Proteomic analysis of formalin-fixed paraffin-embedded colorectal cancer tissue using Tandem Mass Tag protein labeling

Evelyne Maes [1,2,3,*] , Dirk Valkenborg [1,2,4], Inge Mertens [1,2] , Valérie Broeckx[3], Geert Baggerman [1,2] , Xavier Sagaert[5] , Bart Landuyt [3], Hans Prenen[6] and Liliane Schoofs [3]

[1] Flemish Institute for Technological Research (VITO), Mol, Belgium

[2]CFP-CeProMa, University of Antwerp, Antwerp, Belgium

[3]Research Group of Functional Genomics and Proteomics, University of Leuven, Leuven, Belgium

[4]Interuniversity Institute for Biostatistics and Statistical Bioinformatics, Hasselt University, Diepenbeek, Belgium

[5]Centre for Translational Cell and Tissue Research, Leuven, Belgium

[6]Department of Gastro-Enterology, Digestive Oncology Unit, University Hospital Gasthuisberg, Leuven, Belgium.

Corresponding author:

*Email: evelyne.maes@bio.kuleuven.be

Postal address: Research group of Functional Genomics and Proteomics, KU Leuven, Naamsestraat 59, Leuven, Belgium Phone: +32 16 323913 Fax: +32 16 323902

Abstract

In clinical research, repositories of biological samples form a rich source of clinical material for biomarker studies. Banked material however, is often not stored in optimal conditions regarding the technology used for biomarker research. A case in point is formalin-fixed paraffin-embedded (FFPE) tissue that could be used to obtain large cohorts of samples over a short period of time, as these tissues are routinely prepared for pathological analysis. However, in the context of mass spectrometry based peptide-centric proteomics, protein extraction and identification can be hampered by formalin-induced crosslinking. Furthermore, the molecular formalin crosslinks might be entangled differently across various samples, making it more difficult to reproducibly extract the same proteins from different samples. In this study, we establish the crosslink variability using Tandem Mass Tag (TMT) protein labeling followed by protein digestion, separation, identification and quantification of proteins extracted from FFPE colorectal cancer and paired healthy tissues. Moreover, by applying *de novo* interpretation of tandem mass spectra and subsequent analysis by Peaks PTM, unspecified modifications could be elucidated, leading to increased protein and proteome coverage. This approach might be useful for future FFPE proteomics studies.

Keywords: FFPE tissue - modifications - proteomics – Tandem Mass Tags - protein labeling

Introduction

In clinical research, fresh sample material, such as, e.g., tissues are often banked for future investigation by emerging technologies. These repositories of residual clinical specimens (e.g. biobanks, tumorbanks, and pathology archives) are of tremendous interest for biomarker discovery research, as they form a rich source of clinical material. The advantage of using banked sample material is that pathological, clinical and outcome information exists in these collections of tissues [1], which allows for retrospective biomarker research. A disadvantage of using banked sample material is that the sample is often not stored in optimal conditions regarding the technology used for the biomarker research. A case in point is formalin-fixed paraffin-embedded (FFPE) tissue that could be used to obtain large cohorts of samples over a short period of time, as these tissues are routinely prepared for pathological analysis [2]. Moreover, it is the standard method for long-term preservation of clinical specimens. Although often denied in biomarker studies, these millions of archived FFPE tissues represent thus an unexploited treasure of samples[3].

The proteomic analysis of FFPE tissue often employs technologies, such as Liquid Chromatography (LC) and Mass Spectrometry (MS), which are considered as important tools in protein biomarker research. Therefore, a substantial need exists to develop methods and procedures for this technology to discover new protein biomarker candidates with diagnostic or therapeutic potential. However, fixation of tissue samples by formalin leads to extensive inter- and intramolecular crosslinking among proteins in these tissues, which hampers the proteome analysis of these samples. Due to these crosslinks, not only protein extraction is a major challenge, also protein identification might be hindered, as the reaction of proteins with formaldehyde will deliver both known and unknown modifications. Moreover, a small percentage of the formaldehyde-reactive amino acids e.g. arginine, histidine, cysteine and lysine residues will form irreversible crosslinks, which might be difficult to assess by mass spectrometry [4]. So far, it is clear that this event of crosslinking is not completely understood yet.

To date, several research groups have demonstrated that protein extraction from FFPE tissue sections is feasible using heat-induced antigen retrieval and specialized extraction buffers combined with downstream gel-based or gel-free designs [5-8]. To quantify FFPE extracted proteins, label-free approaches are mostly applied [9, 10]. However, in this manuscript, we adopt an approach that uses labels. Because up to eight samples are pooled together and processed simultaneously by LC and MS, they are affected by the same amount of instrument variability, which facilitates a rigorous comparison. Quantification through chemical labeling with isobaric tags for relative and absolute quantification (iTRAQ) has been used before in the context of FFPE tissue proteomics [11, 12]. However, in this study, we will use Tandem Mass Tags (TMT) for quantification purposes and apply a combination of TMT protein labeling and 1D electrophoresis combined with LC-MS/MS (= GeLC-MS/MS). Performing gel electrophoresis is advantageous in formalin crosslinked protein samples, as not only the complex sample is fractionated, also the excess of paraffin and other MS-incompatible buffer reagents are removed. Labeling proteins instead of peptides is beneficial in a GeLC-MS/MS approach, as in this way, we exclude the circuitous labeling of several in-gel digested peptide fractions.

To our knowledge, this is the first report which performs a quantitative proteome analysis using TMT protein labeling in combination with FFPE tissues.

Materials and Methods

Human FFPE tissue samples

Human clinical tissues from patients with colorectal cancer were obtained from biopsies and resected tumor material at the university hospital of Leuven. These samples were collected under informed consent of all patients and were approved by the local ethical committee. Both tumor tissue and adjacent control colon mucosa from surgical resection specimens were collected. Diagnosis of colorectal cancer was made as defined by the criteria of the WHO classification. The surgical resection specimens were fixed in 6% formalin for 24 to 48 hrs and dehydrated before impregnation by paraffin. Afterwards, the FFPE samples were stored at room temperature. None of the samples were stored for longer than one month. More information about the paired colon mucosa samples is summarized in Table 1.

