against any individual TAA is generally low, typically ranging from 1 to ~15%; autoantibody repertoire is heterogeneous and to some extent resemble the response to tissue damage by viral infections or autoimmune diseases; autoantibodies against a number of TAAs, such as CTAG1B, TP53, c-MYC etc. are found in patients with different types of cancer (6-9). To overcome these limitations, a number of novel proteomic approaches including the serological proteome analysis (SERPA) (10), immunoprecipitation of antigens followed by mass spectrometric analysis (11) and antigen microarrays (12-16) have been applied to explore the autoantibody profiles in cancer patients resulting in the discovery of autoantibody signatures with diagnostic significance.

Although the application of the classical SEREX and proteomics techniques to gastric cancer has resulted in the identification of a variety of TAAs (17-20), to our best knowledge, the autoantibody repertoire in GC has not been comprehensively characterized and the diagnostic significance of the autoantibodies has not been evaluated so far. Moreover, it is not clear whether the production of autoantibodies is related to the metastatic spread, size of the primary tumor, its histological type, localization or grade. To address these issues, we applied T7 phage display-based SEREX technique (21) to identify a representative set of antigens eliciting humoral responses in GC patients, produced phage-displayed antigen microarrays and determined the autoantibody profiles in patients with gastric cancer, gastritis, gastric ulcer and healthy individuals and examined the correlation of the autoantibody signatures with the clinicopathological features.

Materials and Methods

Tissue specimens and study population.

Eleven gastric cancer tissue specimens were macroscopically dissected by a histopathologist during surgery at Latvian Oncology Center and stored in RNALater® (Applied Biosystems, USA) at -20°C till processing. Tissue sections were evaluated by an experienced pathologist and the diagnosis was established according to standard histopathological criteria. Five of the specimens were diagnosed as intestinal, 6 – as diffuse type adenocarcinomas, including 3 signet-ring cell adenocarcinomas; 2 were grade II, 5 were grade III and 4 were grade IV cancers.

Pre-treatment serum samples from 232 GC patients were collected at Latvian Oncology Center, aliquoted and stored at -80°C. Another cohort of 134 GC serum samples was received from Clinic of Gastroenterology, Hepatology and Infectious Diseases, Germany. Serum samples from 313 age and gender matched cancer-free healthy individuals and 98 patients with gastritis, including 58 patients with endoscopically detected atrophy, and 52 patients with gastric ulcer were provided by the Genome Database of Latvian Population. Characteristics of the study population are provided in Table 1.

The tissue and serum specimens were collected after the patients' informed consent was obtained in accordance with the regulations of Committee of Medical Ethics of Latvia and the ethical committee of the Otto-von-Guericke University Magdeburg.

Construction of T7-StrepII tag phage display vectors and cDNA expression libraries.

A set of three T7-StrepII tag 1-3 vectors was constructed by cloning DNA sequence encoding StrepII tag (Trp-Ser-His-Pro-Gln-Phe-Glu-Lys) into HindIII and NotI sites of T7Select 10-3b vector DNA (Novagen). Three different oligonucleotides carrying one or two nt insertions after the HindIII site were used in order to position StrepII tag in the three possible reading frames relatively to N-terminus of T7 10B coat protein and cDNA insert. The oligonucleotide duplexes were digested with HindIII and NotI, ligated into the T7Select 10-3b vector and the ligation mixture was subjected to *in vitro* packaging using T7 Select Packaging extract (Novagen). Three individual phage clones carrying the StrepII tag in different reading frames were selected, amplified, vector DNA was isolated using phenol/chloroform extraction, ethanol precipitated and digested with EcoRI and HindIII.

The obtained T7-StrepII tag 1-3 vector set was used for the construction of gastric cancer cDNA expression libraries as described previously (21). Briefly, total RNA was extracted from 11 tumor tissue specimens using TRIzol reagent (Invitrogen) according to the manufacturer's instructions. mRNA was isolated from 300 µg total RNA pooled from 5 or 6 patients to produce GCP5 and GCP6 libraries, respectively, using Dynabeads mRNA purification kit (Invitrogen) and converted to cDNA using HindIII Random primers (5'-TTNNNNNN-3') (Novagen). Then cDNA was ligated to directional EcoRI and HindIII linkers, digested with the corresponding restriction enzymes, size fractioned by gel electrophoresis to isolate fragments of 200 – 1000 bp in length and then ligated into pre-digested T7-StrepII tag 1-3 vectors (0.5 µg each). The ligation mixtures were packaged *in vitro* resulting in GCP5 and GCP6 libraries of 5×10^6 pfu and 8×10^6 pfu, respectively. The libraries were amplified once in IPTG-induced BLT5615 cells.

Selection of serum-reactive phage clones.

The obtained GCP5 and GCP6 cDNA expression libraries were enriched for ORFs by incubating ~5×10^{10}pfu from each library with 40 µl of *Strep*-Tactin coated Magnetic Beads (IBA GmbH, Germany) and the phage particles expressing Strep II tag were purified and eluted according to the manufacturer's instructions. ORF enriched libraries were titrated and used for the biopanning with negative and positive selection followed by the immunoscreening with patients' sera as described previously (21). Briefly, approximately 3-5×10^5pfu from each ORF enriched library were at first

incubated with 100 µl of Protein G coated magnetic beads (Pierce), coupled with IgGs from 5 healthy individuals and then the unbound phage was incubated overnight with pools of sera from 5 or 6 GC or gastritis patients (0.5 µl each). Sera were preabsorbed with BLT5615 and T7 phage lysate coupled to CNBr-Sepharose 4B before adding to the phage libraries. A hundred µl of Protein G coated magnetic beads were washed twice with blocking solution (5% milk powder in TBS, 0.05% Tween 20), added to the phage - serum mixture and incubated for 2 h with agitation. The beads were washed 10 times with 1 ml TBS, 0.05% Tween and all bound phages were used for the immunoscreening. BLT5615 cells grown in LB supplemented with 1×M9 salts, 0.4% glucose, 1mM $MgSO_4$ and carbenicillin (50µg/ml) to OD_{600}=0.5 and induced with IPTG for 30 min were infected with the phages and plated on LB/carbenicillin agar plates at density ~10^3pfu per 150 mm plate. After ~2h incubation at 37°C when the plaques reached ~1mm in diameter, plates were overlaid with Protan nitrocellulose (NC) filters (Whatman) and incubated for 1 h at 37°C. The filters were blocked with 5% (w/vol) milk powder in TBS, 0.05% Tween 20 for 1h, and then incubated overnight with 1:200 diluted patients' serum that has been preabsorbed with *E.coli/* phage lysates immobilised on CNBr-Sepharose 4B. The serum-reactive clones were detected by incubating the filters with alkaline phosphatase conjugated anti-human IgG, Fcγ specific secondary antibody (Pierce) and NBT/BCIP (Fermentas), isolated and purified to monoclonality. The inserts of serum-reactive phages were amplified by 35-cycle PCR using primers flanking the insert and 1 µl of phage solution as a template. PCR products were purified and sequenced using ABI Prism BigDye Terminator v3.1 cycle sequencing kit and 3130 Genetic Analyser (Applied Biosystems). DNA sequences were analysed using BLAST tool at www.ncbi.nlm.nih.gov, Translate tool at www.expasy.org and compared against sequences available at Cancer Immunome Database (www2.licr.org/CancerImmunomeDB).

Production and processing of phage displayed antigen microarrays.

For the production of 1150-feature antigen microarray, a panel of all different serum-reactive phage clones selected from GC cDNA libraries, phage clones previously selected from T7 phage-displayed testis, melanoma and prostate cancer cDNA expression libraries and non-recombinant control phages was assembled and simultaneously amplified to high titre (~5×10^8-1×10^9 pfu/µl) in *E. coli* BLT 5616 cells using 96 well culture plates (Whatman). The lysates were clarified by centrifugation, supplemented with 5% glycerol and 0.1%NaN_3 and arrayed in quadruplicates onto nitrocellulose-coated 2-pad FAST slides (Whatman) using a QArray Mini microarrayer (Genetix). The microarray slides were blocked with 5% (w/vol) milk powder in TBS, 0.05% Tween 20, incubated with 0.9 ml of 1:200 diluted patients' sera that were preabsorbed with 15 µl of UV-inactivated *E.coli*- phage lysates, washed 4 times in TBS, 0.5% Tween 20 for 15 min, and then incubated with monoclonal anti-T7 tail fiber antibody (Novagen). After 3 washes in TBS, 0.5% Tween 20, the microarrays were incubated with Cy5 labelled goat anti-human IgG antibody (1:1500) and Cy3 labelled goat anti-mouse IgG antibody (1:3000) (Jackson ImmunoResearch) for 1 h, then washed thrice in TBS, 0.5% Tween 20, rinsed with distilled water and dried by centrifugation. A reference serum was included in each series of experiments. The arrays were scanned at 10 µm resolution in PowerScanner (Tecan) with 532 and 635 nm lasers, the results were recorded as TIFF files and the data were extracted using GenePix software. The obtained data were further analyzed using an ad hoc program composed in R language.

For the production of 96-feature antigen microarray, the selected phage clones were amplified from the low-titre stocks as described before, quality-controlled by PCR and spotted onto 16-pad FAST slides in duplicates and the arrays were processed as described above.

Microarray data processing and statistical analysis.

For each spot the mean Cy5 and Cy3 signals were background subtracted, averaged between replicates, and the Cy5/Cy3 ratios were calculated for each antigen. Spots that did not pass the quality criteria (morphologically heterogeneous spots and spots that differed by more than 50% between replicates) were excluded from the analysis. A two-step normalization strategy was used for the fluorescent signal ratios in order to eliminate variations introduced by the custom production of microarrays and variable background intensities of different sera. At first, the values in each slide (each serum) were normalized by the median of the middle 80% of all measurements for each fluorescent channel separately. Then the distribution of data across an array was centered by equalizing the standard deviation of the middle 80% of values to 1. Next, for the inter-slide normalization, the Cy5 and Cy3 signal intensities for each spot were divided by the median of the middle 80% of the values for this spot in slides within one batch and the distribution was centered across slides in the batch. The highest and lowest 10% of the values were excluded from the SD calculation in order not to dismiss the outliers that may represent serum-positive antigens.

The cutoff value (T) for each antigen was calculated as follows:

$$T = mean\ (I_{HD}) + 3 \times SD(I_{HD})$$

, where I_{HD} is the signal intensities in healthy controls. Then the antigens were ranked, taking into account the signal intensity and frequency of reactivity with GC patient sera compared to healthy donor sera, using the following formula:

$$R_i = \left(\Sigma I_{GC_i} / N_{GC_i}\right) - 2\left(\Sigma I_{HD_i} / N_{HD_i}\right).$$

Finally, a score for each serum was calculated as follows:

.

$$S = \sum_{i=1}^{n} \sqrt{R_i} \times I_i$$

The non-parametric Mann-Whitney U test was used to compare the serum scores between two independent groups of samples. The receiver operating characteristic (ROC) curve was constructed and the area under the curve (AUC) was calculated to evaluate the diagnostic performance of the serum scores. Leave one-out cross validation (LOOCV) as described by Laxman B et al, 2008 was used to validate the biomarker models to eliminate overestimated values (22). To define cutoff points on the ROC curves with the maximal sum of sensitivity and specificity Minimal misclassification cost term (MCT) approach (23) was used as follows:

$$MCT = (1 - prevalence) \times (1 - Sp) + \left(cost(FN) / cost(FP) \times prevalence \times\right)(1 - Se)$$

, where the cost(FN)/cost(FP) was set at 0.5.

Results and Discussion

Discovery of gastric cancer antigens.

In order to identify a representative set of antigens eliciting humoral immune responses in GC patients, two phage-displayed cDNA expression libraries, called GCP5 and GCP6, were constructed from pools of total RNA isolated from 5 intestinal type and 6 diffuse type gastric adenocarcinoma specimens, respectively, and the serum-reactive phage clones were selected with sera pooled from 27 GC patients. The workflow of this study is shown in Figure 1. A modified T7 Select 10-3b phage display system carrying the Strep II tag was used for the construction of the expression libraries. In the standard version, cDNAs are expressed as C-terminal fusion proteins with the phage coat protein 10B resulting in the display of a large proportion of out-of-frame peptides. In fact, we and others have previously observed that only ~2 to 10% of the phage clones selected from T7 phage display libraries expressed proteins in their natural reading frame, while the remaining clones contained DNA fragments translated in non-natural reading frames that most likely represent mimotopes (13, 21). To increase the proportion of in-frame antigens, we constructed a set of three T7-StrepII tag vectors that allow selecting the phages displaying in-frame antigens by ORF enrichment step using Strep-Tactin coated magnetic beads prior to the selection of serum-reactive clones. After the negative selection step with healthy donors' IgGs, both GC cDNA expression libraries were screened with pooled autologous sera and with 3 pools of allogeneic sera (5 diffuse type and 5 intestinal type adenocarcinomas of various stages and 6 stage I and IIA adenocarcinomas of various types) (Table 1) by performing a single round of biopanning followed by immunoscreening of the enriched libraries. This resulted in the identification of 316 different serum-reactive phage clones. Sixty eight of them (21.5%) encoded proteins in their natural reading frames thus showing that the application of T7-StrepII tag vector system substantially increased the proportion of the identified in-frame antigens. Among the in-frame antigens were 4 known cancer-testis (CT) antigens (CTAG1B/NY-ESO1, CTAG2/LAGE-1, DDX53, HORMAD1), 9 antigens that have been previously identified by applying conventional SEREX to various tumour types (NOL8, UBR2, SC65, KTN1, RPLP1, TPM3, ZNF282, HSPA4 and EEF1A1), a number of ribosomal proteins and heat-shock proteins known to elicit humoral response both, in cancer patients and patients with autoimmune diseases, but the rest of the antigens have not been previously implicated in autoimmune responses. The remaining 248 serum-reactive clones contained cDNAs fused to 10B in a different reading frame, 5' or 3' UTRs, ribosomal RNA genes or mitochondrial DNA, thus expressing 4 to 80 aa long peptides that likely represent mimotopes. The nature of the antigens they represent is not known as they may mimic protein as well as non-protein antigens of cancer or normal cells or various pathogens.

