Serum Antibodies in Lung Cancer

A proteomics approach in the NELSON trial

ISBN: 978-94-6169-406-5

Serum Antibodies in Lung Cancer
A proteomics approach in the NELSON trial
Thesis, Erasmus University Rotterdam, the Netherlands

© 2013 D. de Costa, Rotterdam, the Netherlands

Cover design: Jacco de Haan
Printed by: Optima Grafische Communicatie, Rotterdam, the Netherlands

The research described in this thesis was funded by Roche Diagnostics and the Netherlands Organization of Health Research and Development (ZonMw).

This thesis was financially supported by the J.E. Jurriaanse Stichting and the Department of Pulmonology, Erasmus MC, Rotterdam, the Netherlands.

Serum Antibodies in Lung Cancer

A proteomics approach in the NELSON trial

Serum antilichamen in longkanker

Een proteomics benadering in de NELSON trial

Proefschrift

ter verkrijging van de graad van doctor aan de
Erasmus Universiteit Rotterdam
op gezag van de
rector magnificus

prof.dr. H.G. Schmidt

en volgens besluit van het College voor Promoties.
De openbare verdediging zal plaatsvinden op
6 september 2013 om 13.30uur

door

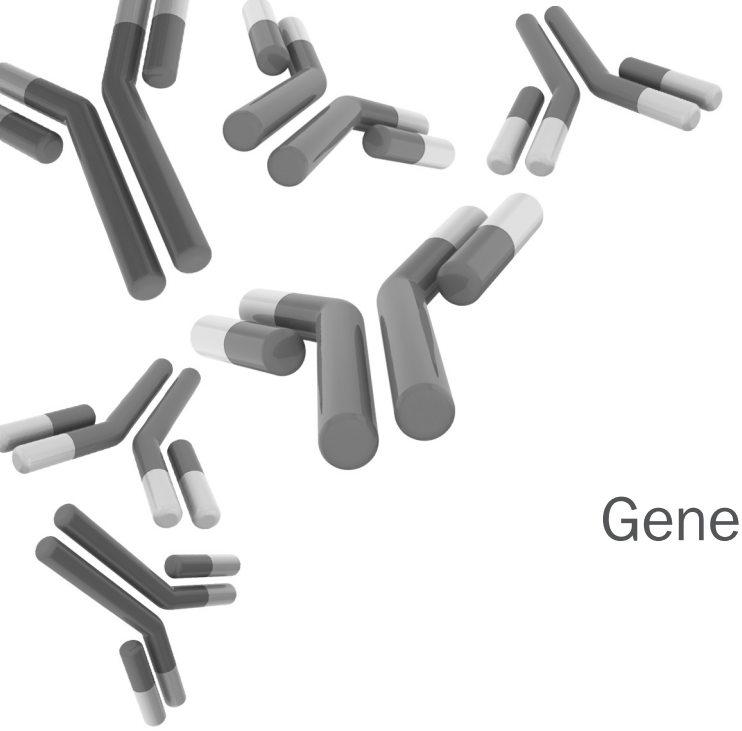
Dominique de Costa
geboren te Rotterdam

Promotiecommissie:

Promotoren:	Prof. dr. H.C. Hoogsteden Prof. dr. P.A.E. Sillevis Smitt
Copromotor:	Dr. Th.M. Luider
Overige leden:	Prof. dr. R.W. Hendriks Prof. dr. H. Hooijkaas Prof. dr. R.Q. Hintzen

voor Opa

Contents



Chapter 1

General introduction

Lung cancer

Incidence and etiology

With approximately 1.6 million cases in 2008 according to GLOBOCAN 2008, lung cancer is currently the cancer with the highest mortality rate (28%) in the World.[1-2] Although this type of cancer is easier to detect compared to other cancers it is difficult to cure, because of poor response to systemic therapy.

The vast majority (80-90%) of lung cancer cases is due to cigarette smoking.[1] Other etiological factors include asbestos, radon gas, ionizing radiation and certain industrial agents and compounds (carcinogens) such as chloromethyl, arsenic, ether, nickel-cadmium and chromium. Tobacco smoking is thought to be synergistic with many of these carcinogens.[3-4]

Lung cancer death among women continues to increase slowly, like the increase of cigarette smoking in women. Passive smoking accounts for 3-5% of all lung cancer cases.[4-5]

The 5-year survival rate for all patients is about 16%, but survival is related to tumor stage and presentation.[3]

The occurrence of lung cancer varies between male and female. The male to female ratio is 3.47 in patients over 45 years of age and 1.7 in patients younger than 45 years.[4-5]

Pathology

Lung cancer is a cancer that forms in tissues of the lung, usually in the cells lining air passages.[6] Primary lung neoplasms are for 95% of epithelial origin (carcinoma), which comprises two different main types: non-small cell lung cancer (NSCLC) and small cell lung cancer (SCLC). These carcinomas can be divided based on clinical and biological features and are diagnosed based on how the cells appear under a microscopic view.

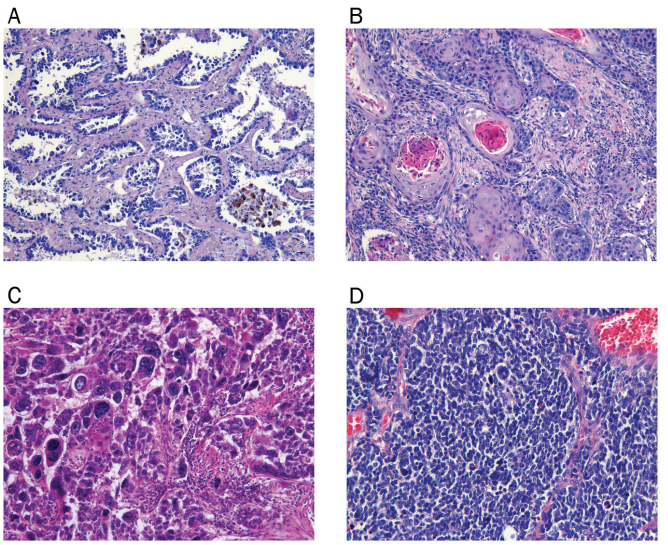


Figure 1. Different types of lung cancer. A) Adenocarcinoma. B) Squamous cell carcinoma. C) Large cell carcinoma. D) Small cell lung carcinoma.

Non-small cell lung cancer consists of three main subtypes, adenocarcinoma (Figure 1a), squamous cell carcinoma (Figure 1b) and large cell carcinoma (Figure 1c). NSCLC comprises 75% of all lung cancers and can be divided in 35% adenocarcinoma, 30% squamous cell carcinoma and 10% large cell carcinoma. Adenocarcinoma originates in two-third of cases from peripheral airways and alveoli (Figure 2). The other third of cases arise centrally in large bronchi (Figure 2) from either the surface epithelium or the submucosal glands. Adenocarcinoma can also spread to other organs with a metastatic frequency of 50-82%.[5]

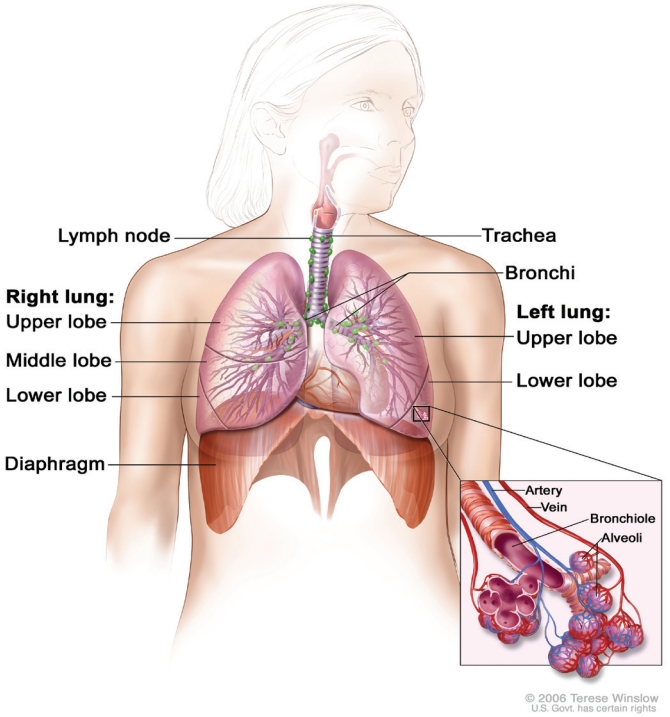


Figure 2. Anatomy of the lung. This figure is adapted from the National Cancer Institute with permission of the illustrater Teresa Winslow.

Most of the squamous cell carcinomas are arising in the central or proximal tracheal-bronchial tree in areas of squamous cell metaplasia and dysplasia. Squamous cell carcinomas are slowly growing tumors and one third of the carcinomas is poorly differentiated.[5] Metastasis can occur in squamous cell carcinoma, but with a lower frequency (25-54%) compared to adenocarcinoma.[5]

Large cell carcinoma is characterized by large cells. Most patients with large cell carcinoma show large, bulky, peripheral tumors. At an early stage metastasis can occur (48-86%) and has a 5-year survival rate of less than 5%.[5]

Small cell lung cancer (SCLC) occurs in approximately 10-20% of all lung cancers (Figure 1d). This type of lung cancer is an extremely fast-growing aggressive cancer that forms in tissue of the lung and is often associated with distant metastases (74-96%).[5] The cancer cells are small and oval-shaped. The prognosis is very poor.[5, 7]

Diagnosis and Staging

Clinical manifestations of lung cancer can differ among patients. Patients can be asymptomatic or symptomatic. The asymptomatic lung cancer patients (5-10%) are most of the time detected during evaluation of an unrelated medical problem or on a chest radiograph at an early stage of cancer development. The symptomatic lung cancer patients are often detected at the time the cancer is already at an advanced stage. The most occurring symptoms are cough, dyspnea, weight loss, chest pain, hemoptysis, bone pain and fatigue.[4] If a patient has any of these symptoms the clinician will ask for a chest X-ray. This X-ray may reveal suspicious areas but from this information a clinician is unable to decide if these areas are cancerous or not. CT scans can be performed when X-rays did not show an abnormality or sufficient information about the size or location of the tumor. Also, for precise location of the tumor an MRI scan can be performed, as CT and MRI scans are focused on the anatomical structures of the tumor, PET (positron emission tomography) scans measures the activity and function of the tissues. They can determine if a tumor is growing and determine the type of cells within the tumor. Besides performing different scans, confirmation of malignant cells is required to make a diagnosis of lung cancer. A pathologist will examine tumor cells for diagnosis. This can be done by sputum cytology, bronchoscopy or needle biopsy through the skin.[4, 6]

To select the most effective treatment the physician needs to know the stage of the tumor. Staging is very important to find out if the cancer has spread to other organs. Lung cancer spreads most often to lymph nodes, brain, bones, liver and adrenal glands. To check if the lung cancer cells have spread to other parts of the body different techniques can be used: CT scan, bone scan, MRI or PET scan.

The International Staging System (ISS) is the most common used system using TNM (primary Tumor; regional lymph Nodes; distant Metastases) categories in staging lung cancer. This system categorizes patients into four stages, I-IV, each having a more progressive and poor survival rate.[8]

High risk population and screening

The risk of dying from lung cancer is associated with cigarette smoking. Eighty to ninety percent of all lung cancer patients are attributable to smoking.[4, 9] Such active cigarette smoking populations have a higher chance to develop lung cancer. Besides high risk of developing lung cancer in these populations, diagnosis occurs most of the time at an advanced stage and complicates even more adequate therapy. Therefore high risk population screening is of great importance. High risk screening populations can be obtained by selecting individuals based on risk factors such as cigarette smoking (smoking history) and non-smoking factors such as professional exposure to asbestos or family history of lung cancer.[10]

Studies have been performed on finding a tool that can be used for early detection of lung cancer and thereby reducing the lung cancer mortality rate.[11-14] All studies have failed until the results of the NLST (National Lung Screening Trial) study were published recently.[15] They showed a lung cancer mortality reduction rate of 20.3% higher in high-risk participants who were screened with low-dose spiral CT compared with participants who were screened by chest X-ray, all with a follow-up of 7 years. Why such a screening tool has not been found much earlier might be due to low sensitivity in detecting a curable stage of lung cancer.

NELSON trial

The NELSON (Nederlands-Leuvens Longkanker Screenings Onderzoek) trial, -the Dutch-Belgian Lung Cancer Screening trial-, is one of the largest randomized controlled lung cancer screening trials. The trial started in 2003 with its purpose to establish a reduction in lung cancer mortality of at least 25% by lung cancer screening using low-dose spiral CT-scan in a high risk population. The second aim was to estimate the cost-effectiveness of lung cancer screening.

Recruitment has started in 2003 by sending questionnaires (smoking history and demographic data) to 548,489 males and females between 50-75 years of age from population registries of 7 public health districts in the Netherlands and population registries of 14 municipalities around Leuven in Belgium. Participants had to be current or former smokers for at least 25 years, smoking at least 15 cigarettes per day or smoking at least 30 years, smoking at least 10 cigarettes per day. Participants with a moderate or bad self-reported health who were unable to climb 2 flights of stairs and weighted over 140 kg were excluded from the trial. Furthermore, people with current or past renal or breast cancer or melanoma were also not included. Individuals with a diagnosis of lung cancer or treatment related to lung cancer within the last five years were excluded and persons who had a chest CT scan less than one year before the first NELSON questionnaire was filled in were also excluded from the trial. Eligible responders were sent a second questionnaire, a NELSON trial information brochure and an informed consent form. From the 548,489 males and females 15,822 participants met the selection criteria and signed the informed consent. These participants were randomized to the screen or control arm. The screen arm received CT screening in years 1, 2 and 4. The control arm participants received no screening (usual care).[16] The CT scans were interpreted by a novel nodule management strategy based on volume and volume doubling time (VDT).[17] Participants with a positive test result were referred to a pulmonologist. If the diagnosis lung cancer was established the patient was treated and went off screening. Participants with an indeterminate test result underwent a follow-up scan three months later. If a negative test result was obtained the second-round CT scan was scheduled for 12 months later. From the screen arm (n=7,915) 70 participants have been diagnosed for lung cancer at baseline. More than 6,600 serum samples have been collected at baseline. For the studies described in this thesis we used serum samples from cases and controls of the screen arm at baseline.

Ongoing studies

Besides the NELSON and NLST studies as described above, more lung cancer screening trials are ongoing. A small overview of the main large-scale randomized lung cancer screening trials is shown in Table 1. The ITALUNG trial started in 2003 and included 3.206 participants.[18] In 2005 the DANTE trial was started with 2.472 participants.[19] The DLCST (Danish Lung Cancer Screening Trial) trial started in 2004 and is comparable with the NELSON trial, but with a smaller number (4.104) of participants.[20] In 2007 the LUSI trial was started with 4.000 participants in Germany and the UKLS (UK lung cancer screening trial) trial has just started (2011-2012) in the UK with 4.000 participants in their pilot study.[21-22] All these trials have the same main objective reducing the lung cancer mortality rate to 20-25%.

All these trials are comparing CT screening vs. chest X-ray or usual care. For CT screening a low-dose spiral CT scan is most of the time used. Different studies have shown

Table 1. Main large-scale randomized controlled lung cancer screening trials.

Trial	Country	Year started	N	Comparison	Years of screening	Age range (years)
NLST	USA	2002	53,456	CT vs. CXR	3	50-74
ITALUNG	Italy	2003	3,206	CT vs. uc*	5	55-69
NELSON	Netherlands/Belgium	2004	15,822	CT vs. uc*	5	50-75
DLCST	Denmark	2004	4,104	CT vs. uc*	5	50-70
DANTE	Italy	2005	2,472	CT vs. clinical review	4	60-74
LUSI	Germany	2007	4,000	CT vs. uc*	5	50-69
UKLS	UK	2011-2012	4,000	CT vs. uc*	Pilot study	50-75

NLST: National Lung Screening Trial; NELSON: Dutch-Belgian lung cancer screening trial; DLCST: Danish Lung Cancer Screening Trial; UKLS: UK lung cancer screening trial; *uc: usual care.

that this type of scan is an appropriate tool for identifying lung cancer at an early stage.[23-27] Some drawbacks of using low-dose spiral CT scan for lung cancer screening is the high rate of false-positive scan results. This results in unnecessary follow-up, additional tests or even surgery, which results in high costs and inconvenience for the patient. Therefore, the American College of Chest Physicians guideline does not recommend CT screening unless it is used as part of a clinical trial.[28]

Biomarkers

Research has been and is still performed on finding a biomarker panel in blood to detect lung cancer at an early stage. Finding a panel that could distinguish lung cancer patients from lung cancer-free patients would be ideal as CT screening has still drawbacks as mentioned above. In the search of a biomarker for lung cancer it is important to know the biology of lung cancer. Lung cancer cells have defects in the pathways which generate normal cell proliferation and homeostasis.[29] Lung cancers are insensitive to growth-inhibitory signals and show limitless replication, tissue invasion, evasion of apoptosis and metastasis. Transformation of normal cells to malignant lung cancer cells occurs in multiple steps in time, initiating tumor genesis from mutations followed by additional/different mutations and epigenetic alterations during clonal expansion where cell growth becomes dominant. These (pre)neoplastic changes in the epithelial cells of the bronchial epithelium cause lung cancer. It is not known if all these cells are sensitive to malignant transformation or if only a subset of these epithelial cells is sensitive for this transformation.[30-31] Identifying these molecular changes is important for prevention, early detection and treatment of the disease.

Lung cancer is highly heterogeneous at the clinical, biological, histological and molecular level. Why this cancer is such a heterogeneous type of cancer is still unclear. This heterogeneity and molecular complexity make it difficult to unravel the pathogenesis of lung cancer. Multiple oncogenes, signaling pathway components, tumor suppressor genes and other cellular processes are involved in the pathogenesis of lung cancer.

Cancer immunology, also known as cancer immunosurveillance, is associated with cancer growth and progression. The hypothesis of immunosurveillance is that the immune system will recognize malignant cells as foreign cells and has the aim to eliminate these cells. This elimination of the cells occurs in the elimination phase. Tumor cells that survive this elimination phase will move into the equilibrium phase. In this phase the tumor cells are kept up or they change to produce new tumors. These

tumors will keep on growing in an uncontrolled way and will finally be detected in the escape phase.[32-33] Therefore, it would be interesting to investigate which antigens or antibodies are involved in failing the immunosurveillance for lung cancer.

Biomarker research

A routine blood test for lung cancer does not exist, yet. Cancer patients produce autoantibodies to certain tumor-associated antigens (TAA's). In Table 2 a list of TAAs is shown that have been found in lung cancer by different groups until now.[34-49]

Table 2.Single TAAS and panels of TAAs recognized by (auto)antibodies that distinguish lung cancer patients from controls.

Single TAA / Panel of TAA	Method	Sensitivity	Specificity	Reference
α-enolase	SERPA + ELISA	28%	98%	He et al.
Annexin I	SERPA	30%	NA	Brichory et al.
Annexin II	SERPA	33%	NA	Brichory et al.
LAMR1	Protein microarray	NA	NA	Qiu et al.
Livin	ELISA	51%	NA	Yagihashi et al.
PGP9.5	SERPA	14%	NA	Brichory et al.
PRKCB1	Phage-display	NA	NA	Leidinger et al.
Prx-1	Western Blot	47%[1] 34%[2]	92%[1] 98%[2]	Chang et al.
ROCK1	Phage-display	NA	NA	Leidinger et al.
SOX2	ELISA	33%	97%	Maddison et al.
Survivin	ELISA	NA + 58%	NA + NA	Rohayem et al.+ Yagihashi et al.
14-3-3 theta, Annexin I, PGP9.5	Western blot + protein microarray	55%	95%	Pereire-Faca et al.
Annexin I, 14-3-3 theta, LAMR1	Protein microarray	51%	82%	Qiu et al.
c-myc, cyclin A, cyclin B1, cyclin D1, CDK2, survivin	ELISA	81%	97%	Rom et al.
HSP70, HSP90, p130, GAGE, BMI-1	Phage-display + ELISA	82%	83%	Zhong et al.
P53, NY-ESO-1, CAGE, GBU4-5, Annexin 1, SOX2	ELISA	39%[3] 37%[4]	89%[3] 90%[4]	Boyle et al. + Lam et al.
P53, c-myc, HER2, NY-ESO-1, CAGE, MUC1, GBU4-5	ELISA	76%	92%	Chapman et al.
Paxillin, SEC15L2, RP11-499F19, XRCC5, MALAT1	Phage-display	91%[3] 88%[4]	91%[3] 83%[4]	Zhong et al.
ROCK1, PRKCB1, KIAA0376	Phage-display	93%	93%	Leidinger et al.
Survivin, Livin	ELISA	71%	100%	Yagihashi et al.

TAA: Tumor associated antigens; ELISA: Enzyme-Linked Immuno Sorbent Assay; SERPA: Serological Proteome Analysis. [1] Tested against autoantibodies. [2] Tested against antigens. [3] Discovery. [4] Validation.

Although, research has been performed, yet a reliable marker has not been found. The limitations of autoantibodies as a marker are low reproducibility and low sensitivity and specificity of the test. A single autoantibody as marker lacks high sensitivity and specificity. For example, Brichory and coworkers showed a sensitivity of 14%, 30% and 33% for PGP 9.5, annexin I and II, respectively.[35-36, 50] Moreover, specific TAAs, e.g. p53, are present in different cancers. Therefore, a panel of markers might be of interest.[34, 38, 40, 43-45, 47-48, 51-58] The group of Robertson developed the EarlyCDT-Lung test that is currently used to aid early detection of lung cancer and has been approved by Clinical Laboratory Improvement Amendments (CLIA). This test measures autoantibodies against a panel of seven TAAs (p53, NY-ESO-1, CAGE, GBU4-5, SOX2,

HuD and MAGE A4). This panel gave a sensitivity of 41% and a specificity of 93%. After analyzing the performance of this panel in an independent clinically relevant sample set the sensitivity and specificity were 47% and 90%, respectively.[51]
Koziol et al. were able to distinguish lung cancer patients from normal individuals with a panel of seven TAAs. A sensitivity of 80% and a specificity of 90% was observed.[59] Unfortunately no validation was performed on these initial findings.

Antibodies as biomarker

Auto-antibody profiling could be a powerful tool for early detection when incorporated into a comprehensive screening strategy. In 1955, Robert Baldwin was the first to ascertain the presence of an immune response to solid tumors.[60] From then an increasing number of reports describe the presence of a humoral immune response in the form of autoantibodies that target specific tumor-associated antigens (TAAs) in lung cancer and other solid tumors.[61-66] This immune response to TAAs destructs precancerous lesions at an early stage of carcinogenesis.[67-68] Tumors are thought to induce the release of many TAAs into the blood. They can be overexpressed, aberrantly expressed, mutated, misfolded or aberrantly degraded such that an auto-reactive immune response is induced.[61-62] Post-translational modifications (PTM) of TAAs could also induce an immune response by generating a neo-epitope or by enhancing self-epitope presentation and affinity to the major histocompatibility complex or T-cell receptor.[61-62, 64] Many of the aberrantly expressed proteins (e.g. HER2/neu, P53, MAGE, NY-ESO-1, SSX2) that trigger an immune response in cancer patients contribute to carcinogenesis processes.[69-72]

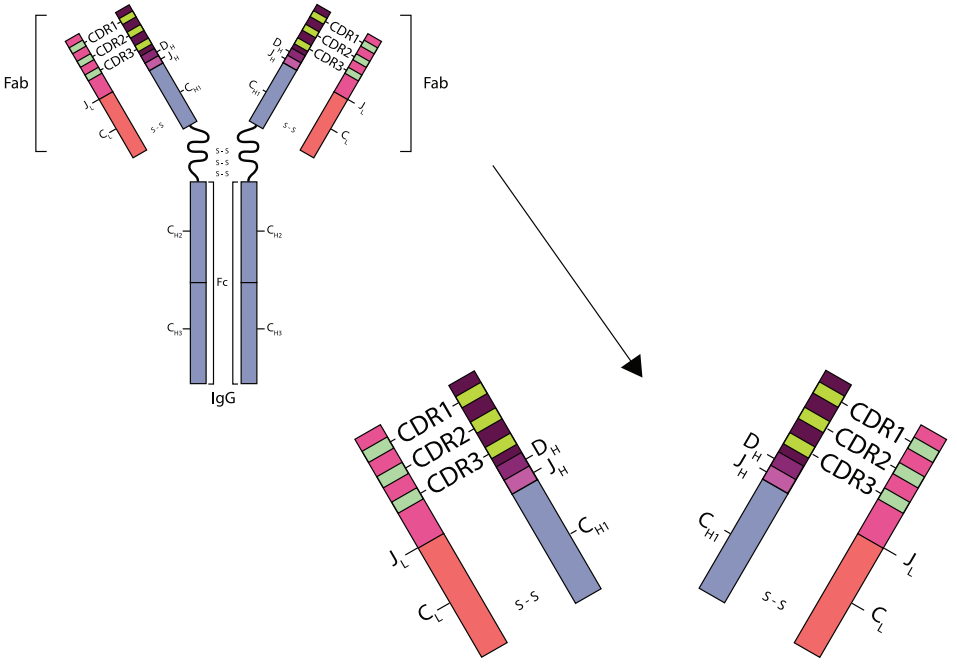


Figure 3. Structure of an immunoglobulin (IgG) molecule. Fab: Fragment antigen binding; CDR: complementarity determining region; J_L Joining segment of the light chain; C_L: Constant region of the light chain; D_H: Diversity segment of the heavy chain; J_H: Joining segment of the heavy chain; C_{H1} Constant region no. 1 of the heavy chain; C_{H2}:Constant region no. 2 of the heavy chain; C_{H3}:Constant region no. 3 of the heavy chain; Fc: Fragment crystallisable.

Recently Liu et al. showed that the concentration of circulating IgG autoantibodies against ABCC3 transporter was significantly higher in female adenocarcinoma patients than in female controls.[73] Therefore, diagnosing cancer based on serum profiling is an attractive concept especially because cancer autoantibodies can be detectable up to 5 years before radiological detection.[48, 74] Secondly, autoantibodies are inherently stable and persist in serum for a relatively long period of time at considerably higher concentrations than TAAs because they are usually not subject to extensive proteolysis and clearance.

A human immunoglobulin molecule consists of four chains, two identical heavy chains and two identical light chains. Every light chain has a variable (V_L) and constant (C_L) part. The heavy chains have three different constant parts (C_H1, C_H2 and C_H3) and a variable part (V_H) (Figure 3). The first constant part and variable part of the heavy chain together with the constant and variable part of the light chain form the antigen binding fragment (Fab). The remaining two constant parts of the heavy chain form the Fc region. Within the Fab three complementarity determining regions (CDR1, CDR2 and CDR3) are located between frameworks (Figure 3). These CDRs determine the antigen specificity and form a surface complementary to a shape that is part of the antigen.[75-76] CDRs are the hypervariable regions of the immunoglobulin molecule. The enormous diversity in antibodies originates from B-cell development when rearrangements of the V-, D-, and J-genes (Figure 3) occur and by somatic hypermutations during affinity maturation.[75, 77-79] Rearranged immunoglobulins specific to an antigen can have a very specific amino acid sequence, may be present among different patients and could be used as a marker for different cancers or autoimmune diseases.

Autoantibody identification

Various methods for identifying autoantibodies have been described in the literature, each of them with its specific advantages and disadvantages (Table 3). All these methods have in common that the autoantibodies are identified either by screening against a panel of known tumor antigens or to lysates of autologous tumor tissue or cancer cell lines.

Serological analysis of tumor antigens by recombinant cDNA expression cloning (SEREX) was first developed in 1995 by Sahin et al.. With this approach the TAAs are identified by screening patient sera against a cDNA expression library obtained from autologous tumor tissues.[71] In 1998 Gure et al. first applied SEREX to lung cancer patients.[63]

With the phage display method a cDNA phage display library is constructed from tumor tissue or cancer cell line. The autoantibodies in patient sera are captured by the phage display library by consecutive rounds of immunoprecipitation. The corresponding antigens are identified by sequencing. Leidinger et al. applied this method to squamous cell lung carcinoma patients.[41]

The Serological Proteome analysis (SERPA) approach was developed by Klade et al. in 2001.[80] This approach is based on a combination of 2D electrophoresis, western blotting and mass spectrometry (MS). Proteins from tumor tissues or cell lines are separated by 2D electrophoresis, transferred onto membranes by western blotting

Table 3. Overview of different techniques used for identification of autoantibodies.

	Principle	Source of candidate autoantigens	Comments	Reference
ELISA or Western-blot	Hypothesis-driven immunomics Candidate-based approach	Known antigens from literature	Availability of the antigens is required	Engvall et al. + van Weemen et al.
SEREX	Discovery-driven immunomics cDNA expression library	Autologous tumour tissue or cancer cell lines	Only linear epitopes detectable Bias towards highly expressed TAA's antigens; low abundance TAA's missed PTM's missed High throughput impossible	Sahin et al. + Gure et al.
Phage Display	Discovery-driven immunomics cDNA phage display library. Individual phage library plaques arrayed on nitrocellulose membranes	Tumour tissue or cell lysate	Each positive phage clone must be individually sequenced Epitopes of native antigens must be precluded Post-translational modifications missed High throughput possible	Leidinger et al.
SERPA	Discovery-driven "top-down" immunomics Combination of 2D electrophoresis and Western blotting followed by "top-down" immunomics	Tumour tissue or cell lysate	PTM's detectable No cDNA constructs needed Bias to abundant proteins Difficulty in producing reproducible 2D gels Only linear epitopes detectable	Klade et al. + Brichory et al.
Protein microarray	Discovery-driven "top-down" immunomics Microarrays on different platforms (2D or 3D)	Known and unknown purified or recombinant proteins or fractionated proteins from tumour or cell lysate	PTM's detectable Identification of unknown TAA's possible Modifications of epitopes on the array surface possible High throughput feasible	Leuking et al. + Joos et al. Qiu et al. + Madoz-Gurpide et al.
MAPPing	Discovery-driven "top-down" immunomics Affinity-based enrichment by 2D immunity chromatography	Tumour tissue or cell lysate	High throughput feasible	Caron et al.

ELISA: Enzyme-Linked Immuno Sorbent Assay; SEREX: Serological analysis of tumor antigens by Recombinant cDNA Expression cloning; SERPA: Serological Proteome Analysis; MAPPing: Multiple Affinity Protein Profiling; TAA: Tumor associated antigens; PTM: Post-translational modification; CDR: Complementarity determining regions.

and probed with sera from healthy individuals or patients with cancer.
The immunoreactive profiles are screened and the cancer associated spots are identified by mass spectrometry. Brichory and coworkers applied SERPA to lung cancer patients.[35]
Lueking et al. and Joos et al. were the first who described the development of antigen arrays for analyzing autoantibodies.[81-82] Purified or recombinant proteins, or fractionated proteins from tumor or cancer cell lysates are spotted onto microarrays and incubated with sera. Qiu et al. and Madoz-Gurpide et al. applied this method to lung cancer patients.[44, 83]

Caron et al. developed the multiple affinity protein profiling (MAPPing) approach in 2005. This approach is aimed at purification of putative autoantigens, digestion of this purification product followed by nano-LC-MS/MS analysis.[84]
Moreover, the group of Zhong combined some approaches. They introduced the application of T7 phage display libraries to the identification of circulating antibodies to non-small cell lung cancer antigens and in 2005 they implemented a protein microarray to this method.[49, 58]
All these methods were able to detect autoantibodies against TAAs which may be used as biomarkers for lung cancer. But none of these autoantibodies are applied in the clinic yet. Why? First of all, most of the studies performed with these techniques were not validated in a new independent sample set and/or not tested in different cancers beside lung cancer as many autoantibodies occur in different cancers and autoimmune diseases in which cancer is not a characteristic. Furthermore, these methods were not able to provide sensitivity and specificity comparable or even higher than observed with CT scan. Another problem is that most of these methods are time-consuming and are not applicable as high-throughput methods. Also, a disadvantage of these methods is that one needs an antigen panel beforehand. Therefore, development of a sensitive and specific autoantibody method where the antigens itself are not known beforehand would be of clinical interest.

Proteomics

The aim of lung proteomics is to characterize proteins and/or peptides and to obtain a more detailed view of the molecular biology of lung cancer in relation to other cancers or diseases. Proteins are responsible for the function of the majority of biology systems. Many crucial proteins are primary regulated by posttranslational modifications. Therefore, full knowledge of the alterations in the expression, modification and function of the protein in cancer cells is very important. New proteomic techniques are developed to analyze thousands of proteins in cancer cells and to understand their structure, function and interactions with other proteins. These techniques can be used in lung cancer to obtain new insights in the biology of lung cancer and new therapeutic targets or to identify novel biomarkers.
For early detection strategies for cancer, proteomic analysis of different complex mixtures can be used such as serum, plasma, sputum or exhaled breath condensate.
Mass spectrometry (MS) is one of the most frequently - used proteomics tools for biomarker discovery, yielding an enormous number of proteins identified in a very short time from clinical samples, including all kinds of body fluid. It enables not only identification of proteins but also quantification of proteins. A mass spectrometer consist of three different basic elements, 1) an ion source for converting proteins/peptides to gaseous ions, 2) a mass analyzer for separating the ions by mass and 3) a detector for detecting the ionized proteins/peptides.
The proteome is probably orders of magnitude more complex than the genome. Reducing this complexity is a necessity and can be achieved by separation and enrichment of specific proteins from the sample and their subsequent identification. The identification can be performed by MS/MS. To reach this, proteins are first digested with proteases, e.g. trypsin, which results in a mixture of peptides. In general, mass spectrometers are not able to identify amino acid sequences without reducing the size of proteins into peptides. The proteins digested into peptides are analyzed in such a way that the peptide ions are separated at a first stage, then each peptide is

fragmented in a so-called collision cell. These fragmented peptides are then separated by a second analyzer and the amino acid sequence of the peptide is determined. The advantages of mass spectrometry include high sensitivity, accuracy and speed. Dekker at al. showed that immunoglobulin peptides of a spiked antibody were detectable in serum at attomole levels.[85] These advantages make MS an attractive approach for biomarker discovery. Besides advantages there are also limitations of mass spectrometry based biomarker discovery approaches. The limitations can be divided in three categories 1) pre-analytical, 2) analytical and 3) post-analytical. Pre-analytical limitations such as storage and sample preparations have influence on the analyses by causing bias. Analytical limitations include poor reproducibility between institutions or even between different runs or limited sensitivity due to the presence of abundant proteins. This last limitation occurs when no extra sample preparation is applied. In this case the detection limit of most proteins is around 1 µg/ml, while known biomarkers are approximately 1,000 times less concentrated in serum. Bioinformatics/biostatistics is the main post-analytical technique used for analyzing the acquired data. Small numbers of samples generate large numbers of spectra which increases the risk of over-fitting the data. Therefore, these techniques need more attention. Furthermore, more than half of the proteins/peptides found in a sample cannot be identified by current computational methods and databases.[86] Besides these limitations, some research groups were able to discover some proteins in blood for lung cancer detection by proteomics with reasonable sensitivity and specificity. Patz et al. found a panel of four proteins that could distinguish lung cancer cases from controls with a sensitivity and specificity of 89% and 85%, respectively. In a new independent validation set they found a sensitivity of 78% and a specificity of 75%. The group of Yildiz observed a panel of seven proteins with sensitivity and specificity of 67% and 89%, respectively in a sample set of NSCLC patients and controls. The validation of the panel gave a sensitivity of 58% and a specificity of 86%.[57, 87] Recently, the research group of Carbone published an article were they showed new differentially expressed proteins in tissue which are of interest as diagnostic biomarkers. These proteins were confirmed by multiple reaction monitoring (MRM) mass spectrometry. Furthermore, a subset of these proteins was differentially expressed in plasma samples from lung cancer patients and matched controls. They observed an AUC of 0.72 for the prediction of the diagnosis of squamous cell carcinoma and for adenocarcinoma an AUC of 0.59 was observed.[88]

As mentioned before, specific amino acid sequences of rearranged immunoglobulins specific to an antigen may be used as marker among patients for different cancers or autoimmune diseases. Arentz et al. recently published that uniquely mutated V regions peptides as surrogates detected anti-Ro52 autoantibodies in sera from primary Sjögren's syndrome patients by mass spectrometry. They observed high sensitivity and specificity, 87.5% and 92.9% respectively, compared to ELISA. This study has not been validated yet, but provides a proof-of-concept for targeted mass spectrometry using autoantibodies as a marker for different autoimmune diseases and may also be used for cancer.89 Although, many studies have found markers by proteomic approaches which were able to distinguish patients from controls, these markers are still not used in the clinic. A reason why they are not applied yet is that most of the studies were not validated and if they were validated the sensitivity and specificity was not high enough compared to for example CT screening or they were not evaluated parallel to the already applied approaches in the clinic.

Aim and outline of the thesis

As CT screening is not recommended unless it is used as part of a clinical trial, biomarkers can be useful complementary to CT screening. Many studies have been performed on different antigens as blood biomarkers for lung cancer using known antibodies. We instead hypothesized that disease – related antibodies without a required prior knowledge of the antigen could function as a biomarker for lung cancer. Therefore, the aim of this thesis is to determine if similar or identical CDR sequences of antibodies are involved in lung cancer and to develop an immunomics method with high sensitivity and specificity.

To accomplish the aim we started to develop an immunomics method. The study of the immune system using genome-wide approaches is called immunomics.[90-91] In this thesis we extend immunomics with a proteomics approach. In **Chapter 2** we describe the IgG Fab purification mass spectrometry approach. We discuss the reproducibility of the approach and the suitability to determine qualitative and quantitative differences in IgG Fab peptides of healthy individuals and if the identification of a CDR signature as biomarker for (lung) cancer or autoimmune diseases is feasible.
As an immunoglobulin molecule is a very complex molecule it would be ideal to reduce the complexity of this molecule for mass spectrometry. In **Chapter 3** we describe molecular dissection of IgG to reduce the complexity for mass spectrometry. We show the benefits of molecular dissection of IgG into Fab-κ, Fab-λ, κ and λ as an addition to the Fab approach mentioned in **Chapter 2**. Furthermore, we discuss the likelihood of finding lung cancer-related CDR sequences by the use of this approach.
Finally, we would like to apply our IgG fab purification method on a case-control study. In **Chapter 4** we discuss the involvement of antibodies in lung cancer. We show a panel of immunoglobulin peptides that might be of interest in distinguishing lung cancer patients from controls. This study shows the proof-of concept that identical sequences of specific antibodies are produced by lung cancer patients detected by IgG Fab purification mass spectrometry approach. For this study we used lung cancer patients and controls from the NELSON trial.
In **Chapter 5** we discuss the new version of the open source software package Peptrix for Orbitrap LC-MS data. We compared this software package with three other open source and commercially available software packages for Orbitrap LC-MS data.