Protein extraction

Per sample, ten slices of 10 µm FFPE sections were collected in a Lobind Eppendorf tube. The FFPE samples were deparaffinised in xylene for 10 min, followed by centrifugation at 10,000 g for 3 min. The tissue pellets were then rehydrated with a graded series of ethanol (100%, 95% and 70%). To extract the proteins used for TMT labeling, three paired control mucosa and colorectal cancer mucosa FFPE samples were suspended in 20mM Tris HCl pH 8,8, 200mM DTT, 2% SDS and 1% protease inhibitor (Complete cocktail, Roche, Penzberg, Germany) and incubated for 20 min at 98°C, followed by an incubation at 80°C for 2 hrs. After centrifugation of the samples at 14,000 g, for 30 min at 4°C, the supernatant was transferred to a new tube and stored at -80°C until further use.

TMT protein labeling

All products in this section are supplied with the TMT labeling kit ((Pierce, Thermo Scientific, Waltham, MA) unless stated differently. In order to achieve optimal labeling conditions, the samples were transferred into 30 kDa centrifugal filter units (Amicon Ultra, 30K, Millipore, Billerica, MA). After centrifugation at 13 000 g for 10 min, the proteins were resuspended in 100 µl of 200 mM triethylammoniumbicarbonate (TEAB) (pH 8,0). Next, the samples were transferred into a mini dialysis tube (cut-off 1kDa, GE Healthcare, Uppsala, Sweden) and dialysed for 2 hrs at 4°C. Afterwards, the protein concentration was determined using the Qubit method (Invitrogen, Carlsbad , CA).

Before labeling the proteins, 2x 10 µg proteins of each sample were reduced using 2 µl of 50 mM tris(2-carboxyethyl) phosphine in a volume of 100 µl 100 mM TEAB , and incubated for 1 h at 55°C. Next, the samples were alkylated with 0,5 µl of a 375 mM iodoacetamide solution for 30 min at ambient temperature in the dark. For the reconstitution of the tags, the TMT labels were dissolved in 41 µl acetonitrile according to the manufacturer's protocol. Subsequently, proteins were labelled with the TMT reagents as follows: the control colon mucosa samples were labeled with TMT 126, TMT 128 and TMT 130 and the paired colon tumor samples with TMT 127, TMT 129 and TMT 131. From every sample, 10 µg was labeled with 4,1 µl of a TMT tag dissolved in acetonitrile. The labeling reaction was stopped by adding 5 µl 5% hydroxylamine. After 15 minutes, a pooled sample was prepared based on the labeled samples with a protein concentration ratio of 1:1:1:1:1:1.

1D electrophoresis and trypsin digestion

The proteins in the mixture were separated by 1D SDS-PAGE using NuPAGE Novex 4-12% Bis-Tris precast gel according to the manufacturer's protocol (Invitrogen, Carlsbad, CA). In each lane, 20 µg of proteins were loaded and SeeBlue 2 Plus (Invitrogen) was used as a pre-stained standard. The gel was stained with Coomassie Brilliant Blue (SimplyBlue SafeStain, Invitrogen) according to the manufacturer’s instructions. Subsequently, a lane was subdivided in three parts in order to pre-fractionate the sample. To destain the gel pieces, they were suspended in a mixture of 50% acetonitrile and 50% 50 mM ammonium bicarbonate for 30 min at 37°C. Afterwards, the gel pieces were hydrated with MilliQ, followed by rehydration in 100% acetonitrile and then dried in the speedvac concentrator. Next, trypsin (Promega, Fitchburg, WI), dissolved in 50 mM ammonium bicarbonate and 5 mM calciumchloride, was added at an enzyme to protein ratio of 1:20 and the sample was incubated overnight at 37°C. The next day, the tryptic peptides were extracted using 50 mM ammonium bicarbonate followed by an extraction with 50% acetonitrile and 1% formic acid. The pooled extracts were vacuum dried and the peptides were stored at -20°C. Prior to mass spectrometric analysis, the samples were desalted and concentrated using C18 ZipTips (Millipore, Billerica, MA) according to the manufacturer’s instructions. The eluted peptides were vacuum dried and stored at -80°C until further analysis.

Nano reverse phase liquid chromatography and mass spectrometry

The peptide mixture was separated by reverse phase chromatography on an Eksigent nano-UPLC system using a Pepmap100 C18 precolumn (200µm x 20mm, 5µm particle size) coupled to an acclaim C18 column (75µm x 15cm, 3µm particle size) (Thermo Scientific, San Jose, CA). Before loading, the sample was dissolved in mobile phase A, containing 2% acetonitrile and 0,1% formic acid and spiked with 20 fmol Glu-1-fibrinopeptide B (Glu-fib, Protea biosciences, Morgantown, WV). A linear gradient of mobile phase B (0,1% formic acid in 98% acetonitrile) in mobile phase A (0,1% formic acid in 2% acetonitrile) from 2 to 40% in 50 min followed by a steep increase to 95% mobile phase B in 2 min was used at a flow rate of 350 nl/min. The nano-LC was coupled online with the mass spectrometer using the Triversa NanoMate (Advion, Ithaca, NY) with LC-coupler.

The LTQ Orbitrap Velos (Thermo Scientific, San Jose, CA) was set up in a MS/MS mode where a full scan spectrum (350 – 5000 m/z , resolution 60 000) was followed by a maximum of five dual CID/HCD tandem mass spectra (100 to 2000 m/z). Peptide ions were selected for further interrogation by tandem MS as the five most intense peaks of a full scan mass spectrum. Collision induced dissociation (CID) scans were acquired in the linear ion trap of the mass spectrometer, High Energy collision activated dissociation (HCD) scans in the orbitrap, at a resolution of 7500. The normalized collision energy used was 35% in CID and 55% in HCD. We applied a dynamic exclusion list of 30 sec for data dependent acquisition.

Data analysis

Proteome discoverer (1.3) software (Thermo Scientific, San Jose, CA) was used to perform database searching against the IPI Human 3.87 database using both Sequest and Mascot algorithms, and following settings: precursor mass tolerance of 10 ppm, fragment mass tolerance of 0.8 Da. Trypsin was specified as digesting enzyme and 2 miscleavages are allowed. Regarding fixed or variable modification settings, carbamidomethylation at methionine was always set as fixed modification, and TMT-sixplex labels at N-terminus and lysine residues, in combination with methionine oxidation were variable modifications. Several extra prespecified modifications e.g. methylation, acetylation,... were also applied in order to gain better results. Only medium and high confident peptides with a global FDR at < 0,05 were included in the results. A summary of these settings and their respective FDR values is given in Table 2.