Identification of autoantibody signature with the diagnostic significance.

A panel of all 316 serum-reactive phage clones identified by screening of GC cDNA expression libraries, 70 non-recombinant phage clones and 764 phage clones previously

selected from T7 phage-displayed testis, melanoma and prostate cancer cDNA expression libraries (21, 24) was assembled and used for the production of 1150-feature phage-antigen microarray. The microarray comprised 46 antigens representing 16 CT antigen families (CTAG1B/CTAG2, CTAGE, MAGE-A, MAGE-C, SSX, GAGE, HORMAD, DDX53, CSAG, PAGE, BAGE, SPANX, CT45, THEG, LDHC and SPAG), 152 in-frame antigens including known tumor-associated antigens such as ANXA11 and TYR, autoantigens such as AKAP12 and LMOD1 and previously uncharacterized antigens, and 882 out-of-frame peptides. In order to define autoantibodies with a potential diagnostic significance, the microarray was tested with sera from 100 patients with gastric cancer of various histological types and stages and 100 age and gender matched healthy individuals as the training set (Table 1). After excluding low-quality spots and correcting for the variations in the phage quantity and differences across the print runs, and allowing inter-slide (inter-serum) comparisons by the two-step normalization, an individual cutoff discriminating between sero-positive and negative samples was calculated for each antigen. The cutoff was experimentally validated by plaque assay for 3 antigens – CTAG1B, HORMAD1 and SPAG17 using 3 positive and 3 negative sera according to the microarray data (data not shown). In total, 888 antigens reacted with at least one of these sera, however many of them reacted at a similar frequency with cancer patients' and healthy donors' sera. In order to select a panel of antigens eliciting the antibody response that is highly specific to cancer patients, the antigens were ranked considering the signal intensity above the cutoff and higher frequency of reactivity with cancer patients' sera than with healthy individuals' sera as described in *Materials and Methods*. Three hundred sixty antigens reacted preferentially with cancer patients' sera thus receiving a positive rating, and therefore were considered to have a potential diagnostic significance. Next, the "serum score" was calculated for each serum by summing up the signal intensities above the cutoff for all the significant antigens. It ranged from 0 to 1400 (mean 107) in GC patients and from 0 to 150 in healthy individuals (mean 11) and the difference was statistically highly significant ($P=9.5\times10^{-26}$) (Fig. 2A). The diagnostic performance of the serum score was evaluated by the ROC curve analysis and yielded AUC of 0.85 (95% CI 0.81-0.88, $P=9.5\times10^{-26}$), sensitivity (Sn) of 57% and specificity (Sp) of 92% at the cutoff point 35.3 (Fig. 2B). However, this biomarker set included antigens with low diagnostic value (reacting with 1-2% of cancer patients' sera or signal intensities less than 2 fold above the cutoff); moreover, the development of clinically applicable biomarker assay based on 360 antigens would be impractical, therefore we next selected a set of 86 top-ranked antigens with the highest specificity. This set had AUC of 0.81 (95% CI 0.78-0.84, $P=1.9\times10^{-23}$), sensitivity of 69% and specificity of 89% at the cutoff point 3.4 (Fig. 2B).

Diagnostic performance of the autoantibody signature.

The phage clones encoding the selected 86 antigens, along with 10 non-recombinant phage clones, were amplified and used for the production of 96-feature phage-antigen array. To evaluate the reproducibility between 1150-feature and 96-feature arrays, it was tested with randomly selected 50 GC patients' and 20 healthy donors' sera previously used in the training set. It showed 89% concordance in detecting sero-positive signals, with an average CV of

~13%, which is an acceptable inter-assay variability for immunoassays. To establish the diagnostic value of the selected antigens, the array was tested for the reactivity with an independent validation set including sera from 239 GC patients at various stages and of various histological types and 213 healthy individuals with no history of cancer (Table 1). The ranking of antigens was performed as described above and it showed that all of these antigens scored above zero. The mean serum score calculated on the basis of these antigens differed by more than 43-fold between GC patients and healthy donors and could discriminate GC patients and healthy donors with AUC of 0.76 (95% CI 0.74-0.79, P= 6×10^{-25}, Sn 49%, Sp 93%) (Fig. 2C). Next, we made an attempt to further reduce the number of biomarkers by backward elimination approach and found that the minimal set of antigens retaining comparative sensitivity comprised 45 antigens yielding AUC of 0.79 (95% CI 0.76-0.81,P= 6×10^{-31}, Sn 58%, Sp 91% at cutoff 2.45) (Table 2, Supplementary Table S1, patent pending). To validate the performance of 45-autoantibody signature, leave-one-out cross validation (LOOCV) was performed and it yielded AUC of 0.75 (95% CI 0.72-0.78, P=1.5×10^{-24}, Sn 39%, Sp 97% at cutoff 4.92) (Fig. 2C)

In order to evaluate the diagnostic performance of the autoantibody signature in the population in which the assay could be employed – e.g. patients with various inflammatory gastric disorders, the array was tested with sera from 52 patients with peptic ulcer and 98 patients with acute or chronic gastritis and all the calculations were performed as described above. The mean serum score in ulcer group was 1.3 that does not significantly differ from that in healthy individuals (Fig. 2D) and the ROC curve analysis showed that the 45-autoantibody signature could discriminate between the GC and ulcer with AUC 0.76 (Table 2, Fig. 2E). This shows that the identified autoantibody signature does not significantly overlap with B cell response to cancer non-related lesions of gastric mucosa. At the same time, the serum scores in gastritis patients, although still significantly lower than in GC patients, were higher than in patients with ulcer and healthy controls (Fig. 2D) and could discriminate between GC with AUC 0.64 (Table 2, Fig. 2E). As shown in Figure 2F, a specific pattern of reactivity was shared between GC and gastritis that was absent in patients with ulcer and healthy controls. Among these antigens was PRKACA that previously has been shown to induce autoantibody responses in various cancers (25) and NOL8 that has been detected by SEREX in lung cancer, while the others are out-of-frame peptides whose nature is unknown. It is likely that these responses are triggered by cancer non-related inflammation and thus have no relevance for diagnosis. Alternatively, taking into account that gastric atrophy is a well recognized precancerous condition (26), it could be possible that among the gastritis patients were clinically undetectable GC cases, hence these could be cancer-specific antigens recognized at very early stages of cancer.

What are the antigens with the highest diagnostic value?

Five of the top 10 ranked antigens are known CT antigens but the others most likely are mimotopes of antigens whose nature is unknown (Table 3). Spontaneous B cell immune responses against CTAG2 and CTAG1B have been observed in patients with various types of cancer, including GC, while anti-DDX53 antibodies have been detected in patients with

melanoma, colon and endometrial cancers and anti-MAGEA3 and MAGEC1 antibodies – in melanoma patients (27-29). Their immunogenicity is thought to be related to the expression pattern: they are expressed in a wide range of cancers, but normally their expression is restricted to immuno-privileged tissues such as testis, fetal ovary and placenta. Hence they become exposed to the immune system only when expressed by tumors, therefore the immune response against these antigens could serve as a very specific indicator of cancer, but is unlikely to discriminate between different types of cancer. The fact that autoantibodies against these antigens are very rarely detected in cancer-free controls (in the current study we detected anti-CTAG2/CTAG1B, DDX53 and MAGEA3 antibodies in 3 controls: two of them were diagnosed with melanoma and lung cancer 8 and 5 months after the blood draw, respectively, and were excluded from the analysis but no follow-up information was available for the third one), seems to argue against the cancer immunoediting concept proposed by Schreiber RD et al (30). However, it has been demonstrated that the CT antigen expression tends to correlate with an advanced stage and is higher in metastatic than primary tumors (28). Thus it could be possible that tumors start to express CT antigens only after they have undergone elimination and equilibrium stages and their expression represents a feature of immunologically sculpted tumors. Alternatively, it has been demonstrated that B cells have a key role in the initiation of tumor-promoting inflammation (31), hence it could be possible that the appearance of tumor-specific antibodies marks the point-of-no-return in the cancer immunoediting process. In line with this, Willimsky G et al (32) has demonstrated that the initiation of B cell response to a tumor antigen coincides with the development of tumor-specific tolerance and cytotoxic T cell unresponsiveness.

The frequency of cancer-specific autoantibodies in GC patients ranges from 16.3 to 0.84% that is relatively low in comparison with the frequencies reported in other cancer types, such as ovarian, lung or breast cancer, where, for example, anti-p53 antibodies may reach 45.5%, anti-cyclin B1 antibodies - 34.8% and anti-Her2/neu antibodies – 26.6%, respectively (33). This suggests that either GC is generally less immunogenic than the other cancers or the autoantibody repertoire is more heterogeneous that is supported by the large number of rare autoantibodies identified in this study. Although these low frequency antibodies still may have a diagnostic value, the validation by standard statistical analysis is difficult, as this requires very large sample cohorts.

Does the autoantibody production correlate with clinicopathological features?

The questions, when during the tumor progression the autoantibodies appear and what are the pathological stimuli triggering this response, are of paramount importance, when considering them as diagnostic or prognostic biomarkers. To address these issues the serum scores and the frequency of individual autoantibodies was correlated with clinical and pathological features of the GC patients. The serum score and the sensitivity for the detection of stage I patients was not significantly lower than that for patients with more advanced disease (stage II-IV) (Fig. 3A, Table 2-3) thus demonstrating the relevance of the identified autoantibody signature for early detection of GC. No significant correlation between the serum score and the histological type, *H. pylori* status, tumor grade, patient age and gender was observed (Table

2). In regard to tumor localization, a set of antigens that were recognized by sera from patients with distal but not proximal GC was identified (Fig. 3B), however the overall serum scores did not significantly differ between these groups. Most importantly and unexpectedly, the serum score did not correlate with the size of primary tumor showing that the tumor burden by itself is not crucial for triggering the antibody response. Instead, we found that the serum scores were significantly higher in patients with distant metastases than with non-metastatic disease, while the metastatic spread limited to the regional lymph nodes did not affect the autoantibody production (Fig. 3C). However, it is not entirely clear whether this is due to the predominant expression of the antigens on metastatic tumors or the metastatic spread by itself stimulates the B cell response.

Taken together, we have developed a novel approach for the analysis of antigen microarray data and applied it for the survey of the autoantibody repertoire in GC. A similar strategy has been taken by Gnjatic S et al (34, 35) and, in contrary to *t*-tests, regression analysis, pattern-recognition approaches and the artificial neural network-based approaches so far used in the antigen microarray data analysis (36, 37), it is aimed at defining clearly sero-positive signals and analyzing them quantitatively, rather than identifying small but consistent differences in the signal intensity with unknown biological significance. This resulted in the identification of 45- autoantibody signature that could discriminate between GC and healthy control sera with 74.5% accuracy. It has substantially higher specificity than the currently known GC serum markers such as CA 72-4, CA19-9 and CEA (38, 39). Strikingly, stage I GC could be detected with a similar sensitivity to advanced GC suggesting that this autoantibody signature could be exploited for the early detection of GC. However, the sensitivity of the model is not sufficient for the development of screening tests for an asymptomatic population and it remains unclear whether this is due to the low immunogenicity of GC or high complexity of autoantibody repertoire in GC, therefore the next task is to explore the autoantibody repertoire in those GC patients that were negative for the identified autoantibody signature. Furthermore, although this study provided some insight into the pathological processes associated with the cancer-specific autoantibody production, the functional role of autoantibodies in the development and/or progression of cancer and anti-tumor immune response is still elusive and it would be of great interest to explore their predictive and prognostic significance.

Acknowledgements

We are thankful to Dr. Liene Ņikitina-Zaķe and Genome Database of Latvian population for the help in selecting appropriate control cohorts and providing the serum samples.

Grant Support

This study was supported in parts by ERDF project No 2010/0231/2DP/2.1.1.1.0/10/APIA/VIAA/044, ESF project No. 009/0220/1DP/1.1.1.2.0/09/APIA/VIAA/016, grant No 09.1288, Latvian State Research Program and individual fellowships from ESF No. 2009/0138/1DP/1.1.2.1.2/09/IPIA/VIAA/004.

