SCUOLA DI DOTTORATO
UNIVERSITÀ DEGLI STUDI DI MILANO-BICOCCA

Department of Medicine and Surgery

PhD program in Public Health Cycle XXXII
Curriculum in Biostatistics and Clinical Research

APPLICATION OF MALDI-IMAGING PROTEOMICS ANALYSIS ON THYROID BIOPSIES: IDENTIFICATION OF BIOMARKERS FOR CLINICAL DIAGNOSIS

Candidate: CAPITOLI GIULIA

Registration number : 823671

Tutor: Maria Grazia Valsecchi

Co-tutor: Stefania Galimberti

Coordinator: Guido Grassi

ACADEMIC YEAR   2019/2020

Table of contents

Abstract

The actual gold standard to exclude the malignant nature of thyroid nodules in the clinical routine is represented by thyroid Fine Needle Aspirations (FNAs) biopsies. Thyroid FNAs are safe and cost-effective. Approximately the 20-30% of cases have an indeterminate for malignancy final report. These patients undergo diagnostic (and not therapeutic) thyroidectomy, but after surgery the 80% of these thyroid nodules are benign. This overtreatment has of course important consequences in the quality of life of the patients and high healthcare costs. The application of -omics techniques might have a potential role in the research for new diagnostic markers able to discriminate benign from malignant nodules, thus minimizing the challenging cases of indeterminate for malignancy.

Mass spectrometry is one of the most important analytical tools able to obtain information regarding the molecular composition of a sample, the presence of biomolecules and their abundance. Among the different proteomics approaches able to extract the molecular alterations of the different type of specimen's lesion, Matrix-Assisted Laser Desorption/Ionization (MALDI) Mass Spectrometry Imaging (MSI) was strongly emerging. MALDI-MSI represents an ideal technology that enables to explore the spatial distribution of biomolecules within tissue, integrating molecular and traditional morphological information while preserving the integrity of the analysed tissue. Various studies applied MALDI-MSI technology for prognostic purposes and for in real time diagnostic setting, showing the usefulness, advantages and applicability of MALDI-MSI in different fields of pathology. Due to the promising results recently obtained with MALDI-MSI in the identification of proteomic signals able to differentiate between benign and malignant cases from the analysis of thyroid tissue after surgery, the idea was to apply for the first time MALDI-MSI on real thyroid FNAs biopsies.

Preliminary to the clinical study, the protocol for the proteomic MALDI-MSI analysis was optimised to avoid degradation, alteration phenomena, contamination and artefacts formation. The methodological improvement of the protocol in a complicated field as thyroid cytological specimens played an important role in this study. Challenging technical aspects, such as i) the interference of haemoglobin due to the high vascularization of the thyroid organ and ii) the stability of the samples over time before the analysis from a morphological and proteomic point of view, were overcome through two studies that were planned and analysed as part of the thesis.

The clinical study for the detection of the potential cluster of signals with discriminant capability was originally planned to involve a large sample of thyroid nodules, however, due to the slow enrolment rate of malignant cases, the thesis contains only the results of a preliminary analysis. Eighteen subjects contributed to the training set with 9 benign and 9 malignant thyroid nodules. However, the statistical model was based on data of 81 specific regions of interest, according to the morphological triage performed by the pathologist in order to overcome false information deriving from non thyrocytes cells. The validation phase was performed on 11 patients with different type of lesions (i.e. benign, indeterminate and malignant). Results are very promising and highlight the possibility to introduce MALDI-MSI as a complementary tool for the diagnostic characterization of thyroid lesions, but a further analysis on a more consistent sample of patients is required to corroborate these findings.

A methodological aspect that emerged from the peculiarity of the proteomic analysis was also investigated as part of this thesis. A review of the most used statistical indices for the assessment of the similarity between mass spectra profiles was performed and a new measure was proposed. A simulation study was implemented in order to identify the best similarity measure to use in comparing proteomic profiles.

1 Introduction

The diagnosis of thyroid lesions is usually performed using a morphological approach on image-guided fine needle aspiration biopsy (FNAB). Even if thyroid FNAs are safe, cost-effective and represent the gold standard to exclude the malignant nature of thyroid nodules, approximately 20-30% of FNABs are considered as "indeterminate for malignancy" and, in these cases, surgery is commonly recommended. Nevertheless, post-operative histological evaluation highlights that 80% of these uncertain diagnoses are benign lesions. These patients undergo diagnostic and not therapeutic thyroidectomy and this overtreatment has important consequences in their quality of life and high healthcare costs.

Omics techniques play an important role in the research for new diagnostic markers and the use of molecular techniques for the characterization and identification of biomarkers persists as an interesting topic in clinical applications. Different molecular test pointing to gene-expression classifiers had been proposed to improve the pre-operative risk assessment of malignancy on thyroid FNABs, but these methods showed disadvantages in terms of cost-effectiveness that limit their use in clinical routine.

Matrix-assisted laser desorption/ionization mass spectrometry imaging (MALDI-MSI) is a powerful tool to explore the spatial distribution of biomolecules directly on cytological specimens, by integrating molecular and morphological evidence. The identification of a powerful tool for assisting cytopathologists in thyroid lesions diagnosis had clinical, ethical and economical relevance. Preliminary studies showed how MALDI-MSI is able to distinguish benign with respect to malignant cases in different cytological samples taken from surgical thyroid nodule (*ex-vivo* cytological samples). Moving forwards from these first results, the aim of the clinical project that motivated my thesis was to apply for the first time MALDI-MSI on real thyroid FNABs (*in-vivo* cytological samples) and to test its possible complementary role in the diagnosis of thyroid nodules. In particular the main goal of the project was to verify the capability of the MALDI-MSI approach in solving "indeterminate for malignancy" and "suspicious for malignancy" cases.