References

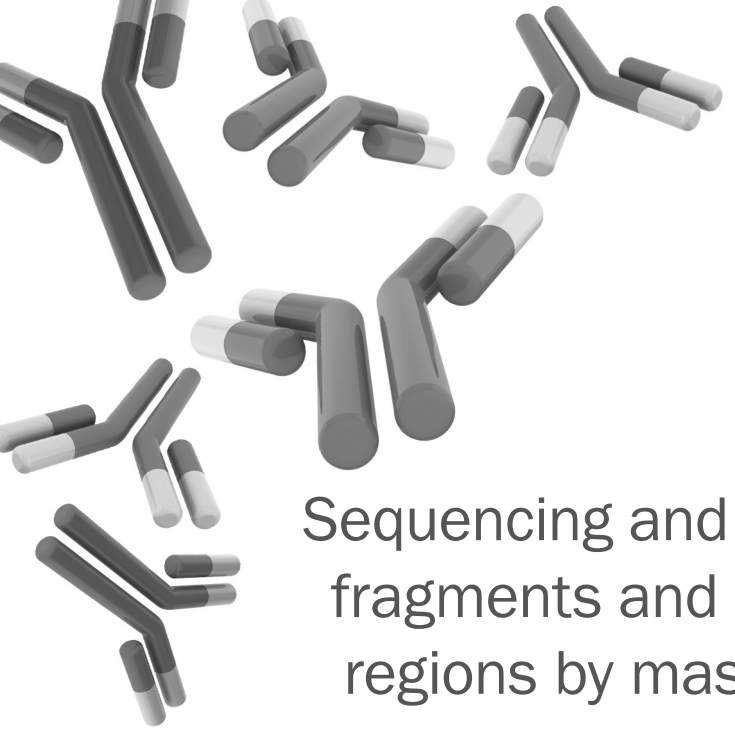
1. IARC: section of cancer information. Globocan. 2008.
2. Jemal A, Siegel R, Xu J, Ward E. Cancer statistics, 2010. CA Cancer J Clin 2010; 60:277-300.
3. American Cancer Society. Cancer facts and figures 2010 American Cancer Society 2010.
4. Pass H.I. CDP, Johnson D.H., Minna J.D., Scagliotti G.V., Turrisi A.T. Principles and practice of lung cancer. The official reference text of the IASLC. 4th ed: Lippincott Williams & Wilkins; 2010.
5. Lorigan P. SAT. Lung cancer: Elsivier; 2007.
6. National Cancer Institute at the the national institute of health
7. Riaz SP, Luchtenborg M, Coupland VH, Spicer J, Peake MD, Moller H. Trends in incidence of small cell lung cancer and all lung cancer. Lung Cancer 2012; 75:280-4.
8. Goldstraw P, Crowley J, Chansky K, Giroux DJ, Groome PA, Rami-Porta R, et al. The IASLC Lung Cancer Staging Project: proposals for the revision of the TNM stage groupings in the forthco ing (seventh) edition of the TNM Classification of malignant tumours. J Thorac Oncol 2007; 2:706-14.
9. WHO. MPOWER. A policy package to reverse the tobacco epidemic. World Health Organization 2008.
10. Wardwell NR, Massion PP. Novel strategies for the early detection and prevention of lung cancer. Semin Oncol 2005; 32:259-68.
11. Flehinger BJ, Kimmel M, Polyak T, Melamed MR. Screening for lung cancer. The Mayo Lung Project revisited. Cancer 1993; 72:1573-80.
12. Frost JK, Ball WC, Jr., Levin ML, Tockman MS, Baker RR, Carter D, et al. Early lung cancer detection: results of the initial (prevalence) radiologic and cytologic screening in the Johns Hopkins study. Am Rev Respir Dis 1984; 130:549-54.
13. Marcus PM, Bergstralh EJ, Zweig MH, Harris A, Offord KP, Fontana RS. Extended lung cancer incidence follow-up in the Mayo Lung Project and overdiagnosis. J Natl Cancer Inst 2006; 98:748-56.
14. Melamed MR, Flehinger BJ. Detection of lung cancer: highlights of the Memorial Sloan-Kettering Study in New York City. Schweiz Med Wochenschr 1987; 117:1457-63.
15. National Lung Screening Trial Research T, Aberle DR, Adams AM, Berg CD, Black WC, Clapp JD, et al. Reduced lung-cancer mortality with low-dose computed tomographic screening. N Engl J Med 2011; 365:395-409.
16. van Iersel CA, de Koning HJ, Draisma G, Mali WP, Scholten ET, Nackaerts K, et al. Risk-based selection from the general population in a screening trial: selection criteria, recruitment and power for the Dutch-Belgian randomised lung cancer multi-slice CT screening trial (NELSON). Int J Cancer 2007; 120:868-74.
17. Xu DM, Gietema H, de Koning H, Vernhout R, Nackaerts K, Prokop M, et al. Nodule management protocol of the NELSON randomised lung cancer screening trial. Lung Cancer 2006; 54:177-84.
18. Lopes Pegna A, Picozzi G, Mascalchi M, Maria Carozzi F, Carrozzi L, Comin C, et al. Design, recruitment and baseline results of the ITALUNG trial for lung cancer screening with low-dose CT. Lung Cancer 2009; 64:34-40.
19. Infante M, Lutman FR, Cavuto S, Brambilla G, Chiesa G, Passera E, et al. Lung cancer screening with spiral CT: baseline results of the randomized DANTE trial. Lung Cancer 2008; 59:355-63.
20. Pedersen JH, Ashraf H, Dirksen A, Bach K, Hansen H, Toennesen P, et al. The Danish randomized lung cancer CT screening trial--overall design and results of the prevalence round. J Thorac Oncol 2009; 4:608-14.
21. Baldwin DR, Duffy SW, Wald NJ, Page R, Hansell DM, Field JK. UK Lung Screen (UKLS) nodule management protocol: modelling of a single screen randomised controlled trial of low-dose CT

screening for lung cancer. Thorax 2011; 66:308-13.
22. Becker N, Delome S, Kauczor HU. LUSI: the German component of the European trial on the efficacy of multislice-CT for the early detection of lung cancer. Onkologie 2008; 31:P0320.
23. Gopal M, Abdullah SE, Grady JJ, Goodwin JS. Screening for lung cancer with low-dose computed tomography: a systematic review and meta-analysis of the baseline findings of randomized-controlled trials. J Thorac Oncol 2010; 5:1233-9.
24. Infante MV, Pedersen JH. Screening for lung cancer: are we there yet? Curr Opin Pulm Med 2010; 16:301-6.
25. Mascaux C, Peled N, Garg K, Kato Y, Wynes MW, Hirsch FR. Early detection and screening of lung cancer. Expert Rev Mol Diagn 2010; 10:799-815.
26. Pastorino U. Lung cancer screening. Br J Cancer 2010; 102:1681-6.
27. van Klaveren RJ, Oudkerk M, Prokop M, Scholten ET, Nackaerts K, Vernhout R, et al. Management of lung nodules detected by volume CT scanning. N Engl J Med 2009; 361:2221-9.
28. Bach PB, Silvestri GA, Hanger M, Jett JR, American College of Chest P. Screening for lung cancer: ACCP evidence-based clinical practice guidelines (2nd edition). Chest 2007; 132:69S-77S.
29. Hanahan D, Weinberg RA. The hallmarks of cancer. Cell 2000; 100:57-70.
30. Nowell PC. The clonal evolution of tumor cell populations. Science 1976; 194:23-8.
31. Wistuba, II, Gazdar AF. Lung cancer preneoplasia. Annu Rev Pathol 2006; 1:331-48.
32. Dunn GP, Old LJ, Schreiber RD. The immunobiology of cancer immunosurveillance and immunoediting. Immunity 2004; 21:137-48.
33. Finn OJ. Cancer immunology. N Engl J Med 2008; 358:2704-15.
34. Boyle P, Chapman CJ, Holdenrieder S, Murray A, Robertson C, Wood WC, et al. Clinical validation of an autoantibody test for lung cancer. Ann Oncol 2011; 22:383-9.
35. Brichory F, Beer D, Le Naour F, Giordano T, Hanash S. Proteomics-based identification of protein gene product 9.5 as a tumor antigen that induces a humoral immune response in lung cancer. Cancer Res 2001; 61:7908-12.
36. Brichory FM, Misek DE, Yim AM, Krause MC, Giordano TJ, Beer DG, et al. An immune response manifested by the common occurrence of annexins I and II autoantibodies and high circulating levels of IL-6 in lung cancer. Proc Natl Acad Sci U S A 2001; 98:9824-9.
37. Chang JW, Lee SH, Jeong JY, Chae HZ, Kim YC, Park ZY, et al. Peroxiredoxin-I is an autoimmunogenic tumor antigen in non-small cell lung cancer. FEBS Lett 2005; 579:2873-7.
38. Chapman CJ, Murray A, McElveen JE, Sahin U, Luxemburger U, Tureci O, et al. Autoantibodies in lung cancer: possibilities for early detection and subsequent cure. Thorax 2008; 63:228-33.
39. He P, Naka T, Serada S, Fujimoto M, Tanaka T, Hashimoto S, et al. Proteomics-based identification of alpha-enolase as a tumor antigen in non-small lung cancer. Cancer Sci 2007; 98:1234-40.
40. Lam S, Boyle P, Healey GF, Maddison P, Peek L, Murray A, et al. EarlyCDT-Lung: an immunobiomarker test as an aid to early detection of lung cancer. Cancer Prev Res (Phila) 2011; 4:1126-34.
41. Leidinger P, Keller A, Ludwig N, Rheinheimer S, Hamacher J, Huwer H, et al. Toward an early diagnosis of lung cancer: an autoantibody signature for squamous cell lung carcinoma. Int J Cancer 2008; 123:1631-6.
42. Maddison P, Thorpe A, Silcocks P, Robertson JF, Chapman CJ. Autoimmunity to SOX2, clinical phenotype and survival in patients with small-cell lung cancer. Lung Cancer 2010; 70:335-9.
43. Pereira-Faca SR, Kuick R, Puravs E, Zhang Q, Krasnoselsky AL, Phanstiel D, et al. Identification of 14-3-3 theta as an antigen that induces a humoral response in lung cancer. Cancer Res 2007; 67:12000-6.
44. Qiu J, Choi G, Li L, Wang H, Pitteri SJ, Pereira-Faca SR, et al. Occurrence of autoantibodies to annexin I, 14-3-3 theta and LAMR1 in prediagnostic lung cancer sera. J Clin Oncol 2008;

26:5060-6.
45. Rohayem J, Diestelkoetter P, Weigle B, Oehmichen A, Schmitz M, Mehlhorn J, et al. Antibody response to the tumor-associated inhibitor of apoptosis protein survivin in cancer patients. Cancer Res 2000; 60:1815-7.
46. Rom WN, Goldberg JD, Addrizzo-Harris D, Watson HN, Khilkin M, Greenberg AK, et al. Identification of an autoantibody panel to separate lung cancer from smokers and nonsmokers. BMC Cancer 2010; 10:234.
47. Yagihashi A, Asanuma K, Kobayashi D, Tsuji N, Shijubo Y, Abe S, et al. Detection of autoantibodies to livin and survivin in Sera from lung cancer patients. Lung Cancer 2005; 48:217-21.
48. 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.
49. Zhong L, Peng X, Hidalgo GE, Doherty DE, Stromberg AJ, Hirschowitz EA. Identification of circulating antibodies to tumor-associated proteins for combined use as markers of non-small cell lung cancer. Proteomics 2004; 4:1216-25.
50. Casiano CA, Mediavilla-Varela M, Tan EM. Tumor-associated antigen arrays for the serological-diagnosis of cancer. Mol Cell Proteomics 2006; 5:1745-59.
51. Chapman CJ, Healey GF, Murray A, Boyle P, Robertson C, Peek LJ, et al. EarlyCDT(R)-Lung test: improved clinical utility through additional autoantibody assays. Tumour Biol 2012.
52. Chapman CJ, Thorpe AJ, Murray A, Parsy-Kowalska CB, Allen J, Stafford KM, et al. Immunobiomarkers in small cell lung cancer: potential early cancer signals. Clin Cancer Res 2011; 17:1474-80.
53. Guergova-Kuras M, Kurucz I, Hempel W, Tardieu N, Kadas J, Malderez-Bloes C, et al. Discovery of lung cancer biomarkers by profiling the plasma proteome with monoclonal antibody libraries. Mol Cell Proteomics 2011; 10:M111 010298.
54. Macdonald IK, Allen J, Murray A, Parsy-Kowalska CB, Healey GF, Chapman CJ, et al. Development and validation of a high throughput system for discovery of antigens for autoantibody detection. PLoS One 2012; 7:e40759.
55. Patel K, Farlow EC, Kim AW, Lee BS, Basu S, Coon JS, et al. Enhancement of a multianalyte serum biomarker panel to identify lymph node metastases in non-small cell lung cancer with circulating autoantibody biomarkers. Int J Cancer 2011; 129:133-42.
56. Wu L, Chang W, Zhao J, Yu Y, Tan X, Su T, et al. Development of autoantibody signatures as novel diagnostic biomarkers of non-small cell lung cancer. Clin Cancer Res 2010; 16:3760-8.
57. Yildiz PB, Shyr Y, Rahman JS, Wardwell NR, Zimmerman LJ, Shakhtour B, et al. Diagnostic accuracy of MALDI mass spectrometric analysis of unfractionated serum in lung cancer. J Thorac Oncol 2007; 2:893-901.
58. Zhong L, Hidalgo GE, Stromberg AJ, Khattar NH, Jett JR, Hirschowitz EA. Using protein microarray as a diagnostic assay for non-small cell lung cancer. Am J Respir Crit Care Med 2005; 172:1308-14.
59. Koziol JA, Zhang JY, Casiano CA, Peng XX, Shi FD, Feng AC, et al. Recursive partitioning as an approach to selection of immune markers for tumor diagnosis. Clin Cancer Res 2003; 9:5120-6.
60. Baldwin RW. Immunity to transplanted tumour: the effect of tumour extracts on the growth of homologous tumours in rats. Br J Cancer 1955; 9:646-51.
61. Anderson KS, LaBaer J. The sentinel within: exploiting the immune system for cancer biomarkers. J Proteome Res 2005; 4:1123-33.
62. Caron M, Choquet-Kastylevsky G, Joubert-Caron R. Cancer immunomics using autoantibody signatures for biomarker discovery. Mol Cell Proteomics 2007; 6:1115-22.
63. Gure AO, Altorki NK, Stockert E, Scanlan MJ, Old LJ, Chen YT. Human lung cancer antigens recognized by autologous antibodies: definition of a novel cDNA derived from the tumor suppressor gene locus on chromosome 3p21.3. Cancer Res 1998; 58:1034-41.

64. Hanash S. Harnessing immunity for cancer marker discovery. Nat Biotechnol 2003; 21:37-8.
65. Mintz PJ, Kim J, Do KA, Wang X, Zinner RG, Cristofanilli M, et al. Fingerprinting the circulating repertoire of antibodies from cancer patients. Nat Biotechnol 2003; 21:57-63.
66. 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.
67. Finn OJ. Immune response as a biomarker for cancer detection and a lot more. N Engl J Med 2005; 353:1288-90.
68. Tan EM. Autoantibodies as reporters identifying aberrant cellular mechanisms in tumorigenesis. J Clin Invest 2001; 108:1411-5.
69. Cheever MA, Disis ML, Bernhard H, Gralow JR, Hand SL, Huseby ES, et al. Immunity to oncogenic proteins. Immunol Rev 1995; 145:33-59.
70. Chen YT, Scanlan MJ, Sahin U, Tureci O, Gure AO, Tsang S, et al. A testicular antigen aberrantly-expressed in human cancers detected by autologous antibody screening. Proc Natl Acad Sci USA 1997; 94:1914-8.
71. Sahin U, Tureci O, Schmitt H, Cochlovius B, Johannes T, Schmits R, et al. Human neoplasms elicit multiple specific immune responses in the autologous host. Proc Natl Acad Sci USA 1995; 92:11810-3.
72. Ward RL, Hawkins NJ, Coomber D, Disis ML. Antibody immunity to the HER-2/neu oncogenic protein in patients with colorectal cancer. Hum Immunol 1999; 60:510-5.
73. Liu L, Liu N, Liu B, Yang Y, Zhang Q, Zhang W, et al. Are circulating autoantibodies to ABCC3 transporter a potential biomarker for lung cancer? J Cancer Res Clin Oncol 2012; 138:1737 42.
74. Fernandez Madrid F. Autoantibodies in breast cancer sera: candidate biomarkers and reporters of tumorigenesis. Cancer Lett 2005; 230:187-98.
75. Murphy K. TP, Walport M. Janeway's immunobiology. 7th ed: Garland Science; 2008.
76. Schroeder HW, Jr., Cavacini L. Structure and function of immunoglobulins. J Allergy Clin Immunol 2010; 125:S41-52.
77. de Wildt RM, van Venrooij WJ, Winter G, Hoet RM, Tomlinson IM. Somatic insertions and deletions shape the human antibody repertoire. J Mol Biol 1999; 294:701-10.
78. Tonegawa S. Reiteration frequency of immunoglobulin light chain genes: further evidence for somatic generation of antibody diversity. Proc Natl Acad Sci U S A 1976; 73:203-7.
79. Xu JL, Davis MM. Diversity in the CDR3 region of V(H) is sufficient for most antibody specificities. Immunity 2000; 13:37-45.
80. 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.
81. Joos TO, Schrenk M, Hopfl P, Kroger K, Chowdhury U, Stoll D, et al. A microarray enzyme-linked immunosorbent assay for autoimmune diagnostics. Electrophoresis 2000; 21:2641-50.
82. Lueking A, Horn M, Eickhoff H, Bussow K, Lehrach H, Walter G. Protein microarrays for gene expression and antibody screening. Anal Biochem 1999; 270:103-11.
83. Madoz-Gurpide J, Kuick R, Wang H, Misek DE, Hanash SM. Integral protein microarrays for the identification of lung cancer antigens in sera that induce a humoral immune response. Mol Cell Proteomics 2008; 7:268-81.
84. Caron M, Joubert-Caron R, Canelle L, Hardouin J. Serological proteome analysis (SERPA) and multiple affinity protein profiling (MAPPING) to discover cancer biomarkers. Mol Cell Proteomics 2005; 4:S142.
85. Dekker LJ, Zeneyedpour L, Brouwer E, van Duijn MM, Sillevis Smitt PA, Luider TM. An antibody based biomarker discovery method by mass spectrometry sequencing of complementarity determining regions. Anal Bioanal Chem 2011; 399:1081-91.
86. Ocak S, Chaurand P, Massion PP. Mass spectrometry-based proteomic profiling of lung cancer.

Proc Am Thorac Soc 2009; 6:159-70.
87. Patz EF, Jr., Campa MJ, Gottlin EB, Kusmartseva I, Guan XR, Herndon JE, 2nd. Panel of serum biomarkers for the diagnosis of lung cancer. J Clin Oncol 2007; 25:5578-83.
88. Kikuchi T, Hassanein M, Amann JM, Liu Q, Slebos RJ, Rahman SM, et al. In-depth proteomic analysis of non-small cell lung cancer to discover molecular targets and candidate biomarkers. Mol Cell Proteomics 2012.
89. Arentz G, Thurgood LA, Lindop R, Chataway TK, Gordon TP. Secreted human Ro52 autoantibody proteomes express a restricted set of public clonotypes. J Autoimmun 2012.
90. Braga-Neto UM, Marques ET, Jr. From functional genomics to functional immunomics: new challenges, old problems, big rewards. PLoS Comput Biol 2006; 2:e81.
91. Sette A, Fleri W, Peters B, Sathiamurthy M, Bui HH, Wilson S. A roadmap for the immunomics of category A-C pathogens. Immunity 2005; 22:155-61.



Chapter 2

Sequencing and quantifying IgG fragments and antigen-binding regions by mass spectrometry

Dominique de Costa, Ingrid Broodman, Martijn M. VanDuijn, Christoph Stingl, Lennard J.M. Dekker, Peter C. Burgers, Henk C. Hoogsteden, Peter A.E. Sillevis Smitt, Rob J. van Klaveren, Theo M. Luider

J Proteome Res 2010; 9:2937-45

Abstract

In cancer and autoimmune diseases, immunoglobulins with a specific molecular signature which could potentially be used as diagnostic or prognostic markers are released into body fluids. An immunomics approach based on this phenomenon relies on the ability to identify the specific amino acid sequences of the complementarity-determining regions (CDR) of these immunoglobulins, which in turn depends on the level of accuracy, resolution and sensitivity that can be achieved by advanced mass spectrometry. Reproducible isolation and sequencing of antibody fragments (e.g. Fab) by high-resolution mass spectrometry (MS) from seven healthy donors revealed 43,217 MS signals: 225 could be associated with CDR1 peptides, 513 with CDR2 peptides, and 19 with CDR3 peptides. Seventeen percent of the 43,217 MS signals did not overlap between the seven donors. The Fab isolation method used is reproducible and fast, with a high yield. It provides only one Fab sample fraction for subsequent characterization by high-resolution MS. In 17% and 4% of these seven healthy donors qualitative (presence/absence) and quantitative (intensity) differences in Fab fragments could be demonstrated, respectively. From these results we conclude that the identification of a CDR signature as biomarker for autoimmune diseases and cancer without prior knowledge of the antigen is feasible.

Introduction

Most immunoglobulin molecules are composed of four polypeptide chains: two identical heavy chains and two identical light chains. A light chain has one variable (V_L) and one constant (C_L) domain. These two domains pair with the variable (V_H) and first constant (C_H1) domain of the heavy chain and are referred to as the antigen-binding fragment (Fab). The remaining two domains of the heavy chain (C_H2 and C_H3) form the Fc region. The amino acid sequence variation in the variable domains (both V_L and V_H) is mainly confined to three small hypervariable loops, which determine antigen specificity by forming a surface complementary to the antigen, and are more commonly termed complementarity-determining regions, or CDRs (CDR1, CDR2, and CDR3). The enormous antibody diversity originates from B-cell development, during which variable gene segments are joined and a large combinatorial diversity is created. This diversity is enhanced by insertions and deletions between the gene segments during rearrangement (junctional diversity), and also by somatic hypermutations in the rearranged immunoglobulin coding sequence.[1-3] Rearranged immunoglobulins specific to an antigen will thus have very specific sequences which could be used as a molecular signature. Antibody sequences from a subject are most commonly characterized on the level of cDNA. This cDNA derived from lymphocytes is characterized by using DNA techniques. However, there are few published articles of the characterization of antibody sequences at protein level, such as techniques for serum or CSF.[1, 4-8]

It has been demonstrated that cancer growth and progression are associated with cancer immunosurveillance and inflammation.[9-12] Not only in autoimmune diseases like multiple sclerosis[8] but also in cancer[13-16], large numbers of immunoglobulins are released into blood.[2, 5, 13-17] The molecular signatures of such immunoglobulins could potentially be used as diagnostic or prognostic markers. Screening for disease-related immune responses is generally performed by testing patients' sera against antigens or antigen libraries. Although successful, techniques such as serological expression cloning (SEREX) are aimed at detecting the targeted antigens, rather than the reactive immunoglobulins.[18-19]

An alternative strategy is to directly compare the amino acid sequences of immunoglobulin molecules between cases and controls, using high-resolution Orbitrap mass spectrometry. The approach depends primarily on the ability to reveal differences between healthy controls in the amino acid sequence of the variable CDRs, which in turn depends on the level of accuracy, resolution and sensitivity that can be achieved by high-resolution mass spectrometry and bioinformatics tools. The differences can be expressed qualitatively, in terms of the presence or absence of specific identified sequenced peptides (CDRs), and quantitatively, by comparing the normalized peak intensities of peptide masses of interest. The next step would be the identification of a single molecular signature or a set of such signatures from a case control training set, followed by independent validation.

Our study had two aims. The first was to investigate whether our IgG Fab isolation approach could reproducibly generate immunoglobulin samples for mass spectrometry analysis. The second was to ascertain whether this is a suitable approach for determining qualitative and quantitative differences in IgG Fab peptides, especially in the mutated CDRs of healthy individuals, obtained by means of high-resolution MS.

Materials and Methods

Purification of IgG Fab

Seven healthy donor serum samples were obtained from the Sanquin Blood Bank Rotterdam, the Netherlands. In accordance with the general guidelines of the Blood Bank, the donors had given written consent for the serum to be used for scientific research. The donors (five males and two females, median age 60 years, range 50-66 years) were not on any medication. Their serum IgG concentrations were analyzed turbidimetrically on the Modular P800 (Roche, Almere, Netherlands) and were found to be within the normal range (7.0-16.0 g/L).

Nine mL venous blood (without additives) from each donor was allowed to clot for one hour at room temperature, centrifuged for 10 min at 2880 g, stored at 4 °C up to 2 hours, distributed into 100 µL aliquots and stored at -80 °C. Serum IgG fractions were purified using the Melon Gel IgG purification kit (Pierce, Rockford, IL), according to the manufacturer's instructions.[20] The concentration of the purified IgG protein was determined by means of the mass extinction coefficient of 1.37 (mg/mL) cm^{-1} at 280 nm on a NanoDrop Spectrophotometer (ND-1000, Nanodrop Technologies, Wilmington, DE).

After purification, 400 µL IgG was digested overnight by papain immobilized on agarose beads according to the manufacturers instructions (Figure 1) (http://www.piercenet.com/ files/0107as4.pdf). After digestion, 2700 µL of the papain digested and purified IgG was concentrated approximately ten times and exchanged in 0.1 M sodium phosphate buffer (Coupling Buffer MicroLink Protein kit, Pierce) (http://www.piercenet.com/files/1509as4.pdf) by an Amicon Ultra 3K centrifugal filter device (Millipore, Amsterdam, Netherlands).

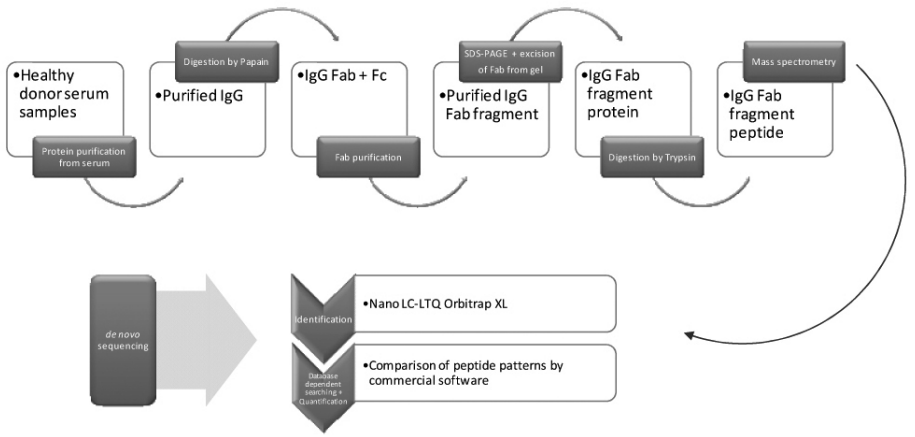


Figure 1. Flow-chart illustrating the different steps in Fab purification and data analysis. The upper part of the chart describes the purification of Fab (including in-solution (papain) digestion) from serum and subsequent in-gel (trypsin) enzymatic digestion of Fab. The lower part describes the mass spectrometry and data analysis used for identification, database searching, and *de novo* sequencing.

To separate Fab from Fc fragments and undigested IgG, protein affinity chromatography was performed using the MicroLink Protein Coupling kit (Pierce) (Figure 1). Briefly, Fc fragment specific ImmunoPure anti-human IgG, (250 µL) was immobilized

on MicroLink Protein support of the AminoLink Plus coupling gel spin column at 4°C overnight with gentle end-over-end mixing. The papain digested purified IgG (200 µL) was loaded onto this column and incubated at 4°C overnight with gentle end-over-end mixing. Finally, the flow-through containing the Fab was collected and analyzed by SDS-PAGE electrophoresis.
After separation by SDS-PAGE under reducing conditions, the proteins were fixed for 10 min in 50% (v/v) methanol and 10% (v/v) acetic acid. The protein bands were visualized with the Colloidal Blue staining kit (Invitrogen, Breda, Netherlands) and gels were destained for a minimum of 4 hours in deionized water.

Sample preparation for mass spectrometry

The intensity of the protein bands on the stained SDS-PAGE gel was determined by scanning on a Molecular Imager GS-800 Calibrated densitometer (Bio-Rad, Veenendaal, Netherlands) with Quantity One® 1-D analysis software (version 4.6.5; Bio-Rad).
After imaging and analysis of the SDS-PAGE gels, the selected protein bands were manually excised from the gels. The gel plugs were washed with 100 µL ultrapure water, destained twice with 200 µL 100 mM NH_2HCO_3 in 70% (v/v) acetonitrile/30% H_2O and washed with 200 µL ultrapure water. Liquid was removed and the gel plugs were dried for 35 min in a vacuum centrifuge (SPD 1010 Speedvac System; Thermo Fisher Scientific Inc., Waltham, MA) until they were completely dry. Digestion of the corresponding gel plugs were performed in 20 µL or up to 40 µL of a solution of 0.1 mg/mL trypsin (Promega, Leiden, Netherlands) in 50 mM Tris-HCl (pH 8.8) to fully submerge the gel plugs. The digestion was performed overnight at room temperature (Figure 1). Subsequently, the tryptic peptides were extracted from the gel by adding three times 50 µL 0.5% (v/v) formic acid in 30% (v/v) acetonitrile/H_2O, mixed by means of an ultrasonic bath (Branson 2510, Danbury, CT). Eventually the three extraction fluids were pooled. The samples were then evaporated for approximately 2 hours in a vacuum centrifuge until they were completely dry. For the nano-LC-MS/MS analysis, the dried peptides were dissolved in 20 µL of an aqueous solution of 0.1 % (v/v) formic acid and 2% (v/v) acetonitrile, using an ultrasonic bath.

NanoLC Orbitrap MS analyses

LCMS measurements were carried out by an Ultimate 3000 nano LC system (Dionex, Amsterdam, Netherlands) online coupled to a hybrid linear ion trap/Orbitrap MS (LTQ Orbitrap XL; Thermo Fisher Scientific, Bremen, Germany). Five µL of the digested Fab were loaded onto a C18 trap column (PepMap C18, 300µm ID × 5mm, 5 µm particle size, 100 Å pore size; Dionex) and desalted for 10 min at a flow rate of 20 µL/min 0.1% TFA (Biosolve, Valkenswaard, Netherlands). Next, the trap column was switched online to an analytic column (PepMap C18, 75µm ID × 150mm, 3 µm particle size and 100 Å pore size; Dionex). Peptides were eluted using a 180 min gradient with the following binary gradient: 0%-25% solvent B in 120 min and 25%-50% solvent B in the next 60 min, where solvent A consists of 2% acetonitrile and 0.1% formic acid in water and solvent B consists of 80% acetonitrile and 0.1% formic acid in water (all solvents used purchased from Biosolve). Column flow rate was set at 300 nL/min. The analytic column was then washed and equilibrated.

To identify the Fab fragments of the seven donors we used a CAD fragmentation. High resolution full scan MS was obtained from the Orbitrap (resolution 30,000; AGC 1,000,000), MS/MS spectra were obtained by CAD fragmentation, and detection was conducted. MS/MS was performed on the top five masses in the full scan spectra. Dynamic exclusion was used, with a repeat count of one; exclusion duration was set at 3 min and exclusion width at +/- 5 ppm.

Data analysis

MS/MS spectra were extracted from raw data files and converted into mgf files using extract msn (part of XCalibur version 2.0.7, Thermo Fisher Scientific Inc.). Mascot (version 2.2.06; Matrix Science Inc., London, UK) was used to perform database searches against the human subset NCBInr database (version nrHuman_database_20090311; Homo sapiens species restriction; 222,071 sequences) of the extracted MS/MS data. The following settings were used for the database search: a maximum of two miss cleavages and methionine as a variable modification of oxidation (15.995); trypsin as enzyme; a permissible peptide mass tolerance of 10 ppm; a permissible fragment mass tolerance of 0.5 Da; an ion score of 25 as a cut-off. A decoy database search was conducted to determine the false discovery rate for the identity threshold ranging from 4.30% - 5.25% for the three replicates and 1.98% - 3.83% for the seven donor samples.

Progenesis software (Version 2.5; Nonlineair Dynamics Ltd, New Castle, UK) was used to calculate the reproducibility and variation of Fab on the basis of peak intensity (peak area) in the donor samples. In addition, the technical reproducibility of three replicate measurements was determined and calculated. The Progenesis software processes the raw data files in two steps: alignment, followed by normalization.[21-22] The data file that yielded most features (Donor 4) was used as reference, towards that the retention time of all other measurements were aligned and intensities (area under the peak) normalized. Correction for experimental variations was done by calculating the robust distribution of all ratios (log(ratio)). The peaks (features) are converted into an intensity list by using the intensities from the raw data files without any converting, except for peaks not observed in the raw data for that specific sample. If such peaks occur Progenesis generates a zero.

The data was filtered using the following criteria: peaks (features) with charge state two to seven and >2 isotopes. In addition, the last 10 minutes were filtered out of the MS run. A minimal threshold of at least 3 isotope peaks per peptide was set.

A matrix of all donor samples was generated, consisting of all masses with corresponding peak intensities (area under the peak) of every donor.

The Mascot search result files from all seven donors were used for the identification of CDR sequences. Regardless of the protein identification, the BLAST algorithm was then used to align all unique peptides from all seven donors identified by Mascot to databases containing human V, J or C-region germline sequences derived from the IMGT database (IMGT®, the international ImMunoGeneTics information system® http://www.imgt.org).[23] Peptides with a sufficient match (>70% sequence alignment) to the V-region database were assigned to a position on the immunoglobulin molecule of the specific germline sequence to which the peptide had been aligned according to the IMGT numbering system. This allows for consistent residue numbers in molecules

with varying CDR lengths. For CDR3 identification, high score (>70% sequence alignment) V-region and J-region sequences with high matching BLAST scores to V-region and J-region germline sequences were traced.
De novo sequencing was performed on the raw data files using the Peaks Studio 5.1 software package (Bioinformatics Solutions Inc., Waterloo, ON, Canada). In combination with a blast and IMGT alignment search against the Human J-region database (IMGT) the CDR3 peptide sequence was found as described above. The *de novo* score was assigned on the basis of the reliability of the amino acid (b- and y-ions; ALC%).

Statistical analysis

Coefficient of variance (CV%) was used to measure the reproducibility of the three repetitive injections of an identical Fab sample (technical variation). It was defined as the ratio of the standard deviation to the mean.
The seven healthy donors were randomly divided into two groups: group 1 consisted of four donors and group 2 of three. The variation in peak intensities between these groups was determined by calculating the p-value ($p<0.05$) using the ANOVA option (equivalent to two sample t-test) of the Progenesis software. Ten different compositions of the two groups were created randomly before analysis. There was no identical grouping. A matrix was generated for each of these ten compositions; each matrix comprised all masses with corresponding peak intensities of every donor. From each matrix, the difference in peak intensities (based on assigned p-value; $p>0.05$) between group 1 and 2 was determined.

Results

Fab isolation

Densitometry showed a recovery rate of ≥95% for IgG and ≥60% for Fab. The lower yield of Fab is mainly attributable to incomplete papain cleavage. The densitometry intensity of 5 purified Fabs had a coefficient of variation (CV) of 7.5% (n=5).

Technical reproducibility of Orbitrap FT mass spectrometry

To determine the variability between MS measurements, Fab isolated from one donor was measured three times on an LTQ Orbitrap XL. The Progenesis software revealed a total of 23,654 MS signals (average 23,645; SD 0.58) in the three injections. When the CV of the intensities of all these peaks observed in the three injections was calculated, 82% of all MS signals were found to have a CV of <20% (Table 1).

Table 1. Technical reproducibility of Orbitrap FT mass spectrometry.

Range of CV%	Number of masses	Cumulative %
>0-10%	13206	55.83%
>10-20%	6242	82.22%
>20-30%	2205	91.54%
>30-40%	917	95.42%
>40-50%	439	97.27%
>50-60%	208	98.15%
>60-70%	152	98.80%
>70-80%	69	99.09%
>80-90%	63	99.35%
More	153	100.00%
Total	23654	

Technical reproducibility of Fab purification injected 3 times on the LTQ Orbitrap XL. 82% of all the MS signals had a CV < 20%.

A database search against the NCBI database revealed that the 23,654 MS signals contained 1,515 peptide masses obtained from all three injections. These identified peptides included 1,458 different Ig peptide sequences with a range of 1,072-1104 per individual sample. From these 1,458 peptide sequences 768 sequences (53%) occurred in all three injections. Of the remaining sequences, 19% occurred in two injections, and 29% in one injection.

Variation in individual IgG molecules

All MS signals (43,217) obtained by the Orbitrap method in the seven donor samples were used to generate a Table (Table 2), which shows that 83% of the MS signals occurred in all seven donor samples. Figure 2a shows the distribution of all observed peaks of the seven healthy donors and Figure 2b shows the distribution of the database-dependent identified peaks of the seven healthy donors.

Table 2. Variation between 7 healthy donors in IgG Fab derived peptides.

mass occurrence	Fab 1	Fab 2	Fab 3	Fab 4	Fab 5	Fab 6	Fab 7	Total	%	Avg	SD
7x	35726	35726	35726	35726	35726	35726	35726	35726	82.7	35726.00	0.00
6x	3995	3985	3879	3099	3958	3864	2006	4131	9.6	3540.86	746.92
5x	1189	1212	1067	685	1134	1135	758	1436	3.3	1025.71	213.89
4x	479	490	390	262	484	481	398	746	1.7	426.29	83.93
3x	246	257	193	114	245	262	249	522	1.2	223.71	53.45
2x	107	106	70	63	122	157	169	397	0.9	113.43	39.97
1x	17	27	3	31	21	35	125	259	0.6	37.00	40.20
Total	41759	41803	41328	39980	41690	41660	39431	43217	100.0	41093.00	973.12

Variation between the 7 healthy donors in the number of Fab derived MS signals obtained by Progenesis. Avg: Average MS signals; SD: Standard deviation.