Peaks studio software (Version 6, Bioinformatics solutions Inc., Waterloo, ON, Canada) was also used to analyse the MS/MS spectra for unexpected modifications. Data were refined in precursor mass and four different analysis steps were used for protein identification. The following analysis steps were comprised in the procedure: *de novo* interpretation of peptides, Peaks DB search for database driven peptide identification, Peaks PTM search for detecting frequently occurring post-translational modifications and the spider search module to align the *de novo* identification on the database.

All MS/MS spectra were searched using Peaks DB against the IPI human 3.87 database. The search parameters were as follows: precursor mass tolerance of 5 ppm, fragment tolerance of 0,5 Da. Trypsin was selected as digestion enzyme, and 3 missed cleavages were allowed. Furthermore, carbamidomethylation (+57,021 Da, C) was set as fixed modification, TMT 6-plex (+229,163 Da, at N-terminus and K) and oxidation (+ 15,995 Da, M) were dynamic or equivalently, variable modifications. We applied a peptide identification filter at FDR<5% and protein identification was based on at least one unique peptide.

The quantitative analysis was conducted by the Peaks software to retrieve information on the TMT 6-plex reporters. TMT 6-plex (N-terminus and K) was used as quantification type, a quantification mass tolerance of 10 ppm was allowed and the threshold for peptide score was set at 20. In the quantification module of Peaks, global intensity normalization was performed to correct for unequal mixing of the TMT labeled samples. The expression patterns of clustered proteins are shown in a heatmap (Figure S1). The normalized log ratios of the reporter ion intensities were shown.

Results

The identification and quantification of proteins extracted from formalin-fixed material is still a challenge. Although several research groups used label-free designs, label-based methods are often discarded in the context of FFPE proteomics [5], as most chemical labeling strategies tag amino acid residues which might be involved in crosslinking. However, using an isobaric labeling approach has several advantages including multiplexing and more precise quantification [13], because the sources of variation are constant for the multiplexed samples. Therefore, we extracted proteins from FFPE samples of three human colon carcinoma tissues and their paired healthy colon mucosa tissues and applied isobaric labeling. We used the direct tissue proteomics strategy, in which proteins of the whole FFPE slices were extracted. Although we removed paraffin using wash steps with xylene and ethanol graded series, some paraffin was still present in the samples, disturbing the LC analysis. However, by applying 1D gel electrophoresis, not only this excess of paraffin can be removed, also fractionation of the sample can be achieved. In spite of the popularity of the peptide labeling protocol, we choose protein labeling, as labeling and multiplexing the samples before 1D gel separation is more reproducible and less laborious than labeling samples after in-gel digestion. The resulting gel lane was subdivided into 3 gel fractions and analyzed separately using LC-MS/MS. For identification and quantification purposes, the three LC-MS/MS runs are combined in one data repository.

Peptide and protein identification

In FFPE proteomics research, several 'unknown' modifications can hamper the identification of proteins. In traditional database search engines, like SEQUEST or MASCOT, the major drawback is that one needs to specify all the expected modifications (fixed and variable) before database searching is performed. Adding more dynamic modifications will combinatorially increase the search space, which increases the number of chance findings. This drawback makes the database search strategy impractical for FFPE data. Using a combination of SEQUEST and MASCOT search engines, we could identify 79 proteins (FDR<5%) when applying standard settings including carbamidomethylation as fixed modification and TMT sixplex and oxidation at methionine as variable modification. Adding extra modifications, like phosphorylation (STY) or hydroxymethylation (K), did not provide substantially better results. A comparison between the different settings can be found in Table 2.

Because the database search strategy is suboptimal in the FFPE proteomic approach, we opted for an alternative identification strategy. In this case, the use of a *de novo* identification method increased the number of identifications as it accommodates unknown modifications in a flexible and robust manner. Consequently, Peaks 6 software was used to identify and quantify the FFPE extracted proteins. This software performs both *de novo* interpretation and database searching. Moreover, two extra modules: Peaks PTM and SPIDER, are of particular interest [14, 15]. These modules are used to find modifications unforeseen for peptides which obtained a good *de novo* alignment, but which could not be identified by Peaks database searching. For example, in our preliminary dataset, *de novo* searches (average local confidence (ALC) ≥ 30%; total local confidence (TLC) ≥3)) resulted in 7943 peptides, from which 451 proteins could be identified using database searching. Our results were filtered such that proteins should have at least 1 unique peptide per protein identification and only confidently identified peptides (FDR<5%; peptide -10lgP ≥19,9) are included. In the Peaks PTM module, 396 high confident proteins were identified by following peptide identification settings (FDR<5%; peptide -10lgP ≥19) and Peaks Spider could identify 377 proteins (peptide FDR <5%, peptide -10lgP ≥18,9). In total, 713 unique proteins could be identified using the Peaks workflow. A detailed list of all protein identifications can be found in Table S1 in the supplementary material. The identified proteins from the FFPE material were found to be part of a broad range of biological processes and arose from diverse cellular compartments.

In comparison to standard database search engines, Peaks *de novo* sequencing algorithms delivers 6 times more high confident protein identifications and reaches even 10 times more identifications when including all the modules of Peaks. Furthermore, Peaks PTM gives a list of the most common post translational modifications that were found in our datasets. These modifications are summed in Table 3. Due to these extra modifications, higher protein coverage was achieved and up to 5% new confident protein accession numbers could be identified. From this list it can be observed that several post translational modifications (PTMs) take place at basic amino acids (K,R,H or N-term). Most likely, these modifications originate from formalin-induced crosslinking, which preferably is positioned at basic amino acids. Besides, the 5% newly identified proteins are found because a PTM was found in one of the unique peptides, which was missed and filtered out previously. Figure 1A shows a protein coverage view (IPI00010779;TPM4,isoform1) which is identified due to Peaks PTM. The figure gives an overview of all peptide matches found. Two unique peptides are present, and have unspecified modifications (acetylation and methylester). The fragmentation spectrum of one unique peptide is shown in Figure 1B.