Figures

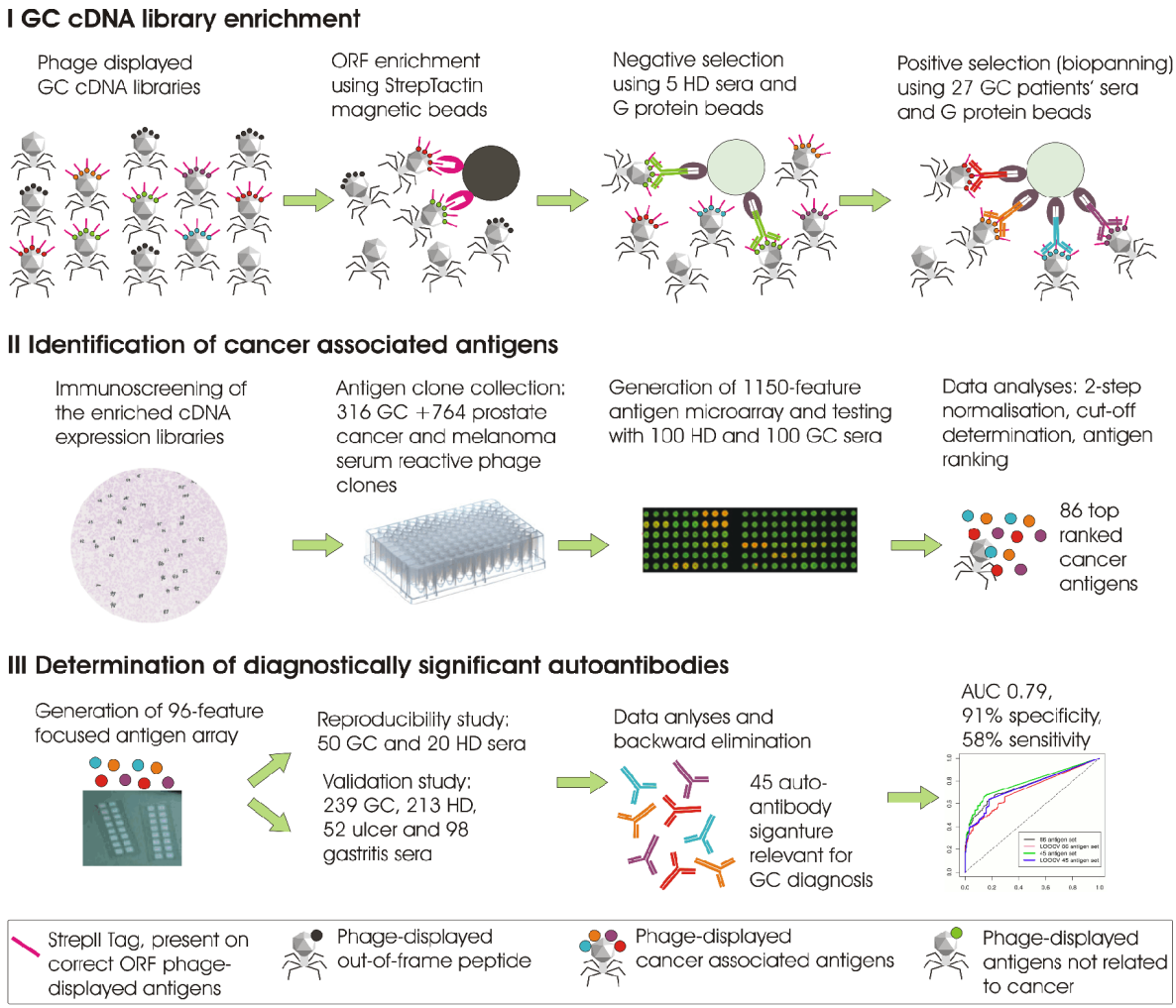

Figure 1. The workflow of antigen discovery, the production of antigen microarrays and the selection of the autoantibody signature with diagnostic relevance.

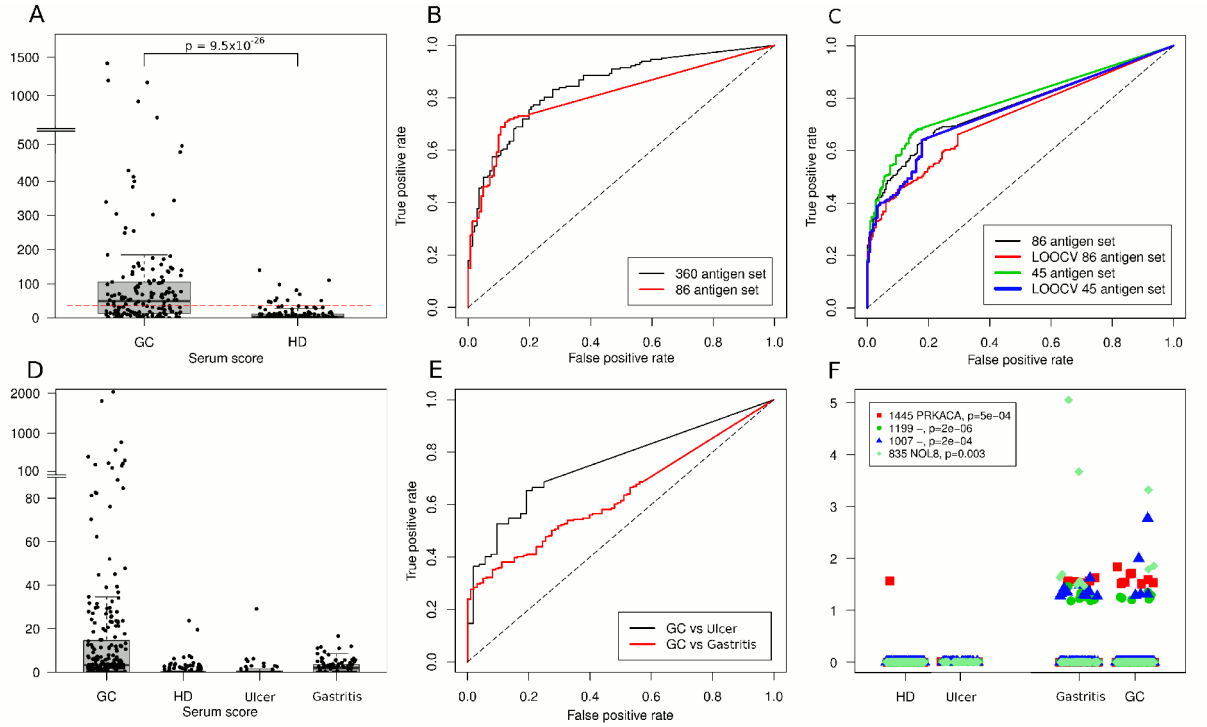

Figure 2. The identification of autoantibody signature with diagnostic relevance and the evaluation of its diagnostic performance. (A) box plot showing serum scores based on the reactivity against the top 360 antigens in the training set consisting of serum samples from 100 GC patients and 100 healthy controls (HD). Boxes represent interquartile range - distance between the 25th and 75th percentiles; whiskers represent most extreme data point, which is no more than 1.5 times the interquartile range from the box and dots represent individual samples. Dotted line represents the cutoff determined by minimal misclassification cost term approach. Statistical significance was calculated using Mann-Whitney *U* test. (B) ROC curves showing the diagnostic performance of the 360-autoantibody and the top 86-autoantibody signatures in the training sample set. (C) ROC curves showing the diagnostic performance (discrimination between GC and HD) of the 86-autoantibody and 45-autoantibody signatures and their LOOCV in the independent validation set consisting of serum samples from 239 GC patients and 213 healthy controls (HD). (D) box plot showing the serum scores based on 45-autoantibody signature in the validation sample set and patients with peptic ulcer and gastritis. (E) ROC curves showing the discrimination between patients with GC, peptic ulcer and gastritis using the 45-autoantibody signature. (F) dot plot showing reactivity pattern shared between patients with GC and gastritis. Statistical significance between groups combing HD and ulcer, and GC and gastrtis was calculated using Mann-Whitney *U* test.

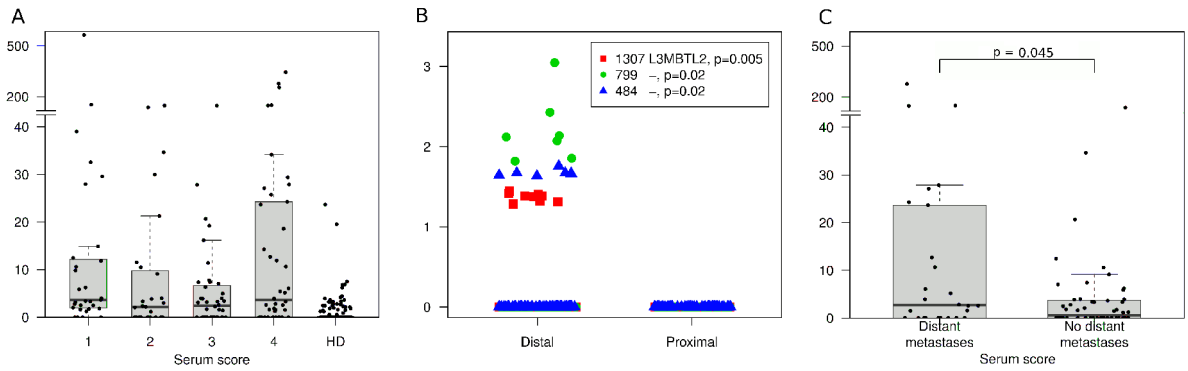

Figure 3. The serum reactivity in various subgroups of patients. (A) box plot showing the serum scores based on 45-autoantibody signature in patients with various stages of GC and HDs. (B) dot plot showing reactivity pattern distinguishing distal and proximal GC. Statistical significance was calculated using Mann-Whitney *U* test. (C) box plot showing the serum scores based on 45-autoantibody signature in GC patients with or without distal metastases (Tx,Nx,M0 vs Tx,Nx,M1).

Table 1. Clinicopathological characteristics of the study population

Variable	Antigen discovery	Training set	Validation set
GC patients			
Number of patients	27 (positive selection)	100	239
Age (mean (yr) ±SD)	64.6 ±9.8	68±13	67±13
Sex:			
Male	16	59	141
Female	11	41	98
Histological type (Lauren classification):			
Intestinal	13	40	88
Diffuse	10	38	80
Mixed	-	9	17
Data not available	4	13	54
Stage:			
I	4	18	31
II	4	12	28
III	3	18	40
IV	5	20	42
Data not available	11	32	98
Grade:			
1 (well differentiated)	1		16
2 (moderately differentiated)	7		46
3 (poorly differentiated)	8		117
4 (undifferentiated)	3		10
Data not available	8	100	50
Localization:			
Cardia	-		43
Fundus	2		-
Corpus	4		102
Antrum/pylorus	7		49
Total	1		3
Data not available	13	100	42
H. pylori status*:			
HP+, CagA+			70
HP+, CagA-			23
HP-, CagA-			41
Data not available	27	100	105
Controls			
Gastritis	-	-	98
Age (mean (yr) ±SD)			68±10
Sex:			
Male			55
Female			43
Gastric ulcer	-	-	52
Age (mean (yr) ±SD)			65±10
Sex:			
Male			28
Female			24
Cancer-free healthy controls	5 (negative selection)	100	213
Age (mean (yr) ±SD)	NA	70±5	71±5
Sex:			
Male	3	52	117
Female	2	48	96

*, *H. pylori* status was serologically determined by analyzing anti.*H. pylori* IgG levels as described in Wex et al., 2010 (40).

Table 2. Diagnostic performance of the 45 biomarker model in various subgroups of patients

Group 1	Mean serum score	ROC curve analysis group 1 vs HD		Group 2	Mean serum score	ROC curve analysis, group 2 vs HD		ROC curve analysis group 1 vs 2	
		AUC	Asympt. signif.			AUC	Asympt. signif.	AUC	Asympt. signif.
GC (n=239)	38	0.79	$6x10^{-31}$	Ulcer (n=52)	1.3	0.54	0.21	0.76	$3x10^{-9}$
				Gastrtis (n=98)	2.6	0.71	$3x10^{-13}$	0.64	$7x10^{-5}$
Stage I GC (n=31)	29	0.88	$2x10^{-18}$	Stage IV GC (n=42)	29	0.80	$1x10^{-15}$	0.53	0.70
				Stage II-IV GC (n=110)	16	0.76	$1x10^{-19}$	0.60	0. 09
Intestinal type GC (n=88)	69	0.78	$6x10^{-20}$	Diffuse type GC(n=80)	14	0.86	$2x10^{-28}$	0.57	0.13
Female GC patients (n=98)	37	0.81	$2x10^{-23}$	Male GC patients (n=141)	36	0.79	$2x10^{-24}$	0.5	0.94
Proximal GC (n=59)	57	0.82	$2x10^{-20}$	Distal GC (n=75)	46	0.91	$6x10^{-35}$	0.58	0.12
H. pylori sero-negative GC (n=41)	23	0.88	$2x10^{-22}$	*H. pylori* sero positive, CagA negative GC (n=23)	21	0.90	$2x10^{-16}$	0.54	0.58
				H. pylori sero positive, CagA positive GC (n=70)	77	0.86	$8x10^{-27}$	0.53	0.60
Primary tumor T4 (n=34)	14	0.69	$9x10^{-7}$	Primary tumor T1-T3 (n=84)	15	0.8	$3x10^{-22}$	0.60	0.09
Grade G1-G2 (n=62)	24	0.8	$2x10^{-19}$	Grade G3-G4 (n=127)	46	0.81	$2x10^{-27}$	0.52	0.70
Age <67 years (n=103)	13	0.81	$8x10^{-25}$	Age >67 years (n=102)	59	0.79	$4x10^{-23}$	0.49	0.73
Lymph node metastases (n=54)	17	0.75	$3x10^{-13}$	No lymph node metastases (n=32)	8.3	0.79	$2x10^{-12}$	0.52	0.74
Distant metastases (n=26)	21	0.77	$1x10^{-9}$	No distant metastases (n=52)	4.4	0.67	$2x10^{-7}$	0.63	0.045

Table 3. Top 10 ranked antigens

Clone No.	Gene symbol	Rating score	Freq. in GC, %			Freq. in HD, %	Odds ratio	Fisher exact test p-value*	Freq. in inflammatory diseases, %	
			All stages	Stage I	Stage II-IV				ulcer	gastritis
1416	CTAG2	197	16.3	16.1	12.7	0.5	41.3	8.4×10^{-11}	0	0
268	CTAG1B	58	8.0	6.5	8.2	0.5	18.3	4.5×10^{-5}	0	0
1428	DDX53	26	6.7	9.7	8.2	0.0	Inf	3.4×10^{-5}	1.9	0
509	-	11	5.0	0	3.6	0.0	Inf	5.2×10^{-4}	0	0
438	-	8	4.2	3.2	5.5	0.0	Inf	0.002	0	0
1256	-	7	2.9	6.5	1.8	0.0	Inf	0.016	3.9	1.0
352	MAGEC1	7	3.4	3.2	1.8	0.0	Inf	0.008	0	0
543	-	7	5.9	3.2	5.5	0.9	6.6	0.0045	0	0
1478	MAGEA3	6	3.4	3.2	0.9	0.0	Inf	0.008	1.9	1.0
799	-	6	4.2	9.7	3.6	0.9	4.6	0.04	0	1.0

* Statistical significance for the frequency of autoantibodies in GC (all stages) and healthy controls (HD) was calculated using Fisher exact test.