When Groups 1 and 2 were compared in terms of the mass intensity of each peptide mass (10 different randomizations), a very high resemblance (96%) was found, though in the remaining 4% there was a significant difference ($p<0.05$), ranging from 1%-8%. These results show that the observed immunoglobulin repertoire of these 7 healthy donors is very similar

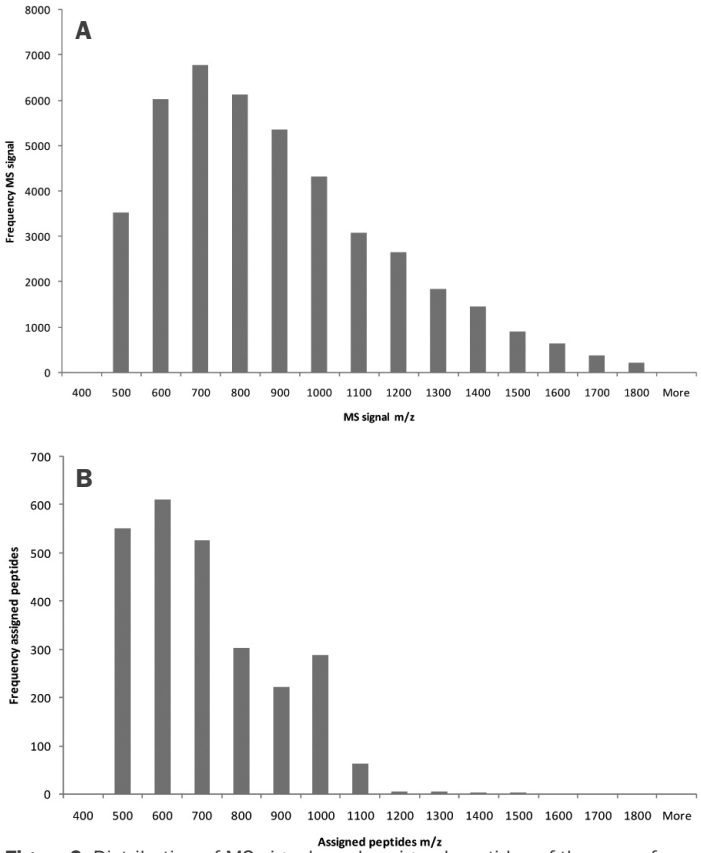


Figure 2. Distribution of MS signals and assigned peptides of the sera of seven healthy donors. The distribution of A) the total MS signals and B) the peptides assigned by the NCBInr database observed by the Progenesis software. On the x-axis bins of m/z are shown and on the y-axis the number of MS signals/ assigned peptides are shown, respectively.

Identification of CDRs

NanoLC Orbitrap MS measurements of the digested Fab of the seven donor serum samples yielded a combined compound list of 43,217 MS signals (average 41,093; SD 973.12) found by Progenesis. According to the IMGT database, 1755 peptide sequences from these MS signals corresponded to the V region, 109 sequences to the J region and 101 sequences to the C region (Figure 3).

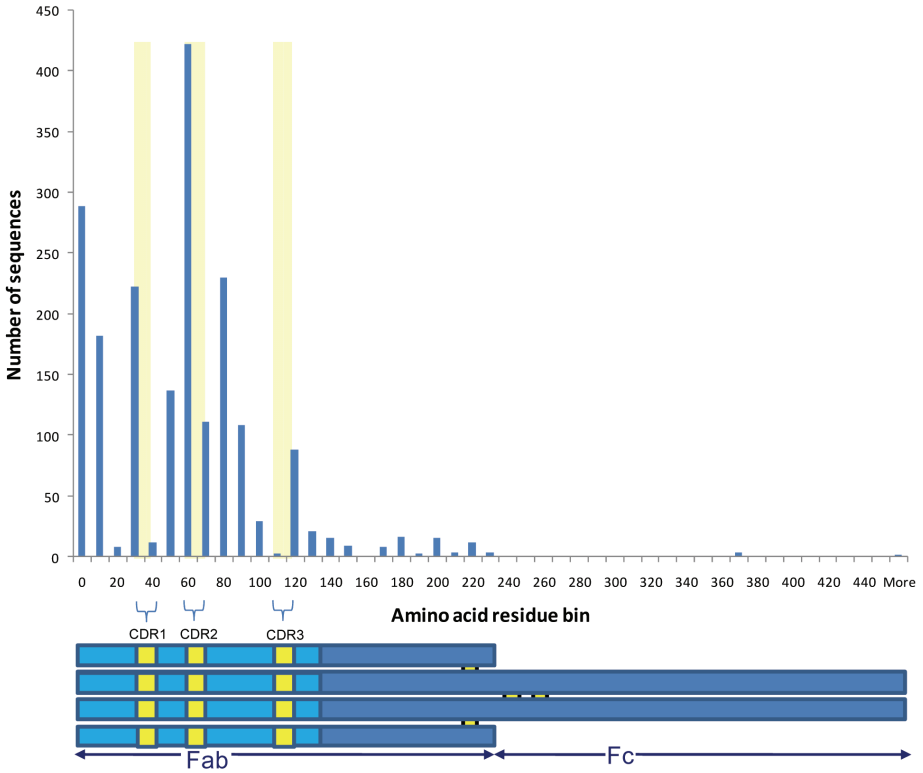


Figure 3. Overview of all immunoglobulin peptides sequenced. The start positions of the sequences are shown on the horizontal axis in bins of 10 amino acids. The blue bar indicates the number of sequences in each bin (vertical axis). The drawing below the graph shows the position of the variable and constant parts of the immunoglobulin molecule. CDRs are shown in yellow in the antigen-binding fragment (Fab) of the heavy and light chains.

The smaller number of C region sequences found is in line with the Fab recovery results shown earlier. According to the IMGT database, 225 peptide sequences corresponded to CDR1, 513 to CDR2 and 19 to CDR3. In Table 3 we show eight examples of sequences with mutations according to IMGT for each CDR. Many peptide sequences have been assigned to a CDR3 that includes parts of the V region as well as the J region. Two CDR3 peptide sequences (DSSGNHVVFGGGTK; DSSGNHLVFGGGTK) including a V region as well as a J region part are shown in Table 3.

Table 3. Overview of identified CDR sequences.

Sequence IMGT aligned CDR1 25.....43	Mutated aa	Subjects	Germline
SSQTIIY..TSNNK..	4	1/7	IGKV4-1*01
SSQSVLHNSDNK.....	3	2/7	IGKV4-1*01
.........SSAVGVGWVR	3	1/7	IGHV2-5*04
ASGYTF....TDYYIHWVR	2	5/7	IGHV1-46*03
SSQSILY..NSNNK	2	4/7	IGKV4-1*01
ASQHI......SNYINWYQQKPGK	2	1/7	IGKV1D-33*01
SSQTVLY..SSNNK	1	7/7	IGKV4-1*01
SSQSLLH...SDGR	1	4/7	IGKV2-29*03
Sequence IMGT aligned CDR2 49.....72	**Mutated aa**		**Germline**
GLEWVAHISPE..GTEEYYADSVK	7	1/7	IGHV3-7*01
GLEYMGLIYPG..DSDTK	3	1/7	IGHV5-51*04
...LLIYQT.......SR	3	1/7	IGLV2-11*01
..............SSALQTGVPSR	3	4/7	IGKV1D-17*02
...LLIHGA.......SNR	2	5/7	IGLV1-40*02
.......INPN..TGGTDYAQK	2	1/7	IGHV1-2*02
........SDG...ETTDYAAPVK	2	3/7	IGHV3-15*04
.......INSD..GSTTNYADSVK	2	1/7	IGHV3-74*01
Sequence IMGT aligned CDR3 97.....132	**Mutated aa**		**Germline**
SDDTAVYYCARTFGAGR.....	ND	1/7	IGHV1-18*01
.....EGWISALNGWGQGTLVTVSSASTK	ND	1/7	IGHJ3*01
..............RFDIWGQGTMVTVSSASTK	1	1/7	IGHJ3*02
..............YYGMDVWGQGTTVTVSSASTK	0	1/7	IGHJ6*02
.....DSSGNHVVFGGGTK	0	7/7	IGLV3-19*01/IGLV3-10*01
.....DSSGNHLVFGGGTK	1	3/7	IGLV3-19*01/IGLV3-10*01
.....IHITVCIDHWGQGTLVTVSSASTK	ND	1/7	IGHJ4
.....DTSGNHLVFGGGTK	2	1/7	IGLV3-19*01/IGLJ3*02

Examples of CDR1, 2, and 3 with germline matches and mutations are shown. Gray bars show the position of the CDRs. The mutated amino acids are shown in red and the amino acids with reliable MS2 spectra are underlined.

Figure 4 shows an example of the spectra of a CDR sequence m/z observed in all donor samples. The retention times generated by the Progenesis software were used from each individual sample.

To check if the CDR sequences assigned by Mascot with a score above 25 can also be confirmed by PEAKS, we performed a Mascot search and used PEAKS to *de novo* sequence the raw data files. We found a 73% overlap between the sequences assigned by PEAKS and Mascot-assigned sequences. In this calculation we did not take into account changes in the assignment of isoleucine and leucine (isobaric amino acids).

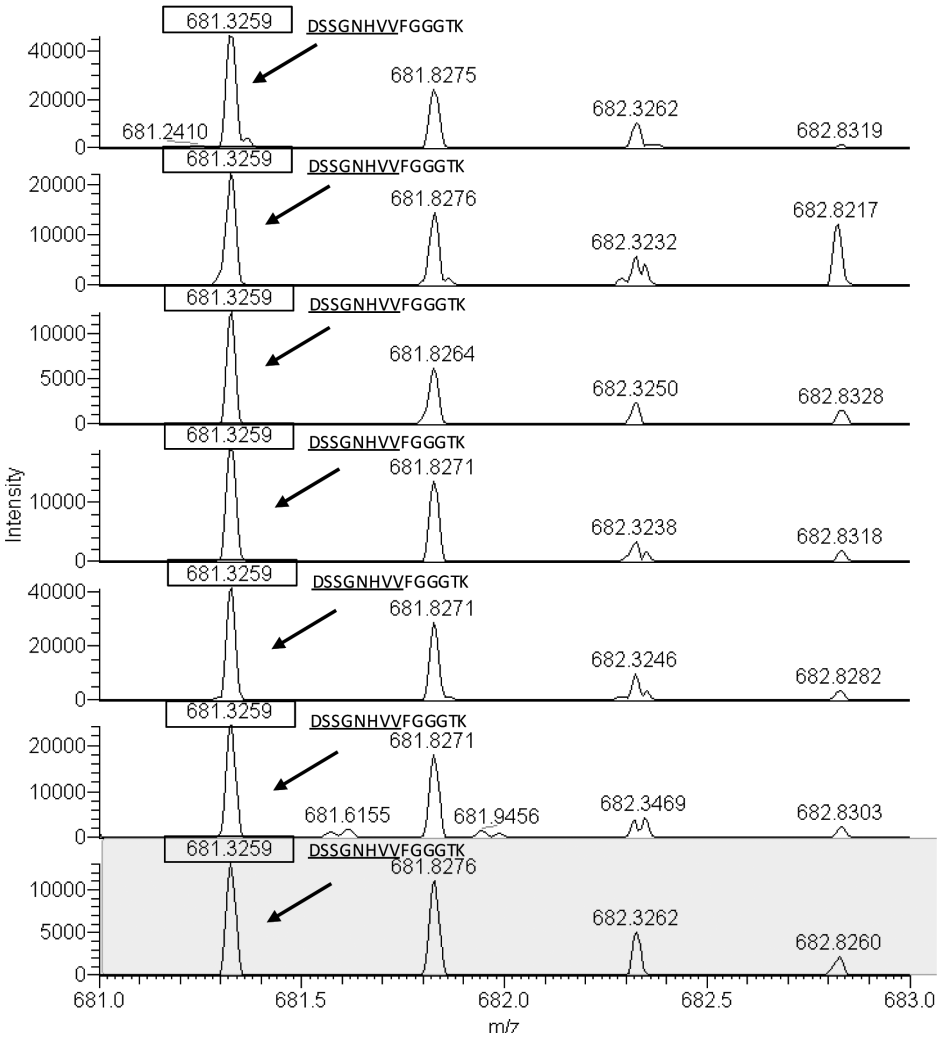
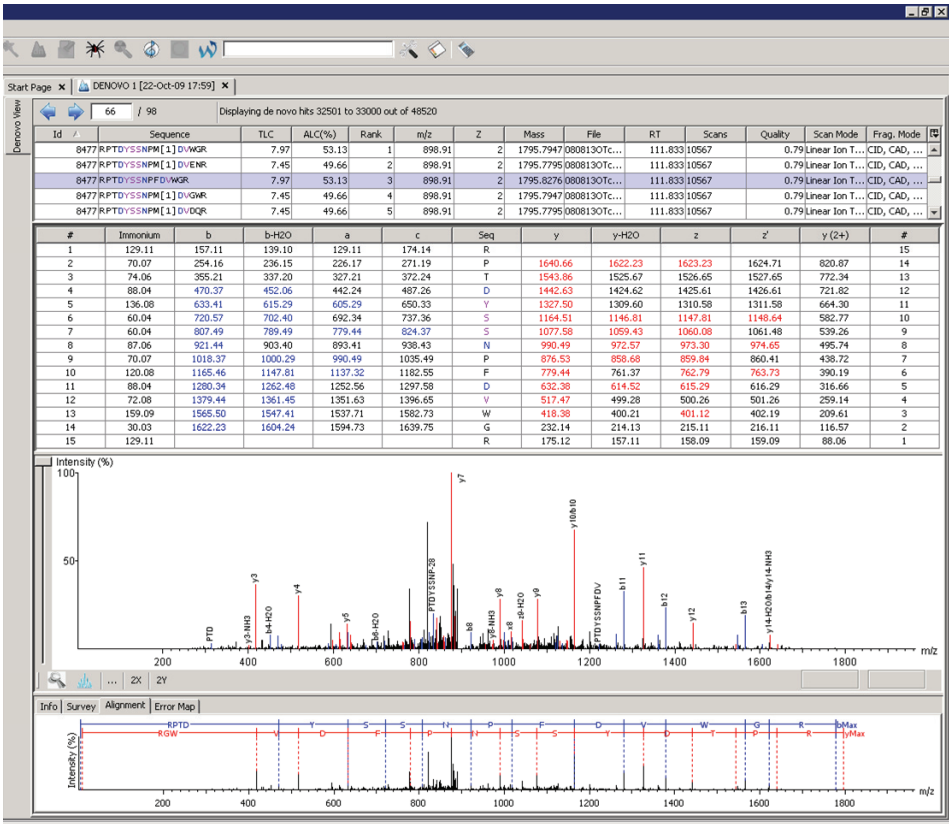


Figure 4. Spectra of an identified CDR3 sequence in various serum samples. The mass spectra of the CDR3 sequence DSSGNHVVFGGGTK are shown with m/z 681.3259 (boxes). This peptide was observed in all sera of the healthy donors.

De novo sequencing of unmatched peptides

Figure 5a shows an example of *de novo* sequencing data of an unidentified mass spectrum from one Fab from one of the seven donors (m/z 898.910, z=2). We screened *de novo* sequences generated by the PEAKS software package by aligning them against a database containing J-regions (IMGT). The example (ALC% 53%) in Figure 5a shows homology to the IGHJ3*01 germline sequence (Figure 5b), and extends into the CDR3. This makes it possible to perform *de novo* sequencing, and to identify the result as a CDR3 based on homology to the conserved region located C-terminal and N-terminal to the highly variable section of the CDR with a 63% identity score.

A



B

de novo sequence: 8 NPFDVWGR 15

FDVWG

Germline: 1 DAFDVWGQ 8

Figure 5. *De novo* sequencing of a non-assigned peptide by the NCBInr database. An example of *de novo* sequencing of a non-assigned peptide with an m/z of 898.910 generated by the PEAKS Studio 5.1 software package. A) The calculated sequences with a confidence score (ALC%) of this peptide are shown; b (blue) and y (red) ions of the sequence with the highest confidence are listed; MS/MS line mass spectrum showing peaks with corresponding masses. B and y ions resulting in an amino acid sequence are shown in the lowest panel. B) Alignment of the *de novo* sequenced peptide with the germline J region sequence (IGHJ3*01). The numbers indicated correspond for the *de novo* sequence to the amino acid position within the total *de novo* sequence. For the germline sequence the numbers indicate the position of the *de novo* sequence aligned to the J region germline. The middle line is the overlap between the two sequences.

Discussion

The immunoglobulin repertoire (immunomics) present in patients' sera can serve as a molecular signature in cancer and autoimmune disorders. The utility of IgG Fab as a marker for cancer and autoimmune diseases depends on two factors: 1) the reproducibility of IgG Fab purification and identification, and 2) the ability to determine qualitative and quantitative differences in IgG Fab between healthy donor serum samples by high-resolution MS.

Using our approach, we were able to demonstrate that 17% of the 43,217 MS signals did not overlap between the seven donor samples and that 4% of the MS signals differed significantly in peak intensity. In addition, we found that the method used for Fab purification is fast, reproducible and leads to high recovery rates and eventually provides only one sample for subsequent sequencing by high-resolution MS.

A large number of masses measured in the Fab samples did not result in a peptide identification after a database (NCBInr) analysis by Mascot. This does not reflect a poor sample quality, but rather the poor coverage in the database of mutated and rearranged immunoglobulins. Only a small fraction of the possible sequences is represented in the (NCBInr) database, and hence the remaining peptide masses will not be identified by such a database approach. In addition, there is need for high quality MS2 spectra in order to be able to identify these peptide masses by *de novo* sequencing. Only in some subjects it was possible to identify a peptide that was present in the database, yet the corresponding parent mass with high mass accuracy actually was present in all samples. Hence, the absence of a signal can be explained by either real absence in a sample or presence below the detection threshold of the instrument. At higher concentrations, the signal will initially show up as a mass at MS1 level, but when the concentration is high enough it can also be triggered for an MS2 spectrum. When a peptide is present in the database, it can be identified by search engine programs like Mascot. Otherwise, *de novo* sequencing can be used to elucidate an entire or partial sequence for a particular mass, potentially assisted by a database of homologous sequences during the *de novo* process and the high mass accuracy of the Orbitrap mass spectrometer. Using peptides identified by Mascot via NCBInr database, 59% had an overlap with the results revealed by PEAKS (Mascot score >25). Peptides sequences obtained by *de novo* sequencing are difficult to assign to a CDR3, especially CDR3 of the heavy chain, as it is a highly diverse region of the immunoglobulin molecule. Due to the combination of V, D, and J genes, and especially the insertions, deletions and frameshifts around the D gene generated by J diversity, it is difficult to *de novo* sequence high quality CDR3 sequences. We should keep in mind that *de novo* sequencing technique is not yet totally reliable because the start and end amino acids are still difficult to *de novo* sequence. *De novo* sequencing will become increasingly successful when better and more MS2 spectra are available for the peptide of interest. Improvement can be achieved by optimizing the MS method by acquiring MS2 spectra with high resolution and high mass accuracy. Furthermore, utilizing developments in column chromatography that allow more loading without reducing the resolution of the separation will result in better quality and more MS2 spectra. In this way, signals of better quality can be acquired by the mass spectrometer.

In our donors, several peptides derived from CDRs were found with mutations from their germline sequence. We assessed the likelihood of finding the same mutated CDR peptide in different donors. For a CDR1 or CDR2 peptide with a length of 15 residues, the number of possible sequences after 1, 2, 3, and 4 mutations on a random position is about 10^2, 10^5, 10^7 and 10^9 respectively. As human blood contains approximately 10^{12} B-cells, and given that many B-cells belong to a clonal subpopulation, the maximum number of different antibodies that can be found in a subject must be less than 10^{12}. Typical examples of CDR1 and CDR2 sequences show up to 4 mutations; these numbers suggest there is a reasonable probability that such sequences are shared by individuals based on statistics alone.

Due to the combination of V, D, and J genes, the CDR3 of the heavy chain is much more random. In particular, the junctional diversity that results in insertions, deletions, and frameshifts around the D gene yields an almost random amino acid sequence within the CDR3.[1] The combinatorial diversity is about 10^6: a random section of 6 amino acids adds enough diversity to surpass the number of B-cells in a human, and is well within the range of CDR3 lengths.[24] Somatic mutations add even more diversity, making it statistically unlikely that a CDR3 peptide is found in multiple subjects by chance alone. Such sharing has nevertheless been described in the literature, which suggests that antibody sequences do not arise totally at random, but rather emerge from a convergent process.[25] The large overlap (83%) in MS signals observed from the seven healthy donors supports the idea that mutated CDRs are shared between individuals more than would be expected by chance. In our small panel of masses that contained a confirmed highly variable CDR3 of the heavy chain (e.g. EGWISALNGWGQGTLVTVSSASTK), we only observed this peptide in the donor in which it was identified, even when looking at the parent mass (MS1). However, we observed identical identified mutated CDR sequences with a limited number of mutations between the seven donors, which suggests that a nonrandom development of antibodies to certain CDRs occurs, rather than a random selection of all possible CDRs. The requirement that antibodies bind a particular antigen structure will severely reduce the sequence diversity. This is illustrated by the phenomenon of repertoire bias, which describes the preferential use of particular germline V genes in response to particular antigens. Other studies have also found similar CDR3 sequences in human subjects, and an extensive DNA sequencing project performed by Weinstein et al. including nucleotide sequencing of the VDJ repertoire of zebrafish confirmed that CDRs are not generated fully at random.[25-26] A study in T-cell TCRβ sequences in mice revealed similarity between different individual animals.[27] These data indicate that the detection and analysis of antibody-derived peptides may be a viable prognostic or diagnostic tool.

The variation between individual IgG molecules in donor sera can be expressed both by the mass identity and concentration (peak intensity) of the peptide masses detected. Based on the fact that we have been able to demonstrate 17% differences in the 43,217 MS signals between the seven healthy donor samples, and that 4% of the MS signals differed significantly in peak intensity, we may conclude that our approach is suitable for identifying differences in IgG variable regions between healthy controls and that despite the variation observed among normal controls, this approach may also be used to detect specific differences between cancer or autoimmune case and

control groups. Furthermore, it may be possible to improve the method by molecular dissection of Fab into CDR parts and column chromatography with the highest resolution characteristics, in order to reduce the complexity of the IgG molecule. As tryptic digestion has its limitations, the use of other cleavage enzymes could be an option to obtain heavy chain CDR3 sequences. For example, at the beginning of most CDR3 sequences is a lysine or arginine, the cleavage sites for trypsin, which makes it difficult to identify the CDR3 regions with part of the flanking sequences. These flanking regions facilitate the sequencing of the CDR regions. Alternative enzymes could help in this respect.

Obermeier et al. used primary antibody structures determined by mass spectrometry in multiple sclerosis B cell samples. They compared the immunoglobulin transcriptomes of B cells with the corresponding immunoglobulin proteomes in CSF of four patients with multiple sclerosis, and created an immunoglobulin subject-specific database from the B-cell cDNA transcripts from the CSF of these four patients. They analyzed the IgG proteome by purifying IgG, which was then separated by isoelectric focusing (IEF), subsequently followed by mass spectrometry.[5] They were able to identify 13-46 peptides related to VDJ, somatic hypermutations or CDR3 regions in the four patients. The positions of these CDRs were established according to the results of the study of Kabat et al..[28] The authors concluded that the immunoglobulin transcriptome of the four patients correlated well with the immunoglobulin proteome.

Unlike Obermeier et al., we used the whole IgG Fab purified method, and in contrast with IEF technology, our approach provides only one Fab sample fraction for analysis, opening ways to compare larger numbers of samples. By choosing the positions of the CDRs in the different IgG variable region proteins according to IMGT, we were able to sequence 757 different CDRs from the purified Fab. The large difference in the number of different CDRs detected is attributable to the use of different databases, different analytical methods (mass spectrometer) and the fact that Obermeier et al. used CSF instead of serum.

Three of the other approaches to identify new antibodies as biomarkers for cancer or autoimmune diseases are 1) serological expression cloning (SEREX), 2) a combination of the phage display method and serological spot assays, and 3) a combination of autoantibody purification and protein microarray.[15, 18-19, 29-32] All three approaches are primarily based on the identification of new disease-specific antigens (present or absent), such as the cancer antigens ROCK1, KIAA1344, SOX2, SART1, MUPP1 and Ubiquilin 1. These newly identified antigens are subsequently used to identify disease-related autoantibodies.

Our goal is to search directly for disease-related antibodies without a required prior knowledge of the antigen. Our technique detects by MS if changes occur at primary amino acid structure level, we can detect them by mass spectrometry. This approach will make it possible to select and identify specific antibodies related to an autoimmune disease or cancer. Using serological assays, one is more restricted to searching for one specific antibody for an antigen.

In autoimmune diseases and cancer, disease-related antibodies are produced by plasma cells and circulate in the blood at relatively high concentrations, which facilitates their detection. In contrast, antigens (including auto antigens) and other proteins and peptides secreted by tumors occur in much lower concentrations in the blood (e.g.

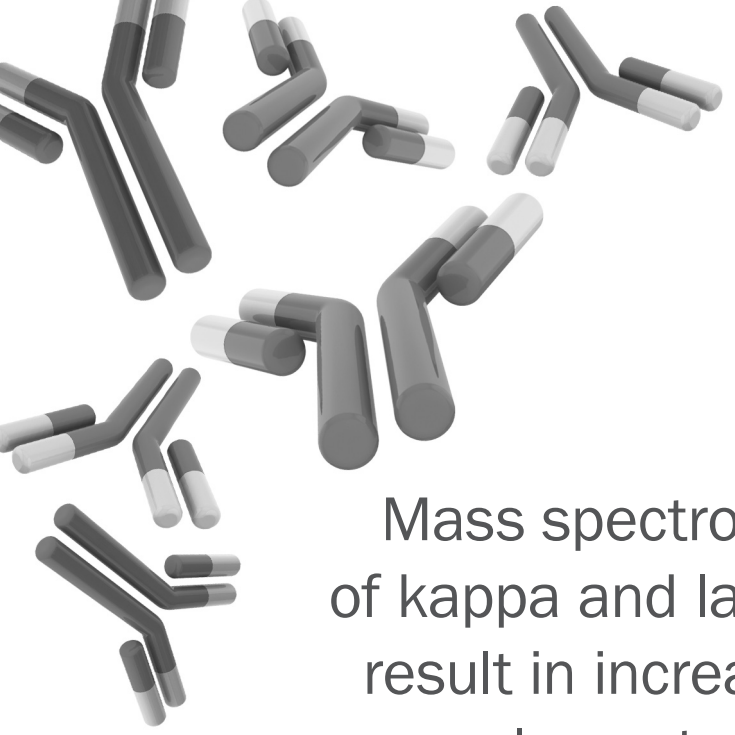
PSA <6.5 µg/L; CA125 <35 kU/L) due to a lower rate of production, degradation and specific clearance.[2, 33] Therefore, antibodies have much more potential as prognostic or diagnostic markers than antigens.

In conclusion, the IgG Fab isolation approach used was not only fast and reproducible but also provided a high yield and produced only one Fab sample fraction for subsequent sequencing by high-resolution Orbitrap MS. Because it identifies qualitative and quantitative differences in Fabs between healthy donors, this method may have an important impact on the prognostic and diagnostic marker discovery of cancer and autoimmune diseases by identifying CDRs of disease-specific antibody fragments without a required prior knowledge of the antigen.

References

1. de Wildt RM, van Venrooij WJ, Winter G, Hoet RM, Tomlinson IM. Somatic insertions and deletions shape the human antibody repertoire. J Mol Biol 1999; 294:701-10.
2. Murphy K. TP, Walport M. Janeway's immunobiology. 7th ed: Garland Science; 2008.
3. Tonegawa S. Reiteration frequency of immunoglobulin light chain genes: further evidence for somatic generation of antibody diversity. Proc Natl Acad Sci U S A 1976; 73:203-7.
4. Hieter PA, Maizel JV, Jr., Leder P. Evolution of human immunoglobulin kappa J region genes. J Biol Chem 1982; 257:1516-22.
5. Obermeier B, Mentele R, Malotka J, Kellermann J, Kumpfel T, Wekerle H, et al. Matching of oligoclonal immunoglobulin transcriptomes and proteomes of cerebrospinal fluid in multiple sclerosis. Nat Med 2008; 14:688-93.
6. Ravetch JV, Siebenlist U, Korsmeyer S, Waldmann T, Leder P. Structure of the human immunoglobulin mu locus: characterization of embryonic and rearranged J and D genes. Cell 1981; 27:583-91.
7. Tomlinson IM, Walter G, Marks JD, Llewelyn MB, Winter G. The repertoire of human germline VH sequences reveals about fifty groups of VH segments with different hypervariable loops. J Mol Biol 1992; 227:776-98.
8. Williams SC, Frippiat JP, Tomlinson IM, Ignatovich O, Lefranc MP, Winter G. Sequence and evolution of the human germline V lambda repertoire. J Mol Biol 1996; 264:220-32.
9. Balkwill F, Mantovani A. Inflammation and cancer: back to Virchow? Lancet 2001; 357:539-45.
10. Dunn GP, Bruce AT, Ikeda H, Old LJ, Schreiber RD. Cancer immunoediting: from immunosurveillance to tumor escape. Nat Immunol 2002; 3:991-8.
11. Dunn GP, Old LJ, Schreiber RD. The immunobiology of cancer immunosurveillance and immunoediting. Immunity 2004; 21:137-48.
12. Kim R, Emi M, Tanabe K. Cancer immunoediting from immune surveillance to immune escape. Immunology 2007; 121:1-14.
13. de Visser KE, Korets LV, Coussens LM. *De novo* carcinogenesis promoted by chronic inflammation is B lymphocyte dependent. Cancer Cell 2005; 7:411-23.
14. Garcia BH, 2nd, Hargrave A, Morgan A, Kilmer G, Hommema E, Nahrahari J, et al. Antibody microarray analysis of inflammatory mediator release by human leukemia T-cells and human non small cell lung cancer cells. J Biomol Tech 2007; 18:245-51.
15. Leidinger P, Keller A, Ludwig N, Rheinheimer S, Hamacher J, Huwer H, et al. Toward an early diagnosis of lung cancer: an autoantibody signature for squamous cell lung carcinoma. Int J Cancer 2008; 123:1631-6.
16. Pan J, Chen HQ, Sun YH, Zhang JH, Luo XY. Comparative proteomic analysis of non-small-cell lung cancer and normal controls using serum label-free quantitative shotgun technology. Lung 2008; 186:255-61.
17. Yildiz PB, Shyr Y, Rahman JS, Wardwell NR, Zimmerman LJ, Shakhtour B, et al. Diagnostic accuracy of MALDI mass spectrometric analysis of unfractionated serum in lung cancer. J Thorac Oncol 2007; 2:893-901.
18. Chen YT, Gure AO, Scanlan MJ. Serological analysis of expression cDNA libraries (SEREX): an immunoscreening technique for identifying immunogenic tumor antigens. Methods Mol Med 2005; 103:207-16.

19. Sahin U, Tureci O, Schmitt H, Cochlovius B, Johannes T, Schmits R, et al. Human neoplasms elicit multiple specific immune responses in the autologous host. Proc Natl Acad Sci U S A 1995; 92:11810-3.
20. Deocharan B, Zhou Z, Antar K, Siconolfi-Baez L, Angeletti RH, Hardin J, et al. Alpha-actinin immunization elicits anti-chromatin autoimmunity in nonautoimmune mice. J Immunol 2007; 179:1313-21.
21. http://www.nonlinear.com/support/progenesis/lc-ms/faq/how-alignment-works.aspx.
22. http://www.nonlinear.com/support/progenesis/lc-ms/faq/how-normalisation-works.aspx.
23. Lefranc MP, Giudicelli V, Ginestoux C, Jabado-Michaloud J, Folch G, Bellahcene F, et al. IMGT, the international ImMunoGeneTics information system. Nucleic Acids Res 2009; 37:D1006-12.
24. Wang X, Wu D, Zheng S, Sun J, Tao L, Li Y, et al. Ab-origin: an enhanced tool to identify the sourcing gene segments in germline for rearranged antibodies. BMC Bioinformatics 2008; 9 Suppl 12:S20.
25. Poulsen TR, Meijer PJ, Jensen A, Nielsen LS, Andersen PS. Kinetic, affinity, and diversity limits of human polyclonal antibody responses against tetanus toxoid. J Immunol 2007; 179:3841-50.
26. Weinstein JA, Jiang N, White RA, 3rd, Fisher DS, Quake SR. High-throughput sequencing of the zebrafish antibody repertoire. Science 2009; 324:807-10.
27. Venturi V, Kedzierska K, Price DA, Doherty PC, Douek DC, Turner SJ, et al. Sharing of T cell receptors in antigen-specific responses is driven by convergent recombination. Proc Natl Acad Sci U S A 2006; 103:18691-6.
28. Kabat EA, Wu TT. Identical V region amino acid sequences and segments of sequences in antibodies of different specificities. Relative contributions of VH and VL genes, minigenes, and complementarity-determining regions to binding of antibody-combining sites. J Immunol 1991; 147:1709-19.
29. Chen G, Wang X, Yu J, Varambally S, Thomas DG, Lin MY, et al. Autoantibody profiles reveal ubiquilin 1 as a humoral immune response target in lung adenocarcinoma. Cancer Res 2007; 67:3461-7.
30. Comtesse N, Zippel A, Walle S, Monz D, Backes C, Fischer U, et al. Complex humoral immune response against a benign tumor: frequent antibody response against specific antigens as diagnostic targets. Proc Natl Acad Sci U S A 2005; 102:9601-6.
31. Ludwig N, Keller A, Comtesse N, Rheinheimer S, Pallasch C, Fischer U, et al. Pattern of serum autoantibodies allows accurate distinction between a tumor and pathologies of the same organ. Clin Cancer Res 2008; 14:4767-74.
32. Qin S, Qiu W, Ehrlich JR, Ferdinand AS, Richie JP, O'Leary M P, et al. Development of a "reverse capture" autoantibody microarray for studies of antigen-autoantibody profiling. Proteomics 2006; 6:3199-209.
33. Traggiai E, Puzone R, Lanzavecchia A. Antigen dependent and independent mechanisms that sustain serum antibody levels. Vaccine 2003; 21 Suppl 2:S35-7.



Chapter 3

Mass spectrometry analyses of kappa and lambda fractions result in increased number of complementarity determining regions identifications

Ingrid Broodman, Dominique de Costa, Christoph Stingl, Lennard J.M. Dekker, Martijn M. VanDuijn, Jan Lindemans, Rob J. van Klaveren, Theo M. Luider

Proteomics 2012; 12:183-91

Abstract

Sera from lung cancer patients contain antibodies against tumor associated antigens. Specific amino acid sequences of the complementarity determining regions (CDRs) in the antigen-binding fragment (Fab) of these antibodies have potential as lung cancer biomarkers. Detection and identification of CDRs by mass spectrometry can significantly be improved by reduction of the complexity of the immunoglobulin molecule. Our aim was to molecular dissect IgG into kappa and lambda fragments to reduce the complexity and thereby identify substantially more CDRs than by just total Fab isolation. We purified Fab, Fab-κ, Fab-λ, κ and λ light chains from serum from 10 stage I lung adenocarcinoma patients and 10 matched controls from current and former smokers. After purification, the immunoglobulin fragments were enzymatically digested and measured by high-resolution mass spectrometry. Finally, we compared the number of CDRs identified in these immunoglobulin fragments with that in the Fab fragments. Twice as many CDRs were identified when Fab-κ, Fab-λ, κ and λ (3330) were combined than in the Fab fraction (1663) alone. The number of CDRs and κ:λ ratio was statistically similar in both cases and controls. Molecular dissection of IgG identifies significantly more CDRs, which increases the likelihood of finding lung cancer-related CDR sequences.

Introduction

Only 15-20% of all lung cancers are detected at an early and potential curable stage today (American Cancer Society, Cancer Facts & Figures 2010 http://www.cancer.org/acs/groups/content/@nho/documents/document/acspc-024113.pdf). An early detection and treatment of lung cancer can reduce the high lung cancer mortality rate. This is currently investigated in several randomized lung cancer CT screening trials [1-4]. At the moment, there is no early detection biomarker for lung cancer available. Biomarkers could be used to stratify people according to their risk to develop lung cancer. The different strata could, dependent on their cancer risk, be invited for baseline CT screening and for subsequent screening rounds. A biomarker for early detection of lung cancer could be used as a complement to CT screening in order to reduce the rate of false-positive tests results and the number of unnecessary biopsies, surgical interventions or serial CT scans [5].

There is increasing evidence that during tumor development a humoral immune response evolves to various tumor types, including lung cancer [6-8]. Immunoglobulins against different tumor associated antigens (TAAs) in lung cancer have been identified by different strategies [9-14] up to 5 years before the tumor was detectable by a CT scan [15-16]. These strategies use immunoglobulins to identify the targeted tumor antigens as potential biomarkers, rather than using the reactive immunoglobulins as potential biomarkers. In contrast to antigens, immunoglobulins are excreted and circulate in the blood at relatively high levels, which support their detection.

We previously described a new approach in which tryptic fragments of the immunoglobulins themselves are used as potential biomarkers [17]. Three hypervariable complementarity determining regions (CDR1, CDR2 and CDR3) in the variable regions of the light and heavy chains of an immunoglobulin form the binding surface complementary to the antigen. As such, these CDRs determine the specificity of the immunoglobulin to the antigen. During immune response and B-cell development, CDRs are generated by somatic rearrangements of different (V, or V, D and J) germline genes to form a specific combination. In both light and heavy chains, the diversity of CDR3 is even further enhanced by the insertions and deletions of nucleotides between the genes. The estimated potential immunoglobulin diversity varies from 10^{13} to more than 10^{50} [18-19]. Despite this large range there is evidence for repertoire bias, which means that certain germline genes are preferentially used in response to a particular antigen [20-21]. Moreover, similar and identical CDR3 sequences have been found in humans and in zebrafish, respectively [22-23]. Our hypothesis is that a specific molecular profile of CDRs may distinguish lung cancer patients from controls and can thus be used as lung cancer biomarker.