TMT protein labeling

As the network of protein crosslinks might be different in each sample, we did establish the crosslink variability between six FFPE samples using Tandem Mass Tag (TMT) protein labeling. In the TMT approach, we receive for each identified peptide quantitative information regarding the sample in which the peptide is found. These TMT isobaric tags consist out of an amine-reactive group, which reacts with N-terminal amine groups and the epsilon-amine groups of lysine, a balancer group and a reporter ion group. In case of TMT, this reporter group has a different mass for each of the six variants of the TMT tags. Demultiplexing the pooled samples occurs in tandem MS mode where the isobaric tags will generate a reporter ion with a unique mass. The relative intensities of each unique reporter will give information about the relative abundance of the peptide in the pooled sample. Moreover, TMT sixplex has the capability to multiplex up to six different samples into one run, which reduces measurement time considerably. As most amine reactive amino acids regain their activity after heat induced extraction, isobaric tag labeling might be feasible in FFPE tissues. Moreover, the crosslink process is not necessarily the same in each sample and in each protein, and thus by multiplexing 6 samples into 1 run, the extraction efficiency and reversibility of the protein crosslinks can be evaluated, without differences in external experimental parameters of both LC and MS.

To evaluate the protein labeling reaction, several factors were checked. First, the presence of TMT modifications in the identified proteins was examined. As these isobaric tags are reactive against free amine reactive groups, only peptides with lysine groups or N-terminal peptides will be labeled. TMT modification at lysine was found in 921 peptide to spectrum matches (PSM) for database searching. By extending the search using Peaks PTM, there were 145 additional peptides found wherein lysine was modified by TMT. TMT sixplex modifications at the N-terminus, however, were present to a lesser extent in the identified proteins, which was expected when applying protein labeling. In total, 471 unique proteins were quantifiable. Secondly, the presence of the TMT reporter ions in the fragmentation spectra of the peptides were evaluated. The presence of 6 reporter ions indicates that the peptide is successfully extracted from six different FFPE samples. Figure 2 shows an example of a protein (IPI00216456, Histone H2A, type 1C) (Figure 2B) which contains several TMT sixplex modifications. In the corresponding fragmentation spectrum, the six reporter ions were found (Figure 2A, 2C). In total, 1629 unique peptides were successfully extracted, identified and quantified in the six samples. Subsequently, the coefficient of variation (CV) values of the peptides extracted from control samples varies between 0,01 and 1,32. The CV values of tumor-extracted peptides ranges from 0,07 until 1,2. Third, protein labeling hinders the activity of trypsin at the lysine residue. As a result, the presence of tryptic peptides containing C-terminal lysine residues should be reduced. When looking at all the identified peptides, about 95% of the peptides are digested at arginine. These observations could also be the result of an extensive crosslinking at the lysine residue, however, most peptides containing lysines, also have TMT modifications. Forth, it is worth mentioning that 66% of the identified proteins using Peaks were also quantified in the six samples (Table S2). All these factors prove that a good efficiency of TMT labeling can be confirmed.

Furthermore, a quality control visualizing the different tumor vs. control ratios was performed using an MA-plot. In this plot, the log intensity ratios (M) of two TMT tags and their average log intensity (A) are plotted against each other for all the identified peptides. The general assumption concerning this plot is that most proteins will not show any change in expression, and should therefore be located around 0, since log (1) = 0. When any deviation is seen, further normalization is needed in order to obtain reliable results regarding quantification. Figure 3 shows the MA-plots of the different TMT pairs (tumor/control) for our TMT protein labeled samples. No abnormalities were seen in these plots, indicating that no interference between FFPE and TMT reporter ions was present. However, deviation from the ‘log(1)=0’ line is observed, which indicates that normalisation might be beneficial. Using the quantitation workflow of Peaks an auto-normalisation was carried out. Also, a clustering analysis on protein expression was performed to detect protein groups which behave similar across the control or tumor group. The resulting heatmap can be found in the Figure S1. As we used paired control and tumor samples, we looked for trends in which the reporter intensities of the 127, 129 and 131 labeled samples have a higher/lower expression than their counterparts, respectively 126, 128 and 130. Several proteins did show a trend in upregulation in tumor samples compared to the paired healthy tissues.

Discussion

It is already known for several decades that FFPE tissue is a treasure for retrospective analysis concerning the amount of samples present in hospital archives, combined with pathological, clinical and outcome information available for every sample. In recent years, several research groups did show that protein extraction and identification is possible. Also, the challenges and pitfalls concerning FFPE proteomics have become clear [5].

Unlocking the proteome of formalin-fixed tissues is still considered a challenge for two reasons: protein extraction might be hindered by crosslinks and protein identifications might be ambiguous due to possible unknown peptide modifications [16]. In this study, we could show that hundreds of proteins can be efficiently extracted, identified and quantified from six FFPE samples which are processed and labeled (with isobaric tags) in parallel and multiplexed to one pool upon LC separation and mass spectrometry detection.

In this study, we applied a combination of TMT protein labeling and 1D electrophoresis followed by LC-MS/MS (= GeLC-MS/MS). The application of 1D gel electrophoresis to FFPE extracted proteins has several advantages: not only will it fractionate the sample and thus reducing the complexity, it also has the possibility to visualize the high abundant proteins, which makes it possible to isolate these proteins from the less abundant ones, rendering more low abundant protein identifications. Moreover, the compatibility with most

FFPE extraction buffers is also major benefit. However, in GeLC-MS/MS label-free approaches, every sample needs to be fractionated and in-gel digested separately before LC-MS/MS analysis is possible, leading to very laborious procedures which might be less reproducible. Therefore, isobaric labeling using TMT, which allows multiplexing up to six samples, might be a good alternative. These labels however, are mostly used in the context of peptide labeling. Performing TMT peptide labeling in combination with GeLC-MS/MS would even more complicate the workflow, as the in-gel digested protein fractions should be reduced, alkylated and labeled separately, before multiplexing is possible. All these additional sample handling steps will introduce a higher overall variability, certainly when working with smaller sample amounts. Therefore, labeling proteins rather than peptides has some major advances in the GeLC-MS/MS setup. The first advantage is that protein-based fractionation is less complex than a peptide-based procedure. Second, labeling of the proteins and multiplexing samples prior to in-gel digestion will benefit the reproducibility. Finally, this protein labeling strategy would also benefit FFPE analysis when using SDS PAGE as fractionation method, as paraffin, which might still be present in low amounts in the extracted protein samples, could otherwise contaminate the LC-run, leading to LC-results of low quality.