In this 3-year project a consecutive series of more than 1000 patients were expected to be enrolled based on the potentiality of recruitment at the San Gerardo Hospital (Monza, Italy). Morphological cytological FNA diagnosis were obtained for all the patient according to 5-tiered system (THY1: unsatisfactory material, THY2: benign, THY3a: low risk indeterminate for malignancy, THY3b: high risk indeterminate for malignancy, THY4: suspicious for malignancy, THY5: malignant). The design was planned to recruit 160 THY2 and 80 THY5 patients for inclusion in the training set for the construction of the classifier based on proteomic data, while additional 40 THY2, 150 THY3, 60 THY4 and 20 THY5 FNABs foreseen for the validation phase of the study.

This thesis is organized as follows. The general clinical context that characterizes the diagnosis of thyroid cancer and a summary of the study protocol is illustrated in **Chapter 2**, where the proteomic analysis by MALDI-MSI is also briefly described from the sample collection of FNABs to the pre-processing of mass spectrometry data. An overview of the main statistical approaches to deal with high dimensional data is presented in **Chapter 3** and the application of these methods to the omics field, in particular the proteomic one, is discussed.

The practical implementation of the clinical project is described in **Chapter 4**. This chapter contains the results of two important steps in the fine tuning of the protocol for the proteomic MALDI-MS analysis that deal with the standardization of the sample preparation workflow of *ex-vivo* and *in-vivo* thyroid FNABs in order to transfer the MALDI-MSI model to routine cytological specimens. Chapter 4 contains also the results of the preliminary analysis of the clinical study where proteomic data of a training set were used for the classification of thyroid lesions in a validation sample.

One of the statistical challenges that originated from this project was the problem of the assessment of the mass spectra similarity. A review of the main approaches existing in the literature to assess this issue is presented **in Chapter 5**, together with a proposal developed specifically for this purpose. Results of a simulation study that was set up to investigate the performance and the reliability of different similarity measures are also reported in Chapter 5. Some final remarks are given in **Chapter 6.**

2 The clinical and proteomic landscape

2.1 Clinical context

An increasing incidence of thyroid cancer had been reported in the last decades due to the primary detection of small tumour nodules in the preclinical stage [1] [2]. The thyroid is a bilobular endocrine gland that is located anteriorly in the lower neck [3]. The main purpose of this organ is to produce, store and secrete the iodine-based hormones triiodothyronine (T3) and thyroxine (T4); they act on fat, protein and carbohydrate metabolism, as well as on the development of central nervous system and general growth. The thyroid hormones are strictly regulated by the Hypothalamus-Pituitary-Thyroid axis (HPT) via the secretion of thyroid regulating hormones (TRH, from hypothalamus) and thyroid stimulating hormone (TSH, from pituitary gland) [4].

Palpable thyroid nodules are present up to the 10% of the adult population. Ultrasound could detect up to 70% of palpable and not palpable nodules that were identified during the execution of an imaging test for other indications [5]. Prevalence is higher in women, elderly and in iodine insufficient areas and the frequency increases with age [6]. In Italy, thyroid nodules are the second most frequent cause of cancer in women under 45 years [7]. Only 5-15% of patients are actually affected by malignant thyroid lesion, so the first purpose is to exclude malignancy [5].

2.1.1 Diagnostic iter and follow-up

The main objective of the ultrasound evaluation of a thyroid nodule is to determine whether the lesion should be evaluated *via* FNA or subjected to ultrasound follow-up [8]. Since ultrasound is an operator-dependent technique and because of the complexity in the interpretation of the ultrasound images, thyroid lesions had been stratified in different risk of malignancy in order to standardise the diagnostic procedure and determine the appropriateness of the execution of FNA [9].

The first diagnostic procedure identified by the guidelines of the American Thyroid Association (ATA) and the American College of Radiology (ACR) involves the evaluation of multiple ultrasound characteristics as: solid aspect, hypo echogenicity, micro calcifications, irregular borders, absence of peripheral halo, intranodular blood flow and shape [5]. Therefore, the ATA and ACR guidelines suggested to evaluate thyroid nodules based on combinations of ultrasound characteristics, stratifying nodules in 5 groups with a different risk of malignancy [10] (Figure1). Biopsies are not recommended for nodules with a diameter lower than 10 mm, since a small nodule is usually not a cancer. In the presence of some echographic criteria of suspicion, small nodules with high growth rate during follow-up are eligible for FNA [9]. Currently, ultrasound technique is able to detect nodules with a diameter under 5 mm without difficulty. These nodules can potentially be biopsied with good results for safety and diagnostic performance for the patients [11].

Although ultrasonography is a good diagnostic tool, the diagnosis of thyroid lesions is performed by image-guided fine needle aspiration biopsy that currently represents the main approach to exclude the malignant lesion of thyroid nodules in patients with echography suspicious features [11]. An experienced pathologist or technician performs all aspirations. Pathologist evaluates 1 to 3 slides prepared as smears for each FNAB needle pass for traditional morphological diagnoses. Samples are then classified according to the 5-tiered Italian Society for Anatomic Pathology and Cytology and the Italian Division of the International Academy of Pathology (SIAPEC-IAP) system [12] (Figure 1). A sample is defined as not representative when the number of cells is insufficient for diagnoses: it is required the presence of at least 6 groups of 10 well preserved cells [13].

The World Health Organization (WHO) divided thyroid neoplasms into benign lesions as follicular thyroid adenomas (FTA), hyperplastic lesions (HP) to differentiated carcinomas, such as papillary thyroid carcinomas (PTC, nearly 90% of thyroid cancer), follicular thyroid carcinomas (FTC), anaplastic thyroid carcinomas (ATC) and medullary thyroid carcinoma (MTC, represents only the 5% of malignant lesions).

A classification of benign (THY 2) is assigned in 60-70% of thyroid FNAs. In the majority of the cases, THY4 and THY5 are diagnosed as classical PTC or as follicular variant of PTC (fvPTC) whereas the THY3 category, defined as "indeterminate for malignancy", may include FTA, non-invasive follicular thyroid neoplasms with papillary-like nuclear features (NIFTP), Hurtle cells carcinoma, PTC and lesion of uncertain malignant potential (UMP) [14]. A surgical approach for the THY3 samples is recommended by the international guidelines [15]. After total thyroidectomy, 80% of these cases result benign [12], with important implications in terms of healthcare costs, operative risks and morbidity, and potential need for a lifelong hormone replacement therapy.