The ability to find differences in CDRs between lung cancer cases and controls depends on the number of CDRs identified, which in turn depends on the accuracy, resolution, sensitivity and reproducibility of the mass spectrometry to identify these very low-abundant CDR peptides. However, ion suppression in the mass spectrometer especially for complex peptide mixtures can reduce the sensitivity [24]. Reduction of this complexity reduces ion suppression and leads to a significantly higher sensitivity to detect CDR peptides. In our previous paper, we presented our method to sequence Fab fragments by using mass spectrometry [25]. To identify as many CDRs as possible, the complexity of the immunoglobulin molecule can be reduced by separating Fab into Fab-κ and Fab-λ, and even further by purifying only the kappa (κ) or lambda (λ) light chain. The normal overall κ: λ ratio in human immunoglobulins is approximately

2 (κ:λ of: IgG 2.34 ± 0.80; IgA 1.59 ± 0.40; IgM 1.86 ± 0.76) with most of the immunoglobulins consisting of IgG [26].
Our aim was to use molecular dissection of IgG into kappa and lambda fragments to identify substantially more CDRs than obtained by the Fab method. To determine if we would be able to identify more CDRs by molecular dissection of IgG in kappa and lambda fragments than of Fab, we designed a pilot study. In this study, we purified Fab, Fab-κ, Fab-λ, κ and λ light chains from serum from 10 stage I lung adenocarcinoma patients and 10 matched controls from current and former smokers of the NELSON trial [3]. After purification, the immunoglobulin fragments were enzymatically digested by trypsin and measured by high-resolution mass spectrometry. Finally, we compared the number of CDRs identified in these immunoglobulin fragments with the number of CDRs identified in the Fab fragments.

Materials and Methods

Cases and Controls from the NELSON Trial

Sera from 20 current and former smokers were obtained from the Dutch-Belgian randomized lung cancer screening trial (NELSON), as described previously [3], and collected under uniform conditions. The subjects were between 53 and 73 years of age (50 % males and 50% females, median age 61 years) and had a smoking history of 3-7 cigarettes per day for 6-11 years. Ten serum samples of stage I lung adenocarcinoma patients without history of other cancer were collected. As non-cancer controls, 10 matched serum samples were taken from participants in the same trial. The controls were matched for gender, smoking status, COPD status, absence of previous cancer and asbestos history. All participants gave written informed consent as approved by the Dutch Minister of Health and the ethics board at each participating center. Samples were blinded and analyzed in random order.

Reference Sample

One reference donor sample (male; 59 years), with a normal serum IgG of 9.75 g/L, was used as a quality control for each analysis step [25]5252. In accordance with the general guidelines of the Sanquin Blood Bank Rotterdam (The Netherlands), the healthy donor gave written consent for the serum to be used for scientific research.

Purification of IgG

Serum IgG (80 µL) was purified using the Melon Gel IgG purification kit (Pierce, Rockford, IL), according to the manufacturer's instructions. The concentration of the purified IgG protein (800 µL) was determined by means of the mass extinction coefficient of 1.37 (mg/mL) cm^{-1} at 280 nm on a NanoDrop Spectrophotometer (ND-1000, Nanodrop Technologies, Wilmington, DE).

Purification of Fab

After purification, purified IgG (400 µL) was digested in Fab and Fc fragments overnight by immobilized papain on agarose beads according to the manufacturer's instructions (Pierce, Rockford, IL). Then this digest (2.800 mL) was concentrated approximately ten times by an Amicon Ultra 3K centrifugal filter device (Millipore, Amsterdam, the Netherlands).
Finally, the Fab fragments were separated from Fc fragments and undigested IgG by SDS-PAGE under reducing conditions [25]. The proteins were fixed and visualized

with the Colloidal Blue staining kit (Invitrogen, Breda, the Netherlands) and gels were washed for a minimum of 4 h in deionized water (Figure 1).

Purification of IgG Fab kappa and IgG Fab lambda

For purification of IgG Fab kappa (Fab-κ) 100 µL λ-specific-anti-human IgG and for purification of IgG Fab lambda (Fab-λ) 200 µL κ-specific-anti-human IgG was immobilized onto the MicroLink Protein support of two AminoLink Plus coupling gel spin column (Pierce, Rockford, IL). Concentrated papain digested IgG (100 µL) was loaded onto the columns and incubated at 4 °C overnight with gentle end-over-end mixing. Finally, the IgG Fab-κ and IgG Fab-λ were individually collected in the flow-throughs and separated from the Fc proteins by SDS-PAGE under reducing conditions and visualized as described above (Figure 1).

Purification of Kappa and Lambda

The Melon Gel purified IgG (100 µL) was loaded onto the λ-specific-anti-human IgG (100 µL) and κ-specific-anti-human IgG (200 µL) columns and incubated at 4 °C overnight with gentle end-over-end mixing. Finally, the IgG-κ and IgG-λ were individually collected in the flow-throughs and the heavy chain (H) and light chain, κ or λ, were separated by SDS-PAGE under reducing conditions and visualized as described above (Figure 1).

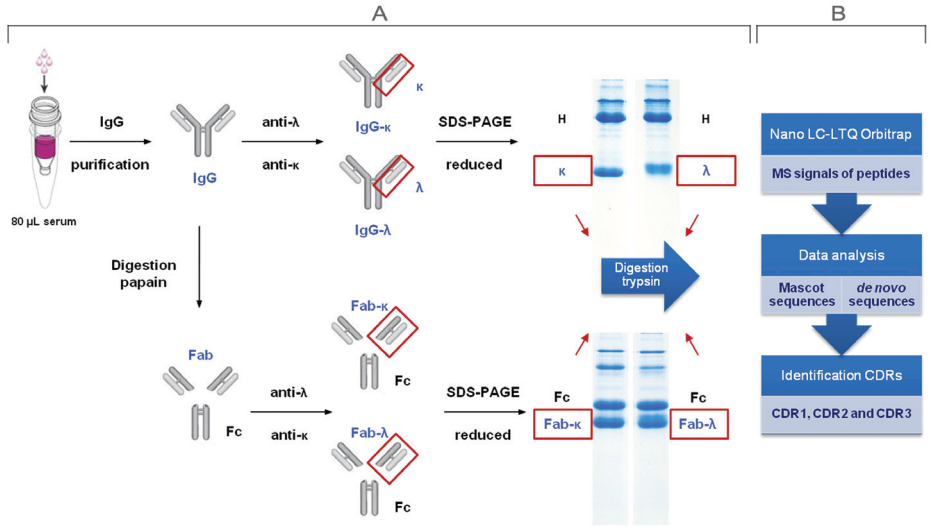


Figure 1. Flow-chart of the different steps in molecular dissection and data analysis. A) Purified IgG and concentrated papain digested IgG were loaded onto anti-λ IgG and anti-κ IgG columns. Subsequently, IgG-κ, IgG-λ, Fab-κ and Fab-λ were individually collected in the flow-throughs. The heavy chain (H) and light chain (κ or λ) of IgG-κ and IgG-λ, and Fab and Fc of IgG Fab-κ and IgG Fab-λ were separated by SDS-PAGE under reducing conditions. B) After in-gel-tryptic digestion, the peptides were measured by nano-LC-LTQ Orbitrap (MS). MS signals were quantified by Progenesis and analyzed by Mascot and *de novo* sequencing. Mascot and *de novo* peptide sequences were used for the identification of CDR sequences.

Sample Preparation for LTQ Orbitrap Mass Spectrometry

Recovery and reproducibility of the purifications were determined by densitometry. Intensities of the protein bands of the reference sample on the stained SDS-PAGE gel were quantified by scanning on a Molecular Imager GS-800 Calibrated densitometer

(Bio-Rad, Veenendaal, the Netherlands) with Quantity One® 1-D analysis software (version 4.6.5; Bio-Rad).
After imaging and analysis of the SDS-PAGE gels, the selected protein bands were excised from the gels and cut into plugs. The in-gel trypsin digestion was performed in Rapigest detergent solution according to the manufacturer's instructions (Waters, Milford, MA).

Nano-LC Orbitrap MS Analyses

LCMS measurements of the tryptic peptides were performed on an Ultimate 3000 nano-LC system (Dionex, Amsterdam, Netherlands) online coupled to a hybrid linear ion trap/Orbitrap MS (LTQ Orbitrap XL; Thermo Fisher Scientific, Bremen, Germany). For identification of the IgG peptides we used CAD fragmentation. High resolution full scan MS was obtained in the Orbitrap (resolution 30,000; AGC 1,000,000) and CAD fragmentation was performed on the five most abundant masses in the full scan spectra.

Data Analysis

Progenesis software (Version 2.5; Nonlinear Dynamics Ltd, Newcastle, UK) was used for the label-free quantification of MS data. In total five Progenesis analyses were performed, one for each individual dissected IgG fraction. The raw data files were aligned by their retention time, features were selected and intensities were normalized (Nonlinear Dynamics http://www.nonlinear.com/support/progenesis/lc-ms/faq/how-alignment-works.aspx, http://www.nonlinear.com/support/progenesis/lc-ms/faq/how-normalisation-works.aspx) [25]. Data matrices containing the feature intensities (area under the peak) were exported for further calculations.
Database searches were performed with Mascot (version 2.2.06; Matrix Science Inc., London, UK) against the NCBInr human database (version nrHuman_database_20090311; Homo sapiens species restriction; 222,0660 sequences). Parameters used for the database search were as follows: a maximum of two miss cleavages; carbamidomethylation of cysteine as a fixed modification and oxidation of methionine as a variable modification; trypsin as enzyme; a peptide mass tolerance of 10 ppm; a fragment mass tolerance of 0.5 Da; an ion score of 25 as a cut-off.
De novo sequencing was used for features not identified by a Mascot search against the database (NCBInr). Therefore, raw data files were processed by the Peaks Studio 5.1 software package (Bioinformatics Solutions Inc., Waterloo, ON, Canada). The Average Local Confidence score (ALC%) was assigned on the basis of the positional confidence for each amino acid in the peptide sequence divided by the total number of amino acids.
Peptide identifications from both Mascot and Peaks were imported into Progenesis, which keeps the best scoring sequence for each MS signal. To this end, Peaks data were manually converted to Mascot XML format for import into Progenesis, and the ALC% scores were divided by a factor of 100. By doing so Mascot scores obtained from the data-dependent search always overruled the ALC% scores. Finally, all intensities and sequences from Mascot and Peaks were combined in a single Progenesis file per individual fraction for further analysis.
Mascot and *de novo* peptide sequences were used for the identification of CDR sequences. Irrespective of the protein identification, the BLAST algorithm was subsequently used to align all peptides to databases containing human V, D, J or C-region germline sequences obtained from the IMGT database (IMGT®, the international

ImMunoGeneTics information system® http://www.imgt.org). All peptides with a bit score of at least 12.5 were assigned to these germline sequences and selected for further analysis. Peptides aligned to a V-region germline sequence were also aligned using the IMGT/DomainGapAlign tool. This tool positions the peptide to the germline sequence in the IMGT unique residue numbering system and helps to identify the peptide as a framework or CDR in the immunoglobulin molecule. Only peptides with an identity score of at least 70% were assigned to a CDR sequence. Total numbers of CDRs were calculated based on the CDRs found by Mascot and *de novo* sequencing.

Statistical Analysis

Coefficient of variation was used to measure the reproducibility of three replicate purifications of the reference sample for each individual IgG fraction. For each individual IgG fraction and each combination of IgG fractions, descriptive summary statistics (number of measurements (N), mean, standard deviation (SD) and Confidence Interval (95% CI)) were provided for the number of CDRs identified in the cases and controls.

The two sample t-Test (two-sided) was used to compare differences in the κ:λ ratio in Fab molecules between cases and controls. We used Microsoft Excel 2007 for the descriptive summary statistics and the t-Tests. Pearson Chi-square tests were performed to establish the existence of significant differences between cases and controls in the number of CDRs identified in each specific molecular dissected IgG fraction and each combination of IgG fractions compared with that of Fab. These tests, odds ratios and 95% confidence intervals were calculated by the application of Vassarstats software (http://faculty.vassar.edu/lowry/VassarStats.html). The non-parametric Kruskal-Wallis test (two-sided) was performed to compare the CDR3 ratio in the three Fab fractions (Fab, Fab-κ, Fab-λ) with the CDR3 ratio in the light chain fractions (κ and λ). Analyses were done using STATA, version 11 (StataCorp, Texas, US).

To determine if the number of significant different CDRs increases by molecular dissection of IgG, we used the Anova available in the Progenesis program and the two sample t-Test (two-sided). By a permutation test that was repeated 20 times we determined the random chance on such an event. Two standard deviations of this permutation test were used as a threshold to determine if the number of CDRs identified was significantly different. For all statistical tests a p value <0.01 was considered statistically significant.

Results

Fab, Fab-κ, Fab-λ, Kappa and Lambda Purification

To calculate recovery and reproducibility of protein band intensities, triplicate purifications of the reference sample were quantified by densitometry. A total recovery of 91% for IgG of total IgG-κ (heavy + light chain) and total IgG-λ (heavy + light chain) combined and, ≥95% for Fab of Fab-κ and Fab-λ combined was calculated. Coefficients of variation of the densitometry intensity of triplicate purifications of the reference sample were 2.1% for κ, 4.8% for λ, 3.1% for Fab-κ and 4.4% for Fab-λ.

Kappa to Lambda Ratio

Calculated serum IgG concentration of controls and cases were on average 9.5 g/L (95% CI 7.4-11.5 g/L) and therefore within the normal range (7.0-16.0 g/L). To determine the κ:λ ratio of the κ and λ purification, Fab, Fab-κ, Fab-λ, κ and λ fractions from the reference donor sample were purified in duplicate and each fraction was measured twice on an LTQ Orbitrap XL. Figure 2 represent the distribution of all Mascot peptide sequences corresponding to the V,D, J, and C region of the κ, λ and heavy chain (BLAST identity score ≥ 70%) and their normalized intensities in the Fab and the total IgG light chain (κ and λ). The κ:λ ratio was calculated by counting all V, D, J and C spectra, with a normalized intensity >0. We found a normal κ:λ ratio of 2.0 in the Fab and 2.1 in the total light chain of IgG (κ and λ) fraction in this healthy donor sample. In addition, we calculated the κ:λ ratio in the Fab of all controls and cases and observed a normal mean ratio of 1.9 (SD 0.1) and 2.0 (SD 0.0), respectively. The unpaired two sample t-Test revealed no statistically significant difference (p=0.10) between the κ:λ ratio in cases and controls.

Enrichment of Kappa and Lambda Peptides

Kappa and lambda enrichment of the cases and controls was calculated by the κ:λ ratio in the Fab-κ, Fab-λ, κ and λ fractions, as described above. After purification we obtained a 7-fold enrichment of κ in the Fab-κ (κ:λ ratio 14:1) and a 3-fold enrichment of λ in the Fab-λ fraction (κ:λ ratio 2:3). An 8-fold increase in enrichment factor was observed in the κ and a 4-fold increase in the λ fractions of IgG. In addition, less than 9% peptides of the heavy chain were found in the light chain fractions of IgG.

Replicate measurements of Reference Sample

In mass spectrometry, replicate measurements can increase the number of peptides identified. To determine whether the total number of CDRs identified in the multiple IgG fractions might increase due to the multiple MS measurements of the sample, we compared the number of CDRs identified in four replicate measurements of the Fab, Fab-κ, Fab-λ, κ, λ fractions of the reference sample. Although, the number of CDRs reached a maximum at each third or fourth replicate measurement, the combined fractions revealed more CDRs than the individual fractions (Figure 3). We found 617 CDRs in the Fab and 1238 CDRs in the combined fractions.

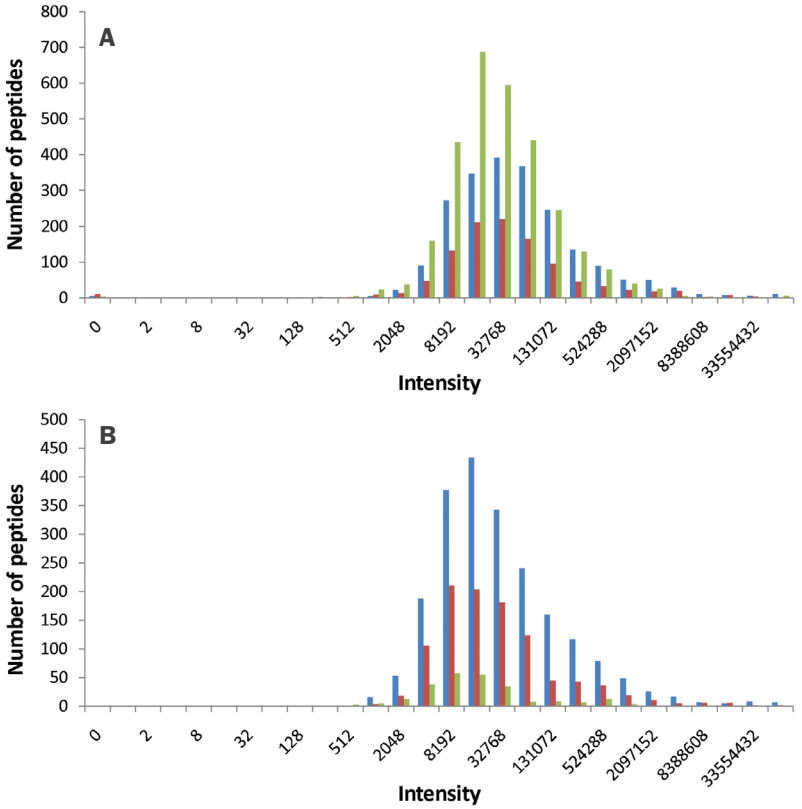


Figure 2. Distribution of VDJC peptides in Fab and IgG light chain of reference sample. All peptides sequences corresponding to the V region, D region, J region and C region of the κ (Blue), λ (Red) and heavy chain (Green) found by Mascot in the IMGT database and their normalized intensities in the Fab (a) and the IgG light chain (b).

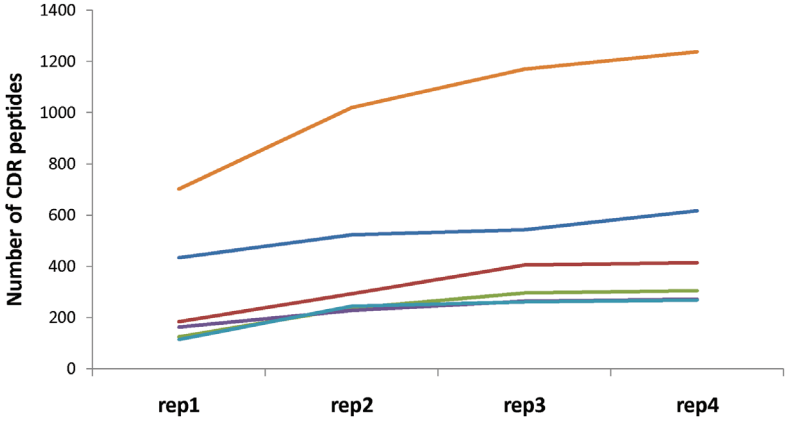


Figure 3. Number of CDRs peptides identified in replicate MS measurements of the fractions individually and combined of the reference sample. Fab (Blue), Fab-κ (Red), Fab-λ (Green), kappa (κ) (Purple), lambda (λ) (Light Blue) fraction and all fractions combined (Orange).

Number of CDRs Identified in Cases and Controls

Nano-LC-LTQ Orbitrap MS measurements of the respective digested Fab, Fab-κ, Fab-λ, κ and λ of the cases and controls yielded a combined peak list of 13061, 12441, 8246, 10294 and 11853 features reported by Progenesis. Automatic alignment in Progenesis was not possible for one Fab, one Fab-λ and one κ sample of the controls and one Fab-λ sample of the cases. Therefore they were excluded from data analysis. Figure 4 shows the number of features (MS signals) and the number of CDR peptide sequences identified by Mascot and *de novo* sequencing according to IMGT for each individual IgG fraction. In addition, we have listed the V, D, J and C regions (VDJC) and the V-region-related peptides.

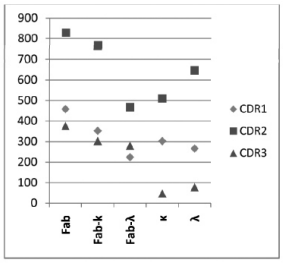
To compare the number of CDRs identified between the fractions, only samples analyzed by Progenesis for each fraction were selected. Redundant peptide sequences of these 16 samples were counted once in order to reveal the number of unique CDRs per fraction. Alignment to the IMGT database showed that 1663 peptide sequences of Fab, 1422 of Fab-κ, 971 of Fab-λ, 859 of κ and 991 of λ corresponded to CDRs. The numbers of the three types of CDR are shown in Figure 5. In all three types of Fab (Fab, Fab-κ, Fab-λ) we observed a mean CDR1:CDR2:CDR3 ratio of approximately 1.0:2.0:1.0 and in the light chains different CDR ratios of approximately 1.4:2.4:0.2 for κ and 1.0:2.7:0.3 for λ were seen. The CDR3 ratio in the three Fab fractions (Fab, Fab-κ, Fab-λ) was significantly ($p < 0.001$) higher than the CDR3 ratio in the light chain fractions (κ and λ).

The mean number of CDRs identified in the individual and combined IgG fractions compared with that of the Fab fraction of lung cancer cases and controls are listed in Table 1. We found 1.73 times more CDRs in the combination of Fab-κ, Fab-λ, κ and λ (Comb 6, Table 1), than in Fab for both cases and controls. Pearson Chi-square tests with odds ratios were performed to measure the association between cases and controls in the number of CDRs identified in the individual and combined IgG fractions compared with that of Fab (Table 1). We found no statistically significant difference ($p > 0.50$) between cases and controls for these numbers.

We calculated the mean additional number of unique CDRs (mean %, Confidence Interval 95%) found in the different individual and combined IgG fractions to the mean number of CDRs of the Fab fraction. Kappa gave 320 CDRs (23.1%, 20.6-25.5%), lambda 501 CDRs (36.0%, 34.2-37.8%), Fab-κ 679 CDRs (48.7%, 46.2-51.2%) and Fab-λ 315 CDRs (22.7%, 20.1-25.4%) additional to Fab. Combined κ and λ fractions resulted in 804 additional CDRs (57.8%, 54.4-61.3%), and combined Fab-κ and Fab-λ fractions in 978 additional CDRs (70.3, 66.8-73.7%). In addition, these four fractions combined showed an additional 1683 unique CDRs (121.0%, 115.5-126.5%) compared with the original Fab.

Figure 6 shows a Venn diagram of all the fractions of the cases and controls and the total number of CDRs found by Mascot and *de novo* sequencing. We found a total of 1663 CDRs in the Fab and an additional 2441 unique CDRs (146.8%) in all the other fractions combined.

	Fab		
	13061 MS signals of controls and cases (9/10)		
Peptides	Total	Mascot	*de novo*
VDJC-region	8013	3087	4926
V-region	5318	2567	2751
V-region unique	3191	1890	1301
CDRs unique	**1663**	1017	646
CDR1	**458**	228	230
CDR2	**829**	517	312
CDR3	**376**	272	104



	Fab-k		
	12441 MS signals of controls and cases (10/10)		
Peptides	Total	Mascot	*de novo*
VDJC-region	7552	2466	5086
V-region	4966	1986	2980
V-region unique	2881	1375	1506
CDRs unique	**1422**	594	828
CDR1	**352**	120	232
CDR2	**767**	421	346
CDR3	**303**	53	250

	Fab-λ		
	8246 MS signals of controls and cases (9/9)		
	Total	Mascot	*de novo*
	4950	1845	3105
	3193	1522	1671
	1885	1134	751
	971	618	353
	224	110	114
	467	295	172
	280	213	67

	kappa (κ)		
	10294 MS signals of controls and cases (9/10)		
Peptides	Total	Mascot	*de novo*
VDJC-region	5492	1893	3599
V-region	3274	1355	1919
V-region unique	1462	710	752
CDRs unique	**859**	422	437
CDR1	**302**	118	184
CDR2	**510**	290	220
CDR3	**47**	14	33

	lambda (λ)		
	11853 MS signals of controls and cases (10/10)		
	Total	Mascot	*de novo*
	6387	1651	4736
	3804	1295	2509
	1774	789	985
	991	437	554
	266	74	192
	647	359	288
	78	4	74

Figure 4. Number of MS signals and number of peptide sequences identified by Mascot and *de novo* sequencing for each individual IgG fraction. Redundant peptides corresponding to the VDJC-region and the V-region, and non-redundant (unique) peptides corresponding to the V-region and CDR (CDR1, 2, 3) region germline sequences from the IMGT database are shown. The graph illustrates the total number of CDR1, CDR2 and CDR3 identified for each individual IgG fraction. *(Published as supplementary Figure 4.)*

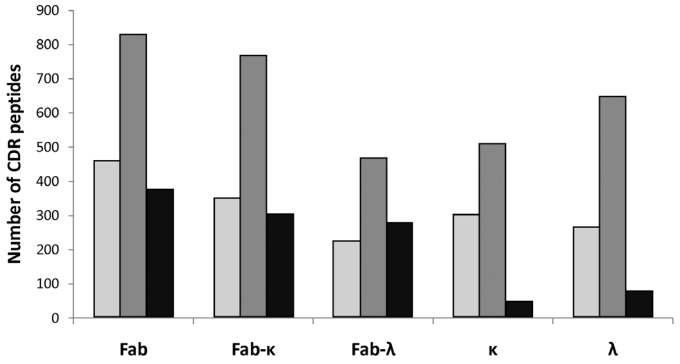


Figure 5. Numbers of CDR1, CDR2 and CDR3 identified in the individual IgG fractions. The total number of CDR1 (light grey), CDR2 (grey) and CDR3 (black) identified in each individual IgG fraction.

Table 1. Number of CDRs identified in individual and combined IgG fractions compared with Fab: Lung cancer cases vs. controls.

Fractions	Number of CDRs							OR	95% CI	p
	Lung cancer cases				Controls					
	N	mean (±SD)		odds	N	mean (±SD)		odds		
Fab	9	1412	(± 72)	ND	7	1370	(± 80)	ND	ND	ND	ND
Fab-κ	9	1119	(± 115)	0.79	7	1072	(± 97)	0.78	1.01	0.91-1.13	0.82
Fab-λ	9	715	(± 90)	0.51	7	706	(± 113)	0.52	0.98	0.86-1.12	0.79
κ	9	585	(± 82)	0.41	7	582	(± 83)	0.42	0.98	0.85-1.12	0.72
λ	9	748	(± 55)	0.53	7	700	(± 57)	0.51	1.04	0.91-1.18	0.58
Comb 1	32	1211	(± 100)	0.86	32	1172	(± 91)	0.86	1.00	0.90-1.12	1.00
Comb 2	32	1652	(± 127)	1.17	32	1609	(± 121)	1.17	1.00	0.90-1.10	0.92
Comb 3	32	1724	(± 85)	1.22	32	1690	(± 83)	1.23	0.99	0.90-1.09	0.84
Comb 4	32	2105	(± 144)	1.49	32	2032	(± 125)	1.48	1.01	0.91-1.11	0.92
Comb 5	48	2400	(± 146)	1.70	48	2335	(± 117)	1.70	1.00	0.91-1.10	1.00
Comb 6	64	2449	(± 159)	1.73	64	2376	(± 170)	1.73	1.00	0.91-1.10	1.00
Comb 7	80	3119	(± 165)	2.21	80	3024	(± 167)	2.21	1.00	0.92-1.09	1.00

Odds ratios between lung cancer cases and controls of the number of CDRs identified in individual and combined IgG fractions compared with Fab. Abbreviations: ND, the value was not determined; OR, odds ratio; CI, confidence interval; p, p-value of Pearson Chi-square test; Comb 1, κ+λ; Comb 2, Fab-κ+Fab-λ; Comb 3, Fab+Fab-λ; Comb 4, Fab+Fab-κ; Comb 5, Fab+Fab-κ+Fab-λ; Comb 6, Fab-κ+Fab-λ+κ+λ; Comb 7, Fab+Fab-κ+Fab-λ+κ+λ.

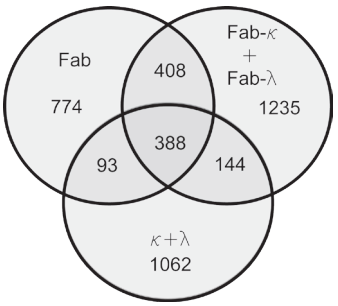


Figure 6. Venn diagram of all fractions of the cases and controls and the total number of CDRs found by Mascot and *de novo* sequencing.

Number of significantly different CDRs between cases and controls

We analyzed the CDR identified sequences of cases and controls obtained via database-dependent and *de novo* sequencing with Anova and the two sample t-Test. We observed that in Fab-κ the number of significantly identified CDR ($p < 0.01$) was significantly increased compared to random chance as determined by a permutation test (data not shown).

Discussion

In this study, we demonstrated that molecular dissection of IgG into kappa and lambda fragments (Fab-κ, Fab-λ, κ and λ) identifies approximately twice as many CDRs

than the Fab method.
In all fractions CDRs were identified exclusive to the specific fraction resulting in a total of 4104 CDR sequences. Results from all fractions and of combinations of them were evaluated. Although multiple MS measurements of an individual fraction also increase the number of CDRs identified, the combination of the various molecular fractions exceeds this significantly. In addition, the method we used to isolate Fab-κ, Fab-λ, κ and λ chains of IgG is reproducible and shows high recovery rates. The technical variation in MS measurements and variation in individual IgG molecules was described previously [25]. More specifically, the normal κ:λ ratio in the Fab and light chain of the healthy donor sample and the Fab of the controls and cases demonstrated a sufficient purification with preserved κ:λ ratio, and was validated by MS measurement. By means of the κ:λ ratio we were able to determine the κ or λ enrichment for the different IgG fractions.
As the whole Fab fraction includes the heavy and light chains (both κ and λ), this fraction yielded the most features and as a result more CDRs were identified in the Fab fraction than in the other individual IgG fractions. The MS sample of the Fab-κ fraction yielded more features than the Fab-λ fraction, and therefore the Fab-κ fraction revealed more CDRs than the Fab-λ fraction. The lower number of CDRs identified in κ and λ light chain fractions is very likely caused by the fact that CDRs specific to the heavy chain are missing, including the highly diverse CDR3 of the heavy chains [27]. This is supported by the observation of a different ratio for CDR1:CDR2:CDR3 in the light chains, which shows an approximately 4-fold lower ratio of CDR3 in the light chains than in the three types of Fab (Fab, Fab-κ, and Fab-λ). In general, in all fractions the CDR3 peptides were relatively difficult to assign. These peptides are highly diverse and their N-terminal side often contains a cleavage site for trypsin (lysine or arginine). As a result, their tryptic digested peptides often contain a mostly conserved V-region fragment or a highly diverse fragment, which makes it difficult to align to the germline sequence. Interestingly, more significantly different CDRs were observed in the Fab-κ than in the other fractions. This points to the possibility of finding lung cancer-related CDRs and in general of finding tumor-related CDRs.
Peptides from the constant regions gave the highest peak intensities. By choosing the maximum injection volume based on the highest peak intensity in the UV chromatogram we were able to maximize the loading of the CDRs on the C18 trap column. Nano-LC-LTQ Orbitrap MS measurements of additional fractions cause an increase in measurement time. Even though, measuring both Fab and Fab-κ fractions instead of only the Fab fraction requires twice as much measurement time, it makes the effort worthwhile because of the additional 50% CDRs identified. In addition, when measuring twice this is the best combination of all fractions because the immunoglobulin molecules that are occasionally expressed by cancer cells have been reported to consist predominantly of the heavy chains and κ chains [28]. This can be explained by the fact that during B cell differentiation first the heavy chain genes rearrange followed by the κ chain genes. Only if none of the κ chain gene rearrangements leads to a functional κ chain the λ chain genes start to rearrange [18]. Another explanation is that the heavy chain contains the highly diverse CDR3, which plays a prominent role in antigen binding [27]. Both heavy chains and κ chains are present in the Fab and Fab-κ fractions.
Recent studies have shown that antibody specificity is determined by a limited number of amino acid residues of the CDRs. Synthesized small peptides based on these CDRs retained the antigen-binding properties and functions of the intact immuno-

globulin [29-30]. Administration of synthetic CDR peptides inhibits tumor cell growth in mice and thereby increases their survival time [31]. These reports support the hypothesis that a specific molecular profile of CDRs may distinguish lung cancer patients from controls. In agreement, we found an increase in the number of significantly different CDRs in the Fab-κ fraction.

In our study, the lung cancer cases did not differ in their normal κ:λ ratio of Fab and in the number of detected CDRs in all the different IgG fractions from the controls. These findings show that our method is technical suitable to compare CDRs in IgG fractions between lung cancer patients and controls. Our approach revealed more CDRs than the original Fab method, which may enhance the possibility to identify a biomarker model for the early detection of lung cancer. However, there is most probably a larger sample set required to identify such statistically and physiologically relevant model. Sample size calculations [32] based on unpublished data estimate that a sample set of approximately 30 lung cancer cases and 30 controls is required to acquire this.

Improvements in sequence coverage and annotation may help to further increase the number of CDRs that is possible to identify. Alternative proteases could be used to obtain larger sequences coverage for a better alignment to the germline sequence. Other potential improvements are using ultra high pressure chromatography techniques to improve resolution for a better identification of sequences and depletion of constant regions by partial digestion of immunoglobulins to enrich CDR regions. In addition, complementing fragmentation spectra by higher energy collision induced dissociation (HCD) and electron transfer dissociation (ETD) can improve *de novo* peptide sequencing compared to CID fragmentation [33-35].

The ability to detect specific tumor-related CDR peptides by mass spectrometry depends on the proportion of total IgG that has affinity to the tumor antigen. Affinity purification of rat sera revealed that 1-3% of IgG had affinity for the antigen used for the immunizations [17]. Such a polyclonal antibody response to an antigen has been estimated to derive from approximately 100 B cell clones [22]. In another study, an upper limit of 0.1-0.3% of the human B cell population was found to have originated from a particular clone [36]. Based on these data, we estimate that 0.01-0.3% of the total IgG may present a particular immunoglobulin, depending on the degree of the immune response against the antigen and the diversity of the B cell clones. In previously published papers we showed that it is possible to detect CDRs of specific immunoglobulins at these levels [17, 25] and in particular, by our recent paper [35] In this paper, we showed that specific CDR peptides of a spiked antibody could be detected at attomole levels which were 5 orders of magnitude lower than the total IgG serum.

In conclusion, we have demonstrated that molecular dissection of IgG into kappa and lambda fragments is a valuable addition to Fab purification. Molecular dissection of IgG into kappa and lambda fragments identifies significantly more CDRs than Fab purification alone. This approach will increase the likelihood of finding lung cancer-related CDR sequences.

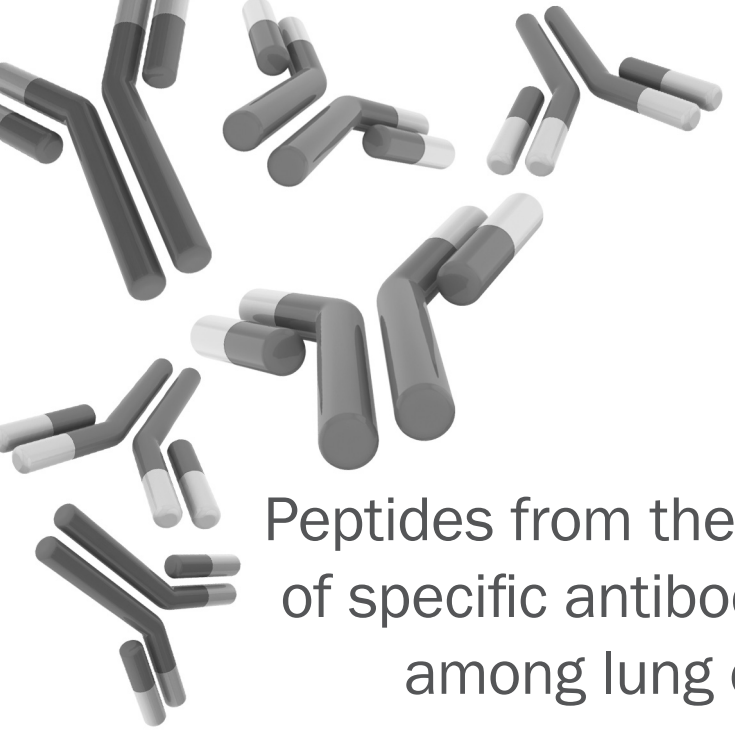
Acknowledgement

The authors acknowledge financial support from an NWO (Nederlandse organisatie voor Wetenschappelijk Onderzoek) Zenith grant 93511034.