In FFPE proteomics, it is known that the efficiency of protein recovery is influenced by the fixation protocol and the fixation time. FFPE tissues which are fixed in high concentrations of formalin (> 10%) or have long fixation times (> 72 hrs) will have a tighter network of molecular crosslinks [17]. The archival time, on the other hand, has only limited influence on the protein extraction efficiency [18]. In our study, samples were fixed in 6% formalin for 48 hrs. Although the crosslinking process thus had the same time to expand, differences in the crosslink network between the different samples are likely to exist. Therefore, it might be possible that proteins are more easily extracted in one sample then in another, depending on the crosslinks formed. To exclude the fact that proteins/peptides might, despite their efficient extraction, not be identified because of unknown or unpredicted modifications resulting from formalin fixation, paraffin embedding or the reversal of these processes, a comparison was made between standard database search engines like SEQUEST or MASCOT and *de novo* interpretation algorithms using Peaks software. A combination of these two standard search engines delivered only a limited amount of confident unique proteins, even when potential formalin-induced modifications were added in the search parameters. In addition, to assess the influence of these degrees of freedom in the search parameters on the number of identifications, several different search parameter settings were tried, all generating these low numbers of identifications (Table 2).

Because of the low identification rate using traditional database search engines, further elucidation of the MS/MS fragmentation spectra was performed using *de novo* interpretation of the data. The Peaks 6 software implements a combined *de novo* sequencing[19], where tandem MS spectra are used to determine peptide sequences based on the obtained fragmentation pattern, and assisted database search for accurate peptide identification[20]. However, this *de novo* DB search does not support the identification of modified peptides to the most possible extent. To elucidate unknown modifications induced by formalin crosslinking, the Peaks PTM module was used. As opposed to traditional database search, where one can specify only a few well-defined modifications, this module considers all known 650 PTMs included in the Unimod database[15]. This way, unspecified modifications due to the crosslinking event can be elucidated with increasing FFPE protein and FFPE proteome coverage as a consequence. To achieve this goal, Peaks PTM includes database filters that allow to consider a rich set of PTMs for every peptide in the filtered database. This reduced database is the list of identified protein candidates proposed by MS/MS data interpretation using Peaks *de novo* and Peaks DB algorithms. From hereupon, an extensive search to find peptides of the protein candidates with one (or more) PTM from the Unimod database can be performed[15]. As expected, several modifications, like acetylations and deamidations were found in the dataset. We noticed that high numbers of methionine oxidations were present. These could be linked to FFPE storage, as it is known that oxidation reactions are more pronounced in FFPE tissues and will only increase as the storage time of the archival samples will increase [18]. Modifications which could be linked to the chemistry of formalin fixation might be (di)hydroxylations at basic amino acids and formylation which were present at lysine and, N-termini as well as at other amino acids. These can be the result from the breakdown of the methylene bridges during heat induced protein extraction. However, FFPE modifications predicted in model peptides, like hydroxymethyl (methylol) groups were not found in real biological samples, although present in the Unimod database. These findings only emphasize that formalin-based crosslinking has a high degree of complexity and that the samples are probably even more heterogeneous than can be assessed by mass spectrometry [16]. This could also explain why a comparison between the proteome of frozen and FFPE tissue only overlapped for 40%-90% [21]. In total, we could identify 713 confident unique proteins using Peaks, which is almost 10 times more than using standard search engines. This emphasizes that Peaks software is a valuable alternative for identification purposes. Moreover, this approach differs from other bioinformatic identification tools dealing with crosslinked peptides [13, 22].

Although identification of several formalin-induced modifications on peptides is possible now using Peaks software, the question remains whether it is possible to extract proteins from FFPE samples in a reproducible way. As far as is understood these days, formalin fixation leads to chemical crosslinks of RNA, DNA and proteins and affects proteins at different levels. First of all, the modification of amino acid residues modifies the primary structure. Metz and colleagues performed experiments on the reaction of formaldehyde with insulin as a model protein, and showed that several amino acids (arginine, asparagine, histidine, glutamine, tryptophan and tyrosine) could react with unstable adducts from the formaldehyde reaction. The position and local environment of each reactive amino acid however, did affect the reactivity[23]. So, in general, each protein in each sample will not have exactly the same environment, possibly leading to variation in modifications and crosslink formation between proteins, DNA and RNA. The variation might be even more extensive when one thinks about the reversal of the protein-induced crosslinks upon protein extraction. Moreover, crosslinking also involves changes in secondary, tertiary and quaternary structures, including the formation of complexes that are hard to unravel by mass spectrometry[16].

To elucidate whether it is even possible to extract the same peptides from 6 different samples, we applied TMT labeling to FFPE extracted proteins. By multiplexing these 6 samples into 1 run, before sample processing was performed, all other technical factors could be ruled out.

The use of TMT labeling as quantification method might seem a challenge in FFPE tissue. We could show that although modifications of lysine and N-termini exists in FFPE tissue, TMT technology showed good performance, as the reporter ions of hundreds of FFPE extracted proteins were found. In this study, we labeled proteins extracted from six different FFPE samples with six different isobaric TMT tags to evaluate the variation in protein extraction and digestion. We could successfully identify and quantify 1629 unique peptides in which the six TMT reporter ratios are present. This indicates that, although the crosslink network might be different in each sample, proteins can efficiently be extracted in these 6 samples using our procedure. A further look at the CV values of the TMT reporter ion intensities showed us that, even though the quantification is successful, several peptides did show CV values above 1. These observations are important for future quantitative proteomics experiments, as more biological replicates will be necessary per experimental group (compared to fresh frozen material) in order to achieve statistically significant results.

Determination of both within and between experimental group variations do also show that the FFPE crosslink ‘background’ is the same for both groups, indicating that biological differences can be observed in FFPE context. Other research groups also used isobaric labeling in formalin-fixed tissues to quantify proteins by applying peptide labeling [11, 12] instead of protein labeling [24]. In this study, we showed that protein labeling can be regarded as a valid alternative and that it also simplifies the labeling protocol in a GeLC-MS/MS setup. Applying isobaric labeling has the advantage that more accurate quantifications can be achieved than in label free (e.g. spectral counting) approaches because it provides information about the six samples simultaneously in one LC and MS run. Also, in the fragmentation spectra, the reporter ions are in a noise-free region of the spectrum which enhances quantification. Moreover, the use of HCD as a second fragmentation method, can ensure that a good quantification can be obtained. Finally, the ability to multiplex up to six samples in one run, has the advantage to save mass spectrometry measurement time. A disadvantage, on the other hand, is that only N-terminal peptides and peptides containing lysine residues are quantifiable.