Reference List

(1) Jemal A, Bray F, Center MM, Ferlay J, Ward E, Forman D. Global cancer statistics. CA Cancer J Clin 2011;61:69-90.

(2) Park JM, Kim YH. Current approaches to gastric cancer in Korea. Gastrointest Cancer Res 2008;2:137-44.

(3) Rosati G, Ferrara D, Manzione L. New perspectives in the treatment of advanced or metastatic gastric cancer. World J Gastroenterol 2009;15:2689-92.

(4) Tureci O, Sahin U, Pfreundschuh M. Serological analysis of human tumor antigens: molecular definition and implications. Mol Med Today 1997;3:342-9.

(5) Preuss KD, Zwick C, Bormann C, Neumann F, Pfreundschuh M. Analysis of the B-cell repertoire against antigens expressed by human neoplasms. Immunol Rev 2002;188:43-50.

(6) Bei R, Masuelli L, Palumbo C, Modesti M, Modesti A. A common repertoire of autoantibodies is shared by cancer and autoimmune disease patients: Inflammation in their induction and impact on tumor growth. Cancer Lett 2009;281:8-23.

(7) Tan HT, Low J, Lim SG, Chung MC. Serum autoantibodies as biomarkers for early cancer detection. FEBS J 2009;276:6880-904.

(8) Kalnina Z, Silina K, Line A. Autoantibody profiles as biomarkers for response to therapy and early detection of cancer. Current Cancer Therapy Reviews 2008;4:149-56.

(9) Preiss S, Kammertoens T, Lampert C, Willimsky G, Blankenstein T. Tumor-induced antibodies resemble the response to tissue damage. Int J Cancer 2005;115:456-62.

(10) 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;1:890-8.

(11) Philip R, Murthy S, Krakover J, Sinnathamby G, Zerfass J, Keller L, et al. Shared immunoproteome for ovarian cancer diagnostics and immunotherapy: potential theranostic approach to cancer. J Proteome Res 2007;6:2509-17.

(12) 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;3:261-7.

(13) 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;66:1181-90.

(14) Wang X, Yu J, Sreekumar A, Varambally S, Shen R, Giacherio D, et al. Autoantibody signatures in prostate cancer. N Engl J Med 2005;353:1224-35.

(15) 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;64:5089-96.

(16) 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;3:2200-7.

(17) Obata Y, Takahashi T, Sakamoto J, Tamaki H, Tominaga S, Hamajima N, et al. SEREX analysis of gastric cancer antigens. Cancer Chemother Pharmacol 2000;46 Suppl:S37-S42.

(18) Line A, Stengrevics A, Slucka Z, Li G, Jankevics E, Rees RC. Serological

identification and expression analysis of gastric cancer-associated genes. Br J Cancer 2002;86:1824-30.
(19) Kalnina Z, Silina K, Bruvere R, Gabruseva N, Stengrevics A, Barnikol-Watanabe S, et al. Molecular characterisation and expression analysis of SEREX-defined antigen NUCB2 in gastric epithelium, gastritis and gastric cancer. Eur J Histochem 2009;53:7-18.
(20) Tsunemi S, Nakanishi T, Fujita Y, Bouras G, Miyamoto Y, Miyamoto A, et al. Proteomics-based identification of a tumor-associated antigen and its corresponding autoantibody in gastric cancer. Oncol Rep 2010;23:949-56.
(21) Kalnina Z, Silina K, Meistere I, Zayakin P, Rivosh A, Abols A, et al. Evaluation of T7 and lambda phage display systems for survey of autoantibody profiles in cancer patients. J Immunol Methods 2008.
(22) Laxman B, Morris DS, Yu J, Siddiqui J, Cao J, Mehra R, et al. A first-generation multiplex biomarker analysis of urine for the early detection of prostate cancer. Cancer Res 2008;68:645-9.
(23) Greiner M. Two-graph receiver operating characteristic (TG-ROC): update version supports optimisation of cut-off values that minimise overall misclassification costs. J Immunol Methods 1996;191:93-4.
(24) Silina K, Zayakin P, Kalnina Z, Ivanova L, Meistere I, Endzelins E, et al. Sperm-associated antigens as targets for cancer immunotherapy: expression pattern and humoral immune response in cancer patients. J Immunother 2011;34:28-44.
(25) Nesterova M, Johnson N, Cheadle C, Cho-Chung YS. Autoantibody biomarker opens a new gateway for cancer diagnosis. Biochim Biophys Acta 2006;1762:398-403.
(26) Hishida A, Matsuo K, Goto Y, Hamajima N. Genetic predisposition to Helicobacter pylori-induced gastric precancerous conditions. World J Gastrointest Oncol 2010;2:369-79.
(27) Stockert E, Jager E, Chen YT, Scanlan MJ, Gout I, Karbach J, et al. A survey of the humoral immune response of cancer patients to a panel of human tumor antigens. J Exp Med 1998;187:1349-54.
(28) Scanlan MJ, Gure AO, Jungbluth AA, Old LJ, Chen YT. Cancer/testis antigens: an expanding family of targets for cancer immunotherapy. Immunol Rev 2002;188:22-32.
(29) Zeng G, Aldridge ME, Wang Y, Pantuck AJ, Wang AY, Liu YX, et al. Dominant B cell epitope from NY-ESO-1 recognized by sera from a wide spectrum of cancer patients: implications as a potential biomarker. Int J Cancer 2005;114:268-73.
(30) Dunn GP, Bruce AT, Ikeda H, Old LJ, Schreiber RD. Cancer immunoediting: from immunosurveillance to tumor escape. Nat Immunol 2002;3:991-8.
(31) Andreu P, Johansson M, Affara NI, Pucci F, Tan T, Junankar S, et al. FcRgamma activation regulates inflammation-associated squamous carcinogenesis. Cancer Cell 2010;17:121-34.
(32) Willimsky G, Czeh M, Loddenkemper C, Gellermann J, Schmidt K, Wust P, et al. Immunogenicity of premalignant lesions is the primary cause of general cytotoxic T lymphocyte unresponsiveness. J Exp Med 2008;205:1687-700.
(33) Reuschenbach M, von Knebel DM, Wentzensen N. A systematic review of humoral immune responses against tumor antigens. Cancer Immunol Immunother 2009;58:1535-44.
(34) Gnjatic S, Ritter E, Buchler MW, Giese NA, Brors B, Frei C, et al. Seromic profiling of ovarian and pancreatic cancer. Proc Natl Acad Sci U S A 2010;107:5088-93.
(35) Gnjatic S, Wheeler C, Ebner M, Ritter E, Murray A, Altorki NK, et al. Seromic

analysis of antibody responses in non-small cell lung cancer patients and healthy donors using conformational protein arrays. J Immunol Methods 2009;341:50-8.
(36) Chen G, Wang X, Yu J, Varambally S, Yu J, Thomas DG, et al. Autoantibody profiles reveal ubiquilin 1 as a humoral immune response target in lung adenocarcinoma. Cancer Res 2007;67:3461-7.
(37) 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;1:513-9.
(38) Schneider J, Schulze G. Comparison of tumor M2-pyruvate kinase (tumor M2-PK), carcinoembryonic antigen (CEA), carbohydrate antigens CA 19-9 and CA 72-4 in the diagnosis of gastrointestinal cancer. Anticancer Res 2003;23:5089-93.
(39) Carpelan-Holmstrom M, Louhimo J, Stenman UH, Alfthan H, Haglund C. CEA, CA 19-9 and CA 72-4 improve the diagnostic accuracy in gastrointestinal cancers. Anticancer Res 2002;22:2311-6.
(40) Wex T, Leodolter A, Bornschein J, Kuester D, Kahne T, Kropf S, et al. Interleukin 1 beta (IL1B) gene polymorphisms are not associated with gastric carcinogenesis in Germany. Anticancer Res 2010;30:505-11.

4. DISCUSSION

The exploitation of autoantibodies as diagnostic, prognostic or predictive biomarkers of cancer for the clinical applications so far is limited by a number of biological factors:

• The frequency of antibodies against any individual antigen in generally relatively low, typically ranging from 1 to ~30% (104-106).

• Autoantibody repertoires, even in patients with the same type of cancer, are very heterogeneous.

• Autoantibodies against a number of TAAs, such as p53, cyclin B1, c-MYC etc, have been found in patients with different cancers (80), thus they have a limited potential to discriminate between different types of cancers.

• The repertoire of cancer-associated autoantibodies partially overlaps with that in inflammatory, autoimmune and viral disorders and partially resemble the immune response induced by tissue damage (9,107,108), therefore the validation of an autoantibody as truly cancer-associated antibody would require testing large cohorts of patients with various disorders.

• The role of autoantibodies in anti-tumour immune response and their clinical significance is still elusive and controversial.

Protein microarray technology is a valuable tool for exploring autoantibody profiles in human sera and defining the repertoire of the humoral immune response to cancer (109). However, the development of such high-throughput technologies raise volume of the obtained data that bring the original matters up and demand more subtle and complex optimization of the data analysis. The original paper I describes the development of phage-displayed antigen microarray (PhD-AM) technology and the comparison of suitability of T7 and λ phage display systems for the production of microarrays.

4.1 . Reproducibility and sensitivity of PhD-MA technique

The protocol for production of PhD-MA was developed by addressing the following issues: choice of slide surface chemistry, choice of method for the amplification of phages, antibody dilutions, serum preabsorbtion, printing conditions and data acquisition. The technique showed a variability that is generally acceptable for the immunoassays. The sensitivity of PhD-MA is comparable with plaque immunoscreening that is the basis of SEREX technique and underlies the identification of the majority of currently known tumour antigens eliciting humoural immune response. Although the immunoscreening is not a quantitative technique, it is perfectly suitable for discrimination between sero-positive and sero-negative signals. The PhD-MA technique maintains the capacity to detected the presence of specific autoantibodies qualitatively and in addition it provides the opportunity to quantify the sero-positive signals. The development of this technique allows us to address a variety of scientifically and clinically relevant questions, such as (i) do the tumour-associated autoantibodies have a diagnostic relevance? (ii) When during the cancer development they appear? (iii) Is the production of autoantibodies related to the metastatic spread, size of the primary tumor, its histological type, localization or grade? (iv) Do they have a prognostic relevance? (v) What do they tell about the status of patient's immune system and are they predictive of response to immunotherapy?

4.2 . Data processing and normalisation

A new normalization approach has been elaborated for our studies described in the original papers II, III and IV.

The systematic or random variability of multiple sources are possible during the manufacturing, processing and scanning of microarrays. It can arise from technical sources: the heterogeneity of the surface of the slide, the differences between printing groups, difference in sample preparation, fluorescence of labels, intensity-dependent variation in two fluorescent dyes, total serum activity and variation of amount of biological material in printed spots. The possibility to allocate close to pure sample when we can guarantee identical quantity of an operating part in all spots of the array not always there. To remove technical variation we should apply normalization to obtained data.

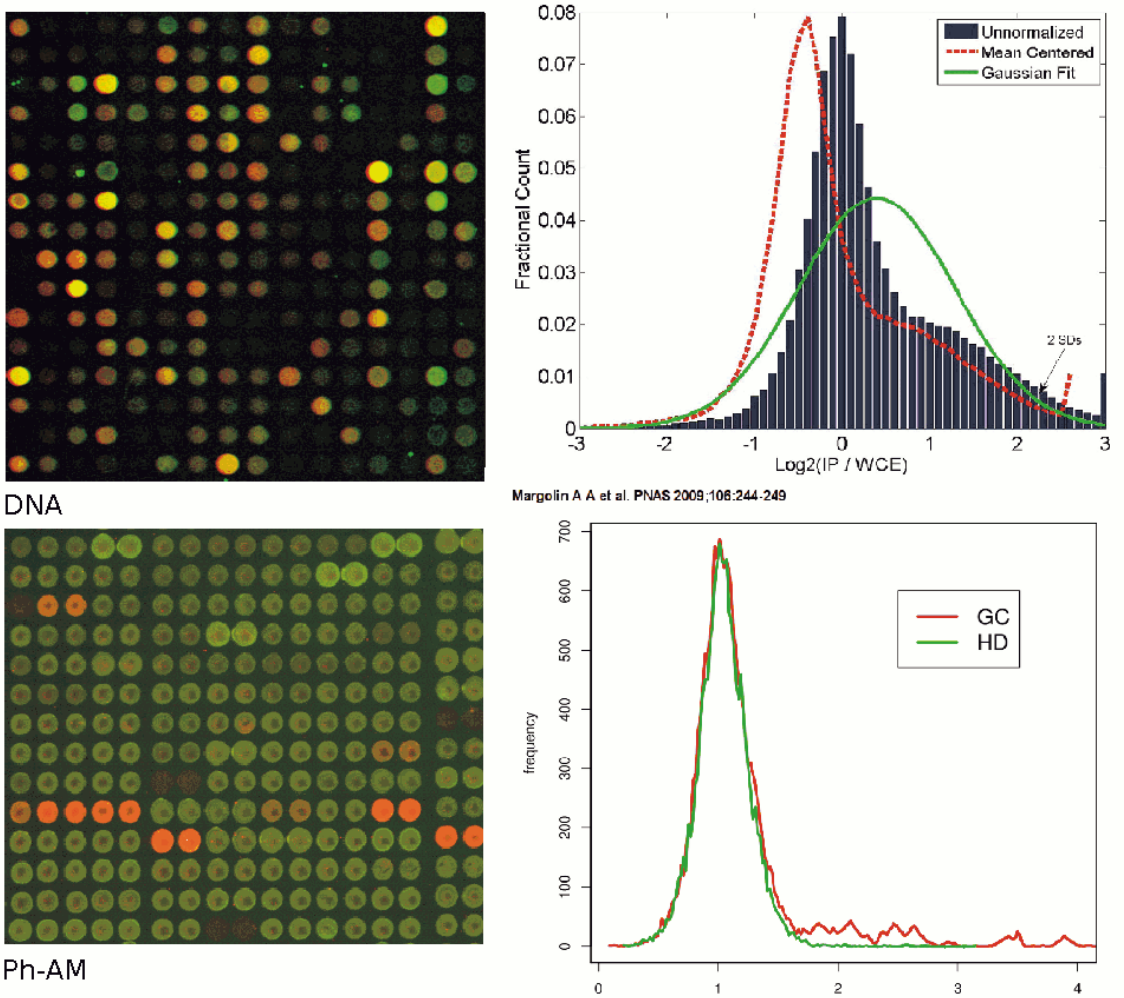

Figure 4. DNA microarray and PhD-AM images and histograms showing the data distribution.