References

1. Lopes Pegna A, Picozzi G, Mascalchi M, Maria Carozzi F, Carrozzi L, Comin C, et al. Design, recruitment and baseline results of the ITALUNG trial for lung cancer screening with low-dose CT. Lung Cancer 2009; 64:34-40.
2. Infante M, Lutman FR, Cavuto S, Brambilla G, Chiesa G, Passera E, et al. Lung cancer screening with spiral CT: baseline results of the randomized DANTE trial. Lung Cancer 2008; 59:355-63.
3. van Iersel CA, de Koning HJ, Draisma G, Mali WP, Scholten ET, Nackaerts K, et al. Risk-based selection from the general population in a screening trial: selection criteria, recruitment and power for the Dutch-Belgian randomised lung cancer multi-slice CT screening trial (NELSON). Int J Cancer 2007; 120:868-74.
4. van Klaveren RJ, Oudkerk M, Prokop M, Scholten ET, Nackaerts K, Vernhout R, et al. Management of lung nodules detected by volume CT scanning. N Engl J Med 2009; 361:2221-9.
5. Reich JM. A critical appraisal of overdiagnosis: estimates of its magnitude and implications for lung cancer screening. Thorax 2008; 63:377-83.
6. Dunn GP, Bruce AT, Ikeda H, Old LJ, Schreiber RD. Cancer immunoediting: from immunosurveillance to tumor escape. Nat Immunol 2002; 3:991-8.
7. Dunn GP, Old LJ, Schreiber RD. The immunobiology of cancer immunosurveillance and immunoediting. Immunity 2004; 21:137-48.
8. Qiu J, Hanash S. Autoantibody profiling for cancer detection. Clin Lab Med 2009; 29:31-46.
9. Chen G, Wang X, Yu J, Varambally S, Thomas DG, Lin MY, et al. Autoantibody profiles reveal ubiquilin 1 as a humoral immune response target in lung adenocarcinoma. Cancer Res 2007; 67:3461-7.
10. Qiu J, Choi G, Li L, Wang H, Pitteri SJ, Pereira-Faca SR, et al. Occurrence of autoantibodies to annexin I, 14-3-3 theta and LAMR1 in prediagnostic lung cancer sera. J Clin Oncol 2008; 26:5060-6.
11. Leidinger P, Keller A, Heisel S, Ludwig N, Rheinheimer S, Klein V, et al. Identification of lung cancer with high sensitivity and specificity by blood testing. Respir Res 2010; 11:18.
12. Rom WN, Goldberg JD, Addrizzo-Harris D, Watson HN, Khilkin M, Greenberg AK, et al. Identification of an autoantibody panel to separate lung cancer from smokers and nonsmokers. BMC Cancer 2010; 10:234.
13. Farlow EC, Patel K, Basu S, Lee BS, Kim AW, Coon JS, et al. Development of a multiplexed tumor-associated autoantibody-based blood test for the detection of non-small cell lung cancer. Clin Cancer Res 2010; 16:3452-62.
14. Wu L, Chang W, Zhao J, Yu Y, Tan X, Su T, et al. Development of autoantibody signatures as novel diagnostic biomarkers of non-small cell lung cancer. Clin Cancer Res 2010; 16:3760-8.
15. 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.
16. Chapman CJ, Murray A, McElveen JE, Sahin U, Luxemburger U, Tureci O, et al. Autoantibodies in lung cancer: possibilities for early detection and subsequent cure. Thorax 2008; 63:228-33.
17. VanDuijn MM, Dekker LJ, Zeneyedpour L, Smitt PA, Luider TM. Immune responses are characterized by specific shared immunoglobulin peptides that can be detected by proteomic techniques. J Biol Chem 2010; 285:29247-53.
18. Murphy KP, Travers P, Walport M, editors. Janeway's Immunobiology. New York: Garland Science; 2008.
19. Saada R, Weinberger M, Shahaf G, Mehr R. Models for antigen receptor gene rearrangement: CDR3 length. Immunol Cell Biol 2007; 85:323-32.
20. Andersen PS, Haahr-Hansen M, Coljee VW, Hinnerfeldt FR, Varming K, Bregenholt S, et al. Extensive restrictions in the VH sequence usage of the human antibody response against the Rhesus

D antigen. Mol Immunol 2007; 44:412-22.
21. Baranzini SE, Jeong MC, Butunoi C, Murray RS, Bernard CC, Oksenberg JR. B cell repertoire diversity and clonal expansion in multiple sclerosis brain lesions. J Immunol 1999; 163:5133-44.
22. Poulsen TR, Meijer PJ, Jensen A, Nielsen LS, Andersen PS. Kinetic, affinity, and diversity limits of human polyclonal antibody responses against tetanus toxoid. J Immunol 2007; 179:3841-50.
23. Weinstein JA, Jiang N, White RA, 3rd, Fisher DS, Quake SR. High-throughput sequencing of the zebrafish antibody repertoire. Science 2009; 324:807-10.
24. Annesley TM. Ion suppression in mass spectrometry. Clin Chem 2003; 49:1041-4.
25. de Costa D, Broodman I, Vanduijn MM, Stingl C, Dekker LJ, Burgers PC, et al. Sequencing and quantifying IgG fragments and antigen-binding regions by mass spectrometry. J Proteome Res 2010; 9:2937-45.
26. Chui SH, Lam CW, Lai KN. Light-chain ratios of immunoglobulins G, A, and M determined by enzyme immunoassay. Clin Chem 1990; 36:501-2.
27. Xu JL, Davis MM. Diversity in the CDR3 region of V(H) is sufficient for most antibody specificities. Immunity 2000; 13:37-45.
28. Chen Z, Qiu X, Gu J. Immunoglobulin expression in non-lymphoid lineage and neoplastic cells. Am J Pathol 2009; 174:1139-48.
29. Eisenhardt SU, Schwarz M, Schallner N, Soosairajah J, Bassler N, Huang D, et al. Generation of activation-specific human anti-alphaMbeta2 single-chain antibodies as potential diagnostic tools and therapeutic agents. Blood 2007; 109:3521-8.
30. Padlan EA, Abergel C, Tipper JP. Identification of specificity-determining residues in antibodies. FASEB J 1995; 9:133-9.
31. Polonelli L, Ponton J, Elguezabal N, Moragues MD, Casoli C, Pilotti E, et al. Antibody complementarity-determining regions (CDRs) can display differential antimicrobial, antiviral and antitumor activities. PLoS One 2008; 3:e2371.
32. Geller NL, Pocock SJ. Interim analyses in randomized clinical trials: ramifications and guidelines for practitioners. Biometrics 1987; 43:213-23.
33. Olsen JV, Macek B, Lange O, Makarov A, Horning S, Mann M. Higher-energy C-trap dissociation for peptide modification analysis. Nat Methods 2007; 4:709-12.
34. Shen Y, Tolic N, Xie F, Zhao R, Purvine SO, Schepmoes AA, et al. Effectiveness of CID, HCD, and ETD with FT MS/MS for degradomic-peptidomic analysis: comparison of peptide identification methods. J Proteome Res 2011.
35. Dekker LJ, Zeneyedpour L, Brouwer E, van Duijn MM, Sillevis Smitt PA, Luider TM. An antibody-based biomarker discovery method by mass spectrometry sequencing of complementarity determining regions. Anal Bioanal Chem 2011; 399:1081-91.
36. Boyd SD, Marshall EL, Merker JD, Maniar JM, Zhang LN, Sahaf B, et al. Measurement and clinical monitoring of human lymphocyte clonality by massively parallel VDJ pyrosequencing. Sci Transl Med 2009; 1:12ra23.



Chapter 4

Peptides from the variable region of specific antibodies are shared among lung cancer patients

Dominique de Costa, Ingrid Broodman, Wim Calame, Martijn M. VanDuijn, Christoph Stingl, Lennard J.M. Dekker, René M. Vernhout, Harry J. de Koning, Henk C. Hoogsteden, Peter A.E. Sillevis Smitt, Rob J. van Klaveren, Theo M. Luider

Submitted

Abstract

Late diagnosis of lung cancer is still the main reason for high mortality rates in lung cancer. Lung cancer is a heterogeneous disease which induces an immune response to different tumor antigens. Several methods for searching autoantibodies have been described that are based on known purified antigen panels. Aim of our study is to find evidence that parts of the antigen-binding-domain of antibodies are shared among lung cancer patients. This was investigated by a novel approach based on sequencing antigen-binding-fragments (Fab) of immunoglobulins using proteomic techniques without the need of previously known antigen panels.

From serum of 93 participants of the NELSON trial IgG was isolated and subsequently digested into Fab and Fc. Fab was purified from the digested mixture by SDS-PAGE. The Fab containing gel-bands were excised, tryptic digested and measured on a nano-LC-Orbitrap-Mass-spectrometry system. Multivariate analysis of the mass spectrometry data by linear canonical discriminant analysis combined with stepwise logistic regression resulted in a 12-antibody-peptide model which was able to distinguish lung cancer patients from controls in a high risk population with a sensitivity of 84% and specificity of 90%.

With our Fab-purification combined Orbitrap-mass-spectrometry approach, we found peptides from the variable-parts of antibodies which are shared among lung cancer patients.

Introduction

Lung cancer is currently the most common cancer with the highest mortality rate (28%) in the World due to late diagnosis at an advanced stage.[1-2] However, with the demonstration of a 20% lung cancer mortality reduction by the NLST trial (National Cancer Screening Trial) low dose CT screening for lung cancer is receiving increasing interest.[3] The NELSON trial (Dutch-Belgian lung cancer screening trial) showed that after three screening rounds 3.6% of all participants of this study had a false-positive screen result.[4] Although, still approximately 27% of the participants were subjected to invasive procedures that revealed benign lung diseases at baseline screening (first round NELSON trial).[5] A good biomarker (panel) will reduce this number of unnecessary invasive procedures. At the moment selection of high risk individuals for screening is done by age and smoking history. A biomarker or biomarker panel would be helpful in selecting high risk individuals for CT screening as this may detect lung cancer at an earlier stage than CT.

Antibodies can be interesting as markers for distinguishing lung cancer patients from lung cancer-free individuals. These antibodies are produced by the immune response that target specific tumor-associated antigens (TAAs) during cancer development, probably at an early stage.[6-12] Recently Liu et *al.* showed that the concentration of circulating IgG autoantibodies against ABCC3 transporter was significantly higher in female adenocarcinoma patients than in female controls.[13]

Antibodies, or immunoglobulins, are highly complex molecules with large variation in their amino acid sequence. The possible diversity in immunoglobulins is estimated between 10^{13} and 10^{50} and therefore the finding of similar or even identical sequences in different individuals by chance is in theory, highly unlikely.[14-15] However, studies of different research groups have recently demonstrated that despite this theoretical small chance to have identical antibodies among individuals, it is possible to identify similar or identical sequences.[16-19] A study performed by us showed that in PNS (paraneoplastic neurological syndrome) patients identical mutated primary amino acid sequences of complementarity determining regions (CDRs) exist. These CDRs are specific for known onconeural antigens, such as HuD and Yo in PNS patients, and most interestingly were shared between different PNS patients.[20]

The aim of this study is to find evidence that specific antibody peptides are shared between lung cancer patients in contrast to lung cancer-free individuals. As lung cancer is a heterogeneous disease and with the variability of an antibody it might be a challenge to detect identical tumor-related antibodies in serum. We experimentally test the hypothesis that specific highly variable regions of an antibody including complementarity determining regions (CDRs) can be shared between lung cancer patients. Our experimental approach to verify this hypothesis is based on sequencing antibody peptides by mass spectrometry. Measurement of serum by a mass spectrometer might be too complex due to the high variability as mentioned above. Purifying IgG Fab from serum will reduce the complexity of the sample from a lung cancer patient and will give the possibility to focus on pure antibody fractions.

Material & Methods

Study population

For this study, we selected 44 lung cancer cases and 49 controls (Figure 1) from the NELSON lung cancer screening trial.[5, 21] For the cases of the discovery set, NELSON 1, only early stage (I and II) squamous cell (n=4) or adenocarcinomas (n=21) were selected. They were carefully matched to the controls by age, gender, smoking status, duration and number of cigarettes smoked per day, chronic obstructive pulmonary disease (COPD) status, asbestos exposure and site of blood sampling (Table 1). The selection criteria for the cases of the NELSON 2 (validation) set (n=19) were similar, except that all non-small cell histology's and disease stages were allowed (Table 1) in order to challenge the results of the discovery phase. On purpose the clinical characteristics of the control patients are dissimilar with the NELSON 1 set in respect to smoking and COPD. Therefore, this NELSON 2 set is not matched with the NELSON 1 set. Serum samples were collected for both NELSON 1 and NELSON 2 obtained from baseline CT screening (first round). The NELSON trial was approved by the Dutch Health Council, the Minister of Health and by the Medical Ethical Committees of all participating centers (clinical trial number ISRCTN63545820). All participants for this study provided written informed consent for the use of their serum samples. The donor of the reference sample used throughout this study provided written consent for the use of his/her serum for scientific purposes according to the guidelines of the Sanquin Blood Bank, Rotterdam.

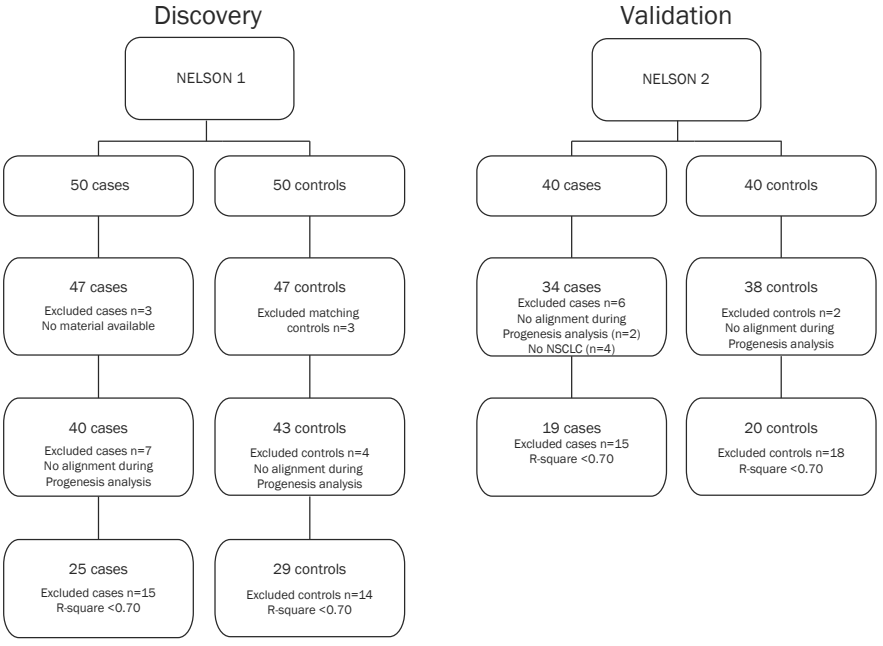


Figure 1. Technical reproducibility of replicate measurements of the reference sample. Reference sample measured at different time points during measurement of the NELSON 1 sample set. A replicate of the reference sample (x-axis) was compared to each other replicate sample based on the raw abundance of each feature. An r-square value was calculated. Each dot represents an r-square (y-axis) value for the comparison of that specific replicate with another replicate. For each replicate the average r-square and standard deviation (SD) is shown.

IgG Fab purification and NanoLC Orbitrap MS analyses

Prior to all sample preparation procedures, all samples were blinded and the key for unblinding was put at the database coordinator of the NELSON trial. IgG Fab purification and nano-LC Orbitrap MS analyses were performed according to the method described before.[22] In brief, IgG was isolated from serum and digested into Fab and Fc. The Fab part was purified from the digested mixture by SDS-PAGE. The Fab containing gel bands were excised and tryptic digested. A blank piece of gel that was not loaded with protein was excised and treated like the excised Fab bands for background assessment.

LCMS measurements were performed on an Ultimate 3000 nano LC system (Thermo Fisher Scientific/Dionex, Amsterdam, the Netherlands) online coupled to a hybrid linear ion trap/Orbitrap MS (LTQ Orbitrap XL; Thermo Fisher Scientific, Bremen, Germany). Four µL of the digested Fab was loaded onto the system. For further settings and solutions we refer to previous published work.[22] All samples were randomized before measurement and were measured in batches of 11 samples including a reference sample. A blank sample was run at the start and end of the measurement to determine background and the existence of carry-over during chromatography.

Data analyses

Raw data files were loaded into the software Progenesis (Version 3.1; Nonlineair Dynamics Ltd, New Castle, UK) and processes as described previously.[22] In addition, we performed a Progenesis analysis where instead of detecting features (peptide masses (m/z)) in all the samples at the same time by the software program, feature detection was performed individually per sample. Features picked thereby were matched to the Progenesis result table containing all samples with a mass tolerance of 5 ppm. This was of advantage, since often features occur with low intensities in one sample and are subsequently matched by Progenesis in all other samples. This result in errors related to background if one takes the respective mass spectra into account. With this relative small adjustment it assures that a feature is detected more correctly throughout the samples. The data acquired by this approach was filtered using the same default settings.[22] A separate data matrix for every case and control was generated consisting of all features with corresponding raw abundance and retention time. To generate one large data matrix that includes all cases and controls from these separate data matrices, we searched masses from the separate data matrices per case or control in the complete data matrix generated from the standard Progenesis analyses. Every mass had to meet three criteria: 1) m/z (±5 ppm), 2) retention time (±1 min) and 3) identical charge. If a mass met these three criteria the raw abundance from the complete matrix (generated by a general procedure[22] recommended by the manufacturer) was used. If a mass did not meet these criteria a zero was generated for the raw abundance.

MS/MS spectra were extracted from raw data files and converted into mgf files using extract-msn (part of Xcalibur version 2.0.7, Thermo Fisher Scientific Inc.). Mascot (version 2.3.01; Matrix Science Inc., London, UK) was used to perform database searches against the human subset NCBInr database (version March 11th, 2009; Homo sapiens species restriction; 222,066 sequences) of the extracted MS/MS data. Database (NCBInr) dependent peptide identification and *de novo* sequencing results (software PEAKS; Version 5.2; Bioinformatics Solutions Inc., Waterloo, Canada) were also included in the Progenesis provided matrix. For settings used for the database search and *de novo* sequencing we refer to previous published work.[22]

Table 1. Clinical characteristics of the NELSON 1 and NELSON 2 sample sets.

	NELSON 1					NELSON 2				
	Cases		Controls			Cases		Controls		
	n=25		n=29		p-value	n=19		n=20		p-value
Gender					0.847					0.589
Male	22	88.0%	26	89.7%		16	84.2%	18	90.0%	
Female	3	12.0%	3	10.3%		3	15.8%	2	10.0%	
Age (years)	62.0	(±6.5)	62.2	(±4.8)	0.754	62.2	(±7.3)	62.2	(±6.8)	0.944
Men	61.8	(±6.6)	62.5	(±5.0)	0.611	63.3	(±7.4)	63.3	(±6.2)	1.000
Women	63.3	(±6.8)	60.0	(±1.7)	0.637	56.7	(±3.5)	52.5	(±0.7)	0.139
Smoking status[a]					0.389					<0.001
Current	15	60.0%	14	48.3%		10	52.6%	0	0.0%	
Former	10	40.0%	15	51.7%		9	47.4%	20	100.0%	
Smoking duration (years)					0.108					0.308
26-40	14	56.0%	11	37.9%		12	63.2%	11	55.0%	
41-45	6	24.0%	15	51.7%		4	21.1%	8	40.0%	
> 45	5	20.0%	3	10.3%		3	15.8%	1	5.0%	
Cigaretes/day					0.891					0.355
0-15	5	20.0%	6	20.7%		3	15.8%	8	40.0%	
16-20	8	32.0%	7	24.1%		2	10.5%	2	10.0%	
21-25	5	20.0%	8	27.6%		4	21.1%	4	20.0%	
> 25	7	28.0%	8	27.6%		10	52.6%	6	30.0%	
COPD[a]					0.641					0.024
Yes	10	40.0%	12	41.4%		5	26.3%	0	0.0%	
No	14	56.0%	14	48.3%		13	68.4%	20	100.0%	
Unknown	1	4.0%	3	10.3%		1	5.3%	0	0.0%	
Asbestos exposure					0.542					0.676
Yes	5	20.0%	4	13.8%		2	10.5%	3	15.0%	
No	20	80.0%	25	86.2%		17	89.5%	17	85.0%	

Table 1. (continued)

	NELSON 1				NELSON 2			
	Cases		Controls		Cases		Controls	
	n=25		n=29	p-value	n=19		n=20	p-value
Center				0.275				0.158
Groningen	4	16.0%	9	31.0%	4	21.1%	7	35.0%
Utrecht	9	36.0%	12	41.4%	4	21.1%	6	30.0%
Haarlem	11	44.0%	6	20.7%	7	36.8%	7	35.0%
Leuven	1	4.0%	2	6.9%	4	21.1%	0	0.0%
Time to cancer (years)[b]								
0-0.5	11	44.0%	-	-	12	63.2%	-	-
0.5-1.5	7	28.0%	-	-	4	21.0%	-	-
1.5-2.5	6	24.0%	-	-	3	15.8%	-	-
2.5-3.5	1	4.0%	-	-	0	0.0%	-	-
Histology								
Adenocarcinoma	21	84.0%	-	-	8	42.1%	-	-
Squamous cell carcinoma	4	16.0%	-	-	2	10.5%	-	-
Other[c]	0	0.0%	-	-	9	47.4%	-	-
Pathological stage								
IA	19	76.0%	-	-	6	31.6%	-	-
IB	3	12.0%	-	-	0	0.0%	-	-
II	3	12.0%	-	-	2	10.5%	-	-
III	0	0.0%	-	-	8	42.1%	-	-
IV	0	0.0%	-	-	3	15.8%	-	-
VDT (Days)								
> 600	0	0.0%	-	-	0	0.0%	-	-
400-600	4	16.0%	-	-	1	5.3%	-	-
<400	7	28.0%	-	-	6	31.6%	-	-
Not applicable[d]	14	56.0%	-	-	13	68.4%	-	-
Volume (mm^3)								
< 50	0	0.0%	-	-	0	15.8%	-	-
50-500	6	24.0%	-	-	3	78.9%	-	-
> 500	19	76.0%	-	-	15	79.4%	-	-
Unknown	0	0.0%	-	-	1	5.3%	-	-

Table 1. (continued)

	NELSON 1				NELSON 2			
	Cases	Controls			Cases	Controls		
	n=25	n=29		p-value	n=19	n=20		p-value
Consistency								
Solid	21 84.0%	- -			17 89.4%	- -		
Partial solid	3 12.0%	- -			1 5.3%	- -		
Non-solid	0 0.0%	- -			0 0.0%	- -		
Unknown	1 4.0%	- -			1 5.3%	- -		
Benign nodules				0.940				0.861
0	14 56.0%	16	55.2%		10 52.6%	12	60.0%	
1	5 20.0%	5	17.2%		4 21.1%	3	15.0%	
>1	6 24.0%	8	27.6%		5 26.3%	5	25.0%	

COPD: Chronic Obstructive Pulmonary Disease; VDT: Volume Doubling Time; significance of all characteristics were tested by chi-square except for age (Mann-Whitney U test); [a]Significantly different in NELSON 2 sample set; [b]interval between blood sampling and diagnosis; [c]NSCLC histology other than adeno- or squamous cell carcinoma; [d]no repeat scan, no growth (decrease in size or resolution). *(Submitted as Supplementary Table 1)*

For *de novo* sequences so far not known from a database, the Peaks software identifies a leucine for the isobaric amino acids leucine and isoleucine. Database dependent peptide identification results or *de novo* sequencing results were included in the matrix based on the highest peptide identity score. All peptide sequences from the cases and controls identified by Mascot or PEAKS were subsequently aligned to databases containing V, D, J or C-region germline sequences derived from IMGT database (IMGT®, the international ImMunoGeneTics information system® http://www.imgt.org) using the BLAST algorithm.[23] Peptides with sufficient match (bitscore ≥12.5 and alignment score ≥70%) to the V-region database were assigned to a position on the immunoglobulin molecule with varying CDR lengths.
Raw data files of the reference samples of each data set were separately loaded into the software Progenesis and followed the standard procedures as mentioned above. To determine the proportion of variation between the reference sample measurements performed on different time points, median r-squares were calculated for each sample. Each sample was compared to all the other reference samples measured in that dataset and a median r-square was calculated for each sample. The comparison was based on the raw abundance of each feature. This was performed separately for both independent datasets, NELSON 1 and NELSON 2. To determine the proportion of variation between the samples (cases and controls) of the two separate datasets, the same calculations were performed as described above for each case and control sample. This analysis was performed separately for the two datasets. Based on the distribution of the median r-squares of each sample, we decided to set a cut-off at r-square >0.70. The cases and controls that obtained a median r-square below 0.70 were excluded from the dataset and further analyses. Calculations were conducted using Microsoft Excel 2007.

Statistical analysis

Two independent data sets have been used, NELSON 1 and NELSON 2. The initial step in the statistical analysis consisted of testing for normality using skewness and kurtosis distribution characteristics on the intensity of the raw abundance of the features.[24]
Subsequently, univariate analysis was performed, applying either an unpaired t-test (parametric) or a Mann-Whitney U-test (non-parametric) to detect significant differences in raw abundance between cases and controls in the NELSON 1 set.[25] The significance limit was set at 0.05 (two-sided). All identified features that were found significantly different were used for the selection of features to distinguish lung cancer patients from controls.
Secondly, we used for multivariate analysis only the significantly identified features that had ≥ 2 triggered MS spectra. We applied a multivariate analysis on features fulfilling these criteria with a (logistic) stepwise regression model ($y= a_1x_1 + a_2x_2 + a_3x_3 \ldots.a_nx_n + c$) in combination with canonical linear discriminant analysis.[26-27] This resulted in a combination of features with high sensitivity and specificity in the NELSON 1 dataset. This combination of features was then tested in the NELSON 2 dataset using the same methodology as described above.[26-27] Note that for the NELSON 2 dataset it was necessary to optimize the coefficients in the model equation in order to optimize the sensitivity and specificity in the NELSON 2 dataset.
To avoid a random-error effect in modeling, we verified the statistical background of the combination of features in a permutated dataset. The background evaluation consisted of the same workflow as used for the model building, except that at the be-

ginning the assignment of cases and controls of NELSON 1 were permutated (Figure 2). This permutation was performed twelve times and the results obtained were tested for significance against the model outcome by z-test (one-sided; $p<0.05$). Since model building was based on the data as provided in NELSON 1 after which validation of this model was done using the data in NELSON 2, the same approach was taken after each individual permutation. Also here, note that for NELSON 2 dataset the coefficients in the model equation were optimized.
All analyses on model building, validation and background evaluation were done using STATA, version 12 (StataCorp, Texas, US). Throughout the study, using two-sided testing (except for one-sided testing for Z-values), p-values of 0.05 or lower were considered to be statistically significant. Statistical analyses of the data shown in Table 1 were generated by SPSS (IBM SPSS Statistics 20).

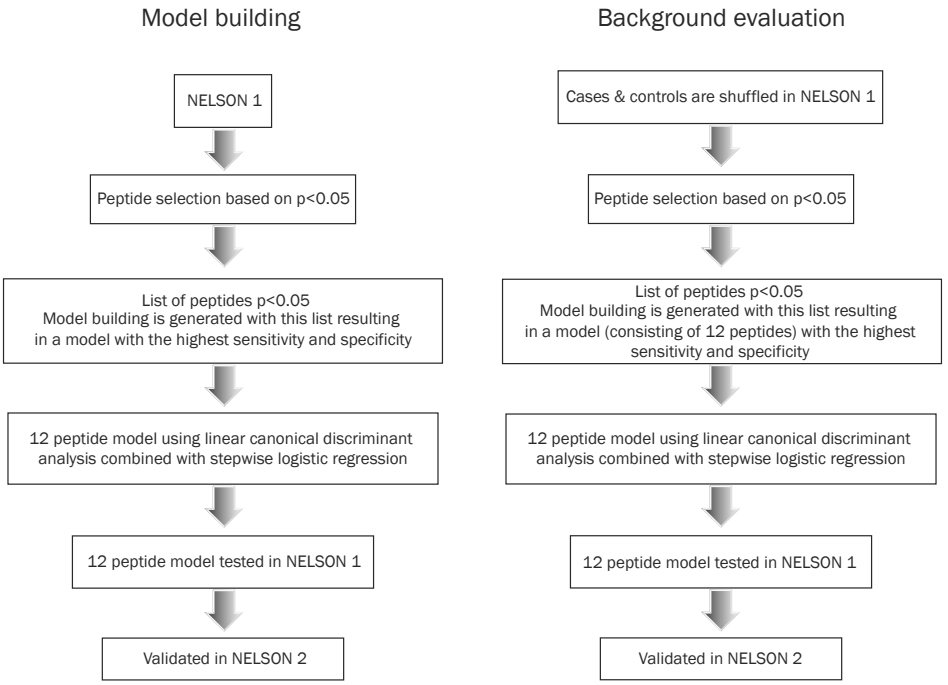


Figure 2. Statistical analysis flow-chart. Before background analysis is performed, cases and controls of the NELSON 1 dataset are shuffled randomly. *(Submitted as Supplementary Figure 2)*

Results

Clinical characteristics of the study population

There was no significant difference in the clinical characteristics between the cases and controls in the NELSON 1 set (Table 1). In the NELSON 2 set, current or former smoker and COPD status differed significantly between cases and controls (Table 1). In 72% and 84% of the cases of the NELSON 1 set, and NELSON 2 set, respectively, the time interval between blood sampling and lung cancer diagnosis was between 0-1.5 years. The median follow-up duration after blood sampling was for the control population 1925 days (range 1075-2086 days) and 1861 days (range 347-2135) in the NELSON 1 set and NELSON 2 set, respectively. None of the controls developed lung cancer during the follow-up period.

Technical variation

During the measurements of the biological samples we measured a reference sample at different time points. R-square values were calculated from the abundances of identified proteins in each reference measurement to show technical reproducibility. The lowest r-square value observed in the different measurements ranged between 0.84 and 0.93 (Figure 3).

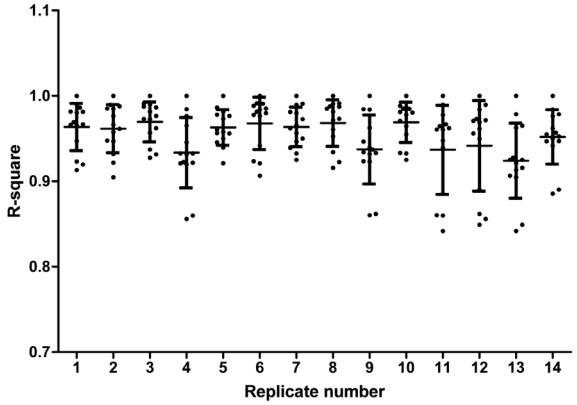


Figure 3. Technical reproducibility of replicate measurements of the reference sample. Reference sample measured at different time points during measurement of the NELSON 1 sample set. A replicate of the reference sample (x-axis) was compared to each other replicate sample based on the raw abundance of each feature. An r-square value was calculated. Each dot represents an r-square (y-axis) value for the comparison of that specific replicate with another replicate. For each replicate the average r-square and standard deviation (SD) is shown.

We performed the same r-square calculation for 5 random biological samples taken from the NELSON 1 set that were measured on two different LC-columns (same batch) at different time points. The technical reproducibility within each column resulted in lowest r-square values ranging from 0.75-0.93, but the technical reproducibility of the five biological samples measured on two independent similar columns was lower. For the two independent similar columns a median r-square of 0.52 was observed. In Figure 4 the correlation between each sample and between columns are shown.
An estimation of the biological variation was performed and resulted in a median r-square of 0.43. This result was much lower than the lowest r-square (0.84) observed for the technical variation. Therefore, the biological variation is higher compared to

the technical variation.
These results show that technical variation should be taken into account and adjustment is needed for comparison of independently measured sample sets since the NELSON 1 and NELSON 2 dataset were measured on two different columns at different time points. To overcome this technical variation, we applied a number of filters on the data before we could start a data analysis as described in the Material & Methods section.

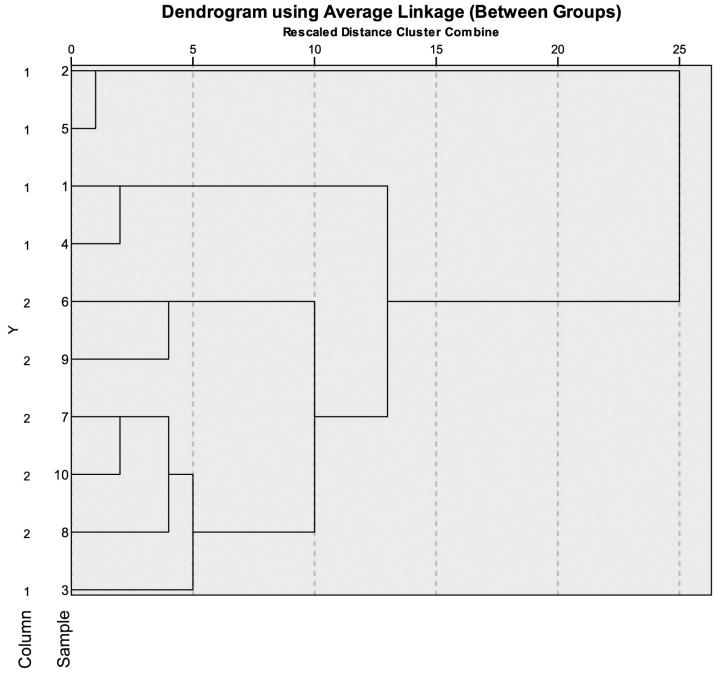


Figure 4. Technical reproducibility of five biological samples measured on two different columns at different time points. This dendrogram shows the correlation between five different biological samples measured on two different columns from same batch, column 1 and column 2 (y-axis). On the y-axis the five different samples are shown. Sample 1-5 are measured on column 1 and 6-10 are measured on column 2. Sample 1 and 6 are from the same individual. This also applies for sample 2 and 7, 3 and 8, 4 and 9 and 5 and 10. On the x-axis the Euclidian distance between each sample is shown. A strong correlation per column is found.

With this data we performed separate univariate analysis on all peptides found in cases and controls from the separate NELSON 1 and NELSON 2 data set. We were able to observe 49 peptides that were significantly different between cases and controls in the NELSON 1 dataset. However, these peptides, with one exception, did not show this difference in the NELSON 2 dataset. There was no trend observed (r-square 0.004) in p-values for the two datasets. Therefore, testing univariately in this manner was either not the right analysis strategy or the process generated randomly selected features (chance). Therefore, the significant peptides from NELSON 1 were analyzed as a next step in a multivariate way.

Antibody peptide model

An optimal combination of 12 peptides was identified by the multivariate statistics used on the NELSON 1 set (discovery set). This combination of peptides could distinguish lung cancer patients from controls with sensitivity and specificity of 96% and 100%, respectively. This antibody peptide model was able to detect lung cancer 373 days on average (range 39-1193 days) before the diagnosis was determined. In Figure 5 we show that the combination of the 12 peptides was able to distinguish cases from controls. The 12 peptides corresponded to 1 sequence overlapping with the CDR2 region, 1 sequence overlapping CDR3 region, 7 sequences overlapping the Framework 1 region and 3 sequences overlapping with the Framework 3 region according to the IMGT database (Table 2).

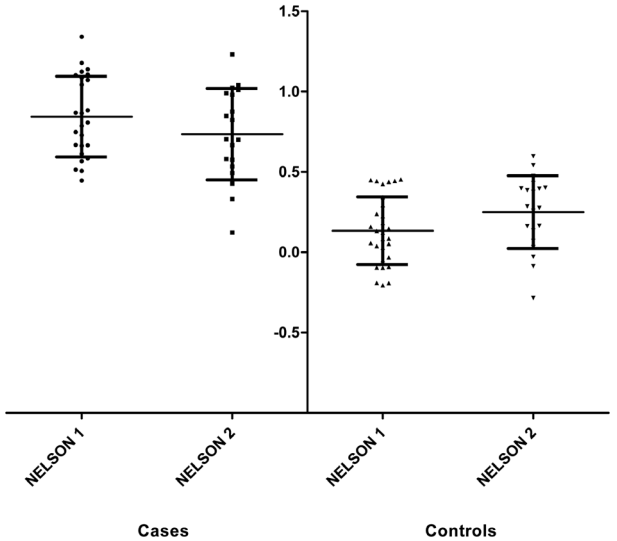


Figure 5. Distribution of the antibody peptide model outcome of the NELSON 1 and NELSON 2 sets. The raw abundances are filled-in in the model equation (y= a1x1 + a2x2 + a3x3 .....anxn + c) of the relevant sample set. On the y-axis (in arbitrary units) the figures generated by the equation are shown.

We performed an external validation in the NELSON 2 (validation) set. When we applied the same 12 peptide model to this set, cases and controls could no longer be distinguished. However, with the same peptides but after re-optimization of the model coefficients, we observed a sensitivity and specificity of 84% and 90%, respectively. As the coefficients of the equation are adjusted we had to check for the chance of overfitting of the data. Therefore, a background evaluation was performed which will be described later. Within the NELSON 2 validation set the combination of peptides was able to detect lung cancer 281 days on average (range 54-777 days) before the diagnosis of lung cancer.

We compared the raw abundance of the 12 peptides between the two NELSON datasets. We observed that the average raw abundance of five peptides was higher in the cases compared to the average abundance of the controls from the NELSON 1 dataset. These data were consistent with the findings from the NELSON 2 dataset. The other seven peptides had a higher average raw abundance in the controls of the NELSON 1 dataset compared to the abundance in the cases of this dataset. For only one of these seven peptides, this difference could be confirmed in the NELSON 2 dataset.

Table 2. Information of the 12 peptides of the antibody peptide model.

Peptide	CDR/ Framework	Sequence	m/z	Charge	Protein description	BLAST bit score	IMGT % Identified	p-value
1	Fr1	GITLSVRP	421.758	2	gi|553734: T-cell receptor	16.8	83.3	0.045
2	Fr3	LMAWLDLK	503.274	2	De Novo Peptide	15.9	75.0	0.009
3	CDR2	IYWDDDKR	555.763	2	gi|39938054: Immunoglobulin heavy chain variable region	32.9	100.0	0.039
4	Fr3	SYPLTFGGGTK	564.288	2	gi|4378188: Immunoglobulin kappa variable region	26.5	100.0	0.014
5	CDR3	LLLYTGGDQR	568.301	2	De Novo Peptide	18.0	75.0	0.030
6	Fr1	EVLLVESGGGLVKPGGSLR	623.025	3	gi|2072264: Immunoglobulin heavy chain	53.2	94.7	0.017
7	Fr3	NTVFLEMNSLR	670.336	2	gi|112699425: Immunoglobulin heavy chain variable region	32.9	81.8	0.040
8	Fr1	HVQLQESGPGLVK	696.386	2	De Novo Peptide	39.7	92.3	0.031
9	Fr1	SYSCQVTHEGSTVEK	827.872	2	gi|16554039: Immunoglobulin heavy chain	50.7	75.0	0.036
10	Fr1	SELTQDPAVSVALGQTVR	936.000	2	gi|87901: Immunoglobulin lambda variable region	57.1	100.0	0.030
11	Fr1	VSSVRCTSGGGLVQPGGSLR	959.501	2	gi|112702369: Immunoglobulin heavy chain variable region	40.1	100.0	0.042
12	Fr1	REMTKPPSVSVSETSHR	964.487	2	De Novo Peptide	29.1	81.8	0.013

Fr: Framework; CDR: Complementarity Determining Region.

Background evaluation of antibody peptide model

In addition to the finding of the optimal combination of peptides which significantly distinguished cases from controls, a background analysis was performed. As the coefficients of the equation of the model were adjusted for each dataset we verified the results for a contribution of random selection of the data and thereby the chance of finding a comparable model by chance. The same workflow was applied for the model building except that at the beginning of the workflow the cases and controls of NELSON 1 were permutated at random. Discovery and validation was performed 12 times in the permutated NELSON 1 and NELSON 2 datasets, each time with 12 different peptides showing the lowest p-value ($p<0.05$) in the NELSON 1 set for that particular permutation. The performance of the multivariate model of the permutated discovery sets (NELSON 1) is shown in Figure 6A and the corresponding power in the validation sets (NELSON 2) in Figure 6B. Also, the performance found for the actual dataset was plotted. It can be observed that the multivariate fitting produces reasonable models even for permutated data in the discovery set.