In the quantification workflow of peaks, hierarchal clustering of normalized quantified proteins was performed (Figure S1). As noticed, patient 1 has lower intensity values for almost all proteins. The protein ratios per patient (127/126 = patient 1, 129/128 = patient 2, 131/130 = patient 3) do show a trend of upregulation of several proteins in colon tumors (Figure S1). Some of them have structural functions (Collagen 4 A2, Histone H2A), some others had functions related with the cancer process: Enolase 1 is involved in the regulation of transcription; Mast cell tryptase beta III on the other hand, is important in the inflammatory process. Moreover, it is known that cancer associated inflammation is an important trigger which promotes cancer progression and metastasis [25]. However, since we are just using a limited number of non-standardized samples, no biomarker candidates will be proposed. First, although all patients had non-metastatic colorectal cancer, further distinction between the cancer types is necessary for biomarker experiments in order to find cancer-stage related differences. Secondly, more biological replicates are needed before any conclusions regarding biomarker candidates can be drawn. Also, using FFPE tissues for biomarker discovery will need more biological replicates compared to fresh or frozen tissues, as the crosslinking event is adding extra technical variation. Finally, further statistical improvements of the Peaks quantification workflow are needed in order to draw conclusions concerning TMT reporter ratios. We want to show, however, that using Peaks software for peptide identification (and quantification) purposes will benefit the FFPE proteome elucidation.

In conclusion, the data reported here show that hundreds of proteins extracted from diverse FFPE tissues could be successfully identified and quantified, despite the variability of the crosslink networks in different samples. The setup using GeLC-MS/MS in combination with TMT protein labeling and *de novo* sequencing algorithms delivers thus reproducible results. However, it should also be mentioned that the event of crosslinking due to formaldehyde is not completely understood, but that software packages, like Peaks including Peaks PTM are a first step in characterizing unexpected peptide modifications. However, further research concerning the crosslink event is necessary in order to obtain a more complete protein and proteome coverage.

Acknowledgements

The authors thank Karin Schildermans for technical assistance. This work was supported by the Flemish Institute for Technological Research (VITO), the Industrial Research Fund (KU

Leuven) and the Institute for the Promotion of Innovation through Science and Technology in Flanders (IWT).

Conflict of interest statement

The authors have declared no conflict of interest.

References

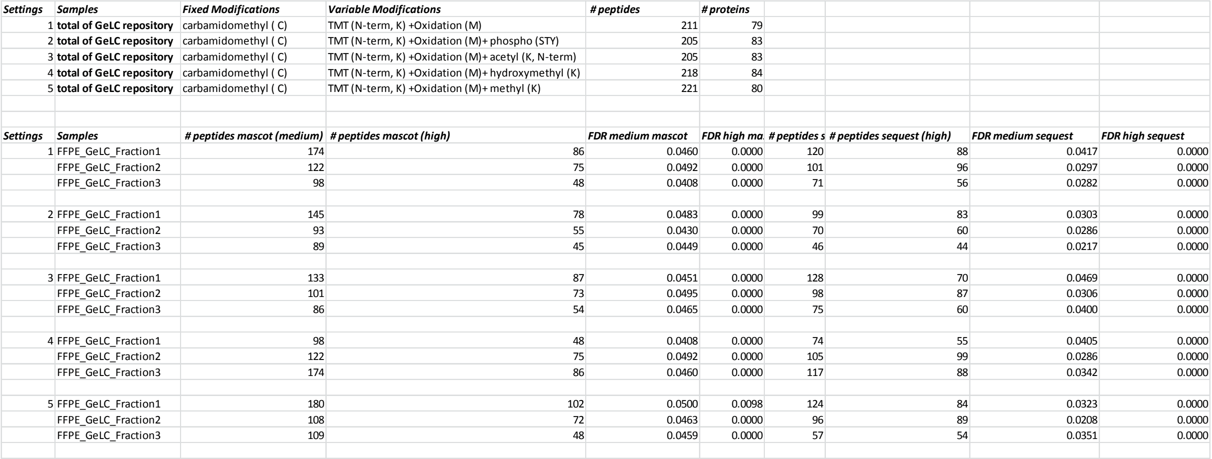
1. S. R. Shi, C. Liu, B. M. Balgley, C. Lee and C. R. Taylor, *J.Histochem.Cytochem.*, 2006, **54**, 739-743.
2. A. Tanca, D. Pagnozzi and M. F. Addis, *Proteomics.Clin.Appl.*, 2012, **6**, 7-21.
3. R. Klopfleisch, A. T. Weiss and A. D. Gruber, *Histol.Histopathol.*, 2011, **26**, 797-810.
4. J. Toews, J. C. Rogalski, T. J. Clark and J. Kast, *Anal.Chim.Acta*, 2008, **618**, 168-183.
5. E. Maes, V. Broeckx, I. Mertens, X. Sagaert, H. Prenen, B. Landuyt and L. Schoofs, *Amino.Acids*, 2013, **45**, 205-218..
6. O. Azimzadeh, Z. Barjaktarovic, M. Aubele, J. Calzada-Wack, H. Sarioglu, M. J. Atkinson and S. Tapio, *J.Proteome.Res.*, 2010, **9**, 4710-4720.
7. A. Tanca, S. Pisanu, G. Biosa, D. Pagnozzi, E. Antuofermo, G. P. Burrai, V. Canzonieri, P. Cossu-Rocca, R. V. De, A. Eccher, G. Fanciulli, S. Rocca, S. Uzzau and M. F. Addis, *Proteomics.Clin.Appl.*, 2013, **7**, 252-263.
8. N. J. Nirmalan, P. Harnden, P. J. Selby and R. E. Banks, *Mol.Biosyst.*, 2008, **4**, 712-720.
9. M. Nomura, T. Fukuda, K. Fujii, T. Kawamura, H. Tojo, M. Kihara, Y. Bando, A. F. Gazdar, M. Tsuboi, H. Oshiro, T. Nagao, T. Ohira, N. Ikeda, N. Gotoh, H. Kato, G. Marko-Varga and T. Nishimura, *J.Clin.Bioinforma.*, 2011, **1**, 23.
10. O. Azimzadeh, H. Scherthan, R. Yentrapalli, Z. Barjaktarovic, M. Ueffing, M. Conrad, F. Neff, J. Calzada-Wack, M. Aubele, C. Buske, M. J. Atkinson, S. M. Hauck and S. Tapio, *J.Proteomics.*, 2012, **75**, 2384-2395.
11. M. R. Jain, T. Liu, J. Hu, M. Darfler, V. Fitzhugh, J. Rinaggio and H. Li, *Open.Proteomics.J.*, 2008, **1**, 40-45.
12. Z. Xiao, G. Li, Y. Chen, M. Li, F. Peng, C. Li, F. Li, Y. Yu, Y. Ouyang, Z. Xiao and Z. Chen, *J.Histochem.Cytochem.*, 2010, **58**, 517-527.
13. W. Li, H. A. O'Neill and V. H. Wysocki, *Bioinformatics.*, 2012, **28**, 2548-2550.
14. Y. F. Han, B. F. Ma and K. Zhang. *J.Bioinform.Comput.Biol.*, 2005, **3**, 697-716
15. X. Han, L. He, L. Xin, B. Shan and B. Ma, *J.Proteome.Res.*, 2011, **10**, 2930-2936.
16. S. Magdeldin and T. Yamamoto, *Proteomics.*, 2012, **12**, 1045-1058.
17. A. Tanca, D. Pagnozzi, G. Falchi, G. Biosa, S. Rocca, G. Foddai, S. Uzzau and M. F. Addis, *J.Proteomics.*, 2011, **74**, 1015-1021.
18. B. M. Balgley, T. Guo, K. Zhao, X. Fang, F. A. Tavassoli and C. S. Lee, *J.Proteome.Res.*, 2009, **8**, 917-925.
19. B. Ma, K. Zhang, C. Hendrie, C. Liang, M. Li, A. Doherty-Kirby and G. Lajoie, *Rapid Commun.Mass Spectrom.*, 2003, **17**, 2337-2342.
20. J. Zhang, L. Xin, B. Shan, W. Chen, M. Xie, D. Yuen, W. Zhang, Z. Zhang, G. A. Lajoie and B. Ma, *Mol.Cell Proteomics.*, 2012, **11**, M111.
21. R. W. Sprung, Jr., J. W. Brock, J. P. Tanksley, M. Li, M. K. Washington, R. J. Slebos and D. C. Liebler, *Mol.Cell Proteomics.*, 2009, **8**, 1988-1998.
22. P. McIlwain S Fau - Draghicescu, P. F. Draghicescu, P. F. Singh, W. S. Goodlett Dr Fau - Noble and W. S. Noble, *J.Proteome.Res.*, 2010, **9**, 2488-2495.
23. B. Metz, G. F. Kersten, G. J. Baart, J. A. de, H. Meiring, H. J. ten, M. J. van Steenbergen, W. E. Hennink, D. J. Crommelin and W. Jiskoot, *Bioconjug.Chem.*, 2006, **17**, 815-822.
24. J. F. Sinclair and J. F. Timms, *Methods*, 2011, **54**, 361-369.
25. A. Mantovani, P. Allavena, A. Sica and F. Balkwill, *Nature*, 2008, **454**, 436-444.