In comparison to DNA microarray protein microarrays are more variable by the complexity of proteomic. The functional protein microarray despite its methodological similarity does not always tolerate the approaches developed for DNA microarray due to additional noise sources (92) - greater chemical diversity of the protein molecules and larger amplitude of the specificity in interaction with the samples. This trend is reinforced by PhD-AM, when total lysate contained the protein of interest in a variable proportion to the total

amount of protein applied to the slide. Since the exact content of spots in DNA microarray is known the elimination of spatial effects always in focus of the within array normalization of DNA data. Some of these effects are excluded at the stage of background removal and the procedure should be preserved for the PhD-AM. The classic methods of normalization of DNA microarray data, such as LOWESS and it most modern variation OLIN, make it by aligning on certain areas.

The following methods for DNA microarrays classically applied for normalization were tested in our study: quantile and global normalizations with or without OLIN. OLIN as well as LOWESS approach, implies a near equal number up- and down- regulated signals on all amplitude of intensities. It is not correct for data with sectors of low density of signals. OLIN and LOWESS clearly show the tendency to loose the same qualitative signals on our dataset. In contrast to DNA case, when these methods provide the reasonable solution, in PhD-AM, where biological variance between printed spots is more higher and in most cases exceed residual spatial effects after background subtraction. Usually global and especially quantile normalization affect very strongly on outliers, which as will be shown below, is an extremely important part of data, bringing them closer to the total normal distribution.

The home-made slides used in the study were printed by portions – print batches due to technical reasons. The "batch effect" issue, when the data from multiple separately produced slides is combined, represents a significant challenge for more monotonous DNA microarray as well (110). Our dataset shows that batch effect is the main reason of systematic variability. Thus, the one of the most difficult type of microarray from the standpoint of normalization was used in our study.

The adopted variant of global normalization and correction by batch mean-centering (111) was developed by us and used for the within array and inter array normalization. It is more careful to the outliers and based on specific knowledge about our dataset. There are no more than 20% coverage of positive serum signals for each of the antigens, and no more than 20% of the antigens react with a single serum, as can be seen by evaluating the preliminary results (Fig. 4). Each microarray batch was hybridized by mixed group of serum (usually about 30% is from healthy donors and other serums from patients with different types of cancer), so we can be sure that no more 10% of signals for each antigen can be positive. One of ways to achieve acceptable reproducibility due spatial effects in DNA microarray is simply increasing the number of local control spots (112). In our proposed method the 80% of spots were used as base group for normalization.

4.3 . Methods of data analysis

The identification of statistically significant changes in the majority of microarray studies is based on the classical methods such as t-test, ANOVA, Bayesian networks or relatively new computationally expensive machine learning algorithms: K nearest neighbor, Linear Discriminant analysis, Artificial Neuron Networks and Support Vector Machines (SVM) (113,114). The latest is one of most powerful modern classifiers, which works by separating virtual high dimensional space by optimal hyper plane that minimizes boundaries between two classes.

Support Vector Machines. We apply SVM with standard RBF kernel and proposed in our study topological rank-based kernel to our PhD-AM data from patients with gastric cancer, melanoma, prostate, gastrointestinal inflammatory diseases and healthy individuals as it described in original papers II. Preliminary analysis showed no statistically significant difference of means for majority antigens. For such the dataset clearly seen one of weakness of standard SVM kernels, as well as many other analytical methods, named "curse of

dimensionality". The common point of these problem for methods that requires statistic significance is that when the dimensionality increases, the volume of the virtual high dimensional space increases so fast that the available data becomes sparse.

Ranking-based topological kernel. Our improvement to conventional linear kernel model of SVM can be characterized as transition from quantitative to qualitative approach, when data during preprocessing were converted from virtually euclidean space to ranked array of antigens for each sample thus mitigates possible overfitting effects. Our proposed SVM approach include Multiple Kernel Learning extension which can improve performance in comparison to single kernel algorithm. Each basis kernel may either use the full set of attributes or subsets of it.

Ranking-based topological kernel significantly increase classification rate in compare to RBF kernel and have good generalization on unseen data as it had been showed by verification sets. Unfortunately it have same disadvantages that is important for development of diagnostic or prognostic signatures of autoantibody.

Our proposed kernel like conventional SVM classifier, as well as classic statistical tests, like t-test or ANOVA based on comparison of mean/median and proposed distribution is looking for significant changes in the majority of samples belonging to two classes being compared even if this changes are very small in absolute scale (115). This approach is not applicable for study, where is preferred biomarkers suitable for diagnostic.

Cutoff and ranking based procedure. An approach more oriented on serological reactivity of autoantibodies has been developed for our study described in original paper IV. We suggested that the serum-reactive signals of cancer patients consist from two groups: first group coincides with serum-reactive signals of healthy patients and second small group received from patients who have the highly reactive specific antibodies to a particular antigen. The data of last patient group can be interpreted as the outliers relatively to the data from healthy individuals and the rest of the cancer patients. Outliers are defined as observations that are far from main part of data. In methods conventionally used for analysis of microarray data, the outliers often are treated as low-quality data and most of the methods offer instruments to remove outliers or replace with acceptable data within the main distribution (116,117). We stress that such outliers as most important part of our dataset (Fig. 5).

The best threshold or "cutoff" for distinguish between positive and negative results typically accepted in practice of biological statistic is mean or median of the negative control group plus two or three times the standard deviation. This coefficient comes from the assumption that the data have a distribution close to the Poisson distribution. In this case, about 95.45% of the values lie within 2 standard deviations. Nearly all (99.73%) of the values lie within 3 standard deviations. We tested our approach with cut-off at coefficient equal to 2, 3, 4 and 5. Threshold at is median of the negative control group plus three times the standard deviation show the smallest error in classification with Leave-One-Out cross-validation method.

We developed a model that have significant advantages for biologists by clear demonstration of most powerful antigens in comparison to SVM that is "black box" system. It gives the classification results and does not provide reasons of classification. Our model based on rank assigned to each antigen accordingly to intensities of positive signals within cancer patients compared to healthy donors. Then scoring of each serum can be calculated. Formula described in section *Materials and Methods*.

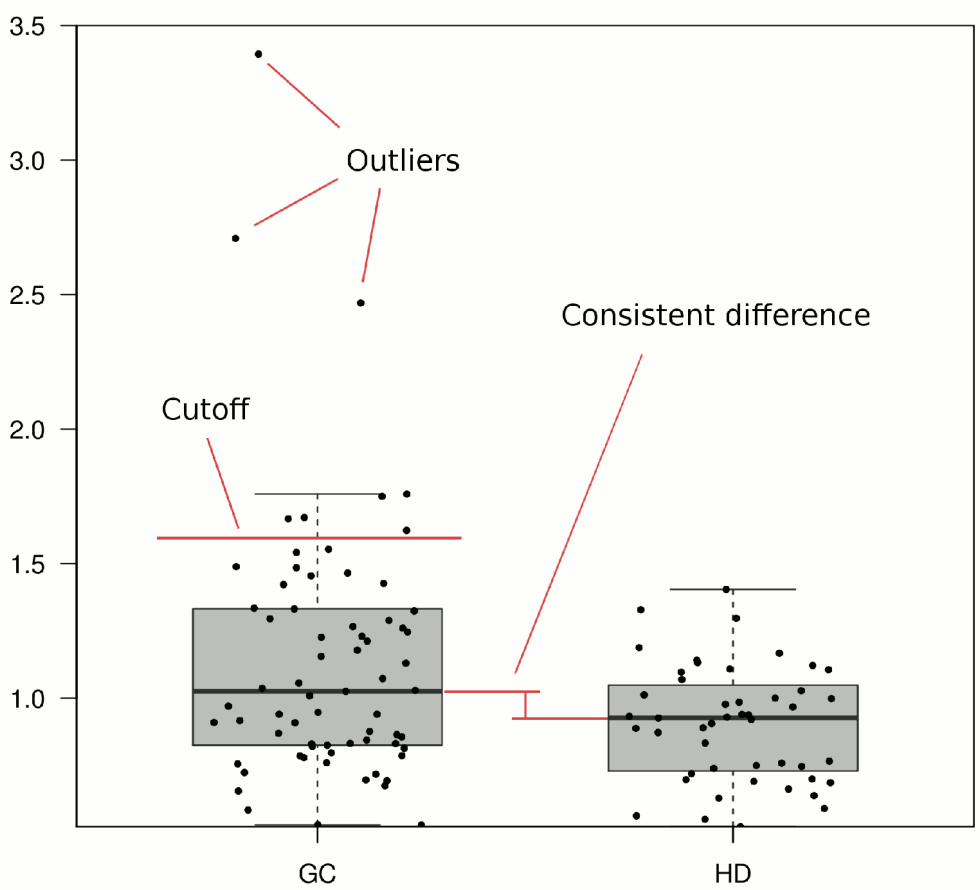

Figure 5. Autoantibody responses in sera of patients with gastric cancer and healthy donors.

Our tests on data of described in original paper IV demonstrate importance of cut-off for SVM with standard and improved Ranking-based topological kernels. Performance of classification were validated with Leave-One-Out method and show 12% difference in performance on data before(AUC 71%) and after(AUC 83%) cutoff applied. Thus stress importance of use of serologically defined cutoff.

4.4 . Analyses of sperm-associated antigens

Among the antigens previously identified by PhD-SEREX approach was a members of the sperm associated antigen (SPAG) group - SPAG8, for which a melanoma associated serum response was demonstrated in our initial microarray-mediated autoantibody profiling of cancer patients' sera (see original paper I). SPAG group antigens share a common expression pattern in male germ cells and infertility-related autoimmunity, as well as recent indications of oncogenic potential (118-129). The germ-cell associated expression, immunogenicity and oncogenicity are the characteristics of a typical cancer-testis (CT) antigen with a potential application in tumour immunotherapy (130). Hence SPAG8 together with the rest of the members of the SPAG group were chosen for in depth analyses with a goal to determine their potential use as novel tumour immunotherapy targets, including mRNA and protein expression analyses in various normal and tumour tissues and the evaluation of the frequency of spontaneous humoral immune response in cancer patients against these antigens. The elaborated PhD-AM production approach was applied to generate a focused SPAG group and

CT antigen microarray on 16-pad nitrocelullose slides by printing the recombinant phages in triplicate. It was used for the screening of 39 breast cancer, 33 colon cancer, 24 lung cancer, 28 thyroid cancer, 28 leukemia, 172 gastric cancer, 52 prostate cancer and 163 melanoma patients' sera. The elaborated 2-step normalization was applied to the obtained data and the seropositivity cutoff was set as four standard deviations above the average signal intensity of 70 non-recombinant phage clones.

Cancer-associated serum response was demonstrated for five of the analysed SPAG antigens (SPAG1, SPAG6, SPAG8 and SPAG17-A1), whose mRNA expression was also elevated in various tumour tissues. All of these were expressed predominantly in testis among normal tissues except for SPAG1 as its mRNA was highly expressed also in normal colon, and ascribe these as novel CT antigens with potential applications in tumour immunotherapy or serodiagnosis.

This study showed that the established autoantibody screening approach can be successfully utilized to analyze humoral immune responses against many cancer types and yields reliable and reproducible results as clearly demonstrated by a further focused microarray generated to validate the specific autoantibody response against SPAG17 novel splice variant SPAG17-A1 (see original paper 3).

4.5 . Identification of autoantibody signature for diagnosis of gastric cancer

The early detection of gastric cancer (GC) is hampered by the lack of specific symptoms before it has spread beyond the original site and the lack of reliable non-invasive screening tests. Detection at late stages leads to application of more radical treatment possessing serious side effects, much lower level of survival rate (less than 20% in 5 years period, in comparison with 75% at detection at early stages)(131,132). Aim of the study described in original paper IV is to discover autoantibody signatures with diagnostic significance and examine the correlation of the autoantibody signatures with the clinicopathological features. We applied T7 phage display-based SEREX technique to identify a representative set of antigens eliciting humoral responses in GC patients. For the production of 1150-feature antigen microarray, a panel of all different serum-reactive phage clones selected from GC cDNA libraries, phage clones previously selected from T7 phage-displayed testis, melanoma and prostate cancer cDNA expression libraries and non-recombinant control phages was assembled and arrayed onto FAST slides. The microarray was tested with sera from 100 patients with GC and 100 cancer-free controls. In total, 888 antigens reacted with at least one of tested sera. Many of them reacted at a similar frequency with cancer patients' and healthy donors' sera.