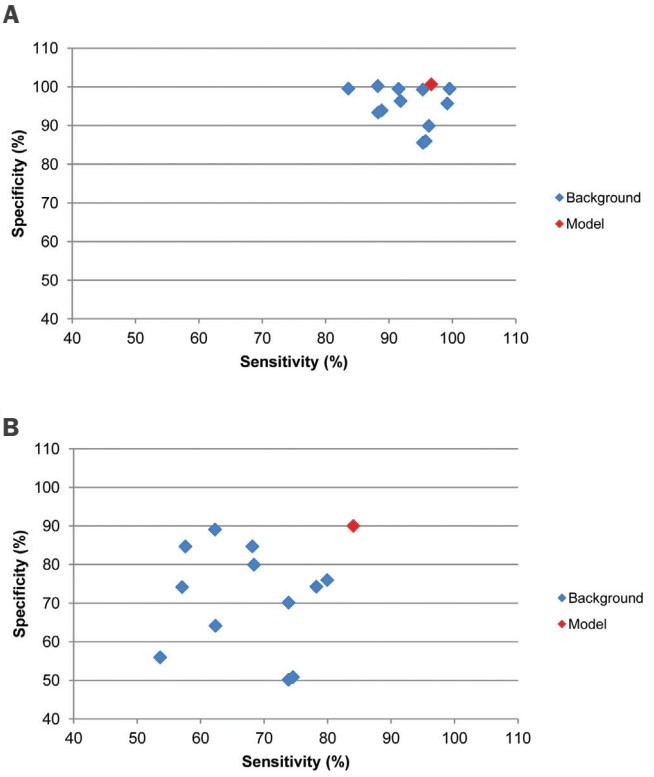


Figure 6. Background determination in NELSON 1 and NELSON 2 datasets. Twelve times a permutation (Background) was performed on the NELSON 1 and NELSON 2 dataset. The sensitivity and specificity of the antibody peptide model are shown in red. Background assessment: A) Twelve permutation runs are shown with the corresponding sensitivity and specificity of the NELSON 1 dataset (blue). The same 12 peptides found in the background evaluation of NELSON 1 were tested in NELSON 2. B) The 12 runs are shown with the corresponding sensitivity and specificity of NELSON 2 dataset (blue). Note, as some results of the background analysis occurred more than once, a random number between -1 and 1 were added to each sensitivity and specificity number to make sure each analysis (blue dot) can be seen in the figure.

However, especially in the validation dataset, the real data performed significantly better ($p<0.05$) than the permutated datasets, suggesting that the immunoglobulin peptides harbor information related to the disease state of the patient. Thus, the results we obtained do not stem from an artifact in the data processing.

CT screening result in NELSON 1 and NELSON 2 dataset

In Figure 7A and 7B the screening results of the baseline CT scans are shown for the NELSON 1 and NELSON 2 set, respectively. According to the screening protocol of the NELSON trial, a repeat CT scan was performed following an indeterminate screening result, approximately 3 months later.

We observed that 68% of the cases had a positive screening result in both the NELSON 1 and NELSON 2 set during the first 3 months of the screening program, the other lung cancers were diagnosed following another repeat CT scan after 3 months or during the second screening round. After on average 367 days (range 39-1193 days) for NELSON 1 and 269 days (range 54-777 days) for NELSON 2, the screening result was positive, i.e. suspect for lung cancer and resulting in clinical work-up by the pulmonologist and eventually finally diagnosis of lung cancer.

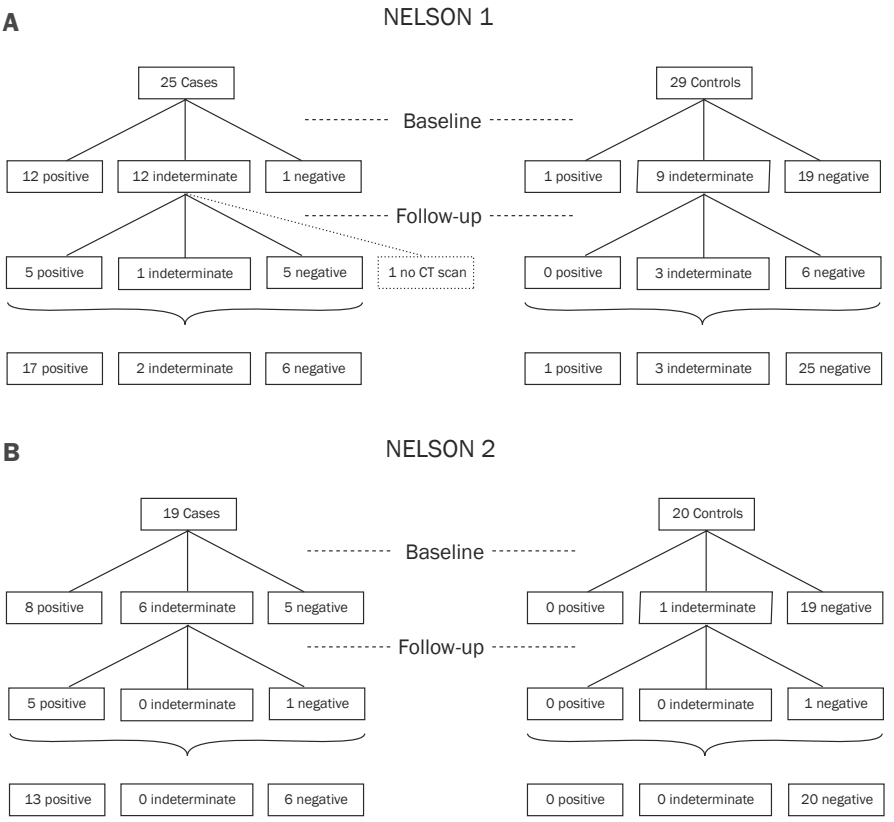


Figure 7. CT scan results of the NELSON 1 and NELSON 2 sample set. CT scan results of the A) NELSON 1 and B) NELSON 2 sample sets are shown at time of blood sampling (Baseline). Also, CT results are shown of the follow-up CT scan after approximately three months (Follow-up). For one case from the NELSON 1 set no Follow-up CT scan result was available. The last row represents the numbers of positive, indeterminate and negative CT scan results of baseline including follow-up results.

Discussion

By mass spectrometry we found evidence that a proportion of peptides of the variable part of antibodies differ between lung cancer patients and controls. A combination of 12 different peptides was able to distinguish lung cancer patients from controls in a high risk population. A sensitivity of 96% and a specificity of 100% were observed in the discovery set. An external validation in an independent case–control set was performed and generated a sensitivity of 84% and a specificity of 90%. The background evaluation showed that the 12 antibody peptide model performed significantly better than a model generated based on permutated data.

Recently, Arentz et al. published that uniquely mutated V regions peptides could be used as a proxy for the detection of anti-Ro52 autoantibodies in sera from primary Sjögren's syndrome patients by mass spectrometry.[28] Why these and other studies were able to identify similar or identical sequences could be explained by repertoire bias and the convergent evolution of antibodies during somatic mutation and selection.[19-20] This selection favors specific alleles and sequences of antibodies with the optimal affinity towards the specific antigens during immune response.[18, 29-30]

We were able to identify peptide sequences which were distributed differently between lung cancer patients and controls. The antibody peptide model consisted not only of peptide sequences positioned at the CDR regions of an immunoglobulin but also at the framework region surrounding the CDRs. It may appear surprising that most of the peptides that are represented in the antibody peptide model derive from framework regions of the immunoglobulin, rather than from the hypervariable CDRs. This may be explained by their abundance in the immunoglobulin pool. Peptides carrying only few mutations relative to the germline are more likely to occur in several antibody clones, and thus have a higher abundance. This favors their detection by the mass spectrometer, especially in samples of high complexity. While technological advances may enable the reliable quantitation of also lower abundant peptides, it may even be that hypermutated CDRs are not as likely to be common among patients sharing an immune response. But moderately mutated peptides strike the best balance between specificity, abundance and sharing for the purposes of a diagnostic marker. The large heterogeneity of lung cancer could also contribute to the presence of fewer CDR peptides shared between lung cancer patients.

We observed that the average raw abundance of 6 from the 12 peptides where distributed differently in the cases and controls between the two datasets. The average raw abundance of these six peptides was higher in the controls in the NELSON 1 set but in the NELSON 2 set the average raw abundance was higher in the cases. Probably this is due to technical problems such as measurements on different LC columns or changes in sample composition over time. The mass spectrometer is probably not able to detect every time the same peptides and at similar intensities. This was also observed with the five different biological samples we measured on two separate LC columns which were manufactured at the same time (same batch).

Beside technical problems we also have to cope with the high variability of immunoglobulins, which make the samples probably too complex for the mass spectrometer. A solution to this problem could be reduction of the complexity of the sample before it is measured on the mass spectrometer. This reduction could be established by fractionation into smaller protein fragments such as Fab-κ and Fab-λ, or by producing immunoglobulin fragments containing just the variable domains of the IgG molecule.

It was our aim to offset biological variation by including a relatively large number of patients in this study, but unfortunately large sample numbers translate to extended

measurement times of up to 8 weeks for a dataset. These measurement times introduce technical variation that counteracts the advantage gained from the number of included patients.

We were not able to distinguish lung cancer cases from controls univariately by one peptide. Instead we needed a panel of different peptides to discriminate significantly between cases and controls. Lung cancer is a very heterogeneous disease which results in high variability between patients and cancer types. This might induce various immune responses to different tumor antigens.[6-12] Therefore, finding only one antibody that is shared between all lung cancer patients is highly unlikely. Brichory et *al.* for instance showed for PGP 9.5, annexin I and II a sensitivity of only 14%, 30% and 33%, respectively.[31-32] Chapman et *al.* tested a panel of seven TAAs and found a sensitivity of 41% and a specificity of 93%. Validation of this panel in an independent sample set showed a sensitivity and specificity of 47% and 90%, respectively.[33] Koziol et *al.* were able to distinguish lung cancer patients from normal individuals with a panel of seven TAAs. A sensitivity of 80% and a specificity of 90% were observed, but no validation was performed.[34] Moreover, Khattar et al. and Zhong et al. were able to identify validated autoantibody peptide panels for lung cancer screening with sensitivity and specificity ranging from 84%-91% and 73%-91%, respectively.[35-36] It is therefore not surprising that no single peptide could be found in the current data set that distinguishes cases from controls.

Using a multivariate model, we were able to distinguish lung cancer patients from controls. However, due to the experimental and biological variation, it was necessary that we recalibrated our model for each group of patients. This limits the current applicability of the method in the clinical practice, at least until significant technical advances enable a more robust quantification and identification of peptides in such complex samples. Still, we conclude from our data that differences exist between the immunoglobulin-derived peptides from early lung cancer patients and controls. This is corroborated by data from earlier studies in our own group as well as in other research groups that showed conservation and sharing of rearranged immunoglobulin sequences in immunoglobulins against a particular antigen.[19-20, 28]

So far, only age and smoking history have been used as selection criteria for enrolment in screening trials, but it is well known that even though over 80% of all lung cancer cases are directly related to smoking, only 11% of female smokers and 17% of male smokers will be diagnosed with lung cancer during their lifetimes.[37-38] Therefore, additional diagnostic tests might select high risk individuals more precise when combined with the selection criteria age and smoking history in screening trials. The cases and controls we used for this study were selected based on their diagnosis of lung cancer within three years (range 39-1193 days) after the baseline CT scan. Therefore, calculation of sensitivity and specificity of CT screening in our subset of cases and controls from the NELSON trial are not applicable in this retrospective study. However, in this study we have demonstrated that 68% of the cases were detectable for lung cancer by CT screening. Eventually after approximately 1 year the screening result of all cases were positive.

In the high risk population of the NELSON trial still approximately 27% of the participants are subjected to invasive and expensive follow-up studies that revealed in benign disease at baseline CT screening.[5] The performance of CT improves after follow-up scans, but only after an amount of time has passed, on average a year for the sets in this study. Thus, there is need for additional diagnostic capabilities that can improve the performance of the current testing at baseline. For example, the group

of Massion recently published their results on a combination of a serum proteomic biomarker panel with clinical and CT data.[39] In the current study, we were able to detect lung cancer with an antibody peptide model in the NELSON 1 and NELSON 2 set with sensitivities of 96% and 84% and specificities of 100% and 90%, respectively at an early stage. This indicates that specific antibodies are present at an early disease stage and that such a panel of antibodies is able to detect lung cancer at an earlier stage than CT. Auto-antibody profiling has the potential to be a tool for early detection when incorporated into a comprehensive screening strategy if technical challenges described in this study can be overcome.
In conclusion, a panel of antibody peptides is identified that discriminates samples of lung cancer patients from controls. This is a first indication that peptides generated from the variable part of antibodies are shared between lung cancer patients and can be used to discriminate lung cancer patients and control groups. More quantitative work is still needed to assess the use of these peptides in clinical settings.

Acknowledgments

The authors would like to thank Frank Santegoets and Roel Faber from the department Public Health, Erasmus Medical Center for providing us the CT scan data of the used cases and controls from the NELSON trial in this study.

References

1. IARC: section of cancer information. Globocan. 2008.
2. Jemal A, Siegel R, Xu J, Ward E. Cancer statistics, 2010. CA Cancer J Clin 2010; 60:277-300.
3. Aberle DR, Adams AM, Berg CD, Black WC, Clapp JD, Fagerstrom RM, et al. Reduced lung-cancer mortality with low-dose computed tomographic screening. N Engl J Med 2011; 365:395-409.
4. Horeweg N, Aalst CMvd, Vliegenthart R, Zhao Y, Xie X, Scholte ET, et al. Volumetric computer tomography screening for lung cancer: three rounds of the NELSON trial. Eur Respir J 2013; in press.
5. van Klaveren RJ, Oudkerk M, Prokop M, Scholten ET, Nackaerts K, Vernhout R, et al. Management of lung nodules detected by volume CT scanning. N Engl J Med 2009; 361:2221-9.
6. Anderson KS, LaBaer J. The sentinel within: exploiting the immune system for cancer biomarkers. J Proteome Res 2005; 4:1123-33.
7. Baldwin RW. Immunity to transplanted tumour: the effect of tumour extracts on the growth of homologous tumours in rats. Br J Cancer 1955; 9:646-51.
8. Caron M, Choquet-Kastylevsky G, Joubert-Caron R. Cancer immunomics using autoantibody signatures for biomarker discovery. Mol Cell Proteomics 2007; 6:1115-22.
9. Gure AO, Altorki NK, Stockert E, Scanlan MJ, Old LJ, Chen YT. Human lung cancer antigens recognized by autologous antibodies: definition of a novel cDNA derived from the tumor suppressor gene locus on chromosome 3p21.3. Cancer Res 1998; 58:1034-41.
10. Hanash S. Harnessing immunity for cancer marker discovery. Nat Biotechnol 2003; 21:37-8.
11. Mintz PJ, Kim J, Do KA, Wang X, Zinner RG, Cristofanilli M, et al. Fingerprinting the circulating repertoire of antibodies from cancer patients. Nat Biotechnol 2003; 21:57-63.
12. 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.
13. Liu L, Liu N, Liu B, Yang Y, Zhang Q, Zhang W, et al. Are circulating autoantibodies to ABCC3 transporter a potential biomarker for lung cancer? J Cancer Res Clin Oncol 2012.
14. Murphy K. TP, Walport M. Janeway's immunobiology. 7th ed: Garland Science; 2008.
15. Saada R, Weinberger M, Shahaf G, Mehr R. Models for antigen receptor gene rearrangement: CDR3 length. Immunol Cell Biol 2007; 85:323-32.
16. Foreman AL, Lemercier B, Lim A, Kourlisky P, Kenny T, Gershwin ME, et al. VH gene usage and CDR3 analysis of B cell receptor in the peripheral blood of patients with PBC. Autoimmunity 2008; 41:80-6.
17. Scheid JF, Mouquet H, Ueberheide B, Diskin R, Klein F, Oliveira TY, et al. Sequence and structural convergence of broad and potent HIV antibodies that mimic CD4 binding. Science 2011; 333:1633-7.
18. VanDuijn MM, Dekker LJ, Zeneyedpour L, Smitt PA, Luider TM. Immune responses are characterized by specific shared immunoglobulin peptides that can be detected by proteomic techniques. J Biol Chem 2010; 285:29247-53.
19. Weinstein JA, Jiang N, White RA, 3rd, Fisher DS, Quake SR. High-throughput sequencing of the zebrafish antibody repertoire. Science 2009; 324:807-10.
20. Maat P, VanDuijn M, Brouwer E, Dekker L, Zeneyedpour L, Luider T, et al. Mass spectrometric detection of antigen-specific immunoglobulin peptides in paraneoplastic patient sera. J Autoimmun 2012; 38:354-60.
21. van Iersel CA, de Koning HJ, Draisma G, Mali WP, Scholten ET, Nackaerts K, et al. Risk-based selection from the general population in a screening trial: selection criteria, recruitment and power for the Dutch-Belgian randomised lung cancer multi-slice CT screening trial (NELSON). Int J Cancer 2007; 120:868-74.

22. de Costa D, Broodman I, Vanduijn MM, Stingl C, Dekker LJ, Burgers PC, et al. Sequencing and quantifying IgG fragments and antigen-binding regions by mass spectrometry. J Proteome Res 2010; 9:2937-45.
23. Lefranc MP, Giudicelli V, Ginestoux C, Jabado-Michaloud J, Folch G, Bellahcene F, et al. IMGT, the international ImMunoGeneTics information system. Nucleic Acids Res 2009; 37:D1006-12.
24. Snedecor GW, W.G. C, editors. Statistical Methods. 7 ed. Iowa: The Iowa State University Press; 1980.
25. Armitage P, Berry G, S. MJN, editors. Statistical Methods in Medical Research. 4 ed. Oxford: Blackwell Scientific.; 2002.
26. Afifi A, Clark V, May S, editors. Computer-aided multivariate analysis. 4 ed. Florida: Chapman & Hall; 2004.
27. Kleinbaum DG, Kupper LL, Muller KE, Nizam A, editors. Applied regression analysis and other multivariable methods. 3 ed. California: Duxbury Press; 1998.
28. Arentz G, Thurgood LA, Lindop R, Chataway TK, Gordon TP. Secreted human Ro52 autoantibody proteomes express a restricted set of public clonotypes. J Autoimmun 2012.
29. Andersen PS, Haahr-Hansen M, Coljee VW, Hinnerfeldt FR, Varming K, Bregenholt S, et al. Extensive restrictions in the VH sequence usage of the human antibody response against the Rhesus D antigen. Mol Immunol 2007; 44:412-22.
30. Baranzini SE, Jeong MC, Butunoi C, Murray RS, Bernard CC, Oksenberg JR. B cell repertoire diversity and clonal expansion in multiple sclerosis brain lesions. J Immunol 1999; 163:5133-44.
31. Brichory F, Beer D, Le Naour F, Giordano T, Hanash S. Proteomics-based identification of protein gene product 9.5 as a tumor antigen that induces a humoral immune response in lung cancer. Cancer Res 2001; 61:7908-12.
32. Brichory FM, Misek DE, Yim AM, Krause MC, Giordano TJ, Beer DG, et al. An immune response manifested by the common occurrence of annexins I and II autoantibodies and high circulating levels of IL-6 in lung cancer. Proc Natl Acad Sci U S A 2001; 98:9824-9.
33. Chapman CJ, Healey GF, Murray A, Boyle P, Robertson C, Peek LJ, et al. EarlyCDT(R)-Lung test: improved clinical utility through additional autoantibody assays. Tumour Biol 2012.
34. Koziol JA, Zhang JY, Casiano CA, Peng XX, Shi FD, Feng AC, et al. Recursive partitioning as an approach to selection of immune markers for tumor diagnosis. Clin Cancer Res 2003; 9:5120-6.
35. Khattar NH, Coe-Atkinson SP, Stromberg AJ, Jett JR, Hirschowitz EA. Lung cancer-associated auto-antibodies measured using seven amino acid peptides in a diagnostic blood test for lung cancer. Cancer Biol Ther 2010; 10:267-72.
36. 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.
37. Pass H.I. CDP, Johnson D.H., Minna J.D., Scagliotti G.V., Turrisi A.T. Principles and practice of lung cancer. The official reference text of the IASLC. 4th ed: Lippincott Williams & Wilkins; 2010.
38. Villeneuve PJ, Mao Y. Lifetime probability of developing lung cancer, by smoking status, Canada. Can J Public Health 1994; 85:385-8.
39. Pecot CV, Li M, Zhang XJ, Rajanbabu R, Calitri C, Bungum A, et al. Added value of a serum proteomic signature in the diagnostic evaluation of lung nodules. Cancer Epidemiol Biomarkers Prev 2012; 21:786-92.

Chapter 5

Label-free peptide profiling of Orbitrap™ full mass spectra

Mark K Titulaer, Dominique de Costa, Christoph Stingl, Lennard J Dekker, Peter AE Sillevis Smitt, Theo M Luider

Adapted from BMC Res Notes 2011;27:4-21

Abstract

We developed a new version of the open source software package Peptrix that can yet compare large numbers of Orbitrap™ LC-MS data. The peptide profiling results for Peptrix on MS1 spectra were compared with those obtained from a small selection of open source and commercial software packages: msInspect, Sieve™ and Progenesis™. The properties compared in these packages were speed, total number of detected masses, redundancy of masses, reproducibility in numbers and CV of intensity, overlap of masses, and differences in peptide peak intensities. Reproducibility measurements were taken for the different MS1 software applications by measuring in triplicate a complex peptide mixture of immunoglobulin on the Orbitrap™ mass spectrometer. Values of peptide masses detected from the high intensity peaks of the MS1 spectra by peptide profiling were verified with values of the MS2 fragmented and sequenced masses that resulted in protein identifications with a significant score. Peptrix finds about the same number of peptide features as the other packages, but peptide masses are in some cases approximately 5 to 10 times less redundant present in the peptide profile matrix. The Peptrix profile matrix displays the largest overlap when comparing the number of masses in a pair between two software applications. The overlap of peptide masses between software packages of low intensity peaks in the spectra is remarkably low with about 50% of the detected masses in the individual packages. Peptrix does not differ from the other packages in detecting 96% of the masses that relate to highly abundant sequenced proteins. MS1 peak intensities vary between the applications in a non linear way as they are not processed using the same method. Peptrix is capable of peptide profiling using Orbitrap™ files and finding differential expressed peptides in body fluid and tissue samples. The number of peptide masses detected in Orbitrap™ files can be increased by using more MS1 peptide profiling applications, including Peptrix, since it appears from the comparison of Peptrix with the other applications that all software packages have likely a high false negative rate of low intensity peptide peaks (missing peptides).

Introduction

High throughput Orbitrap™ (Thermo Fischer Scientific, Germany) mass spectrometry (MS) makes it possible to obtain full MS1-spectra and fragmentation-MS2 (MS/MS) spectra of peptides for comparison and identification purposes. The technique can be applied to compare the differences in quantities of proteins in body fluid and tissue samples. The peptides from enzymatic digested proteins are separated on an LC column. During elution, depending on sample complexity 1-100% of separated peptides detected in the spectra of the MS1 scans can be MS2 triggered by the Xcalibur™ instrument software for MS2 fragmentation [1].

The Peptrix application can handle raw Orbitrap™ files as well as MALDI-TOF and MALDI-FT-ICR mass spectra [2-7]. Peptide profiling requires the following basic steps: 1) peak picking from the raw mass spectra; 2) time alignment of the extracted peak masses between different LC runs; 3) aggregation of masses and corresponding intensities of different sample runs on the Orbitrap™ in a peptide profile matrix; and 4) statistical analysis to highlight masses differentially expressed between different groups. A peptide profile matrix, frequently called Peptide Array or PepArray, is created as an output file. Peptide peak intensities are presented in this matrix for all masses detected in every Orbitrap™ measurement. These MS1 masses can eventually be linked to protein identifiers using MS2 sequence information and available protein databases. Table 1 shows a fragment of such a peptide profile matrix. Replicate measurements from a tryptic digested IgG Fab sample are presented as numbers 1, 2 and 3 in the matrix columns, with the retention time and mass of a peptide in the matrix rows, e.g. peptide mass 1239.259 Da eluting at a retention time of 7969.383 s. The three replicate peak intensities measured for the mass 1239.259 Da are given in the matrix cells, e.g. the values 10005, 13333, 19683 in arbitrary units.

Peptrix is not completely new software, but an extension of already published nameless software. The architecture of Peptrix is described in [7]. The application consists of: 1) a Java™ graphical interface; 2) Mysql database for storage of meta-data; 3) ftp storage of raw data and processed files: and 4) an interface to R for statistical analysis. The software has changed in many aspects with respect to the previously reported version. Firstly, the peptide profile matrix created from LC-MS experiments contains an extra retention time dimension as peptides elute at different time points from the nano-LC column. Peak-picking algorithms over time are implemented combining more Orbitrap™ scans. Time alignment has to be implemented between different LC runs of the sample. Nano-spray ionization from LC-MS also generates multiple charged peptide ions and a different de-isotoping algorithm was implemented than was required for single charged peptides in MALDI-TOF and MALDI-FTICR measurements. Instead of eliminating isotopes from the peak-lists, which is possible in MALDI experiments, mono-isotopes have to be selected from the raw Orbitrap™ spectra by peak-picking algorithms based on expected isotopic intensity distributions.

Other software packages exist for comparing the raw Xcalibur™ MS1 data between samples, possibly converted into mzXML formatted files, e.g. msInspect, MZMine, OpenMS, VIPER, PEPPer, MSight [8-10]. These tools generate peptide profile matrices, in which spectral intensities and retention times of peptide masses from samples belonging to different groups are presented in various ways.

Table 1. A fragment of a peptide profile matrix or PepArray.

MH+	time (s)	Peak intensity in sample 1	Peak intensity in sample 2	Peak intensity in sample 3	Peptide present in sample 1	Peptide present in sample 2	Peptide present in sample 3	Total count of peptides
1238.712	5702.29	30528	25175	23642	1	1	1	3
1238.735	7770.22	12416	9487	7326	1	1	1	3
1238.899	713.267	7848	5629	6229	1	1	1	3
1239.259	7969.383	10005	13333	19683	1	1	1	3
1239.53	4314.73	7110	10243	7283	1	1	1	3
1239.597	8150.09	5207	6428	2798	1	1	1	3
1239.599	4408.91	8264	7158	6992	1	1	1	3
1239.601	7048.683	4542	8373	6982	1	1	1	3
1239.621	1190.17	370540	333496	302810	1	1	1	3
1239.622	6657.29	69391	66874	53379	1	1	1	3
1239.624	5446.54	26198	32726	20632	1	1	1	3
1239.635	4654.07	60855	59416	159055	1	1	1	3
1239.638	6675.558	10973	0	14356	1	0	1	2
1239.64	3143.02	6429	6080	5409	1	1	1	3
1239.642	5808.01	192225	191568	159055	1	1	1	3
1239.692	4271.67	256980	297801	209433	1	1	1	3
1239.734	10051	6161	6481	5449	1	1	1	3
1239.749	7239.35	18034	14470	16265	1	1	1	3
1239.75	6547.471	7043	8459	5901	1	1	1	3
1240.065	5805.98	14427	14851	6499	1	1	1	3
1240.098	5509.82	19378	25322	20168	1	1	1	3
1240.499	2631.25	17863	9718	15101	1	1	1	3
1240.521	4792.05	15576	14008	16506	1	1	1	3

The replicate measurements of an tryptic digested IgG Fab sample are presented as numbers 1, 2 and 3 in the columns of the matrix whereas the retention time and mass of a peptide are presented in the rows of the matrix, e.g. peptide mass 1239.259 Da eluting at a retention time of 7969.383 s. The measured peak intensities for the three replicate measurements of peptide mass 1239.259 Da are presented in the cells of the matrix, 10005, 13333, 19683 in arbitrary units.

Some of these software packages, such as SuperHirn and SpecArray, did at the time of analysis not run on the Windows Operating System (OS) but only on Linux [9]. Other applications required customized data input formats or connection to pre-filled databases with equipment-dependent retention times for sequenced peptide masses. The Orbitrap™ files contains all the necessary MS1 and MS2 information (for time alignment), and full analysis only requires an internet connection to a protein database interface, e.g. Mascot™, as implemented in Progenesis™. Some applications cannot handle the approximately 1.8 GB big mzXML files, processed by readw.exe version 4.2.1 from the raw files [11]. This can be due to the size of the files causing RAM related issues. Another reason might be that readw.exe generates not entirely correct structurized mzXML files. In some files mzXML closing tags are missing. Readw.exe could not process files larger than 2 GB on our hardware; Intel Xeon W3520 Quad-Core 2.67 GHz processor with 3.5 GB RAM.

The result of peptide profiling by Peptrix on MS1 spectra were compared with that obtained from a small selection of open source and commercial Windows software packages, i.e. 1) commercial Sieve™ [1]; 2) open source msInspect [12-13]; and 3) com-

mercial Progenesis™ [14]. The aspects compared were: 1) speed; 2) total number of detected masses in the profile matrix; 3) redundancy of masses; 4) reproducibility of number of masses and CV of intensity; 5) overlap of masses between the selected packages; and 6) differences of peptide peak intensities determined by the software packages. The (basic) workflow of activities for the software tools compared - Peptrix, Sieve™, msInspect, and Progenesis™ - is shown in Figure 1.

To compare the four packages, we analyzed three technical replicates of tryptic digested immunoglobulin G (IgG) Fragment antigen binding (Fab) of human serum. We used Peptrix to compare the output of the triplicate measurements from the software packages. Unless important for interpretation of the results, we will not describe how they actually work in terms of algorithms, time alignment, peak selection by isotopic pattern recognition, using peak maxima, features or framing. For these matters, we refer you to the manufacturer's documentation, the comparison study in [9] or the (basic) workflow of activities for the tools compared - Peptrix, Sieve™, msInspect, and Progenesis™ - depicted in Figure 1.

As a practical example of Peptide profiling by Peptrix, we present the analysis results of Orbitrap™ measurements of in total 40 micro-dissected tissue samples, 10 spectra of glioma blood vessels, 10 spectra of tissue surrounding the glioma vessels, 10 spectra of normal endothelial vessels, and 10 spectra of endothelial tissue surrounding the normal vessels, previously analyzed by FT-ICR MS [15] .

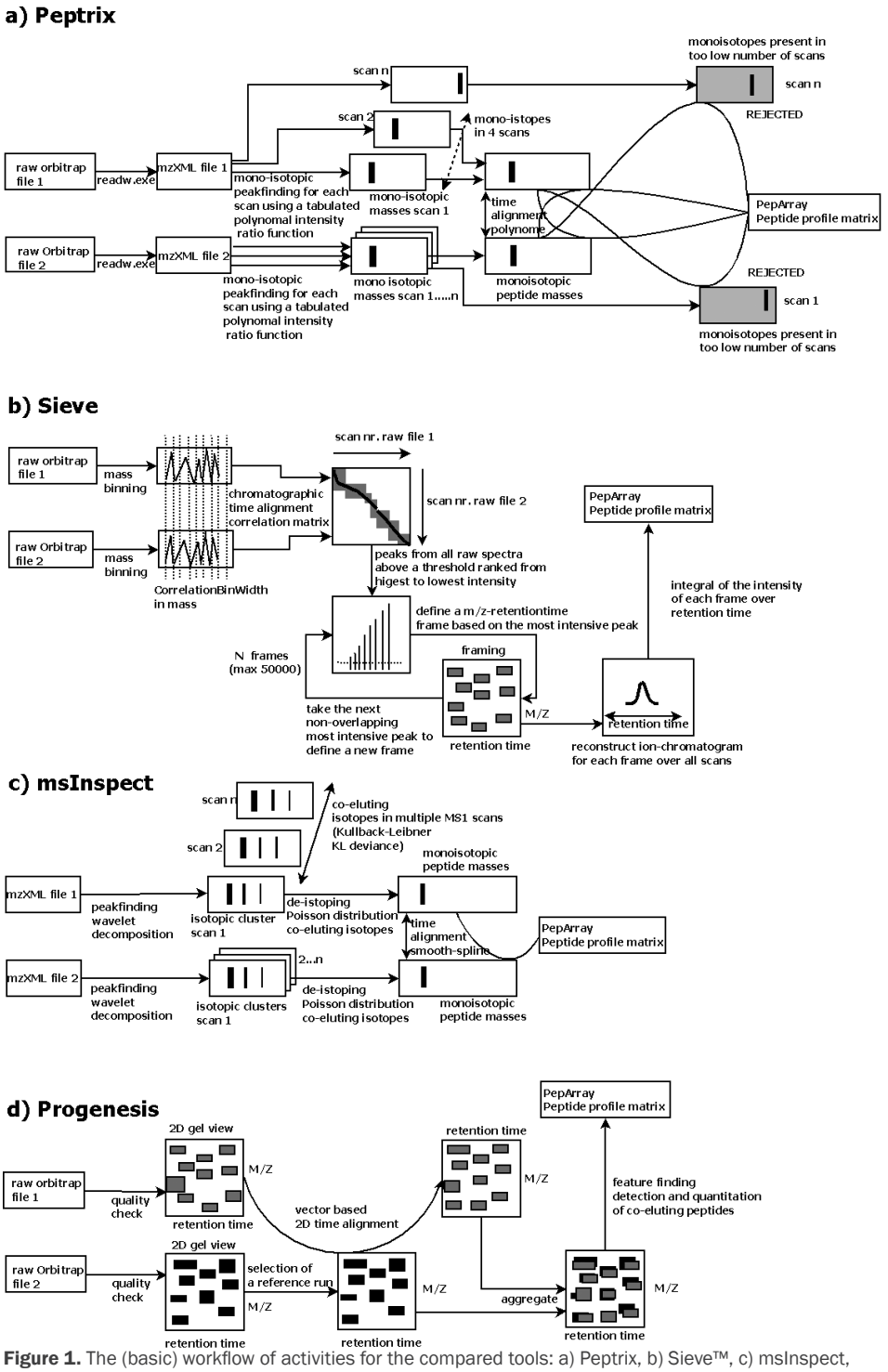


Figure 1. The (basic) workflow of activities for the compared tools: a) Peptrix, b) Sieve™, c) msInspect, and d) Progenesis™.

Materials and Methods

The purification of IgG Fab in human serum sample, tryptic digestion of the isolated Fab for MS and the mass spectrometry measurements are described in previous published work.[16] MS1 Peptide profile matrices of the IgG Fab serum replicate samples were created from raw Orbitrap™ files using Peptrix version 2.4.9, Sieve™ version 1.2, Progenesis™ version 2.0 and msInspect (build 382, 2004-2009) after conversion into mzXML files. The software packages Peptrix and msInspect de-convolute charged masses in isotopic clusters to the mono-isotopic MH^+ values. All peptide profile matrices were created with the software package default settings, such as exclusion of single charged masses.

Mass and retention time window

To prevent software packages recording too much redundant peptide masses and misfits through the use of a retention time window that is too narrow when the peptide profile matrix is being constructed, expected retention time differences of peptides were determined between two consecutive LC runs of replicates.
Based on the maximum expected retention time differences observed (data not shown), we used a conservative time window of 5 min [17] (400 frames for msInspect) and mass window of 0.02 Da (10 ppm for Peptrix) for the software packages Peptrix and Sieve™ to produce the four MS1 peptide profile matrices for the three IgG Fab replicates for mainly double and triple charged peptides. A setting of 50,000 frames was used for Sieve™. Mass and retention time windows could not be set for Progenesis™.
The time window of 5 minutes is used in an additional way in Peptrix. Peptrix has an algorithm that avoids redundancies of peptide masses. When a peptide mass is detected by Peptrix at a specific retention time it is recorded in the Peptide profile matrix. When the same eluting peptide mass is detected again at a later moment within the time window of the previous measurement it is not recorded twice in the Peptide profile matrix.

MS2 sequenced and identified masses

Values of peptide masses found in MS1 spectra by peptide profiling were verified using values of MS2 sequenced masses that resulted in protein identifications with a significant score and a Gene Identifier (GI), using the Mascot™ Daemon interface (Matrix Science, UK) [18]. MS2 triggered means that the Xcalibur™ software selects the peak mass in the MS1 spectrum for MS2 fragmentation and sequencing depending on inclusion settings. The MS2 fragmentation does not necessarily result in peptide identifications of proteins with a GI. The quality of the MS2 spectra may be too low to have a significant score from the search engine used. Also sequences of good quality MS2 spectra are sometimes not found in the in-silico digested protein databases.
The Thermo Fischer Scientific extract_msn.exe [18] program embedded in Mascot™ Daemon version 2.2.2 [18] interface extracts Mascot™ generic files (MGF files). The resulting MGF files contain the precursor masses (m/z), their charge states (z), scan identifiers, and peak lists of all MS2 spectra. The MGF files were then sent to the Mascot™ server and the following settings were used for the NCBI human database: tryptic digestion considering 1 possible missed cleavage, variable modification oxidation of Methionine (M) (mass + 15.9994 Da), 10 ppm precursor and 0.6 Da fragment tolerance.

Comparison matrix

We used Peptrix to compare the matrices from the 4 software packages investigated together with the list of MS2 spectra triggered and MS2 sequenced masses where applicable. It is possible in Peptrix to create a profile matrix of mass-intensity peak-lists from MS experiments [6]. A total of 4 peak-lists containing masses and intensities (1 for each matrix), together with one list of MS2 triggered masses and one list of the MS2 fragmented masses where sequencing succeeded were extracted from the 4 software package peptide profiles. An artificial reference list or grid of 20,275 masses, approximately equal to the number of features in the MS1 profile matrices, was constructed in a mass range between 1,600 and 2,400 Da with fixed distances of 20 ppm between the grid masses. A somewhat greater tolerance than the maximum expected mass inaccuracy of 10 ppm was used to reduce the possibility of slightly different masses being measured in the profile matrices of two different software applications for the same peptide end in two bins.
Peak masses from the generated 6 peak lists: 4 extracted from the matrices produced by the four software packages; the list of MS2 triggered masses; and the list of MS2 triggered masses where fragmentation and sequencing succeeded were matched with the artificial reference list using Peptrix. A mass window setting of plus or minus 10 ppm was used. This forces all the masses between 1,600 and 2,400 Da from the peptide profile matrices to match with at least one of the 20,275 grid points of the reference list. The numbers of overlapping and non-overlapping masses from the software packages were calculated using this constructed comparison matrix (Table 2).

Table 2. A fragment of the comparison matrix.