Before the routine use of FNA, only the 14% of the resected thyroid resulted as malignant, after FNA practice the percentage increased further than 50% [15]. The diagnostic accuracy of FNA was nearly 90%, and the percentage of false positive and false negative was less than 3%, except for FTA [13].

To improve the pre-operative risk assessment of malignancy on thyroid FNAs different molecular tests, as genetic testing (BRAF, N-H-KRAS point mutations and RET/PTC1, RET/PTC3, PAX8/PPAR rearrangements) or gene-expression classifiers (Veracyte, Thyroseq) had been proposed. These methods show several disadvantages in terms of cost-effectiveness that limit their use in routine diagnostics, and sometimes are inconclusive (i.e. 50% of malignant cases were BRAF negative).

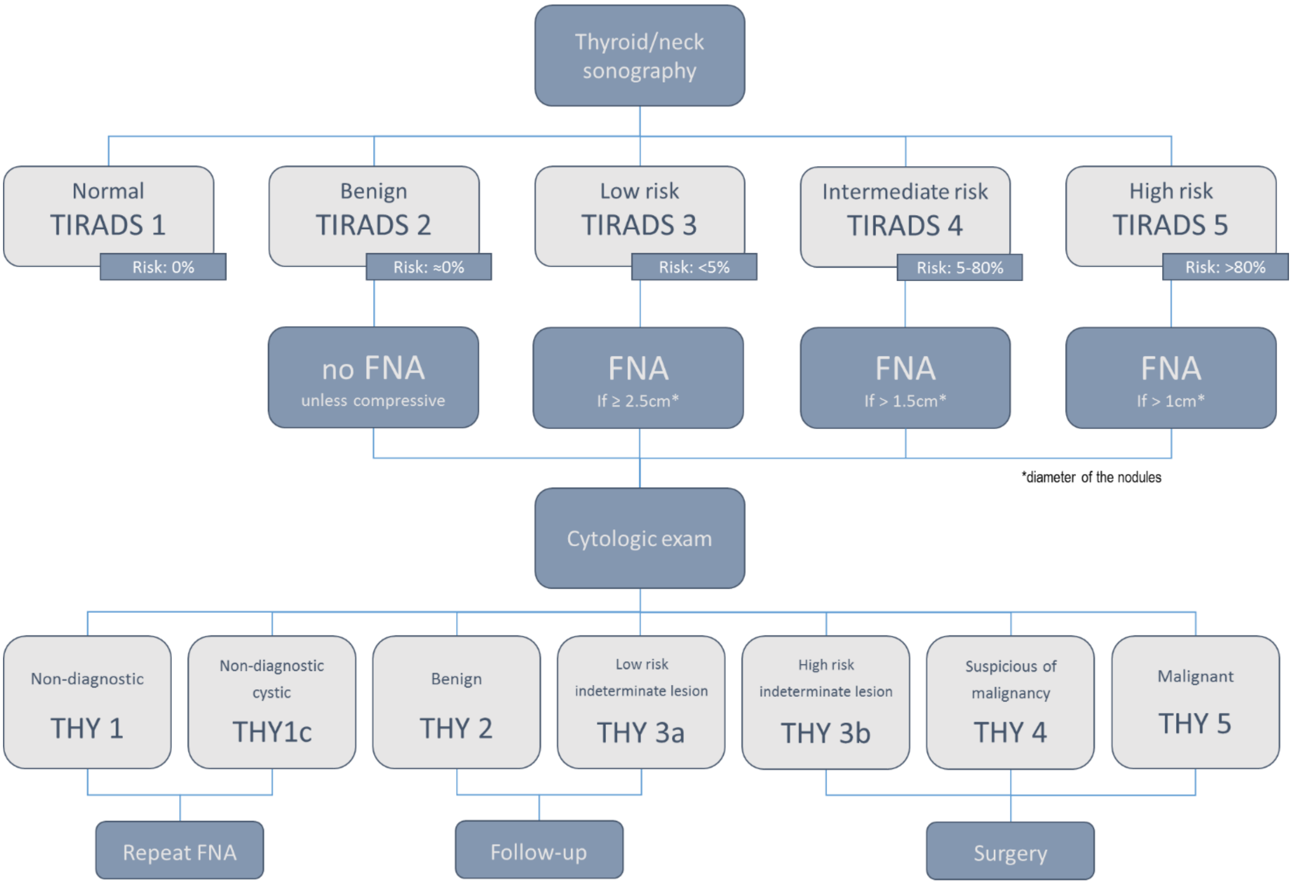


Figure 1 ATA and ACR guidelines for thyroid nodules based on combinations of ultrasound and cytological characteristics, stratifying nodules in groups with a different risk of malignancy

2.1.2 The clinical protocol

Patients population: consecutive subjects admitted to the Ultra Sound (US)-guided FNA ambulatory of the ASST MONZA-BRIANZA (San Gerardo Hospital HSG-UNIMIB, Monza, Italy) from 1 June 2017 to 1 June 2019.

FNA: a standard procedure of US-guideed FNA that includes a minimum of 2 passes for nodule. Only "needle washing" from leftover material of every pass was collected and send to MALDI examination.*Cytologic diagnosis:* a morphological cytological FNA diagnosis obtained from all subjects according to the 5-tiered reporting system of the British guidelines.

Histology and follow-up: the cytological diagnosis had been differentially confirmed. In particular, BENIGN-THY2 cases were certified by performing a US examination 12-months after the first US-guided FNA and confirming: i) absence of new echographic malignant features ii) absence of significant increasing nodule size iii) absence of nodes metastasis iv) no incidence of new suspicious nodules. For malignant cases, histological diagnoses were progressively collected after thyroidectomy to certify the nature of the nodules.