The cutoff selection approach described in the section *Cutoff and ranking based procedure* was applied and experimentally validated by plaque assay for 3 antigens – CTAG1B, HORMAD1 and SPAG17 using 3 positive and 3 negative sera according to the microarray data.

The top-ranked 86 antigens were used for the production of focused array. Test of the reproducibility of the signals made for randomly selected 50 patients with gastric cancer and 20 healthy donors on 1150-feature and 96-feature arrays showed with an average coefficient of variability 13%. It is an acceptable level for immunoasseys (103).

Focused array was tested with an independent validation set comprising serum samples from 239 GC patients, 150 peptic ulcer and gastritis patients and 213 healthy controls.

The mean serum score differed between GC patients and healthy donors by more than 43-fold. The 86 antigen set could discriminate GC patients and healthy donors with AUC of 0.76 (95% CI 0.74-0.79, $P= 6\times10^{-25}$, Sn 49%, Sp 93%).

To reduce the number of biomarkers we applied backward elimination approach and found the minimal set of 45 antigens that keeps the same level of sensitivity. ROC curve analysis showed that 45-autoantibody signature could discriminate GC and healthy controls with AUC of 0.79 (95% CI 0.76-0.81,P= 6×10^{-31}, Sn 58%, Sp 91% at cutoff 2.45), GC and peptic ulcer with AUC of 0.76, and GC and gastritis with AUC of 0.64.

Accuracy of the identified 45-autoantibody signature (0.745) is nearly equal to the accuracy (0.74) of serum pepsinogen test (SPT) that have a sensitivity 77% and specificity 73% (at pepsinogen I level ≤ 70 ng/ml and pepsinogen I/II ≤ 3) and is widely used in Japan (63). Known disadvantage of SPT is a strong correlation with intestinal-type cancer only. We found that the serum score did not correlate with the histological type. It could be of interest to correlate the results of SPT test and autoantibody profiling as the combination of these approaches can open new possibilities for improving diagnostic accuracy.

Evidences on other known gastric cancer biomarkers show for CEA, CA19-9, CA50, CA72-4 highest diagnostic accuracy 0.53, 0.64, 0.59 and 0.75 respectively (64-68,72). However, all this antigens shared between different types of cancer showing higher frequency in other, such as pancreatic or colorectal, cancers (69).

Although this study gave some insight into the pathological processes associated with the cancer-associated autoantibody production, generally their biological significance and the role in the cancer development and anti-tumour immune response is unknown. A large number of wild-type self antigens analysed in this study were shown to elicit autoantibody production at similar frequencies in cancer patients and cancer-free healthy individuals. In the context of cancer immunoediting concept proposed by Robert Schreiber (37), it would be tempting to speculate that these antibodies reflect the immune response during the elimination phase of nascent tumour. However, detailed molecular examination of these antigens performed by Dr. Siliņa (PhD thesis), failed to reveal somatic mutations or cancer-associated expression pattern of these antigens thus making it less likely that these antigens may have served as targets for the protective immune response. Conversely, we and others have rarely observed the autoantibody production against tumour associated antigens, such as CT antigens, in cancer-free controls (104,133). Immunogenicity of CT antigens is thought to be related to their expression pattern: they are expressed in a wide range of cancers, but normally their expression is restricted to immuno-privileged tissues such as testis, fetal ovary and placenta. Hence they become exposed to the immune system only when expressed by tumors, therefore the immune response against these antigens could serve as a very specific indicator of cancer. If the immunosurveillance of nascent tumours is a general phenomenon, it could be expected that the immune responses against tumour-associated antigens, such as CT antigens, could be detected at a relatively high frequency in individuals without clinically detectable tumour. Hence, at the first glance, our data seems to argue against the cancer immunoediting concept. However, it has been demonstrated that the CT antigen expression tends to correlate with an advanced stage of the disease and is higher in metastatic than primary tumors (130). Thus it could be possible that tumors start to express CT antigens only after they have undergone elimination and equilibrium stages and their expression represents a feature of immunologically sculpted/advanced cancers. Alternatively, it could be possible that the elimination and equilibrium phases of immunoediting process rely entirely on T and/or NK cells and that B cells and/or autoantibody production are not involved in these processes and triggering B cell response requires a specific signal (e.g. sufficiently large tumour mass, necrosis of tumour cells etc.) absent at the early phases of tumour development. Another possibility is that B cells may promote the cancer development. In fact, it has been demonstrated that B cells have a key role in the initiation of tumor-promoting inflammation

(134), and that B cells repress antitumour immunity in TNF-α-dependent manner (135), hence it could be possible that the appearance of tumor-specific antibodies marks the point-of-no-return in the cancer development/immunoediting process. In line with this, Willimsky G et al (136) has demonstrated that the initiation of B cell response to a tumor antigen coincides with the development of tumor-specific tolerance and cytotoxic T cell unresponsiveness.

Hence, it would be of great theoretical interest and clinical relevance to explore further the autoantibody profiles in various cohorts of patients, including the samples collected before the clinical diagnosis, as well as in patients undergoing immunotherapy, and to correlate the data with the prognostic information and the response to immunotherapy.

5. CONCLUSIONS

- The technology for production and processing of phage-displayed antigen microarrays was elaborated and it showed similar sensitivity to the phage plaque assay.

- The intra-assay variability of the phage-displayed antigen microarray technology was 7%, while the average CV was ~13% in the inter-assay comparisons, which is generally acceptable variability for immunoassays.

- The application of phage-displayed antigen microarray technology for the characterization of the humoral immune response against sperm associated antigens revealed cancer-associated spontaneous humoral immune response against SPAG1, SPAG6, SPAG8 and a novel testis-restricted splice variant SPAG17-A1 that allowed to classify them as novel CT antigens with potential relevance as immunotherapeutic targets and serological biomarkers.

- Support vector machine using the improved kernel generated the biomarker model with higher classification accuracy for discriminating sera from cancer patients and healthy controls. However, the detailed examination of the biomarker model showed that it is based on the detection of small but consistent differences in the signal intensities between the cases and controls.

- The conventional methods for the normalization and statistical analysis commonly applied for DNA microarray data analysis are not suitable for the analysis of autoantibody profiles.

- The procedures for two-step normalization of microarray data, defining cutoffs and ranking of antigens were elaborated and allowed to discriminate between sero-positive and sero-negative cases and to perform quantitative analysis.

- Application of the phage-displayed antigen microarray technology and the developed data analysis approach for the exploration of autoantibody profiles in patients with gastric cancer gastric, inflammatory diseases and healthy individuals resulted in the identification of 45- autoantibody signature that could discriminate between GC and healthy control sera with 74.5% accuracy (AUC of 0.79, 58% sensitivity and 91% specificity), GC and peptic ulcer with 73.0% accuracy, and GC and gastritis with 63.5% accuracy.

- The identified 45-autoantibody signature could detect stage I GC with equal sensitivity than advanced GC thus demonstrating its relevance for the early detection of GC.

- At least in GC, the autoantibody production does not correlate with histological type, *H. pylori* status, grade, localization and size of the primary tumor while it appears to be associated with the metastatic disease.

MAIN THESIS OF DEFENSE

I. The phage-displayed antigen microarray technology is a valuable tool for the exploration of autoantibody profiles in human sera and it has a potential to revel autoantibody signatures with the diagnostic relevance.

II. The microarray data analysis approach that allows to discriminate between sero-positive and sero-negative cases and to perform quantitative analysis is of paramount importance for the analysis of autoantibody profiles.

III. The identified 45-autoantibody signature has higher specificity than currently known gastric cancer serum biomarkers and is applicable for the early detection of gastric cancer.

ACKNOWLEDGEMENTS

This research was supported by:
ERDF project No 2010/0231/2DP/2.1.1.1.0/10/APIA/VIAA/044, grant No 09.1288;
European Social Fund project "Support for Doctoral Studies at the University of Latvia" Nr.2009/0138/1DP/1.1.2.1.2/09/IPIA/VIAA/004;

I would like to thank to:
My supervisor Dr. Aija Linē for direction, assistance, and guidance.

My colleagues from Latvian Biomedical Research and Study Centre:
Dr. Zane Kalniņa,
Vilens Jumuts,
Dr. Karina Siliņa,
Elīna Zandberga,
Irēna Meistere,
Diāna Andrējeva,
Angelina Pismennaja,
Artūrs Ābols,
Lāsma Ivanova,
Edgars Endzeliņš.

Our collaborators:

Genome Database of Latvian population,

Latvian Oncology Center,

Clinic of Gastroenterology, Hepatology and Infectious Diseases,
Otto-von-Guericke University Magdeburg, Germany.

References

[1] Jemal A, Bray F, Center MM, Ferlay J, Ward E & Forman D. Global cancer statistics. *CA Cancer J Clin* (2011) **61**: pp. 69-90.

[2] Catalona WJ, Smith DS, Ratliff TL & Basler JW. Detection of organ-confined prostate cancer is increased through prostate-specific antigen-based screening. *JAMA* (1993) **270**: pp. 948-954.

[3] Türeci O, Sahin U & Pfreundschuh M. Serological analysis of human tumor antigens: molecular definition and implications. *Mol Med Today* (1997) **3**: pp. 342-349.

[4] Preuss K, Zwick C, Bormann C, Neumann F & Pfreundschuh M. Analysis of the B-cell repertoire against antigens expressed by human neoplasms. *Immunol Rev* (2002) **188**: pp. 43-50.

[5] Anderson KS & LaBaer J. The sentinel within: exploiting the immune system for cancer biomarkers. *J Proteome Res* (2005) **4**: pp. 1123-1133.

[6] Bei R, Masuelli L, Palumbo C, Modesti M & Modesti A. A common repertoire of autoantibodies is shared by cancer and autoimmune disease patients: Inflammation in their induction and impact on tumor growth. *Cancer Lett* (2009) **281**: pp. 8-23.

[7] Tan HT, Low J, Lim SG & Chung MCM. Serum autoantibodies as biomarkers for early cancer detection. *FEBS J* (2009) **276**: pp. 6880-6904.

[8] Kalniņa Z, Siliņa K, Meistere I, Zayakin P, Rivosh A, Abols A, Leja M, Minenkova O, Schadendorf D & Linē A. Evaluation of T7 and lambda phage display systems for survey of autoantibody profiles in cancer patients. *J Immunol Methods* (2008) **334**: pp. 37-50.

[9] Preiss S, Kammertoens T, Lampert C, Willimsky G & Blankenstein T. Tumor-induced antibodies resemble the response to tissue damage. *Int J Cancer* (2005) **115**: pp. 456-462.

[10] Pecorino L. . *Molecular Biology of Cancer* (2005) **2 – 27.**: .

[11] Gil J, Stembalska A, Pesz KA & Sasiadek MM. Cancer stem cells: the theory and perspectives in cancer therapy. *J Appl Genet* (2008) **49**: pp. 193-199.

[12] Hanahan D & Weinberg RA. The hallmarks of cancer. *Cell* (2000) **100**: pp. 57-70.

[13] Cheng N, Chytil A, Shyr Y, Joly A & Moses HL. Transforming growth factor-beta signaling-deficient fibroblasts enhance hepatocyte growth factor signaling in mammary carcinoma cells to promote scattering and invasion. *Mol Cancer Res* (2008) **6**: pp. 1521-1533.

[14] Bhowmick NA, Neilson EG & Moses HL. Stromal fibroblasts in cancer initiation and progression. *Nature* (2004) **432**: pp. 332-337.

[15] Dreesen O & Brivanlou AH. Signaling pathways in cancer and embryonic stem cells. *Stem Cell Rev* (2007) **3**: pp. 7-17.

[16] Curto M, Cole BK, Lallemand D, Liu C & McClatchey AI. Contact-dependent inhibition of EGFR signaling by Nf2/Merlin. *J Cell Biol* (2007) **177**: pp. 893-903.

[17] Partanen JI, Nieminen AI & Klefstrom J. 3D view to tumor suppression: Lkb1, polarity and the arrest of oncogenic c-Myc. *Cell Cycle* (2009) **8**: pp. 716-724.

[18] Collado M & Serrano M. Senescence in tumours: evidence from mice and humans. *Nat Rev Cancer* (2010) **10**: pp. 51-57.

[19] Evan GI & d'Adda di Fagagna F. Cellular senescence: hot or what?. *Curr Opin Genet Dev* (2009) **19**: pp. 25-31.

[20] Ghebranious N & Donehower LA. Mouse models in tumor suppression. *Oncogene* (1998) **17**: pp. 3385-3400.

[21] Adams JM & Cory S. The Bcl-2 apoptotic switch in cancer development and therapy. *Oncogene* (2007) **26**: pp. 1324-1337.

[22] Wertz IE & Dixit VM. Regulation of death receptor signaling by the ubiquitin system. *Cell Death Differ* (2010) **17**: pp. 14-24.

[23] Blasco MA. Telomeres and human disease: ageing, cancer and beyond. *Nat Rev Genet* (2005) **6**: pp. 611-622.

[24] Artandi SE & DePinho RA. Telomeres and telomerase in cancer. *Carcinogenesis* (2010) **31**: pp. 9-18.

[25] Kang HJ, Choi YS, Hong S, Kim K, Woo R, Won SJ, Kim EJ, Jeon HK, Jo S, Kim TK, Bachoo R, Reynolds IJ, Gwag BJ & Lee H. Ectopic expression of the catalytic subunit of telomerase protects against brain injury resulting from ischemia and NMDA-induced neurotoxicity. *J Neurosci* (2004) **24**: pp. 1280-1287.