Mass MH^+	ms2 triggered (for fragmentation)	ms2 sequenced (proteins found)	msInspect Peak intensity (x 10^3)	Peptrix Sum peak intensities (x 10^3)	Progenesis™ Peak intensities (x 10^3)	Sieve™ Peak intensities (x 10^3)
1741.6671	0	0	0	0	0	0
1741.7007	0	0	473	0	0	652985
1741.7447	1	1	30698	73319	133568	73185
1741.7638	1	0	277640	995497	107051	256383
1741.8094	1	0	3816	222852	14080	0
1741.8397	1	0	41071	146578	58422	78170
1741.8707	1	1	26311	92368	74067	2169410
1741.9122	0	0	3738	19527	6889	0
1741.9355	1	0	202031	95472	564808	273129
1741.9807	0	0	0	0	0	0
1742.008	0	0	1547	0	0	0
1742.0503	0	0	0	0	0	0
1742.0852	0	0	0	0	0	0
1742.12	0	0	0	0	0	0
1742.1472	0	0	0	0	0	101289
1742.193	0	0	0	0	0	45510
1742.2281	0	0	0	2691996	0	0
1742.2594	0	0	0	0	0	0
1742.2909	0	0	1407	69951	0	70092
1742.3208	0	0	0	0	0	3561843
1742.3641	0	0	2474	56134	0	106378
1742.3994	0	0	424	0	0	0
1742.4336	0	0	0	0	0	0
1742.4685	0	0	0	0	0	0
1742.5033	0	0	0	0	0	0
1742.5382	0	0	0	0	0	0
1742.573	0	0	0	0	0	0
1742.6079	0	0	0	0	0	0
1742.6389	0	0	0	0	621012	0
1742.6776	0	0	0	0	0	0
1742.7124	0	0	0	0	0	0
1742.7512	0	0	460	17303	0	60179
1742.7796	0	0	242	5676	0	467398
1742.8157	1	0	1002	67850	82288	43121
1742.8449	0	0	15846	10995	5417	56105
1742.8839	1	1	582	102837	26609	27587
1742.9085	0	0	15849	120678	69898	54135
1742.95	0	0	0	25553	8075	263965
1742.9913	0	0	0	0	0	0
1743.0287	0	0	1005	0	0	0
1743.0563	0	0	0	**4248221**	0	0
1743.0959	0	0	0	0	0	0
1743.1307	0	0	0	0	0	0

A total of 4 peak-lists containing masses and intensities (1 for each matrix), together with one list of MS2 triggered masses and one list of the MS2 fragmented masses where sequencing succeeded were matched plus or minus 10 ppm with an artificial reference list or grid of masses with fixed distances of 20 ppm. The comparison matrix contains unoccupied space of MH^+ mass values, roughly separated by the dashed lines. The triple charged peptide mass 581.69 Da, which when recalculated to the MH^+ value of 1743.0563 Da is only detected by Peptrix.

Results

Computation time

Table 3 displays the computation time for Peptrix and the 3 compared software packages. Progenesis™ has the lowest computation time of 1 hour (with 24 MB RAM). The software packages msInspect and Sieve™ need somewhat more time with 2 hours, while Peptrix processes the data in a slightly longer period of 3.5 hours. This is due to: 1) storage of the peak list on an FTP server for every MS1 scan; and 2) the extra comparison steps indicated by the grey boxes in Figure 1a when preparing the matrix. Peptide masses found in at least 4 MS1 scans (file size ~1MB) are also compared with mono-isotopic masses present in less than 4 MS1 scans (file size ~9 MB). These necessary extra comparison steps guarantee reproducibility of peak intensities in the three replicate measurements when working with peak lists.

Table 3. Analysis times of the software packages investigated

Processor	Peptrix Intel Xeon W3520 Quad-Core 2.67 GHz	Sieve™ Intel Xeon X5472 Quad Core 3 GHz	msInspect Intel Xeon 5160 Dual Core 3 GHz	Progenesis™ Intel Xeon E5430 Quad Core 2.66 GHz
RAM (Giga Byte, GB)	3.5	3	2	24
Analysis Time (hours)	3.5	2	2	1

Numbers and reproducibility of peptide masses in the peptide profile matrices

Figure 2 shows a histogram representing the number of peptide masses, recalculated to MH^+ values, detected in 1, 2 or 3 technical replicate measurements of the IgG Fab at a specific retention time in MS1 peptide profile matrices produced by the four software packages. Ideally all masses should be measured with the same intensity in the 3 replicate measurements in the sample. The peptide profile matrices produced by the software packages contain about 20,000 to 70,000 mass-retention time entities (Figure 2). The Peptrix profile matrix contains a total number of 30,986 MH^+ masses mainly detected in double or triple charged peak clusters in the spectra. 86% of all masses are measured in all three replicates.

Sieve™ displays a larger total number of 33,967 peptide masses in the profile matrix detected in 50,000 frames with charge states > 0, of which about 21,000 masses have charge states 2 or 3. The peptide profile matrix produced by Sieve™ contains peak masses that are nearly present in all 3 replicates for all charge states.

The peptide profile matrix produced by msInspect displays the largest number of masses, i.e. 72,895 masses (Figure 2). The most important reason for the large number of masses and the relative lower overlap in msInspect is that it includes peptide masses that are only present in a few Orbitrap™ MS1 scans. The other software packages use more scans, e.g. Peptrix requires a peptide mass in at least 4 consecutive MS1 scans. In msInspect, most masses occur in one replicate, i.e. 76% of the total number. This is due to the fact that msInspect creates the peptide profile matrix in a sequence-dependent way. It matches a mass in the third replicate if it is already measured in the first and second replicate. Therefore, masses that occur only in the second, third or both measurements are not included in the peptide profile matrix.

Progenesis™ measures a low total number of 23,654 masses, of which 19,039 have

charge states 2 or 3. Like for Sieve™, nearly all masses measured in all 3 replicates for all charge states (Figure 2). However, this matrix contains redundant peptide masses deviating less than 10 ppm from each other measured at consecutive retention times.

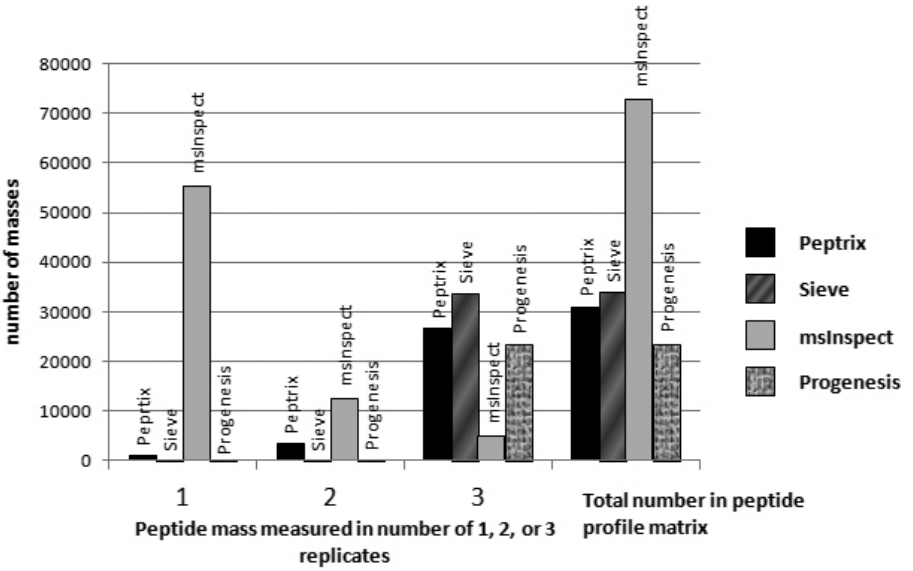


Figure 2. Histogram representing the number of MS1 peptide masses detected in 1, 2, or 3 technical replicate measurements and the total number of peptide masses in 4 different MS1 software packages. Ideally all masses should be measured in the 3 replicate measurements of the sample, since the spectra originate from the same sample.

Repeated measurement of a single peptide mass in the peptide profile matrices

The four software packages show problems with peak tailing of high abundant peaks. Although the intensities in the peak tail are just fractions (≤ 0.1%) of the apex intensities, they still can be detected minutes after the peak, and in extreme cases smear until the end of the run. Such an example is shown for MH^+ peptide mass 1502.756993 Da of Ig kappa chain C region (Table 4) with the sequence DSTYSLSSTLTLSK, which eluted from approximately 85 to 180 minutes. The peptide mass was measured 28 times in the Progenesis™ peptide profile matrix, about 27 times in the Sieve™ profile matrix, 8 times in msInspect, and only 1 time in the Peptrix profile matrix.

Another extreme example is the peptide VYACEVTHQGLSSPVTK with mass MH^+ 1818.9042 Da of Ig kappa chain C region (Table 4). This peptide eluted between approximately 50 and 130 minutes and the mass was measured 19 times in the Progenesis™ peptide profile matrix, 10 times in the Sieve™ profile matrix, about 18 times in msInspect, and 3 times in the Peptrix profile matrix.

Table 4. Peptide masses of in-silico digested Ig kappa chain C region, GI 157838230, either found in the software packages Peptrix, Sieve™, msInspect, Progenesis™ and the Mascot™ Daemon.

		Peptrix				Sieve™				msInspect				Progenesis™				Mascot Daemon			
		Peak intensity (x 10^3)				Peak intensity (x 10^3)				Peak intensity (x 10^3)				Peak intensity (x 10^3)				(GI: gene identifier)			
Mass MH+	# Sequence	1	2	3	§	1	2	3	§	1	2	3	§	1	2	3	§	1	2	3	§
888.49378	1 EAKVQWK	70	68	60	2	40	37	33	0	17	18	0	4	29	27	26	0	0	0	0	
1502.75844	0 K.DSTYSLSSTLTLSK.A	23172	22659	18733	0	15343	14428	12549	1	10100	9391	8288	3	16869	15721	13663	1	157838230	157838230	157838230	0
1740.87377	0 SGTASVVCLLNNFYPR	55	58	45	1	396	366	320	4	19	21	19	3	118	113	83	1	157838230	157838230	157838230	0
1818.90547	0 VYACEVTHQGLSSPVTK	663	588	465	2	177	182	175	0	155	162	142	0	239	218	193	2	157838230	157838230	157838230	0
1946.02696	0 TVAAPSVFIFPPSDEQLK	29952	10994	24525	2	25191	23771	21825	1	3585	3517	5738	2	83933	78466	73157	2	157838230	157838230	157838230	0
2069.04844	1 SGTASVVCLLNNFYPREAK	17	17	13	0	43	36	28	3	30	0	0	3	6	6	4	1	0	0	0	
2084.05934	1 HKVYACEVTHQGLSSPVTK	2	3	4	0	0	0	0		5	0	0	1	(5	4	4)&	0	157838230	1*	157838230	0
2109.02339	1 DSTYSLSSTLTLSKADYEK	593	576	499	1	201	191	163	0	250	247	221	5	280	265	238	0	157838230	157838230	157838230	0
2135.96873	0 VDNALQSGNSQESVTEQDSK	65032	65514	62998	1	20162	19137	18511	1	33000	3100	0	3	43683	35245	41405	1	157838230	157838230	157838230	0
2323.14995	1 VYACEVTHQGLSSPVTKSFNR	3	7	6	3	13	12	13	4	50	52	36	7	10	11	11	8	0	1*	1*	7
2553.22833	1 DIEMTQSPSSLSASVGDRVTITCR	0	0	0		63	60	52	2	0	0	0		0	0	0		0	0	0	
2677.27	1 VQWKVDNALQSGNSQESVTEQDSK	422	381	329	0	142	134	115	2	279	259	6	7	369	341	304	2	157838230	157838230	157838230	3
3619.70933	1 VDNALQSGNSQESVTEQDSKDSTYSLSSTLTLSK	17	15	13	5	0	0	0		31	34	25	4	68	60	50	1	0	0	0	

missed cleavage, § mass accuracy in ppm, * MS triggered but no protein matches, & charge 4+ (charge 4+ not included in Peptrix) The spectral intensities of the triplicate measurements of each peptide mass are presented in the columns, except for Mascot™ Daemon. The corresponding GI numbers of the identified proteins by Mascot™ Daemon are presented in the Mascot™ Daemon columns.

Overlap of peptide masses between the MS1 peptide profile matrices

A number of 1,578 Peptrix peptide masses between 1,600 and 2,400 Da differ more than 20 ppm from each other. This number represents 27% of the grid points in the comparison matrix. In reality, 38% (> 27%) of the total number of grid points match with a Peptrix single peptide. This means that some Peptrix masses within 20 ppm split-up and match with two grid-points. Therefore, the non-matching peptide masses measured between the packages are really significant for 100 * 27/38 ≈ 70%. The other grid points match with more than one mass in the Peptrix peptide profile matrix. Percentages: 36%, 16%, 7%, 2% and 1% of the 5,777 grid points are measured for combinations with 2, 3, 4, 5 and 6 masses of the Peptrix peptide profile matrix respectively. This means that overlap between packages is likely to be overestimated using the comparison matrix grid, not taking the retention time into account.

Figure 3 shows the pair-wise overlap between two packages in descending overlap order, using the comparison matrix. The average number of 5,618 masses in the comparison matrix for each package is about four times lower than the 20,275 reference points between 1,600 and 2,400 Da, since the grid in the comparison matrix contains unoccupied space of MH^+ mass values (see Table 2). The overlap between each time two packages is relatively low with 1/3 of the number of matches with the grid of two software packages together. Most overlap was determined between Peptrix and msInspect, and the least overlap was between Sieve™ and Progenesis™.

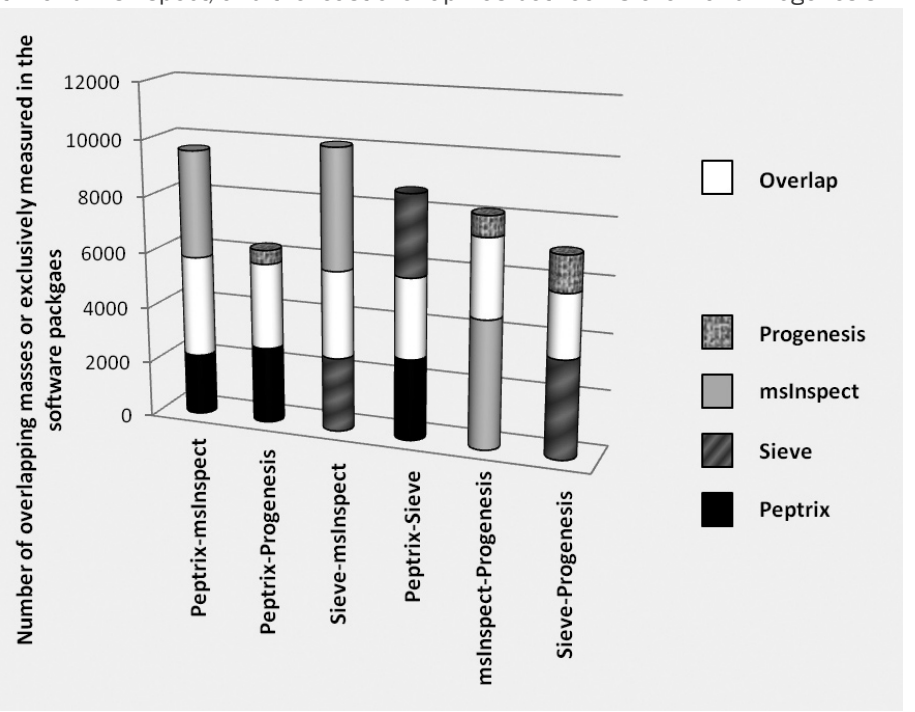


Figure 3. A pair-wise comparison of overlapping peptide masses, each time between the peptide profile matrices of two MS1 software packages. A comparison matrix was constructed by Peptrix using an artificial grid of 20275 masses between 1600 and 2400 Da with distances of 20 ppm between the masses. Peak masses from two peptide profile matrices for the software packages Peptrix, Sieve™, Progenesis™ and msInspect were matched with this grid in the comparison matrix, using a mass window with 10 ppm in two directions and a total distance of 20 ppm.

The numbers of matched masses between more than 2 software packages are presented in the 4-way Venn diagram in Figure 4. The number of non-overlapping masses from the software packages is relatively large for Peptrix, Sieve™ and msInspect, i.e. 1,302 for Peptrix, 1,920 for Sieve™ and 2,791 for msInspect, while 168 is measured for Progenesis™. The number of non-matching MH^+ peptide masses in Figure 4 increases with the size of the peptide profile matrices (Figure 2). Only a small number of masses (1,802) overlap between all software packages. This number represents approximately 32% of the average total number of 5,618 possible matches with the grid for each software package. If the number of masses present in three applications reflects real masses, the same number represents the number of missing masses, since these masses should be detected by the four software packages. In total 1,561 distinct missing masses MH^+ are measured between 1,600 and 2,400 Da; 759+124+258+420 (Figure 4). The ratio between detected and not-detected MH^+ masses for each software application, irrespective of their accuracy, can be estimated at 1,561:1,802 ≈ 1:1.

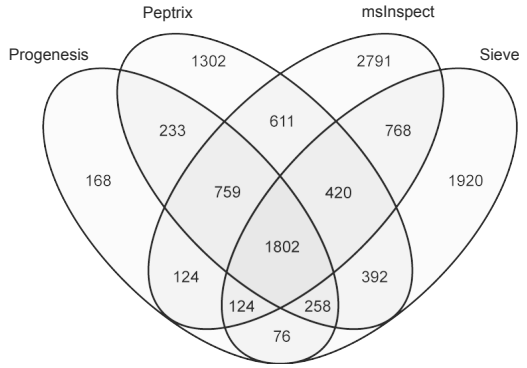


Figure 4. A 4-way Venn diagram representing the numbers of peptide masses in each profile matrix for Peptrix, Sieve™, Progenesis™ and msInspect, which match between 1, 2, 3 or 4 software packages. A comparison matrix was constructed by Peptrix using an artificial grid of 20275 masses between 1600 and 2400 Da with distances of 20 ppm between the masses. Peak masses from the individual software package peptide profile matrices were matched with the grids for the comparison matrix, using a mass window with 10 ppm in two directions and a total distance of 20 ppm.

Overlap of MS2 triggered and sequenced masses

A sub-selection from the comparison matrix in Table 2 was taken for MS2 triggered masses where sequencing succeeded and proteins were identified, using the MGF files from the 3 technical replicates. Figure 5 shows the pair-wise overlap of MS2 triggered, sequenced, and identified masses between two software packages in descending order of overlap. The most overlap (96%) is measured between Peptrix and Progenesis™, with the least overlap (78%) between Sieve™ and msInspect. We find just 260 MS2 precursors identified in a 3 h gradient between 1,600 and 2,400 Da. One major reason for this relatively low number is that we are working with an IgG Fab fragment sample yielding a lower number of identifications, presumably because quite a proportion of the peptides have unknown sequences not present in the protein database, which means they are MS2 triggered and sequenced, but protein identification did not succeeded.

Figure 6 shows the overlap of MS2 triggered and MS2 sequenced and identified peptide masses between all software packages, presented in a 4-way Venn diagram.

When comparing identified MS2 precursors from MS2 spectra with a Mascot™ score > 25, the overlap between all software packages is relatively high with approximately 76% (197/260) of the total number of sequenced and identified masses.

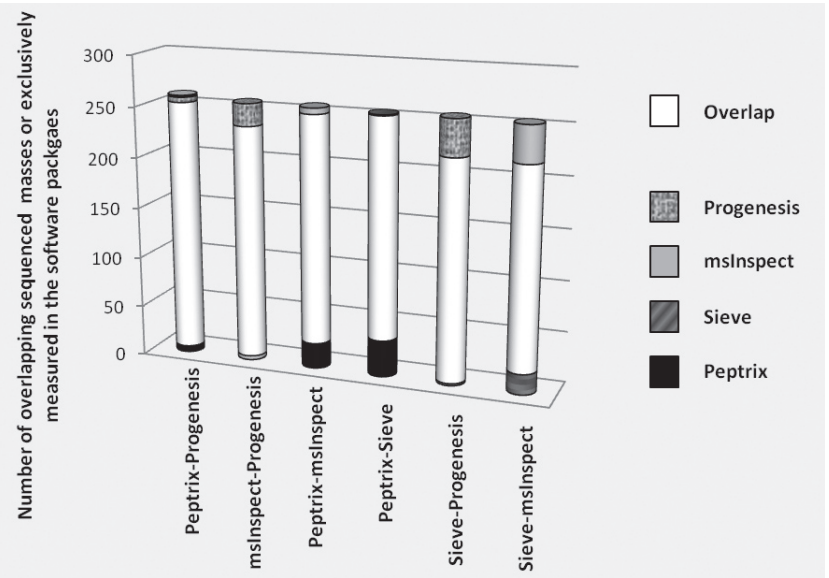


Figure 5. A pair-wise comparison of peptide masses overlapping with identified protein GI's, each time between the peptide profile matrices from two MS1 software packages All masses are sequenced and identified, including the non-overlapping masses in the individual packages. A comparison matrix was constructed by Peptrix using an artificial grid of 20275 masses between 1600 and 2400 Da with distances of 20 ppm between the masses. Peak masses from two peptide profile matrices identified with protein GI's for the software packages Peptrix, Sieve™, Progenesis™ and msInspect were matched with the comparison matrix grid, using a mass window with 10 ppm in two directions and a total distance of 20 ppm.

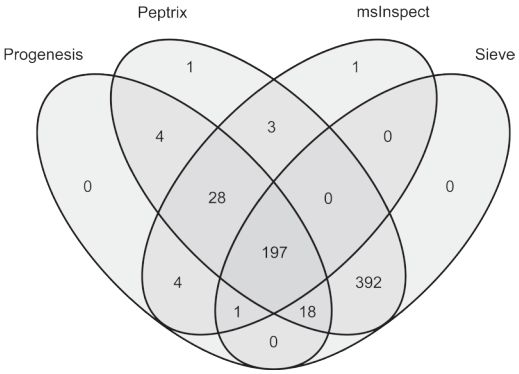


Figure 6. A 4-way Venn diagram representing the numbers of peptide masses identified with protein GI's in each matrix for Peptrix, Sieve™, Progenesis™ and msInspect, which match between 1, 2, 3, or 4 software packages. A comparison matrix was constructed by Peptrix using an artificial grid of 20275 masses between 1600 and 2400 Da and distances of 20 ppm between the masses. Peak masses identified with protein GI's from the individual peptide profile matrices for the software packages were matched with the comparison matrix grid, using a mass window with 10 ppm in two directions and a total distance of 20 ppm.

Differences in peak intensities

Table 4 shows peptide masses MH^+ for the in-silico digested highly abundant, most frequently sequenced protein Ig kappa chain C region, GI 157838230, found in any of the investigated software packages. The corresponding GI protein number is presented in the Mascot™ Daemon columns of Table 4. It appears that the non-sequenced peptide mass 2553.22833 Da is found exclusively by Sieve™, and is absent from the matrices of the other packages. Conversely, the low intensity MS1 masses 888.49378 and 2069.04844 are found by all software applications in one or more replicates, but are never triggered for MS2.

The peak intensities do not vary much between different replicate measurements in one application, but vary greatly between the investigated applications as shown in Table 4. An exception to this is msInspect. Low intensity peaks are not always detected in triplicate by msInspect as was already visible in the histogram in Figure 2, e.g. for mass 2069.04844 Da, which has only one intensity of 30 arbitrary units measured in replicate number 1.

On average, the peak intensity increases in the order: msInspect, Sieve™, Peptrix, Progenesis™. This ranking is not consistent over all peptide masses, however. For example, the high intensity of mass 1502.75844 Da ranks in the order: msInspect, Sieve™, Progenesis™, Peptrix; and the lower intensity of the mass 1740.87377 Da ranks in the order: msInspect, Peptrix, Progenesis™, Sieve™. For the lower peak intensities, with presumably low signal to noise ratios Sieve™ measures a relative high intensity, while for the relative high peak intensities with presumably high signal to noise ratios Progenesis™ and Peptrix measure relative high peak intensities.

Discussion

Peak picking from individual spectra before generation of the profile matrix as implemented in Peptrix has the advantage of parallel processing and scalability to a large number of spectra and distributed computer power. Peptrix, Sieve™, and msInspect run on average computer systems. Comparison of spectral intensities over all samples at once requires a lot of RAM for Progenesis™, which may be a disadvantage with an increasing size of datasets. We determined the reproducibility of the measurements by triplicate LC runs for one sample. We are aware that the number of samples used (3 replicates) does not really reflect an experimental setup for regular proteomics studies, but allows in-depth analysis how these software packages perform technically. The biological replicates used usually far exceed the numbers presented. Therefore we present as a practical example of Peptide profiling by Peptrix, the analysis results of Orbitrap™ measurements of in total 40 micro-dissected tissue samples. Peptrix can analyze the 40 Orbitrap™ raw files of the micro-dissected tissue samples, each of approximately 500 MB in 53 hours, 1 hour and 20 minutes for each file, using a 2.67 GHz Intel Xeon W3520 Quad-Core processor and 3.5 GB of RAM, a relatively low Java memory heap size (XMX) with settings of 1024 Mega Byte (MB).

In particular, the software packages that compare the spectral patterns over all samples directly, such as Progenesis™ and Sieve™, produce very reproducible peak lists with a low CV of intensity as was demonstrated in the histogram of the triplicate measurements for the IgG Fab. A single replicate measurement of a sample in large sample datasets should be sufficient in peptide profiling studies.

When matrices are prepared from peak lists, it is important to also store the masses of rejected peaks into “noise” lists (grey boxes in Figure 1a) to improve reproducibility of the measurements (Figure 2), since low intensity peaks can either just fit or not fit the selection criteria. These additional noise-lists can be used when preparing the peptide profile matrix. First, a peak mass in the list from one sample is matched with the peak mass in the list from another sample. If this mass is not present in the peak list from the other sample, it is searched for in the noise-list from the other sample. An FT-ICR MS example of such an approach was presented in our previous paper [6], and we have extended this approach for LC-MS.

High intensity peptide mass peaks in the MS1 spectra result most frequently in better MS2 fragmentation spectra and lead to more identified proteins after searching for peptide sequences in the protein databases. As expected, the peak picking of the high intensity peaks is more effective since the overlap between the software packages for these sequenced mass lists is higher as been demonstrated in Figures 5 and 6 than for all peptide masses as been demonstrated in Figures 3 and 4. The peptide masses in Figures 3 and 4 include low abundant peptide masses digested from low abundant proteins. It shows that all packages are capable of detecting peptides of high abundant proteins in a reliable way, but that they differ in detection of low concentration peptides. The average overlap for low abundant peptides is ½ / (½ + ½ + ½) ≈ 32% (Figure 3) as approximately ½ of the peaks are not found. Peak finding might be especially difficult for low intensity overlapping isotopic clusters in the mass spectra.

MS2 sequencing and protein identification requires accuracy of the mono-isotopic mass. The applications that perform isotopic pattern recognition, such as Peptrix, msInspect, and Progenesis™, show the largest overlap in Figure 5. The software application msInspect has the largest number of non-overlapping masses. In compar-

ing numbers of peptide features detected alone (Figure 2), one could conclude that Sieve™ works the best, followed by Peptrix and Progenesis™. Sieve™ does not perform isotopic pattern recognition. This has the disadvantage that isotopes of a peptide mass can be wrongly assigned as the mono-isotopic mass. This may explain why Sieve™ measures a relative high intensity for the lower peak intensities, with presumably low signal to noise ratios (Table 4). Some features in the Sieve™ profile matrix also display non integer values for the charge state, because the real MH^+ value is for one reason or another difficult to calculate (for example overlapping peaks). Sieve™ presumably combines different peptides with different charge states in a single frame (Figure 1b).

False Discovery Rates (FDRs) could be calculated, by comparing the peptide masses in the profile matrices with those from hand-picked peaks in a single MS1 scan, for example scanning 9919 at a retention time of 113.48 minutes. However, in this single scan, hundreds and perhaps even thousands of low intensity peptide features can be detected, making a manual FDR calculation impossible. This indicates that the software packages have likely a high false negative rate (missing peptides).

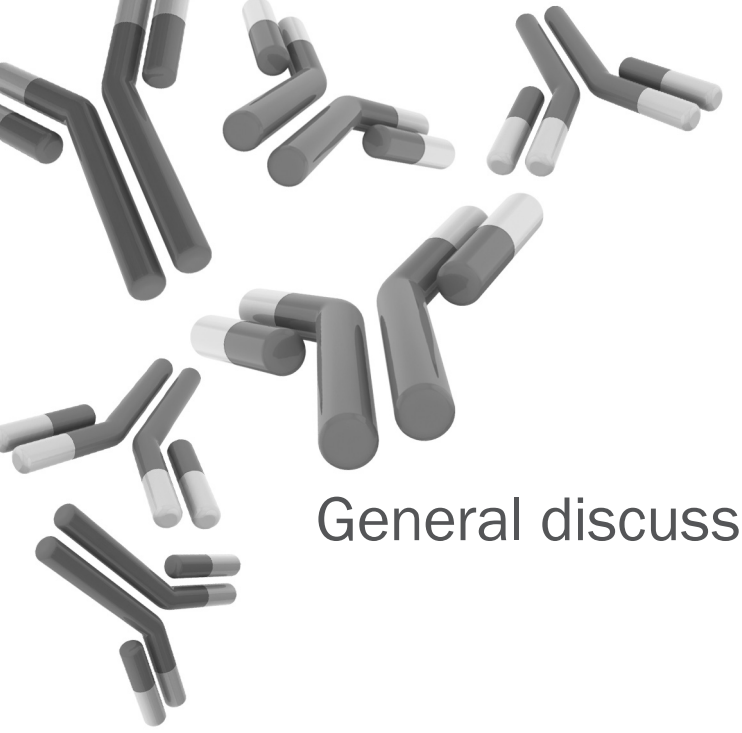
When comparing the peak lists for in-silico digested peptide masses from IgG Fab, it appears that all software packages are capable of extracting almost 100% of the peptide masses, however, with different intensities, and by contrast for msInspect not always in a reproducible way in replicates. Apparently, peak intensities are established by the MS1 software packages investigated in a different way. Peptrix determines the highest intensity of a peptide MH^+ mono-isotope mass in a LC elution profile. The intensities of double and triple charged peptides are combined. Sieve™ takes the integral of intensity under the elution curve of a frame (Figure 1b). The software package msInspect determines the highest intensity of an isotopic mass in an isotopic cluster as a function of LC time. This is not necessarily the mono-isotope. Progenesis™ calculates the integral of intensity in two dimensions, in direction of mass and retention time in a 2-D gel view (Figure 1).

Acknowledgements

The authors thank David Alexander and John Shippey for the critical review of the manuscript. Funding: This study was financially supported by the Virgo Consortium; the Biorange Research program linked with the Netherlands Proteomics Centre (NPC); Top Institute Pharma (TI Pharma) Netherlands (project D4-102-1); and the EU P-mark project.

References

1. www.thermo.com.
2. Alves RD, Eijken M, Swagemakers S, Chiba H, Titulaer MK, Burgers PC, et al. Proteomic analysis of human osteoblastic cells: relevant proteins and functional categories for differentiation. J Proteome Res; 9:4688-700.
3. Dekker LJ, Burgers PC, Charif H, van Rijswijk AL, Titulaer MK, Jenster G, et al. Differential expression of protease activity in serum samples of prostate carcinoma patients with metastases. Proteomics; 10:2348-58.
4. Stoop MP, Singh V, Dekker LJ, Titulaer MK, Stingl C, Burgers PC, et al. Proteomics comparison of cerebrospinal fluid of relapsing remitting and primary progressive multiple sclerosis. PLoS One; 5:e12442.
5. Carvalho-Oliveira IM, Charro N, Aarbiou J, Buijs-Offerman RM, Wilke M, Schettgen T, et al. Proteomic analysis of naphthalene-induced airway epithelial injury and repair in a cystic fibrosis mouse model. J Proteome Res 2009; 8:3606-16.
6. Titulaer MK, Mustafa DA, Siccama I, Konijnenburg M, Burgers PC, Andeweg AC, et al. A software application for comparing large numbers of high resolution MALDI-FTICR MS spectra demonstrated by searching candidate biomarkers for glioma blood vessel formation. BMC Bioinformatics 2008; 9:133.
7. Titulaer MK, Siccama I, Dekker LJ, van Rijswijk AL, Heeren RM, Sillevis Smitt PA, et al. A database application for pre-processing, storage and comparison of mass spectra derived from patients and controls. BMC Bioinformatics 2006; 7:403.
8. Lange E, Tautenhahn R, Neumann S, Gropl C. Critical assessment of alignment procedures for LC-MS proteomics and metabolomics measurements. BMC Bioinformatics 2008; 9:375.
9. Mueller LN, Brusniak MY, Mani DR, Aebersold R. An assessment of software solutions for the analysis of mass spectrometry based quantitative proteomics data. J Proteome Res 2008; 7:51-61.
10. http://omics.pnl.gov/software/.
11. http://sourceforge.net/projects/sashimi/files/.
12. http://proteomics.fhcrc.org/.
13. Bellew M, Coram M, Fitzgibbon M, Igra M, Randolph T, Wang P, et al. A suite of algorithms for the comprehensive analysis of complex protein mixtures using high-resolution LC-MS. Bioinformatics 2006; 22:1902-9.
14. www.nonlinear.com.
15. Mustafa DA, Burgers PC, Dekker LJ, Charif H, Titulaer MK, Smitt PA, et al. Identification of glioma neovascularization-related proteins by using MALDI-FTMS and nano-LC fractionation to microdissected tumor vessels. Mol Cell Proteomics 2007; 6:1147-57.
16. de Costa D, Broodman I, Vanduijn MM, Stingl C, Dekker LJ, Burgers PC, et al. Sequencing and quantifying IgG fragments and antigen-binding regions by mass spectrometry. J Proteome Res 2010; 9:2937-45.
17. Mueller LN, Rinner O, Schmidt A, Letarte S, Bodenmiller B, Brusniak MY, et al. SuperHirn - a novel tool for high resolution LC-MS-based peptide/protein profiling. Proteomics 2007; 7:3470-80.
18. www.matrixscience.com.



Chapter 6

General discussion & summary

General discussion & summary

As lung cancer is most of the time detected at an advanced stage it is one of the cancers with the highest deaths among cancer. The risk of dying from lung cancer is associated with cigarette smoking. Populations of smokers have a higher risk to develop lung cancer. Therefore, screening of high risk populations is of great importance. Low-dose spiral CT scan is used for lung cancer screening. Although, CT screening is not yet recommended unless it is used as part of a clinical trial, because of the drawback of high rate of false-positive scan results. Therefore, biomarkers may be useful complementary to CT screening in a screening strategy or most ideally as an independent biomarker. Autoantibody profiling could be a powerful tool for early detection of lung cancer when incorporated into a screening strategy. In Chapter 2 we described our approach for identification and quantification of IgG Fab and CDRs by mass spectrometry. With this approach we tried to determine the possibility of using immunoglobulins with specific molecular signatures for diagnostic use in different cancers and autoimmune diseases. We were able to demonstrate that our approach is reproducible with a high recovery, fast and provides sequencing information by mass spectrometry. Immunoglobulins, or antibodies, are complex molecules with large variation. The possible diversity in immunoglobulins is estimated between 10^{13} and 10^{50}.[1-2] Therefore, sequence variation is extremely high and one could assume that finding similar or even identical sequences in different individuals is highly unlikely. However, different studies including ourselves have shown the opposite.[3-7] We described in Chapter 2 a similarity of 83%, based on MS signals, between seven healthy individuals by mass spectrometry. The group of Weinstein performed a sequencing project in zebra fish and found identical CDR3 sequences in different zebra fishes.[8] For PNS patients IgG autoantibodies against onconeural antigens such as HuD, Yo, amphiphysin and CV2, are occurring in serum. Maat et al. showed that primary amino acid structures of the antibodies that were specific for these antigens were shared between different PNS patients.[9] Why these studies were able to identify similar or identical CDR sequences could be explained by repertoire bias. During immune response antibodies could be subjected to some kind of selection after rearrangement and affinity maturation.[10-13] VanDuijn et al. showed with their recombinant HuD protein immunized rat study that the development of immunoglobulins is not a random process but selection do occur during immune response and that this selection is shared between different rats.[6]
In Chapter 2 we also showed that we were able to purify and sequence many CDR sequences by mass spectrometry. By using different databases we were able to identify even significantly more CDR sequences compared to Obermeier and colleagues.[14] Besides the use of different databases, it might be possible to improve the method by molecular dissection. This reduces the complexity of the immunoglobulin molecule which may lead to an increase of identifying CDR sequences and therefore more disease - related CDR sequences could possibly be detected. In Chapter 3 we described a novel method to reduce the complexity of the immunoglobulin by molecular dissect IgG into κ and λ fragments. The method developed was reproducible with a high recovery of IgG-κ and IgG-λ for IgG and Fab-κ and Fab-λ for Fab. We observed a higher yield of identified CDRs in the Fab fraction than in the IgG fractions. This can be explained by the fact that IgG κ and λ are missing CDRs of the heavy chain. This was also supported by this study as we found a four times higher CDR ratio (CDR1:CDR2:CDR3) in Fab, Fab-κ and Fab-λ compared to IgG-κ and IgG-λ. We showed a twofold higher yield of identified CDRs when Fab-κ, Fab-λ, κ and λ were combined and compared

to the yield of Fab. Fab-κ consisted of more CDRs than Fab-λ and more significantly different CDRs were observed in Fab-κ compared to the other three fractions. This can be explained by the fact that during B-cell differentiation the heavy chain genes are first rearranged. After rearrangement of the heavy chain genes the κ chain genes are rearranged followed by the λ chain genes.[1] In literature, the antibodies that are described to be expressed during cancer development are the heavy chains and κ chains.[15] Therefore, we advised in this chapter to use Fab-κ next to Fab. Applying two purifications and two different mass spectrometry measurements results in twice as much measurement time, but the effort is worthwhile as one receives an addition of 50% more identified CDRs. These results may imply an increase of the likelihood of finding (lung-) cancer related CDR sequences.

In chapte4 we showed a panel of 12 different antibody peptides that were able to distinguish lung cancer patients from controls in a high risk population by the approach described in Chapter 2. This antibody peptide model consisted not only out of peptide sequences which originated purely from the CDR regions of an immunoglobulin but also from the framework regions. Why we did not obtain an antibody peptide model consisting of only CDR region peptides but also peptides derived from framework regions of an immunoglobulin could be explained by biological issues such as abundance in the immunoglobulin pool. It is more likely that peptides with only few mutations compared to the germline, occur in several antibody clones and therefore having a higher abundance. This results in more chance of being detected by a mass spectrometer.

Moreover, the samples we used in this study might be too complex for a mass spectrometer. Therefore, purifying CDR fragments would be ideally, but unfortunately sample preparation procedures do not exist yet.