Potentiality of recruitment: around 700 subjects per year for a total of 1400 in a 24-months recruitment period.

Sample size: Training-phase: a) 160 cases with a clear-cut benign diagnosis (THY2); b) 80 cases with a clear-cut malignant diagnosis (THY5).

Validation-phase: a) 150 indeterminate (THY3), b) 50 suspicious (THY4), c) 40 benign (THY2) and d) 20 malignant (THY5) Fine Needle Aspiration Biopsies (FNABs) with a subsequent clear-cut diagnosis based either on follow-up or histology.

The sample size was estimated on the basis of the mean difference in peak intensities of protein expression, assuming that the base 2 logarithm of peak intensities has a Gaussian distribution. The MALDI analysis of a total of 240 patients in the training phase, according to a ratio 2:1 between THY2 and THY5, have a 90% power to detect a 1.5-fold change in the mean intensities of the two groups, with an α error of 0.001 and assuming one technical replicate, two-sided test and the variance in the log-peak equal to 0.9 (data from previous experiences on thyroid tissue). Even in presence of a higher level of variability, with a variance of 1, our study would have a power of 80%, to show as statistically significant the same difference. This was an highly powered study for the discovery of new markers based on proteomic profiling that properly controls for the False Discovery Rate due to multiple testing, since around 150 peaks were tested for differences in mean intensities at the biomarker discovery stage.

Data collection: for each recruited patient demographic and clinical information were collected, including: age, sex, number of nodules, echographic parameters, presence of autoimmunity, medical history, concomitant therapies, serological TSH levels, history of radioactive exposure.

Ethics: - the study was approved by the Ethical Committee of the ASST MONZA-BRIANZA;

-subjects provided signed informed consent.

2.2 Mass spectrometry

Despite FNA is considered the standard procedure for thyroid nodules diagnoses, due to the fact that is an easy-to-perform technique, cost-effective and minimally invasive method [16], collected samples usually contain few tumour cells. Limiting the usefulness of cytological analyses could create ambiguity in the diagnoses. To address this problem, several groups had used proteomics approach to find potential markers for thyroid tumours from FNA samples and other techniques, such as fresh-frozen thyroid tissue specimens obtained after thyroidectomy and serum samples [17].

Proteomics represents a possible complementary analytical strategy that was just routinely used in microbiology. In this regard, Matrix assisted laser desorption/ionization (MALDI) mass spectrometry imaging (MSI) is a new proteomics technology that explored the composition of biomolecules and their spatial distribution in-situ [18]. MALDI imaging had already been used to build proteomic signatures of carcinoma in different organs such as oesophagus, breast, colon, liver, kidney, stomach, and thyroid gland using histological tissues [19][20]. The possibility to investigate material collected by FNAB, available before thyroidectomies (in-vivo FNAB), for the preoperative diagnostic phase of thyroid tumours, could reduce the number of unnecessary surgeries.

In the last years the interest in the application of molecular techniques for the diagnoses of thyroid lesions has grown, and MALDI-MSI had been used to analyse the proteomic profile in thyroid tissues and FNA samples [21][22]. Several groups have worked in this field, as shown in the following examples. Because histological analyses on thyroid surgical samples was still the most efficient one due to the high quality of the sample, several groups developed different strategies and optimised protocol to allow a MALDI-MSI proteomic investigation on surgical specimens, fresh frozen (FF) and Formalin-Fixed Paraffin-Embedded (FFPE) tissue specimens based on tryptic peptide extraction after enzymatic digestion [23][24].

In 2017 Pietrowska et al. focused their studies on distinguishing different types of thyroid cancer on thyroid tissue samples, and validated a proper classification of MTC and anaplastic cancers [25]. In the same years, Galli et al. used MALDI-MSI to investigate Tissue Microarrays (TMAs) on different type of thyroid nodules, such as HP, FTA, PTC and fvPTC [21]. A group of proteins able to discriminate between HP and FA or HP and PTC was identified. Moreover, MALDI-MSI showed the possibility to highlight the heterogeneity of those TMA samples that contained both benign and malignant cells. Different groups worked also on rare forms of thyroid cancer, as MTC, and their studies showed an high sensitivity and specificity to detect MTC [25][26].

MALDI-MSI has been performed also on cytological specimens, but the approach on this type of samples was different. Complementary results on FNA smear samples (liquid biopsy was smeared on a glass microscope slide) were obtained in 2016 by Pagni et al.. The application of MALDI-MSI on ex-vivo FNAB (FNAB after thyroidectomies) showed the possibility to correlate morphological information with protein expression in the distinction between malignant (PTC) and benign lesions [22]. MALDI-MSI demonstrated the ability to distinguish not only benign vs malignant lesions, but to reveal significant differences between thyroid lesions with similar pathological behaviour. Potential discriminative features were found in the malignant group, where PTC and MTC lesions showed independent proteomic profiles.

A recent work published in 2019 on thyroid lesions, demonstrated the potential use of metabolomic analysys (by Desorption Electrospray Ionization Mass Spectrometry DESI-MS imaging) on FNA smears to reduce the number of unnecessary diagnostic thyroidectomies [27]. Molecular signatures of benign vs FTC and benign vs PTC were found.

These studies showed how MALDI-MSI, that was an emerging approach, allows to provide specific molecular profiles of the thyroid lesions not only on surgical specimens, but also on standard FNAB samples used for the clinical routinely diagnoses.

The imaging approach on in-vivo FNAB specimens to individuate putative discriminant biomarkers is still relatively new. The addition of MALDI-MSI into the clinical routine, to improve the diagnoses of thyroid nodules especially for indeterminate for malignant cases, is promising. Potentially the number of not therapeutic surgery can be reduced, improving the life (style) of patients and high healthcare costs.

A consistent part of this thesis has been dedicated to solve technical aspects on sample preparation, such as morphological and protein profile stability, and haemoglobin interference problem that suppressed any other protein signature in untreated samples. The technical details reported in the following sections are the results of the studies whose results are reported in Chapter 5.