[26] Masutomi K, Possemato R, Wong JMY, Currier JL, Tothova Z, Manola JB, Ganesan S, Lansdorp PM, Collins K & Hahn WC. The telomerase reverse transcriptase regulates chromatin state and DNA damage responses. *Proc Natl Acad Sci U S A* (2005) **102**: pp. 8222-8227.

[27] Ferrara N. Vascular endothelial growth factor. *Arterioscler Thromb Vasc Biol* (2009) **29**: pp. 789-791.

[28] Baeriswyl V & Christofori G. The angiogenic switch in carcinogenesis. *Semin Cancer Biol* (2009) **19**: pp. 329-337.
[29] Talmadge JE & Fidler IJ. AACR centennial series: the biology of cancer metastasis: historical perspective. *Cancer Res* (2010) **70**: pp. 5649-5669.
[30] Thiery JP, Acloque H, Huang RYJ & Nieto MA. Epithelial-mesenchymal transitions in development and disease. *Cell* (2009) **139**: pp. 871-890.
[31] Qian B & Pollard JW. Macrophage diversity enhances tumor progression and metastasis. *Cell* (2010) **141**: pp. 39-51.
[32] Gupta GP, Minn AJ, Kang Y, Siegel PM, Serganova I, Cordón-Cardo C, Olshen AB, Gerald WL & Massagué J. Identifying site-specific metastasis genes and functions. *Cold Spring Harb Symp Quant Biol* (2005) **70**: pp. 149-158.
[33] Kim M, Oskarsson T, Acharyya S, Nguyen DX, Zhang XH, Norton L & Massagué J. Tumor self-seeding by circulating cancer cells. *Cell* (2009) **139**: pp. 1315-1326.
[34] Hanahan D & Weinberg RA. Hallmarks of cancer: the next generation. *Cell* (2011) **144**: pp. 646-674.
[35] Hsu PP & Sabatini DM. Cancer cell metabolism: Warburg and beyond. *Cell* (2008) **134**: pp. 703-707.
[36] Dunn GP, Old LJ & Schreiber RD. The immunobiology of cancer immunosurveillance and immunoediting. *Immunity* (2004) **21**: pp. 137-148.
[37] Dunn GP, Bruce AT, Ikeda H, Old LJ & Schreiber RD. Cancer immunoediting: from immunosurveillance to tumor escape. *Nat Immunol* (2002) **3**: pp. 991-998.
[38] Smyth MJ, Dunn GP & Schreiber RD. Cancer immunosurveillance and immunoediting: the roles of immunity in suppressing tumor development and shaping tumor immunogenicity. *Adv Immunol* (2006) **90**: pp. 1-50.
[39] Eyles J, Puaux A, Wang X, Toh B, Prakash C, Hong M, Tan TG, Zheng L, Ong LC, Jin Y, Kato M, Prévost-Blondel A, Chow P, Yang H & Abastado J. Tumor cells disseminate early, but immunosurveillance limits metastatic outgrowth, in a mouse model of melanoma. *J Clin Invest* (2010) **120**: pp. 2030-2039.
[40] Vesely MD, Kershaw MH, Schreiber RD & Smyth MJ. Natural innate and adaptive immunity to cancer. *Annu Rev Immunol* (2011) **29**: pp. 235-271.
[41] Willimsky G & Blankenstein T. Sporadic immunogenic tumours avoid destruction by inducing T-cell tolerance. *Nature* (2005) **437**: pp. 141-146.

[42] Visvader JE & Lindeman GJ. Cancer stem cells in solid tumours: accumulating evidence and unresolved questions. *Nat Rev Cancer* (2008) **8**: pp. 755-768.
[43] Feinberg AP, Ohlsson R & Henikoff S. The epigenetic progenitor origin of human cancer. *Nat Rev Genet* (2006) **7**: pp. 21-33.
[44] Esteller M. Epigenetics in cancer. *N Engl J Med* (2008) **358**: pp. 1148-1159.
[45] Coffelt SB, Lewis CE, Naldini L, Brown JM, Ferrara N & De Palma M. Elusive identities and overlapping phenotypes of proangiogenic myeloid cells in tumors. *Am J Pathol* (2010) **176**: pp. 1564-1576.
[46] Bonnet D & Dick JE. Human acute myeloid leukemia is organized as a hierarchy that originates from a primitive hematopoietic cell. *Nat Med* (1997) **3**: pp. 730-737.
[47] Reya T, Morrison SJ, Clarke MF & Weissman IL. Stem cells, cancer, and cancer stem cells. *Nature* (2001) **414**: pp. 105-111.
[48] Wicha MS, Liu S & Dontu G. Cancer stem cells: an old idea--a paradigm shift. *Cancer Res* (2006) **66**: p. 1883-90; discussion 1895-6.
[49] Galli R, Binda E, Orfanelli U, Cipelletti B, Gritti A, De Vitis S, Fiocco R, Foroni C, Dimeco F & Vescovi A. Isolation and characterization of tumorigenic, stem-like neural precursors from human glioblastoma. *Cancer Res* (2004) **64**: pp. 7011-7021.
[50] Singh SK, Hawkins C, Clarke ID, Squire JA, Bayani J, Hide T, Henkelman RM, Cusimano MD & Dirks PB. Identification of human brain tumour initiating cells. *Nature* (2004) **432**: pp. 396-401.
[51] Harper LJ, Piper K, Common J, Fortune F & Mackenzie IC. Stem cell patterns in cell lines derived from head and neck squamous cell carcinoma. *J Oral Pathol Med* (2007) **36**: pp. 594-603.
[52] Ricci-Vitiani L, Lombardi DG, Pilozzi E, Biffoni M, Todaro M, Peschle C & De Maria R. Identification and expansion of human colon-cancer-initiating cells. *Nature* (2007) **445**: pp. 111-115.
[53] Quintana E, Shackleton M, Sabel MS, Fullen DR, Johnson TM & Morrison SJ. Efficient tumour formation by single human melanoma cells. *Nature* (2008) **456**: pp. 593-598.
[54] Yeung TM, Gandhi SC, Wilding JL, Muschel R & Bodmer WF. Cancer stem cells from colorectal cancer-derived cell lines. *Proc Natl Acad Sci U S A* (2010) **107**: pp. 3722-3727.
[55] Gupta PB, Chaffer CL & Weinberg RA. Cancer stem cells: mirage or reality?. *Nat Med* (2009) **15**: pp. 1010-1012.

[56] Diamandis EP, Fritche HA, Lilja H, Chan DW & Schwartz MK. Tumor Markers: Physiology, Pathobiology, Technology and Clinical Applications. . eds. Washington (Ed.). AACC Press, 2002.
[57] Kohn EC, Mills GB & Liotta L. Promising directions for the diagnosis and management of gynecological cancers. *Int J Gynaecol Obstet* (2003) **83 Suppl 1**: pp. 203-209.
[58] Rustin GJ, Nelstrop AE, Tuxen MK & Lambert HE. Defining progression of ovarian carcinoma during follow-up according to CA 125: a North Thames Ovary Group Study. *Ann Oncol* (1996) **7**: pp. 361-364.
[59] Kuriyama M, Wang MC, Papsidero LD, Killian CS, Shimano T, Valenzuela L, Nishiura T, Murphy GP & Chu TM. Quantitation of prostate-specific antigen in serum by a sensitive enzyme immunoassay. *Cancer Res* (1980) **40**: pp. 4658-4662.
[60] Thompson IM, Pauler DK, Goodman PJ, Tangen CM, Lucia MS, Parnes HL, Minasian LM, Ford LG, Lippman SM, Crawford ED, Crowley JJ & Coltman CAJ. Prevalence of prostate cancer among men with a prostate-specific antigen level < or =4.0 ng per milliliter. *N Engl J Med* (2004) **350**: pp. 2239-2246.
[61] Catalona WJ, Smith DS, Ratliff TL, Dodds KM, Coplen DE, Yuan JJ, Petros JA & Andriole GL. Measurement of prostate-specific antigen in serum as a screening test for prostate cancer. *N Engl J Med* (1991) **324**: pp. 1156-1161.
[62] Chou R, Croswell JM, Dana T, Bougatous C, Blazina I, Fu R, Gleitsmann K & Koenig H. Screening for Prostate Cancer - A Review of the Evidence for the U.S. Preventive Services Task Force.. *United States Preventive Services Task Force.* (2011) : .
[63] Miki K. Gastric cancer screening using the serum pepsinogen test method. *Gastric Cancer* (2006) **9**: pp. 245-253.
[64] Carpelan-Holmström M, Louhimo J, Stenman UH, Alfthan H & Haglund C. CEA, CA 19-9 and CA 72-4 improve the diagnostic accuracy in gastrointestinal cancers. *Anticancer Res* (2002) **22**: pp. 2311-2316.
[65] Micali B, Florio MG, Venuti A, Artemisia A, Caputo G & Brancato U. Usefulness of carcinoembryonic antigen measurement in gastric juice of patients with gastric disorders. *J Clin Gastroenterol* (1983) **5**: pp. 411-415.
[66] Guadagni F, Roselli M, Amato T, Cosimelli M, Perri P, Casale V, Carlini M, Santoro E, Cavaliere R, Greiner JW & et al.. CA 72-4 measurement of tumor-associated glycoprotein 72 (TAG-72) as a serum marker in the management of gastric carcinoma. *Cancer Res* (1992) **52**:

pp. 1222-1227.
[67] Ishigami S, Natsugoe S, Hokita S, Che X, Tokuda K, Nakajo A, Iwashige H, Tokushige M, Watanabe T, Takao S & Aikou T. Clinical importance of preoperative carcinoembryonic antigen and carbohydrate antigen 19-9 levels in gastric cancer. *J Clin Gastroenterol* (2001) **32**: pp. 41-44.
[68] Marrelli D, Pinto E, De Stefano A, Farnetani M, Garosi L & Roviello F. Clinical utility of CEA, CA 19-9, and CA 72-4 in the follow-up of patients with resectable gastric cancer. *Am J Surg* (2001) **181**: pp. 16-19.
[69] Ni XG, Bai XF, Mao YL, Shao YF, Wu JX, Shan Y, Wang CF, Wang J, Tian YT, Liu Q, Xu DK & Zhao P. The clinical value of serum CEA, CA19-9, and CA242 in the diagnosis and prognosis of pancreatic cancer. *Eur J Surg Oncol* (2005) **31**: pp. 164-169.
[70] Takahashi Y, Takeuchi T, Sakamoto J, Touge T, Mai M, Ohkura H, Kodaira S, Okajima K, Nakazato H. The usefulness of CEA and/or CA19-9 in monitoring for recurrence in gastric cancer patients: a prospective clinical study. *Gastric Cancer* (2003) **6**: pp. 142-145.
[71] Pålsson B, Masson P & Andrén-Sandberg A. Tumour marker CA 50 levels compared to signs and symptoms in the diagnosis of pancreatic cancer. *Eur J Surg Oncol* (1997) **23**: pp. 151-156.
[72] Harrison JD, Stanley J & Morris DL. CEA and CA 19.9 in gastric juice and serum: an aid in the diagnosis of gastric carcinoma?. *Eur J Surg Oncol* (1989) **15**: pp. 253-257.
[73] Guadagni F, Roselli M, Cosimelli M, Ferroni P, Spila A, Casaldi V, Cavaliere F, Carlini M, Garofalo A, Rinaldi G & et al.. Correlation between positive CA 72-4 serum levels and lymph node involvement in patients with gastric carcinoma. *Anticancer Res* (1993) **13**: pp. 2409-2413.
[74] Barua A, Bradaric MJ, Kebede T, Espionosa S, Edassery SL, Bitterman P, Rotmensch J & Luborsky JL. Anti-tumor and anti-ovarian autoantibodies in women with ovarian cancer. *Am J Reprod Immunol* (2007) **57**: pp. 243-249.
[75] Plotz PH. The autoantibody repertoire: searching for order. *Nat Rev Immunol* (2003) **3**: pp. 73-78.
[76] Suzuki H, Graziano DF, McKolanis J & Finn OJ. T cell-dependent antibody responses against aberrantly expressed cyclin B1 protein in patients with cancer and premalignant disease. *Clin Cancer Res* (2005) **11**: pp. 1521-1526.
[77] Trivers GE, Cawley HL, DeBenedetti VM, Hollstein M, Marion MJ, Bennett WP,

Hoover ML, Prives CC, Tamburro CC & Harris CC. Anti-p53 antibodies in sera of workers occupationally exposed to vinyl chloride. *J Natl Cancer Inst* (1995) **87**: pp. 1400-1407.

[78] Trivers GE, De Benedetti VM, Cawley HL, Caron G, Harrington AM, Bennett WP, Jett JR, Colby TV, Tazelaar H, Pairolero P, Miller RD & Harris CC. Anti-p53 antibodies in sera from patients with chronic obstructive pulmonary disease can predate a diagnosis of cancer. *Clin Cancer Res* (1996) **2**: pp. 1767-1775.

[79] Kalnina Z, Zayakin P, Silina K & Linē A. Alterations of pre-mRNA splicing in cancer. *Genes Chromosomes Cancer* (2005) **42**: pp. 342-357.

[80] Zhang J, Casiano CA, Peng X, Koziol JA, Chan EKL & Tan EM. Enhancement of antibody detection in cancer using panel of recombinant tumor-associated antigens. *Cancer Epidemiol Biomarkers Prev* (2003) **12**: pp. 136-143.

[81] Jäger E, Stockert E, Zidianakis Z, Chen YT, Karbach J, Jäger D, Arand M, Ritter G, Old LJ & Knuth A. Humoral immune responses of cancer patients against "Cancer-Testis" antigen NY-ESO-1: correlation with clinical events. *Int J Cancer* (1999) **84**: pp. 506-510.