Tables

Table 1: Information on the paired colon mucosa samples used in the labeling experiment

Healthy tissue		**Cancer tissue**		
Sample	*Labeling tag*	*Sample*	*Cancer stage*	*Labeling tag*
Sample 1a	TMT 126	Sample 1b	T4N0	TMT 127
Sample 2a	TMT 128	Sample 2b	T1N0	TMT 129
Sample 3a	TMT 130	Sample 3b	T2N1	TMT 131

Table 2: Overview of search parameters and identifications using standard search engines



Settings	*Samples*	*Fixed Modifications*	*Variable Modifications*	*# peptides*	*# proteins*
1	**total of GeLC repository**	carbamidomethyl (C)	TMT (N-term, K) +Oxidation (M)	211	79
2	**total of GeLC repository**	carbamidomethyl (C)	TMT (N-term, K) +Oxidation (M)+ phospho (STY)	205	83
3	**total of GeLC repository**	carbamidomethyl (C)	TMT (N-term, K) +Oxidation (M)+ acetyl (K, N-term)	205	83
4	**total of GeLC repository**	carbamidomethyl (C)	TMT (N-term, K) +Oxidation (M)+ hydroxymethyl (K)	218	84
5	**total of GeLC repository**	carbamidomethyl (C)	TMT (N-term, K) +Oxidation (M)+ methyl (K)	221	80

Settings	*Samples*	*# peptides mascot (medium)*	*# peptides mascot (high)*	*FDR medium mascot*	*FDR high ma*	*# peptides s*	*# peptides sequest (high)*	*FDR medium sequest*	*FDR high sequest*
1	FFPE_GeLC_Fraction1	174	86	0.0460	0.0000	120	88	0.0417	0.0000
	FFPE_GeLC_Fraction2	122	75	0.0492	0.0000	101	96	0.0297	0.0000
	FFPE_GeLC_Fraction3	98	48	0.0408	0.0000	71	56	0.0282	0.0000
2	FFPE_GeLC_Fraction1	145	78	0.0483	0.0000	99	83	0.0303	0.0000
	FFPE_GeLC_Fraction2	93	55	0.0430	0.0000	70	60	0.0286	0.0000
	FFPE_GeLC_Fraction3	89	45	0.0449	0.0000	46	44	0.0217	0.0000
3	FFPE_GeLC_Fraction1	133	87	0.0451	0.0000	128	70	0.0469	0.0000
	FFPE_GeLC_Fraction2	101	73	0.0495	0.0000	98	87	0.0306	0.0000
	FFPE_GeLC_Fraction3	86	54	0.0465	0.0000	75	60	0.0400	0.0000
4	FFPE_GeLC_Fraction1	98	48	0.0408	0.0000	74	55	0.0405	0.0000
	FFPE_GeLC_Fraction2	122	75	0.0492	0.0000	105	99	0.0286	0.0000
	FFPE_GeLC_Fraction3	174	86	0.0460	0.0000	117	88	0.0342	0.0000
5	FFPE_GeLC_Fraction1	180	102	0.0500	0.0098	124	84	0.0323	0.0000
	FFPE_GeLC_Fraction2	108	72	0.0463	0.0000	96	89	0.0208	0.0000
	FFPE_GeLC_Fraction3	109	48	0.0459	0.0000	57	54	0.0351	0.0000

Table 3: Overview of PTM profile of FFPE samples using PeaksPTM

Name	*Δ Mass (Da)*	*# PSM*	*Position*
TMT6-plex	229,16	1066	K,N-term
Oxidation	15,99	367	M
Hydroxylation	15,99	279	DKPR
Dihydroxy	31,99	166	FKPRW
Deamidation	0,98	105	NQ
Carbamidomethylation	57,02	104	C
Acetylation	42,01	71	N-term
Acetylation	42,01	58	Protein N-term
Methyl ester	14,02	37	DE,C-term

Deoxy	-15,99	31	T
Cation:Fe	53,92	28	DE
Acetylation	42,01	26	K
Dehydration	-18,01	21	DST
Carbamidomethyl	57,02	11	DEHK,N-term
Sulfone	31,99	11	M
Formylation	27,99	11	K,N-term
TMT	224,15	8	N-term
Formylation	27,99	7	Protein N-term
Didehydro	-2,02	6	T,C-term
Oxidation	15,99	5	W

Figure Legends

Figure 1: Identification and quantification of proteins

Using PeaksPTM, several unspecified modifications were found. Panel A shows the sequence coverage of tropomyosin 4, isoform 1, in which two modifications were found in two unique peptides which made it possible to confidently identify this protein. Panel B shows a fragmentation spectrum of a unique protein.