2.2.1 Sample Preparation

The design of a robust and simple protocol, by focusing on the morphological and protein stability of the sample and the repeatability of the workflow, was of great importance.

Samples were collected by performing 3 or 5 needle passes with a 25-Gauge needle and immediately transferred into a falcon tube with preservative mediums, such as CytoLyt solution. This solution allowed to preserve the morphology of the samples during transportation and over time, to prevent protein precipitation and the lysis of red blood cells.

Cytological samples deposited into CytoLyt solution were centrifuged for 10 minutes at room temperature (RT) to separate cells and aggregates from the cyst fluid, then the pellet was re-suspended in CytoLyt solution again and the supernatant was discarded and the procedure iterated [28][29][30][31]. Finally, samples were transferred onto indium tin oxide (ITO) conductive slides as small cytospin spot and measurement of total protein concentration was performed by using a spectrophotometer.

The amount of material was usually scarce; a limiting number of needle passes can be performed due to the aspiration being performed in living patients who do not undergo anaesthesia. When the amount of cellular material was enough the specimen was equally divided into multiple spot in order to obtain multiple replicates. A maximum of eight cytospin spots can be positioned onto one ITO-conductive slide.

Finally, dry and washing steps were performed. Cytospin samples onto ITO-slides were dried under vacuum two times, respectively for 30 and 15 minutes, interspersed by a consecutive washing steps of 30 seconds each, with increased concentration of ethanol (70%, 90% and 95%) in order to remove salt and lipid contamination that could unfavourably affect the quality of MALDI-MSI data [32][33]. Then ITO-slides were stored at -80°C until the day of the analyses at the spectrophotometer.

Before MALDI-MSI analyses, cytospin spots were stabilised to room temperature, dried under vacuum for 30 minute, and the MALDI-matrix sinapinic acid was uniformly deposited, using the iMatrixSpray automated spraying system. Different types of matrix exist, depending on the nature of the analyte to analyse and the mass range of the analyses.

The analysis reported in this study were performed with a MALDI- time of flight (MALDI-TOF) mass spectrometer. This type of instrument is one of the most commonly used, since it allows to analyse a wide variety of molecules such as proteins, peptides and lipids. After the solvent evaporates, the matrix co-crystallises with the molecules of the sample. When the laser of the MALDI instruments hits the sample, the matrix absorbs the energy and transfer it from the laser to the analyte molecules, which are now ionised, causing their detachment from the ITO-slides. Ions extracted from the sample, enter and move in a drift space free from electromagnetic fields under vacuum.

The mechanism of operation of a TOF analyser is very simple: it separates ions according to their velocity and examines the time that an ion took to run across the flight tube and the amount of ions hitting the detector concurrently. Ions with different m/z will arrive separately to the detector because the lighter ions of the same charge fly faster than the heavy ones and arrive to the detector earlier. The detector records the impacts of the ions and transform them into electric signals, obtaining mass spectra for each sample. Signal abundance is directly proportional to the quantity of the same ions hitting the detector plate.

Concluding, the molecules present in the sample are ionised in the source and separated in the TOF analyser depending on their mass to charge (m/z) ratio and finally a mass spectrum is generated, with the m/z ratio in the x-axis and the relative aboundance (intensity) on the y-axis.

The TOF analyser can work in two different ways: linear and reflectron (Figure 2). In the first one, the analytes are detected at the end of the flight tube. In the reflected mode, ion mirrors reflect the ions and an electrical field is applied before the detector to increase the resolution of the mass analyser, i.e. incrementing the ability of the analyser of separating two ions with small difference in their mass to charge ratios (m/z), generating two different peaks in the spectrum. In this work linear MALDI-MSI was used.

After the MALDI analyses, the matrix was removed with 70% concentration of ethanol and the ITO-slides were stained and digitally scanned in order to directly correlate the molecular information to the morphological data.