Besides the high variability of an antibody, lung cancer is a heterogeneous disease which results in high variability between patients and may induce immune responses to various tumor antigens.[16-22] Therefore, it is not surprising that we were not able to find only one antibody peptide that could distinguish lung cancer patients from controls. A study performed by Brichory et al. showed low sensitivities for single TAAs.[23-24] Instead of using a single TAA, Khattar et al. and Zhong et al. tested a panel of TAAs and validated the panel. They showed sensitivities ranging from 84%-91% and specificities from 73%-91%.[25-26] We observed in the NELSON 1 (discovery) and NELSON 2 (validation) set a sensitivity of 96% and 84% and specificities of 100% and 90%, respectively. Due to technical problems we had to recalibrate our model for each patient group. The background evaluation showed that our antibody peptide model performed significantly better than a model generated based on just permutated data.

Until now only age and smoking history have been used as selection criteria for enrolment in screening trials. Additional diagnostic test might select high risk individuals more precise when combined with age and smoking history. CT screening has demonstrated its ability to detect lung cancer with high sensitivities and specificities at different screening rounds (baseline and one year later).[27] However, in the NELSON trial 27% of the participants are subjected to invasive and expensive follow-up studies that revealed in benign disease at baseline CT screening. The performance of CT improves after follow-up scans, but only after a long period of time, on average a year. Therefore, additional diagnostic tests are needed. Massion et al. showed their results on a combination of a serum proteomic biomarker panel with clinical and CT data.[28] In Chapter 4 we were able to detect lung cancer with an antibody peptide model at an early stage. Our results indicate that specific antibodies are able to detect lung can-

cer at an earlier stage than CT screening. Auto-antibody profiling has the potential to be a powerful additional test for early detection of lung cancer in a screening strategy. But still technical challenges have to be improved before this method is applicable in the clinical practice.
Besides correct sample handling and preparation and mass spectrometry measurement, proper data analysis is very important. In Chapter 5 we discuss the software package developed by our group and compared it with three other open source and commercial available software packages. Peak picking is a very crucial step in data analysis. The reproducibility of finding peptide masses in different replicates should be high and correct. For example we showed that 86% of all peptides masses were observed by Peptrix in all the three replicates of an IgG Fab sample. In contrast, msInspect, one of the compared software packages, had less overlap in peptide masses, 76% of the total number of masses was observed in only one replicate. This is related to the fact that this software package only matches a mass in a second or third replicate if it was detected in the first replicate. Therefore, a mass observed in the second and third replicate but not in the first replicate will not be observed in the final peptide profile matrix, the matrix that contains all the results of the detected masses from all samples. This can lead to loss of interesting peptides for biomarker discovery.
A solution could be storage of rejected peaks in to a list, as is included in the Peptrix software. In this way a peak that was detected as described above in the second and third sample could be included in the peptide matrix when this peak was stored on the list of the first sample. This will increase the reproducibility of the peak lists between different samples.
The four software packages differ in detection of low abundant peptides. One has more difficulties in detecting them than the other. This results in large numbers of non-matching peptides that do not overlap between the four software packages. The more MS1 spectra the higher the number of non-matching peptides was observed. For example, Progenesis had less MS1 spectra compared to msInspect and had less non-matching peptides than msInspect. The overlap of peptides between the four software packages was therefore only 32%.
Also, peak intensity is an important feature during data analysis. The intensities did not differ between replicates for each software package, but the intensity differed between the software packages. This may be related to the different way of establishing peak intensities by the software package.
In conclusion, using more software packages can increase the number of detected peptide masses and give understanding of how, especially commercial software packages work.

Future research

Our Fab purification combined Orbitrap mass spectrometry approach is well suited for the discovery of an antibody panel for lung cancer. It reaches high levels of accuracy, resolution and sensitivity. A limitation of this method is that it is not a high-throughput approach. Therefore, we are interested to validate our panel with techniques like MRM (multiple reaction monitoring) or SRM (selective reaction monitoring).
Validation of peptides or proteins that are of interest for a specific disease is still a critical point. Antibodies or ELSIA kits are not available for the majority of interesting peptides or proteins. Techniques such as MRM and SRM are relatively new proteomic techniques to quantify proteins at the ng/ml level and are interesting alternatives to measure panels of peptides or proteins. SRM can be used for monitoring and quan-

tification of roughly 10 to 50 specific peptides within complex mixtures.[29-30] Besides applying this technique for validation it is also suitable for biomarker discovery.[29, 31] The development of SRM might result in using proteomics as a biomarker discovery tool but also as a diagnostic method itself.

With the studies included in this thesis we have shown that one can obtain an increase of CDR sequences by molecular dissection of IgG. Besides this, molecular dissection improvements in sequence coverage may also be an option. Improvement of the resolution by using ultra high-pressure chromatography techniques, better sequence identification and depletion of constant regions of immunoglobulins can increase the enrichment of specific CDRs that are interesting for lung cancer detection. Secondly, increasing the spectra quality and mass accuracy, identification of *de novo* sequencing peptides can be improved. A combination of CID (collision-induced dissociation) and HCD (higher energy collision-induced dissociation) spectra will improve the identification of *de novo* peptides. With HCD low mass regions can be measured because of the additional fragments that can be obtained. Furthermore, ETD (electron-transfer dissociation) can also increase the identification of *de novo* peptides. By ETD fragmentation, longer peptides can be detected and compared to CID to gain more sequence information.[32-34]
In conclusion, with the research performed for this thesis we have accomplished a proof of concept that shows that specific antibody related peptides exist which shows that lung cancer patients and controls can be discriminated. This research work needs further evaluation with relative high-throughput techniques such as immunoassay or SRM to determine the value for clinical use.

References

1. Murphy K. TP, Walport M. Janeway's immunobiology. 7th ed: Garland Science; 2008.
2. Saada R, Weinberger M, Shahaf G, Mehr R. Models for antigen receptor gene rearrangement: CDR3 length. Immunol Cell Biol 2007; 85:323-32.
3. Foreman AL, Lemercier B, Lim A, Kourlisky P, Kenny T, Gershwin ME, et al. VH gene usage and CDR3 analysis of B cell receptor in the peripheral blood of patients with PBC. Autoimmunity 2008; 41:80-6.
4. Poulsen TR, Meijer PJ, Jensen A, Nielsen LS, Andersen PS. Kinetic, affinity, and diversity limits of human polyclonal antibody responses against tetanus toxoid. J Immunol 2007; 179:3841-50.
5. Scheid JF, Mouquet H, Ueberheide B, Diskin R, Klein F, Oliveira TY, et al. Sequence and structural convergence of broad and potent HIV antibodies that mimic CD4 binding. Science 2011; 333:1633-7.
6. VanDuijn MM, Dekker LJ, Zeneyedpour L, Smitt PA, Luider TM. Immune responses are characterized by specific shared immunoglobulin peptides that can be detected by proteomic techniques. J Biol Chem 2010; 285:29247-53.
7. Foreman AL, Van de Water J, Gougeon ML, Gershwin ME. B cells in autoimmune diseases: insights from analyses of immunoglobulin variable (Ig V) gene usage. Autoimmun Rev 2007; 6:387-401.
8. Weinstein JA, Jiang N, White RA, 3rd, Fisher DS, Quake SR. High-throughput sequencing of the zebrafish antibody repertoire. Science 2009; 324:807-10.
9. Maat P, VanDuijn M, Brouwer E, Dekker L, Zeneyedpour L, Luider T, et al. Mass spectrometric detection of antigen-specific immunoglobulin peptides in paraneoplastic patient sera. J Autoimmun 2012; 38:354-60.
10. Baranzini SE, Jeong MC, Butunoi C, Murray RS, Bernard CC, Oksenberg JR. B cell repertoire diversity and clonal expansion in multiple sclerosis brain lesions. J Immunol 1999; 163:5133-44.
11. Loh DY, Bothwell AL, White-Scharf ME, Imanishi-Kari T, Baltimore D. Molecular basis of a mouse strain-specific anti-hapten response. Cell 1983; 33:85-93.
12. Andersen PS, Haahr-Hansen M, Coljee VW, Hinnerfeldt FR, Varming K, Bregenholt S, et al. Extensive restrictions in the VH sequence usage of the human antibody response against the Rhesus D antigen. Mol Immunol 2007; 44:412-22.
13. Sehgal D, Schiaffella E, Anderson AO, Mage RG. Analyses of single B cells by polymerase chain reaction reveal rearranged VH with germline sequences in spleens of immunized adult rabbits: implications for B cell repertoire maintenance and renewal. J Immunol 1998; 161:5347-56.
14. Obermeier B, Mentele R, Malotka J, Kellermann J, Kumpfel T, Wekerle H, et al. Matching of oligoclonal immunoglobulin transcriptomes and proteomes of cerebrospinal fluid in multiple sclerosis. Nat Med 2008; 14:688-93.
15. Chen Z, Qiu X, Gu J. Immunoglobulin expression in non-lymphoid lineage and neoplastic cells. Am J Pathol 2009; 174:1139-48.
16. Anderson KS, LaBaer J. The sentinel within: exploiting the immune system for cancer biomarkers. J Proteome Res 2005; 4:1123-33.
17. Baldwin RW. Immunity to transplanted tumour: the effect of tumour extracts on the growth of homologous tumours in rats. Br J Cancer 1955; 9:646-51.
18. Caron M, Choquet-Kastylevsky G, Joubert-Caron R. Cancer immunomics using autoantibody signatures for biomarker discovery. Mol Cell Proteomics 2007; 6:1115-22.
19. Gure AO, Altorki NK, Stockert E, Scanlan MJ, Old LJ, Chen YT. Human lung cancer antigens recognized by autologous antibodies: definition of a novel cDNA derived from the tumor suppressor gene locus on chromosome 3p21.3. Cancer Res 1998; 58:1034-41.
20. Hanash S. Harnessing immunity for cancer marker discovery. Nat Biotechnol 2003; 21:37-8.

21. Mintz PJ, Kim J, Do KA, Wang X, Zinner RG, Cristofanilli M, et al. Fingerprinting the circulating repertoire of antibodies from cancer patients. Nat Biotechnol 2003; 21:57-63.
22. 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.
23. Brichory F, Beer D, Le Naour F, Giordano T, Hanash S. Proteomics-based identification of protein gene product 9.5 as a tumor antigen that induces a humoral immune response in lung cancer. Cancer Res 2001; 61:7908-12.
24. Brichory FM, Misek DE, Yim AM, Krause MC, Giordano TJ, Beer DG, et al. An immune response manifested by the common occurrence of annexins I and II autoantibodies and high circulating levels of IL-6 in lung cancer. Proc Natl Acad Sci U S A 2001; 98:9824-9.
25. Khattar NH, Coe-Atkinson SP, Stromberg AJ, Jett JR, Hirschowitz EA. Lung cancer-associated auto-antibodies measured using seven amino acid peptides in a diagnostic blood test for lung cancer. Cancer Biol Ther 2010; 10:267-72.
26. 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.
27. van Klaveren RJ, Oudkerk M, Prokop M, Scholten ET, Nackaerts K, Vernhout R, et al. Management of lung nodules detected by volume CT scanning. N Engl J Med 2009; 361:2221-9.
28. Pecot CV, Li M, Zhang XJ, Rajanbabu R, Calitri C, Bungum A, et al. Added value of a serum proteomic signature in the diagnostic evaluation of lung nodules. Cancer Epidemiol Biomarkers Prev 2012; 21:786-92.
29. Anderson L, Hunter CL. Quantitative mass spectrometric multiple reaction monitoring assays for major plasma proteins. Mol Cell Proteomics 2006; 5:573-88.
30. Kuhn E, Wu J, Karl J, Liao H, Zolg W, Guild B. Quantification of C-reactive protein in the serum of patients with rheumatoid arthritis using multiple reaction monitoring mass spectrometry and 13C-labeled peptide standards. Proteomics 2004; 4:1175-86.
31. Luna LG, Williams TL, Pirkle JL, Barr JR. Ultra performance liquid chromatography isotope dilution tandem mass spectrometry for the absolute quantification of proteins and peptides. Anal Chem 2008; 80:2688-93.
32. Dekker LJ, Zeneyedpour L, Brouwer E, van Duijn MM, Sillevis Smitt PA, Luider TM. An antibody-based biomarker discovery method by mass spectrometry sequencing of complementarity determining regions. Anal Bioanal Chem 2011; 399:1081-91.
33. Shen Y, Tolic N, Purvine SO, Smith RD. Improving collision induced dissociation (CID), high energy collision dissociation (HCD), and electron transfer dissociation (ETD) fourier transform MS/MS degradome-peptidome identifications using high accuracy mass information. J Proteome Res 2012; 11:668-77.
34. Shen Y, Tolic N, Xie F, Zhao R, Purvine SO, Schepmoes AA, et al. Effectiveness of CID, HCD, and ETD with FT MS/MS for degradomic-peptidomic analysis: comparison of peptide identification methods. J Proteome Res 2011; 10:3929-43.

Appendices

Nederlandse samenvatting

Dutch Summary

Longkanker is op dit moment één van de meest voorkomende vormen van kanker met het hoogste sterftegetal (28%) in de wereld. Het risico om longkanker te ontwikkelen is geassocieerd met het roken van sigaretten. Tachtig tot negentig procent van alle longkankerpatiënten zijn te wijten aan roken. Slechts 15% van deze patiënten wordt daadwerkelijk in een vroeg stadium gediagnosticeerd en kunnen geopereerd worden. Het is daarom van belang om longkanker in een vroeg stadium te kunnen detecteren. Het is bewezen dat CT screening longkanker in een vroeg stadium kan opsporen, maar helaas wordt er nog een percentage van ongeveer 27% van de patiënten foutief gediagnosticeerd voor longkanker wat onnodige stress bij deze patiënten geeft en/of dat er zelfs onnodige operaties uitgevoerd worden wat extra kosten met zich mee brengt. Hoog risico screening op longkanker is daarom van groot belang. Op dit moment zijn alleen leeftijd en rookgedrag selectiecriteria voor het opnemen van individuen in een screening trial. Naast het includeren van individuen waarbij bekend is dat er longziekten in de familie voorkomt zou een biomarker ook hoog risicogroepen voor longkanker kunnen detecteren. Er zijn al een aantal studies uitgevoerd waarbij gezocht is naar biomarkers voor longkanker, maar ondanks dat een aantal van deze studies goede sensitiviteit en specificiteit opbrengen zijn deze biomarkers niet gevalideerd of waren ze niet reproduceerbaar. Dit proefschrift laat een andere aanpak zien dan die in andere studies gebruikt zijn.

Zoals gezegd zijn er al een aantal studies die getracht hebben een biomarker of biomarkerpanel te vinden voor longkanker. Deze studies zijn voornamelijk gefocusseerd op het vinden van een specifiek antigeen, waarbij men op voorhand al weet naar welk antigeen men op zoek gaat. Verschillende studies hebben aangetoond dat er een humoraal immuunrespons optreedt in longkanker maar ook in andere typen kanker. Tumoren zouden ervoor zorgen dat er veel TAAs (tumor geassocieerde antigenen) worden vrijgelaten in het bloed, waardoor er mogelijk autoantilichamen worden geproduceerd tegen deze antigenen. In hoofdstuk 2 laten wij een methode zien die mogelijk toegepast zou kunnen worden op het vinden van een biomarker voor longkanker of andere typen kanker of auto-imuunziekten. Bij deze methode hoeft men niet op voorhand te weten naar welk antigeen gezocht gaat worden. Deze methode is gebaseerd op het vinden van antilichamen die mogelijk een rol bij kanker of auto-imuunziekten spelen. Doordat men niet gericht op zoek gaat naar één of meerdere antigenen vergroot men de kans op het vinden van een marker die ook na validatie zowel een goede sensitiviteit als specificiteit zal geven. De methode die ontwikkeld is, is gebaseerd op het sequencen en kwantificeren van IgG Fab en CDRs (complementarity determining regions) met behulp van massaspectrometrie. Antilichamen, ook wel immuunglobulines genoemd, bestaan uit constante gedeelten en variabele gedeelten. Een IgG Fab is een fragment van een immuunglobuline type G waarin het variabele gebied van het molecuul zich bevindt in één derde van het constante gebied. Binnen het variabele gebied bevinden zich drie verschillende CDRs, CDR1, CDR2 en CDR3. Deze CDRs zijn hypervariabel en zorgen voor de specificiteit en binding van het antigeen aan het antilichaam. De ontwikkelde methode bestaat uit een opzuiveringsstap van IgG Fab waarna deze fragmenten worden geanalyseerd op een massaspectrometer. De zuiveringsstap gaf ons reproduceerbare resultaten. Daarnaast is het een snelle methode die maar één sample geeft per individu om gemeten te worden op een massaspectrometer. Immuunglobulines zijn complexe

moleculen met een grote variatie. Theoretisch ligt de diversiteit van een immuunglobuline tussen de 10^{13} en 10^{50} mogelijkheden. Daarom zou het erg moeilijk kunnen zijn om vergelijkbare of zelfs identieke aminozuur sequenties te kunnen vinden met de massaspectrometer. Wij hebben kunnen aantonen in hoofdstuk 2 dat 83% van de resultaten die wij gevonden hebben met de massaspectrometer overeenkomen tussen zeven gezonde individuen.
Omdat het immuunglobuline zeer variabel is hebben we in hoofdstuk 3 geprobeerd om de complexiteit van het immuunglobuline te reduceren zodat de variatie van het immuunglobuline lager zou kunnen worden gebracht voor de massaspectrometer. Hierdoor zouden we de kans op het vinden van CDR sequenties die longkanker gerelateerd zijn kunnen vergroten ten opzichte van de methode die in hoofdstuk 2 is beschreven. Deze methode verlaagd de complexiteit van het immuunglobuline door uit IgG en Fab de kappa (κ) en lambda (λ) fragmenten op te zuiveren. Ook deze methode liet reproduceerbare resultaten zien met goede zuiveringsopbrengsten voor vier verschillende fragmenten, IgG-κ en IgG-λ gezuiverd uit IgG en Fab-κ en Fab-λ gezuiverd uit Fab. Om uit te zoeken of deze methode een hogere opbrengst van CDR's kan genereren hebben we de opbrengsten van CDR's vergeleken tussen de twee methoden. Het aantal CDR's dat in de Fab-κ, Fab-λ, IgG-κ en IgG-λ werd gevonden was twee keer zoveel als dat er aan CDR's gevonden werd bij de Fab's. Onze conclusie uit dit hoofdstuk was dat Fab-κ samen met Fab de meeste opbrengst aan CDR's opleverde. Dit konden wij concluderen, aangezien Fab-κ meer CDR sequenties opleverde dan Fab-λ en meer significante CDR's opleverde ten opzichte van de andere drie fragmenten. Deze methode kost weliswaar twee keer zoveel tijd ten opzichte van de Fab methode uit hoofdstuk 2, maar het levert wel 50% meer geïdentificeerde CDRs op wat de kans op het vinden van longkanker gerelateerde CDR sequenties verhoogt.
In hoofdstuk 4 laten wij een antilichaam-peptide model zien bestaande uit 12 verschillende peptiden. Dit model was in staat om longkanker patiënten van controles te onderscheiden in een hoog-risico populatie. Hierbij is gebruik gemaakt van de methode die beschreven is in hoofdstuk 2.
In de studie beschreven in hoofdstuk 4 hebben we zowel technische problemen als biologische problemen ondervonden. Zo zou het technisch zeer moeilijk kunnen zijn voor de massaspectrometer om vergelijkbare of identieke CDR peptide sequenties in elk individu te kunnen detecteren, omdat CDRs hypervariabele gebieden van een antilichaam zijn en er daarom een grote diversiteit aan mogelijkheden is. Dit zou een verklaring kunnen zijn voor het feit dat wij geen model hebben kunnen vinden dat alleen uit CDRs bestond, maar ook uit framework gebieden van het immuunglobuline. Daarnaast zouden CDR peptiden minder gemeenschappelijk kunnen voorkomen in longkanker patiënten door hun hoge mate van diversiteit. De CDR-peptiden zouden ook lager in concentratie aanwezig kunnen zijn ten opzichte van specifiek gemuteerde framework gebieden die op hun beurt effect hebben op de detectie van CDRs door de massaspectrometer.
Naast de hoge variabiliteit van een antilichaam, is longkanker een zeer heterogene ziekte. Dit resulteert in hoge variabiliteit tussen patiënten en zou een immuunresponse tegen verschillende tumor antigenen kunnen opwekken.
Daarom waren wij net als andere onderzoekers niet in staat om één enkel antilichaam-peptide te vinden dat longkanker patiënten van controles kon scheiden. In plaats van het zoeken naar één enkel antilichaam-peptide, zijn wij op zoek gegaan naar een model dat uit verschillende antilichamen-peptiden bestaat. Hiervoor hebben wij gebruik gemaakt van longkanker patiënten en controles (NELSON 1) uit de NELSON

trial. Wij waren in staat om een antilichaam-peptide model te vinden dat een sensitiviteit van 96% en een specificiteit van 100% gaf. Dit model hebben wij getest in een nieuwe set met longkanker patiënten en controles (NELSON 2) uit de NELSON trial en vonden een sensitiviteit van 84% en een specificiteit van 90%.
Doordat wij een aantal technische problemen hebben ondervonden hebben wij het model moeten recalibreren voor elke patiëntengroep. Om te controleren op kans op random selectie van data hebben wij een achtergrond evaluatie uitgevoerd. Deze evaluatie liet zien dat het antilichaam-peptide model dat wij gevonden hadden significant beter was dan een model dat bij de achtergrond evaluatie was gevonden door permutatie van de data.
Tot op heden zijn alleen leeftijd en rookhistorie selectiecriteria voor het opnemen van deelnemers in een screeningtrial. Het selecteren van hoog risico individuen zou meer nauwkeurig kunnen door additionele testen te combineren met leeftijd en rookhistorie.
Longkanker screeningtrials waarbij gebruik gemaakt wordt van CT-screening hebben laten zien dat longkanker aangetoond kan worden op baseline (beginpunt van een screeningtrial). Daarnaast is ook aangetoond dat circa 27% van de deelnemers aan een CT screeningtrial onnodige invasieve en dure follow-up studies moeten ondergaan die resulteren in niet-kwaadaardige longziekten. Follow-up CT-scans verlagen dit aantal, maar helaas duurt deze follow-up ongeveer een jaar. Daarom zijn additionele testen waardevol.
In hoofdstuk 4 laten wij zien dat wij longkanker in een vroeg stadium kunnen detecteren met een antilichaam-peptiden model. Onze resultaten wijzen in de richting dat specifieke antilichamen aanwezig zijn in een vroeg stadium van longkanker en dat ze mogelijk longkanker kunnen detecteren in een vroeger stadium dan CT-screening. Autoantilichamen hebben de potentie om een toegevoegde waarde te hebben als additionele test voor vroegtijdige detectie van longkanker in een screeningstrategie. Maar de technische problemen die wij hebben ondervonden moeten wel verholpen worden voordat deze methode toepasbaar is in de kliniek.
Juiste samplebehandeling en correcte massaspectrometrie-metingen zijn zeer belangrijke aspecten voor een goed resultaat. Daarnaast is juiste data-analyse zeer belangrijk. In hoofdstuk 5 van dit proefschrift wordt het softwarepakket Peptrix besproken dat door onze onderzoeksgroep ontwikkeld is om massaspectrometrie-data te analyseren. Dit softwarepakket is vergeleken met drie andere ‘open source’ en commercieel verkrijgbare softwarepakketten.
Verschillende aspecten zijn belangrijk tijdens de data-analyse van massaspectrometrie-data. Peak picking is een cruciaal punt in de data-analyse. Het vinden van de massa van peptiden moet zo reproduceerbaar mogelijk zijn tussen verschillende metingen van hetzelfde sample. Met het softwarepakket Peptrix konden wij 86% van alle peptiden massa’s terugvinden in drie metingen van één IgG Fab sample. In vergelijking tot de andere softwarepakketten was dit een zeer hoge reproduceerbaarheid.
Een tweede belangrijk punt is het kunnen detecteren van low abundant peptiden. Het ene softwarepakket heeft er meer moeite mee dan het andere. Dit veroorzaakt dat er grote hoeveelheden niet geïdentificeerde peptiden worden gevonden die niet teruggevonden worden in alle vier de softwarepakketten. Er blijkt een relatie te bestaan tussen het aantal MS1 spectra en het aantal niet geïdentificeerde peptiden. Het softwarepakket Progenesis detecteerde bijvoorbeeld minder MS1 spectra wat in minder niet geïdentificeerde peptiden resulteerde vergeleken met het softwarepakket MS Inspect. Daarom werd er een laag percentage van 32% gevonden van peptiden die in

alle vier de softwarepakketten gevonden konden worden.
Een derde belangrijk punt is piekintensiteit. Tussen de verschillende metingen van één IgG Fab sample was er geen verschil te meten in intensiteit, maar wanneer de metingen tussen de verschillende softwarepakketten vergeleken worden, wordt er wel een verschil in intensiteit opgemerkt. Dit komt doordat er mogelijk een verschil zit tussen de softwarepakketten in de manier waarop de piekintensiteit bepaald wordt.
De conclusie van dit proefschrift is dat specifieke antilichaam-gerelateerde peptiden bestaan die longkanker patiënten van controles kunnen onderscheiden in een hoog-risicopopulatie. Verder onderzoek moet gedaan worden naar de methode die hiervoor is gebruikt om de toepasbaarheid in de kliniek te kunnen realiseren. Hierbij moet men denken aan high-throughput-technieken zoals immunoassays of SRM (Selective reaction monitoring).

List of abbreviations

CAD	collisionally activated dissociation
CDR	complementarity determining regions
CID	collision-induced dissociation
COPD	chronic obstructive pulmonary disease
CT	computed tomography
DLCST	Danish lung cancer screening trial
ELISA	enzyme-linked immuno sorbent assay
ETD	electron-transfer dissociation
Fab	fragment antigen binding
Fc	fragment crystallisable
HCD	higher energy collision-induced dissociation
IEF	isoelectric focusing
IgG	immunoglobulin G
ISS	international staging system
LC	liquid chromatography
MRI	magnetic resonance imaging
MRM	multiple reaction monitoring
MS	mass spectrometry
NELSON	Dutch-Belgian lung cancer screening
NLST	national lung screening trial
NSCLC	non-small cell lung cancer
PET	positron emission tomography
PTM	post-translational modifications
SCLC	small cell lung cancer
SDS-PAGE	sodium dodecyl sulfate polyacrylamide gel electrophoresis
SEREX	serological expression cloning
SERPA	serological proteome analysis
SRM	selective reaction monitoring
TAA	tumor-associated antigens
TNM	tumor, node and metastases
UKLS	UK lung cancer screening

Dankwoord (Acknowledgment)

Het is gelukt, mijn proefschrift is klaar! Als ik terugdenk aan de afgelopen jaren van mijn promotie periode heb ik heel veel bijzondere mensen leren kennen waar ik zoveel van geleerd heb zowel op wetenschappelijk gebied als persoonlijk vlak. Hiervoor wil ik iedereen graag bedanken en een aantal mensen in het bijzonder.

Als eerste wil ik de deelnemers van de NELSON studie bedanken voor hun medewerking aan deze studie. Zonder hun toestemming voor het afstaan van hun bloed voor de zijstudies binnen de NELSON studie, waar dit project er één van is, had een groot deel van dit proefschrift niet tot stand kunnen komen. Daarnaast wil ik ook de donoren van de Sanquin Bloedbank Rotterdam bedanken voor het afstaan van bloed, waarmee we onze methode hebben kunnen opzetten.

Dr. R.J. van Klaveren, beste Rob, ik weet nog goed dat je mij belde met het goede nieuws dat ik als OIO bij jou en Theo mocht beginnen. Je bent altijd zeer betrokken geweest bij het project, wat ook wel bleek uit de regelmatige werkbesprekingen die we hadden. Ik wil je heel erg bedanken voor het vertrouwen in mij, je steun en de kansen die je mij gegeven hebt. Bedankt voor je adviezen en je snelle feedback op de manuscripten.

Dr. Th.M. Luider, beste Theo, ook jou wil ik bedanken voor je vertrouwen in mij en je steun. Ik stond en sta er nog iedere keer weer versteld van als ik weer een hele donkere grijze wolk zag hangen, omdat de resultaten weer eens niet waren wat we hadden gehoopt, jij er weer een mooi fel oranje zonnetje doorheen wist te laten breken. Ontzettend bedankt voor jouw positiviteit. Daarnaast heb ik ook heel veel van je mogen leren door de kansen die je me gegeven hebt.

Prof. Dr. H.C. Hoogsteden, beste Henk, graag wil ik je bedanken voor de mogelijkheden die ik van je heb gekregen om mijn proefschrift te schrijven en de feedback die ik van je heb mogen ontvangen op een aantal stukken in dit proefschrift. Het was fijn dat je altijd, indien nodig, tijd voor mij vrij kon maken. Daarnaast wil ik je ook bedanken voor je hulp bij het versnellen van de afronding van mijn proefschrift.

Prof. Dr. P.A.E. Sillevis Smitt, beste Peter, ook jou wil ik graag bedanken voor de mogelijkheden die je me gegeven hebt om dit proefschrift te kunnen schrijven. Ik wil je heel erg bedanken voor je kritische feedback op de manuscripten en tijdens de werkbesprekingen en je to-the-point-heid, geen gedraai eromheen, dat heb ik altijd zeer gewaardeerd.

Graag wil ik de leden van de kleine commissie, Prof. Hendriks, Prof, Hooijkaas en Prof. Hintzen bedanken voor de tijd en moeite die jullie hebben genomen om dit proefschrift te beoordelen en mij te voorzien van goede adviezen.

Lieve Linda en Lennard, bedankt dat jullie mijn paranimfen willen zijn en voor jullie mentale support voor vandaag. Bedankt voor alle gezellige momenten op congres, borrels na het werk en tijdens onze culinaire etentjes. Hopelijk volgen er nog vele.

Graag wil ik mijn collega's Peter, Lennard, Linda, Lona, Marcel, Coskun, Azadeh, Christoph, Ingrid, Martijn, Vaibhav, Giovanni, Evert-Jan, Nick en alle oud-collega's Halima, Gero, Karin, Roland, Mark en Henk bedanken. Ik wil Christoph en Martijn in het bijzonder bedanken, Christoph bedankt voor al je hulp en medewerking als ik weer met een hele serie van samples aankwam die gemeten moesten worden op de Orbitrap en Martijn, bedankt voor al je hulp wat betreft antibodies. Zonder jouw scriptjes en trucjes had het allemaal nog langer geduurd en natuurlijk al je kennis die je met mij gedeeld hebt over antibodies. Lieve (oud-)collega's, bedankt voor de gezellige werksfeer en de nodige hulp die ik van jullie altijd heb gekregen. Bedankt voor alle gezelligheid tijdens de borrels na het werk en tijdens de congressen.
Collega's van de neuro-oncologie bedankt voor jullie gezelligheid toen ik nog bij jullie op het lab mijn Fab-jes aan het opzuiveren was en voor jullie input tijdens de werkbesprekingen.

Beste Wim, natuurlijk kan ik jou niet vergeten te bedanken. Het eerste model dat je gebouwd had noemde je " Model Dominique", maar ze had niet de juiste maten en er moest nog wel wat aan de knopjes gedraaid worden. Na een lange weg hebben we dan uiteindelijk een goed en betrouwbaar model gevonden. Daarom bedankt voor je geduld en doorzettingsvermogen. Bedankt voor alles.

Beste René Vernhout, Ton de Jongh, Roel Faber en Frank Santegoets, bedankt voor alle klinische data die ik van jullie netjes in Excel documentjes heb gekregen die bij NELSON deelnemers hoorden waarvan ik samples heb gebruikt. Bedankt voor de fijne samenwerking.

Lieve familie en vrienden, bedankt voor jullie interesse in mijn onderzoek en de nodige afleiding zoals gezellige etentjes, feestjes, relax-momentjes, shop-dagjes en vakanties.

Lieve pap en mam. Waar moet ik mee beginnen. Ik heb zoveel aan jullie te danken. Bedankt voor alle steun, begrip en vertrouwen op welk front dan ook die ik al heel mijn leven van jullie krijg. Het was voor mij niet altijd even gemakkelijk de afgelopen jaren, maar jullie waren er altijd voor mij als ik weer even mijn ei kwijt moest. Bedankt voor de nodige afleiding in de afgelopen drukke jaren van mijn promotieperiode.

Lieve Jacco, wat zou ik zonder jou moeten. Bedankt voor je liefde, steun, positiviteit en enorme geduld met mij. Bedankt voor de nodige afleiding zodat ik thuis mijn werk even van me af kon zetten. En natuurlijk voor je hulp bij het maken van de kaft en de lay-out van dit boekje. Het was leuk om er samen aan te werken. Op deze manier was Indesign minder frustrerend ;-)!

List of publications

de Costa, D., Broodman, I., Calame, W. Vanduijn, M.M., Stingl, C., Dekker, L.J., Vernhout, R.M., de Koning, H.J., Hoogsteden, H.C., Sillevis Smitt, P.A.E., van Klaveren, R.J., Luider, T.M., Peptides from the variable region of specific antibodies are shared among lung cancer patients, submitted.

Broodman, I., de Costa, D., Stingl, C., Dekker, L.J., Vanduijn, M.M., Lindemans, J., van Klaveren, R.J., Luider, T.M., Mass spectrometry analyses of kappa and lambda fractions result in increased number of complementarity determining regions identifications, Proteomics, 2012, 12(2), 183-91

Titulaer, M.K., de Costa, D., Stingl, C., Dekker, L.J., Sillevis Smitt, P.A., Luider, T.M., Label-free peptide profiling of Orbitrap TM full mass spectra, BMC Res Notes. 2011, 4:21.

de Costa, D., Broodman, I., Vanduijn, M.M., Stingl, C., Dekker, L.J., Burgers, P.C., Hoogsteden, H.C., Sillevis Smitt, P.A., van Klaveren, R.J., Luider, T.M., Sequencing and quantifying of IgG fragments and antigen-binding regions by mass spectrometry, J Proteome Res., 2010, 9(6), 2937-45.

Smedts, H.P., Isaacs, A., de Costa, D., Uitterlinden, A.G., van Duijn, C.M., Gittenberger-de Groot, A.C., Helbing, W.A., Steegers, E.A., Steegers-Theunissen, R.P., VEGF polymorphisms are associated with endocardial cushion defects: a family-based case-control study, Pediatr Res., 2010, 67(1), 23-8.

van den Boogaard, M.J., de Costa, D., Krapels, I.P., Liu, F., van Duijn, C.M., Sinke, R.J., Lindhout, D., Steegers-Theunissen, R.P., The MSX1 allele 4 homozygous child exposed to smoking at periconception is most sensitive in developing nonsyndromic orofacial clefts., Hum Genet., 2008, 124(5), 525-34.

PhD Portfolio

Name PhD student:	Dominique de Costa
Erasmus MC Department:	Pulmonology/Neurology
Research School:	Erasmus Postgraduate school Molecular Medicine
Promotors:	Prof. dr. H.C. Hoogsteden
	Prof. dr. P.A.E. Sillevis Smitt
Supervisor:	Dr. T.M. Luider

PhD training

	Year	Workload (Hours/ECTS)
General courses		
- Biomedical English Writing and Communication, Erasmus MC	2008/2009	4.0 ECTS
Specific courses		
- Principles of Research in Medicine, NIHES	2008	0.7 ECTS
- Introduction to Data-analysis, NIHES	2008	1.0 ECTS
- 3200/4000 QTRAP®LC/MS System Proteomics, Applied Biosystems, Warrington, UK	2008	24 hours
- Molecular Immunology, Erasmus MC	2009	3.0 ECTS
- Topics in Meta-analysis, NIHES	2009	0.7 ECTS
- Regression Analysis, NIHES	2009	1.9 ECTS
- Case Studies in Quantative Proteomics, 58th Conference of the American Society for Mass Spectrometry, Salt Lake City, USA	2010	16 hours
- Photoshop & Illustrator CS5, Erasmus MC	2011	0.3 ECTS
- Bioinformatics for Protein Identification, 59th Conference of the American Society for Mass Spectrometry, Denver, USA	2011	16 hours
- InDesign CS5, Erasmus MC	2011	0.3 ECTS
- Biomedical Research Techniques, Erasmus MC	2011	32 hours
- Next Generation Sequencing Data Analysis	2011	1.4 ECTS
Presentations		
- 12th Molecular Medicine Day of the Erasmus Postgraduate school Molecular medicine, *Biomarker identification for early detection of lung cancer by proteomics techniques in the NELSON lung cancer screening study*, poster presentation, Rotterdam, the Netherlands	2008	40 hours
- 99th American Association for Cancer Research Annual Meeting, *Biomarker identification for early detection of lung cancer by proteomics techniques in the NELSON lung cancer screening trial*, Poster presentation, San Diego, USA	2008	40 hours
- NELSON Meeting, *Sequencing and quantification of IgG fragments by mass spectrometry: A proteomic approach in the NELSON lung cancer screening trial*, Oral presentation, Schiphol, the Netherlands	2009	32 hours
- 58th Conference of the American Society for Mass Spectrometry, *Sequencing and quantification of IgG fragments and antigen binding regions by mass spectrometry*, Oral presentation, Salt Lake City, USA	2010	40 hours
- Refereeravond " Het pulmonologisch jaar 2010", *Antistoffen als biomarker voor vroegtijdige detectie van longkanker*, Oral presentation, Rotterdam, the Netherlands	2010	32 hours
- 59th Conference of the American Society for Mass Spectrometry, *CDRs as an early detection marker for non-small cell lung cancer using mass spectrometry*, Poster presentation, Denver, USA	2011	40 hours
- 14th World Conference on Lung Cancer, *CDRs as an early detection marker for non-small cell lung cancer using mass spectrometry*, Poster presentation, Amsterdam, the Netherlands	2011	16 hours
- NELSON Meeting, *Longkankerscreening en biomarkers: Antistoffen als biomarker voor vroegtijdige detectie van longkanker*, Oral presentation, Rotterdam, the Netherlands	2011	32 hours

(inter)national conferences		
- 12th Molecular Medicine Day of the Erasmus Postgraduate school Molecular medicine, Rotterdam , the Netherlands	2008	8 hours
- 99th American Association for Cancer Research Annual Meeting, San Diego, CA, USA	2008	40 hours
- 100th American Association for Cancer Research Annual Meeting, Denver, CO, USA	2009	40 hours
- NELSON Symposium, Utrecht, the Netherlands	2009	8 hours
- 58th Conference of the American Society for Mass Spectrometry, Salt Lake City, UT, USA	2010	32 hours
- 59th Conference of the American Society for Mass Spectrometry, Denver, CO, USA	2011	32 hours
- 14th World Conference on Lung Cancer, Amsterdam, the Netherlands	2011	32 hours