[82] van Rhee F, Szmania SM, Zhan F, Gupta SK, Pomtree M, Lin P, Batchu RB, Moreno A, Spagnoli G, Shaughnessy J & Tricot G. NY-ESO-1 is highly expressed in poor-prognosis multiple myeloma and induces spontaneous humoral and cellular immune responses. *Blood* (2005) **105**: pp. 3939-3944.

[83] Fosså A, Berner A, Fosså SD, Hernes E, Gaudernack G & Smeland EB. NY-ESO-1 protein expression and humoral immune responses in prostate cancer. *Prostate* (2004) **59**: pp. 440-447.

[84] Tsai-Turton M, Santillan A, Lu D, Bristow RE, Chan KC, Shih I & Roden RBS. p53 autoantibodies, cytokine levels and ovarian carcinogenesis. *Gynecol Oncol* (2009) **114**: pp. 12-17.

[85] Lubin R, Schlichtholz B, Teillaud JL, Garay E, Bussel A & Wild CP. p53 antibodies in patients with various types of cancer: assay, identification, and characterization. *Clin Cancer Res* (1995) **1**: pp. 1463-1469.

[86] Tang R, Ko MC, Wang JY, Changchien CR, Chen HH, Chen JS, Hsu KC, Chiang JM & Hsieh LL. Humoral response to p53 in human colorectal tumors: a prospective study of 1,209 patients. *Int J Cancer* (2001) **94**: pp. 859-863.

[87] Tangkijvanich P, Janchai A, Charuruks N, Kullavanijaya P, Theamboonlers A, Hirsch P & Poovorawan Y. Clinical associations and prognostic significance of serum anti-p53

antibodies in Thai patients with hepatocellular carcinoma. *Asian Pac J Allergy Immunol* (2000) **18**: pp. 237-243.
[88] Haidopoulos D, Konstadoulakis MM, Antonakis PT, Alexiou DG, Manouras AM, Katsaragakis SM & Androulakis GF. Circulating anti-CEA antibodies in the sera of patients with breast cancer. *Eur J Surg Oncol* (2000) **26**: pp. 742-746.
[89] Yang XF, Wu CJ, McLaughlin S, Chillemi A, Wang KS, Canning C, Alyea EP, Kantoff P, Soiffer RJ, Dranoff G & Ritz J. CML66, a broadly immunogenic tumor antigen, elicits a humoral immune response associated with remission of chronic myelogenous leukemia. *Proc Natl Acad Sci U S A* (2001) **98**: pp. 7492-7497.
[90] Vural B, Chen L, Saip P, Chen Y, Ustuner Z, Gonen M, Simpson AJG, Old LJ, Ozbek U & Gure AO. Frequency of SOX Group B (SOX1, 2, 3) and ZIC2 antibodies in Turkish patients with small cell lung carcinoma and their correlation with clinical parameters. *Cancer* (2005) **103**: pp. 2575-2583.
[91] Schena M, Shalon D, Davis RW & Brown PO. Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science* (1995) **270**: pp. 467-470.
[92] Sboner A, Karpikov A, Chen G, Smith M, Mattoon D, Freeman-Cook L, Schweitzer B & Gerstein MB. Robust-linear-model normalization to reduce technical variability in functional protein microarrays. *J Proteome Res* (2009) **8**: pp. 5451-5464.
[93] Quackenbush J. Microarray data normalization and transformation. *Nat Genet* (2002) **32 Suppl**: pp. 496-501.
[94] Statnikov A, Wang L & Aliferis CF. A comprehensive comparison of random forests and support vector machines for microarray-based cancer classification. *BMC Bioinformatics* (2008) **9**: p. 319.
[95] Menjoge RS & Welsch RE. Chapter 2: Comparing and Visualizing Gene Selection and Classification Methods for Microarray Data. In , Machine
Learning in Bioinformatics (pp. 47-68).
. Zhang YQ & Rajapakse JC (Eds.). Hoboken, New Jersey: John Wiley & Sons, Inc., 2009.
[96] R Development Core Team. R: A language and environment for statistical computing.. *R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org/.* (2011) : .
[97] Smyth GK. Limma: linear models for microarray data. In: Bioinformatics and Computational Biology Solutions using R and Bioconductor. Gentleman R, Carey V, Dudoit

S, Irizarry R & Huber W (Eds.). Springer, New York, pages 397--420., 2005.
[98] Futschik ME & Crompton T. OLIN: optimized normalization, visualization and quality testing of two-channel microarray data. *Bioinformatics* (2005) **21**: pp. 1724-1726.
[99] MATLAB. version 7.10.0 (R2010a). . The MathWorks Inc., 2010.
[100] Anderson E, Bai Z, Dongarra J, Greenbaum A, McKenney A, Du Croz J, Hammerling S, Demmel J, Bischof C & Sorensen D. LAPACK: a portable linear algebra library for high-performance computers. In *Proceedings of the 1990 acm/ieee conference on supercomputing*. 1990.
[101] Chang C & Lin C. LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology* (2011) **2**: p. 27:1-27:27.
[102] Rakotomamonjy A, Rouen UD, Bach F, Canu S & Grandvalet Y. Y.: SimpleMKL. *Journal of Machine Learning Research 9* (2008) : .
[103] Bastarache JA, Koyama T, Wickersham NE, Mitchell DB, Mernaugh RL & Ware LB. Accuracy and reproducibility of a multiplex immunoassay platform: a validation study. *J Immunol Methods* (2011) **367**: pp. 33-39.
[104] Stockert E, Jäger E, Chen YT, Scanlan MJ, Gout I, Karbach J, Arand M, Knuth A & Old LJ. A survey of the humoral immune response of cancer patients to a panel of human tumor antigens. *J Exp Med* (1998) **187**: pp. 1349-1354.
[105] Zhang J. Tumor-associated antigen arrays to enhance antibody detection for cancer diagnosis. *Cancer Detect Prev* (2004) **28**: pp. 114-118.
[106] Soussi T. p53 Antibodies in the sera of patients with various types of cancer: a review. *Cancer Res* (2000) **60**: pp. 1777-1788.
[107] Scanlan MJ, Gout I, Gordon CM, Williamson B, Stockert E, Gure AO, Jäger D, Chen YT, Mackay A, O'Hare MJ & Old LJ. Humoral immunity to human breast cancer: antigen definition and quantitative analysis of mRNA expression. *Cancer Immun* (2001) **1**: p. 4.
[108] Ludewig B, Krebs P, Metters H, Tatzel J, Türeci O & Sahin U. Molecular characterization of virus-induced autoantibody responses. *J Exp Med* (2004) **200**: pp. 637-646.
[109] Gnjatic S, Ritter E, Büchler MW, Giese NA, Brors B, Frei C, Murray A, Halama N, Zörnig I, Chen Y, Andrews C, Ritter G, Old LJ, Odunsi K & Jäger D. Seromic profiling of ovarian and pancreatic cancer. *Proc Natl Acad Sci U S A* (2010) **107**: pp. 5088-5093.
[110] Johnson WE, Li C & Rabinovic A. Adjusting batch effects in microarray expression

data using empirical Bayes methods. *Biostatistics* (2007) **8**: pp. 118-127.
[111] Sims AH, Smethurst GJ, Hey Y, Okoniewski MJ, Pepper SD, Howell A, Miller CJ & Clarke RB. The removal of multiplicative, systematic bias allows integration of breast cancer gene expression datasets - improving meta-analysis and prediction of prognosis. *BMC Med Genomics* (2008) **1**: p. 42.
[112] Anderson T, Wulfkuhle J, Liotta L, Winslow RL & Petricoin E3. Improved reproducibility of reverse-phase protein microarrays using array microenvironment normalization. *Proteomics* (2009) **9**: pp. 5562-5566.
[113] Smyth GK. Linear models and empirical bayes methods for assessing differential expression in microarray experiments. *Stat Appl Genet Mol Biol* (2004) **3**: p. Article3.
[114] Chu F & Wang L. Applications of support vector machines to cancer classification with microarray data. *Int J Neural Syst* (2005) **15**: pp. 475-484.
[115] Gnjatic S, Wheeler C, Ebner M, Ritter E, Murray A, Altorki NK, Ferrara CA, Hepburne-Scott H, Joyce S, Koopman J, McAndrew MB, Workman N, Ritter G, Fallon R & Old LJ. Seromic analysis of antibody responses in non-small cell lung cancer patients and healthy donors using conformational protein arrays. *J Immunol Methods* (2009) **341**: pp. 50-58.
[116] McClintick JN, Liu Y & Edenberg HJ. Mapping of trans-acting regulatory factors from microarray data. *BMC Proc* (2007) **1 Suppl 1**: p. S155.
[117] Shieh AD & Hung YS. Detecting outlier samples in microarray data. *Stat Appl Genet Mol Biol* (2009) **8**: p. Article 13.
[118] Beech DJ, Madan AK & Deng N. Expression of PH-20 in normal and neoplastic breast tissue. *J Surg Res* (2002) **103**: pp. 203-207.
[119] Buechler S. Low expression of a few genes indicates good prognosis in estrogen receptor positive breast cancer. *BMC Cancer* (2009) **9**: p. 243.
[120] Ceriani RL, Sasaki M, Sussman H, Wara WM & Blank EW. Circulating human mammary epithelial antigens in breast cancer. *Proc Natl Acad Sci U S A* (1982) **79**: pp. 5420-5424.
[121] Ceriani RL & Blank EW. Experimental therapy of human breast tumors with 131I-labeled monoclonal antibodies prepared against the human milk fat globule. *Cancer Res* (1988) **48**: pp. 4664-4672.
[122] Garg M, Chaurasiya D, Rana R, Jagadish N, Kanojia D, Dudha N, Kamran N, Salhan

S, Bhatnagar A, Suri S, Gupta A & Suri A. Sperm-associated antigen 9, a novel cancer testis antigen, is a potential target for immunotherapy in epithelial ovarian cancer. *Clin Cancer Res* (2007) **13**: pp. 1421-1428.

[123] Garg M, Kanojia D, Khosla A, Dudha N, Sati S, Chaurasiya D, Jagadish N, Seth A, Kumar R, Gupta S, Gupta A, Lohiya NK & Suri A. Sperm-associated antigen 9 is associated with tumor growth, migration, and invasion in renal cell carcinoma. *Cancer Res* (2008) **68**: pp. 8240-8248.

[124] Garg M, Kanojia D, Salhan S, Suri S, Gupta A, Lohiya NK & Suri A. Sperm-associated antigen 9 is a biomarker for early cervical carcinoma. *Cancer* (2009) **115**: pp. 2671-2683.

[125] Guinn B, Bland EA, Lodi U, Liggins AP, Tobal K, Petters S, Wells JW, Banham AH & Mufti GJ. Humoral detection of leukaemia-associated antigens in presentation acute myeloid leukaemia. *Biochem Biophys Res Commun* (2005) **335**: pp. 1293-1304.

[126] Kanojia D, Garg M, Gupta S, Gupta A & Suri A. Sperm-associated antigen 9, a novel biomarker for early detection of breast cancer. *Cancer Epidemiol Biomarkers Prev* (2009) **18**: pp. 630-639.

[127] Madan AK, Yu K, Dhurandhar N, Cullinane C, Pang Y & Beech DJ. Association of hyaluronidase and breast adenocarcinoma invasiveness. *Oncol Rep* (1999) **6**: pp. 607-609.

[128] Neesse A, Gangeswaran R, Luettges J, Feakins R, Weeks ME, Lemoine NR & Crnogorac-Jurcevic T. Sperm-associated antigen 1 is expressed early in pancreatic tumorigenesis and promotes motility of cancer cells. *Oncogene* (2007) **26**: pp. 1533-1545.

[129] Vazquez-Ortiz G, García JA, Ciudad CJ, Noé V, Peñuelas S, López-Romero R, Mendoza-Lorenzo P, Piña-Sánchez P & Salcedo M. Differentially expressed genes between high-risk human papillomavirus types in human cervical cancer cells. *Int J Gynecol Cancer* (2007) **17**: pp. 484-491.

[130] Scanlan MJ, Gure AO, Jungbluth AA, Old LJ & Chen Y. Cancer/testis antigens: an expanding family of targets for cancer immunotherapy. *Immunol Rev* (2002) **188**: pp. 22-32.

[131] Park J & Kim YH. Current approaches to gastric cancer in Korea. *Gastrointest Cancer Res* (2008) **2**: pp. 137-144.

[132] Rosati G, Ferrara D & Manzione L. New perspectives in the treatment of advanced or metastatic gastric cancer. *World J Gastroenterol* (2009) **15**: pp. 2689-2692.

[133] Reuschenbach M, von Knebel Doeberitz M & Wentzensen N. A systematic review of

humoral immune responses against tumor antigens. *Cancer Immunol Immunother* (2009) **58**: pp. 1535-1544.

[134] Andreu P, Johansson M, Affara NI, Pucci F, Tan T, Junankar S, Korets L, Lam J, Tawfik D, DeNardo DG, Naldini L, de Visser KE, De Palma M & Coussens LM. FcRgamma activation regulates inflammation-associated squamous carcinogenesis. *Cancer Cell* (2010) **17**: pp. 121-134.

[135] Schioppa T, Moore R, Thompson RG, Rosser EC, Kulbe H, Nedospasov S, Mauri C, Coussens LM & Balkwill FR. B regulatory cells and the tumor-promoting actions of TNF-α during squamous carcinogenesis. *Proc Natl Acad Sci U S A* (2011) **108**: pp. 10662-10667.

[136] Willimsky G, Czéh M, Loddenkemper C, Gellermann J, Schmidt K, Wust P, Stein H & Blankenstein T. Immunogenicity of premalignant lesions is the primary cause of general cytotoxic T lymphocyte unresponsiveness. *J Exp Med* (2008) **205**: pp. 1687-1700.