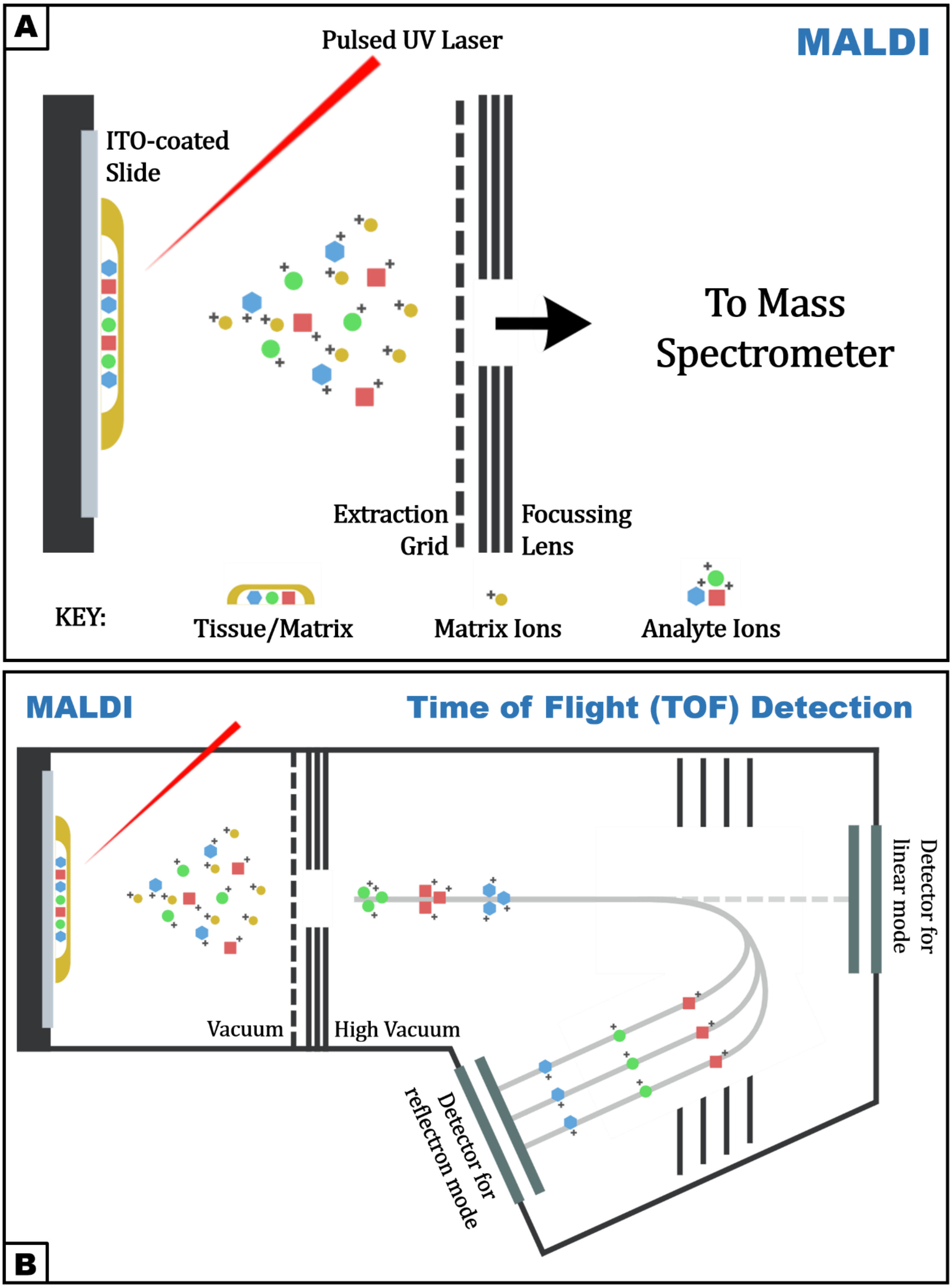


Figure 2 MALDI mass analyser can work in two different ways: in linear mode (A) or in reflectron mode (B) which add an electrical field and a mirror to reflect the flight of the ions incrementing the ability to separate them.

2.2.2 Data acquisition

MSI technique maps the biological molecules as proteins, peptides, lipids and metabolites visualizing their distribution in the biological sample [34]. To generate MSI data with a MALDI instrument, the sample was divided into multiple pixels depending on the laser characteristic. The distance between two consecutive shots (raster) depends on the laser size, a reduction of the raster increases the spatial resolution leading to the visualisation of single cells [35].

Mass spectra (MS) were acquired for each single point (pixel) in which the laser beams hit the surface of the sample. Once we arranged the mass spectra data by their pixel positions, we obtained a set of MSI data structured as a data cube. The acquisition of one single spectrum for each pixel led to the possibility to reconstruct the spatial distribution of a given analyte (at a specific m/z value) on the sample. This made up an intensity image, which maps the distribution of the specific molecules on the specimens, colouring image pixels according to the abundance in each spectra of the *m/z* chosen (Figure 3). The advantage of MSI instead of MS data is that MSI also incorporate spatial information in MSI proteomic data.

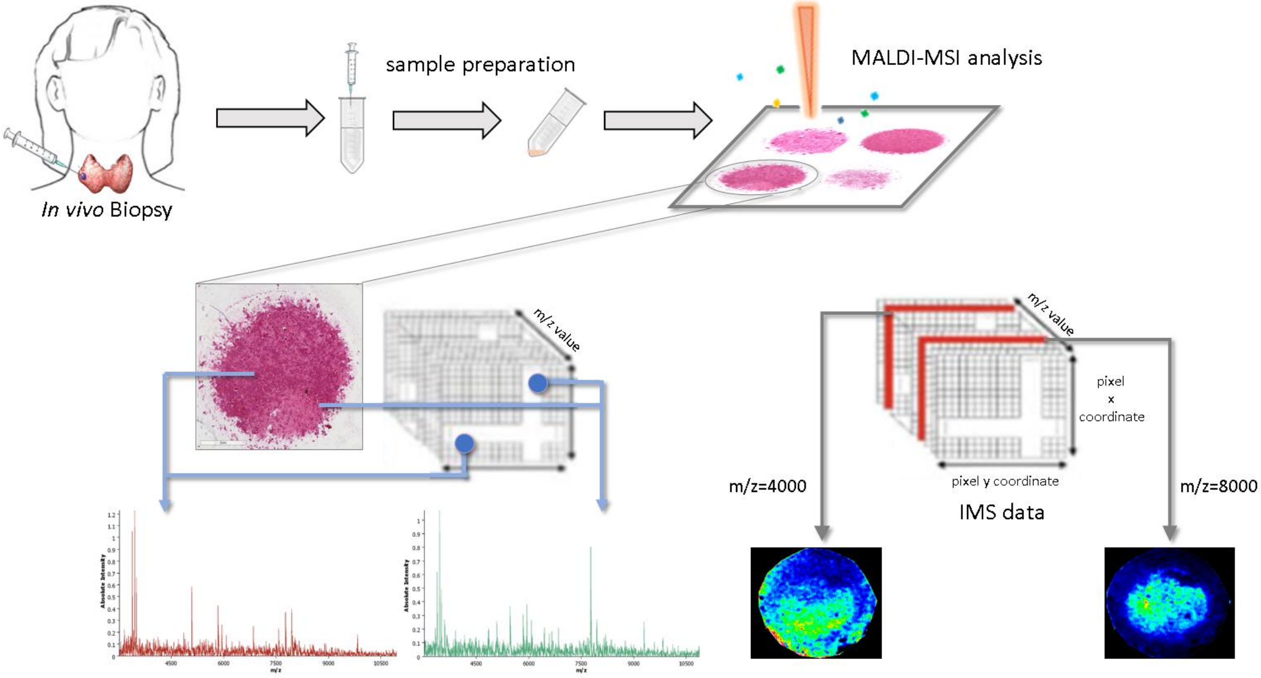


Figure 3 Experimental plan: from thyroid biopsies to MALDI-MSI analysis, spectra generation and imaging

For each sample, an overall average spectrum of the entire specimen, single spectra from each pixel of the sample or average spectra of regions of interest (ROIs) could be obtained. ROIs contain pathological areas of interest, annotated by the pathologist, to reduce the bias made by the not informative pixels, only the part of cancer and non-cancer cell were selected.