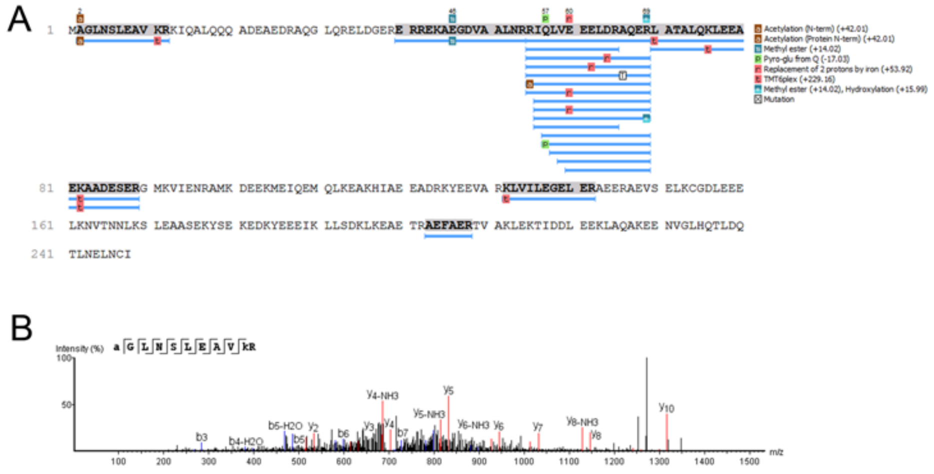


Figure 2: Fragmentation spectra with six reporter ions

In panel B, the sequence coverage of histone H2A, type 1C is shown. Several peptides have TMT modifications, pointing out that the TMT labeling was efficient. Moreover, in their corresponding fragmentation spectra (panel A), all six reporters ions were found, which makes quantification feasible. Panel C shows some high energy collision dissociation spectra, in which the efficiency of reporter ion formation is higher.

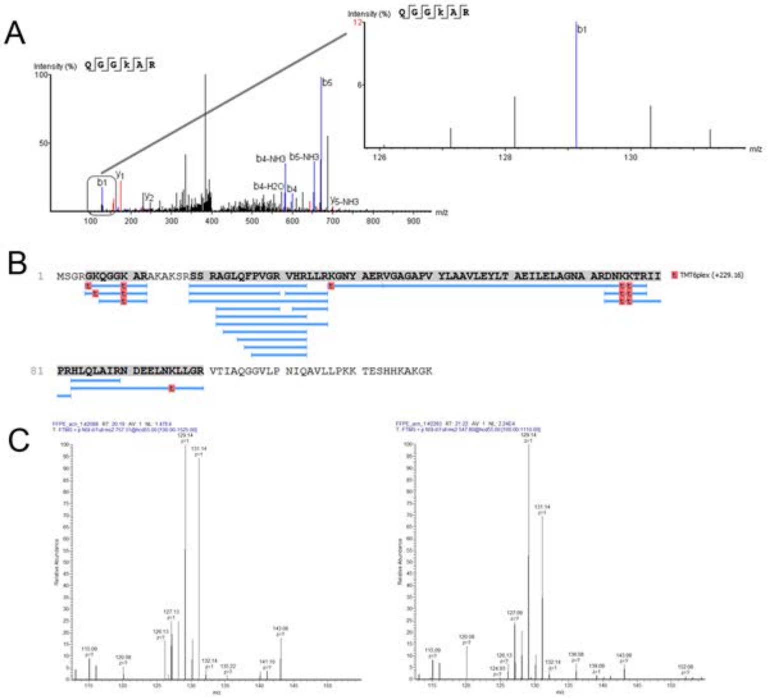
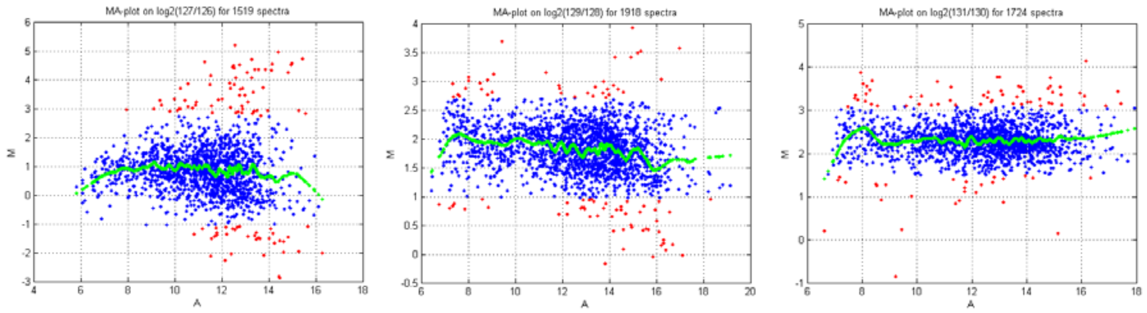


Figure 3: Quality control of TMT protein labeling

The MA-plots of the different paired tumor versus control samples (127/126, 129/128 and 131/130) are shown. Although the center of the ‘cloud’ should be found around 0, assuming that most peptides are not differentially expressed, several deviations can be seen, which means normalisation has to be performed.



Supplemental data

Table S1: Overview of all identified proteins found in the ACN dataset. All identified proteins contain at least one unique peptide and the overall protein score (-10logP) must be higher than 20 (FDR< 5%) to be selected as a confident identification.

Table S2 : Overview of all quantified proteins found in the ACN dataset. In order to be selected as quantified protein, reporter intensities of each of the six reporter ions should be available.

Figure S1: Heatmap of quantified proteins. Trends in upregulation of tumor samples (TMT 127,129,131) can be seen compared to control colon mucosa (TMT 126,128,130.