The average spectrum was obtained calculating the mean value of intensity of each *m/z*. The average evens the spatial molecular information, specific signals (peaks) of a defined area might not be represented in the average spectrum. It would have been suitable if the samples were homogeneous and all single spectrum had similar peak intensities.

2.2.3 Pre-processing:

Before statistical analyses, data have to be visualised and pre-processed in order to smooth the intra and inter-sample variability in intensities and *m/z* localisation due to technical variations as sample preparation and mass spectrometric instrumentation [36][37].

The elaboration phase was divided in five steps: baseline, smoothing, normalisation, alignment and peak picking.

Baseline:

The baseline is a line that connects all the lowest value of the spectrum, usually with an exponential shape; it is related to electrical noise and chemical impurities in the sample. The baseline noise is estimated and subtracted from the original spectrum, bringing all the intensities to start from zero. A high baseline leads to false intensity values. Several algorithms could be performed. The SNIP method replaces for each data point the minimum between the data itself and the mean of the extreme of a define window, centred in the data point observed. Small window can lead to a strength erosion of peaks intensity, so the width of the window has to be greater than the base of shape of the peak considered; large window leads to a loss of small peaks and intensity of small m/z values, cutting the base of peaks. The Top-Hat method applies a two erosion filter, a moving minimum erosion filter and subsequently a moving maximum dilation filter. As the SNIP method, it depends on a moving window, whose size could lead to artefacts or loss of information. The Convex Hull method connects the extremities of the spectrum with a convex curve. It does not take into account the local variations of the spectrum, losing informative portions of the mass spectrum. The Median filter substitutes each data point with the median of the define window of which is the central point. Width window has to be larger than the base of the peak shape. Iterative Convolution algorithm uses a Gaussian filter to estimate the normal shape of each peak and estimates the baseline as an interpolation of all the minimum of this Gaussian curve.

Smoothing:

Smoothing process removed false positive peaks corresponding to artefacts, smooth out fluctuations and highlight the shape of the spectrum. Common approaches are the Savitzky-Golay filter, the Moving Average, the Gaussian and the wavelet transform. Savitzky-Golay filter interpolates data point with low-degree polynomial function, returning a polynomial curve. It preserves the shape of the spectrum, the intensity value of peaks and their position along m/z axis. The Moving Average filter moves a define window along the spectrum, calculates the average of extremes data and replaces these values to the original ones. Large windows could lead to a loss of intensity value of the peaks, while small windows are computationally expensive. Gaussian filter smooths the spectra using a Gaussian kernel, assuming that each peak follows a Gaussian distribution.

Normalization:

Another task in the pre-processing stage was the normalization. Normalization divides all the intensities of a spectrum for a fix scaling factor. It is indispensable to bring all the spectra to the same intensity range in order to compare spectra not only within the same analyses but also among different ones. One of the most frequently used approaches is the Total Ion Current (TIC) approach. It divides each intensity by the sum of all the intensities in the mass spectrum. It is the most suitable method for MSI data pre-processing, but it is not robust in the presence of an high intense peak. The TIC corresponds entirely to that signal suppressing all the other intensities. To overcome this problem, TIC could be performed with the exclusion of that peak, or the Median method could be used; it divides each intensity by the median intensities of the entire spectrum, being robust in the presence of high intense peaks. However, it is sensitive to the noise variability and a not symmetrical noise profile could lead this method to generate significant artefacts. Other alternative to TIC normalization is the Root Mean Square (RMS), which scales intensity by the sum of the squares of the intensities. It is appropriate in presence of small variations in peak intensities but, as for the TIC, it is not robust in presence of prominent peaks. Other studies perform a normalisation step respect to the largest peak or performing a linear scaling with the smallest and the largest peak intensities.

Alignment:

Consequently, spectra must be aligned; slightly differences in *m/z* values had to be recognised as the same and aligned with the same *m/z* names, so peaks in different spectra that represent the same protein species were matched. Peak alignment was performed firstly by extracting the most suitable peaks from a reference spectrum and then matching peak maxima of all the spectra to the reference ones. Other less frequently used methods regarding peak alignment have been proposed by Kim and Zhang that took advantage from mass spectral similarity measure, such as correlation functions, to affect the performance of peak matching-based alignment, increasing the alignment accuracy [38]. Another one is the complete linkage hierarchical clustering performed on the *m/z* axis: the same peaks that are shifted each other's in different spectra were grouped under the same tight cluster [39].

Peak picking:

Finally, peaks detection identified peaks in the mass spectrum. Peak picking extracts the information regarding the only true informative peaks. In some Peak Picking algorithm spectrum noise is estimated, and only peaks with a Signal-to-Noise ratio (S/N) higher than an arbitrary threshold are retained for statistical analyses. Other processes, as the Orthogonal Matching Pursuit, estimate how much a peak looks like a Gaussian curve and detect as peak the *m/z* value that corresponds to the global maximum of each Gaussian shape [40][37].

Only the information regarding the intensity and *m/z* value of the informative peaks were retained for the statistical analyses.

Bibliography

[1] H. Liu and F. Lin, “Application of immunohistochemistry in thyroid pathology,” *Arch. Pathol. Lab. Med.*, vol. 139, no. 1, pp. 67–82, 2015.

[2] J. P. Zevallos, C. M. Hartman, J. R. Kramer, E. M. Sturgis, and E. Y. Chiao, “Increased thyroid cancer incidence corresponds to increased use of thyroid ultrasound and fine-needle aspiration: A Study of the veterans affairs health care system,” *Cancer*, vol. 121, no. 5, pp. 741–746, 2015.

[3] P. W. Flint *et al.*, *Cummings Otolaryngology Head and Neck Surgery*. 2015.

[4] N. Stathatos, “Anatomy and Physiology of the Thyroid Gland,” in *Thyroid Cancer*, Totowa, NJ: Humana Press, 2006, pp. 3–7.

[5] S. Paschou, A. Vryonidou, and D. G. Goulis, “Thyroid nodules: A guide to assessment, treatment and follow-up,” *Maturitas*, vol. 96, pp. 1–9, 2017.

[6] H. R. Hemmati, B. Shahnazari, and M. Foroutan, “The effect of fine needle aspiration on detecting malignancy in thyroid nodule,” *Biomol. Concepts*, vol. 10, no. 1, pp. 99–105, 2019.

[7] L. Dal Maso *et al.*, “Incidence of thyroid cancer in Italy, 1991-2005: Time trends and age-period-cohort effects,” *Ann. Oncol.*, vol. 22, no. 4, pp. 957–963, 2011.

[8] W. D. Middleton *et al.*, “Multiinstitutional analysis of thyroid nodule risk stratification using the American College of radiology thyroid imaging reporting and data system,” *Am. J. Roentgenol.*, vol. 208, no. 6, pp. 1331–1341, 2017.

[9] G. Russ, S. J. Bonnema, M. F. Erdogan, C. Durante, R. Ngu, and L. Leenhardt, “European Thyroid Association Guidelines for Ultrasound Malignancy Risk Stratification of Thyroid Nodules in Adults: The EU-TIRADS,” *Eur. Thyroid J.*, vol. 6, no. 5, pp. 225–237, 2017.

[10] F. N. Tessler *et al.*, “ACR Thyroid Imaging, Reporting and Data System (TI-RADS): White Paper of the ACR TI-RADS Committee,” *J. Am. Coll. Radiol.*, vol. 14, no. 5, pp. 587–595, 2017.

[11] W. K. Dong *et al.*, “Ultrasound-guided fine-needle aspiration biopsy of thyroid nodules smaller than 5 mm in the maximum diameter: Assessment of efficacy and pathological findings,” *Korean J. Radiol.*, vol. 10, no. 5, pp. 435–440, 2009.

[12] M. Bongiovanni *et al.*, “Comparison of 5-tiered and 6-tiered diagnostic systems for the reporting of thyroid cytopathology,” *Cancer Cytopathol.*, vol. 120, no. 2, pp. 117–125, 2012.

[13] F. Nardi *et al.*, “Italian consensus for the classification and reporting of thyroid cytology.,” *J. Endocrinol. Invest.*, vol. 37, no. 6, pp. 593–9, Jun. 2014.

[14] F. Basolo, A. Bondi, O. Maggiore, C. A. Pizzardi Di Bologna, and G. Bussolati, “(No Title),” 2010.

[15] E. S. Cibas, S. Z. Ali, S. J. Mandel, and served W. as moderators Zubair Baloch, “The Bethesda System for Reporting Thyroid Cytopathology,” *Am J Clin Pathol*, vol. 132, pp. 658–665, 2009.

[16] S. A. Polyzos and A. D. Anastasilakis, “Clinical complications following thyroid fine-needle biopsy: A systematic review,” *Clin. Endocrinol. (Oxf).*, vol. 71, no. 2, pp. 157–165, 2009.

[17] I. Piga *et al.*, “Update on: proteome analysis in thyroid pathology - part II: overview of technical and clinical enhancement of proteomic investigation of the thyroid lesions.,” *Expert Rev. Proteomics*, vol. 15, no. 11, pp. 937–948, 2018.

[18] F. P. Y. Barré *et al.*, “Derivatization strategies for the detection of triamcinolone acetonide in cartilage by using matrix-assisted laser desorption/ionization mass spectrometry imaging,” *Anal. Chem.*, vol. 88, no. 24, pp. 12051–12059, 2016.

[19] S. Meding *et al.*, “Tumor classification of six common cancer types based on proteomic profiling by MALDI imaging,” *J. Proteome Res.*, vol. 11, no. 3, pp. 1996–2003, 2012.

[20] F. Magni *et al.*, “Proteomics imaging and the kidney,” *J. Nephrol.*, vol. 26, no. 3, pp. 430–436, 2013.

[21] M. Galli *et al.*, “Proteomic profiles of thyroid tumors by mass spectrometry-imaging on tissue microarrays,” *Biochim. Biophys. Acta - Proteins Proteomics*, vol. 1865, no. 7, pp. 817–

827, 2017.

[22] F. Pagni *et al.*, “Proteomics in thyroid cytopathology: Relevance of MALDI-imaging in distinguishing malignant from benign lesions,” *Proteomics*, vol. 16, no. 11–12, pp. 1775–1784, 2016.

[23] O. J. R. Gustafsson, G. Arentz, and P. Hoffmann, “Proteomic developments in the analysis of formalin-fixed tissue,” *Biochim. Biophys. Acta - Proteins Proteomics*, vol. 1854, no. 6, pp. 559–580, 2015.

[24] G. De Sio *et al.*, “A MALDI-Mass Spectrometry Imaging method applicable to different formalin-fixed paraffin-embedded human tissues,” *Mol. Biosyst.*, vol. 11, no. 6, pp. 1507–1514, 2015.

[25] M. Pietrowska *et al.*, “Molecular profiles of thyroid cancer subtypes: Classification based on features of tissue revealed by mass spectrometry imaging,” *Biochim. Biophys. Acta - Proteins Proteomics*, vol. 1865, no. 7, pp. 837–845, 2017.

[26] A. Smith *et al.*, “Molecular signatures of medullary thyroid carcinoma by matrix-assisted laser desorption/ionisation mass spectrometry imaging,” *J. Proteomics*, vol. 191, no. March 2018, pp. 114–123, 2019.

[27] R. J. DeHoog *et al.*, “Preoperative metabolic classification of thyroid nodules using mass spectrometry imaging of fine-needle aspiration biopsies,” *Proc. Natl. Acad. Sci.*, p. 201911333, Oct. 2019.

[28] I. Piga *et al.*, “The management of haemoglobin interference for the MALDI-MSI proteomics analysis of thyroid fine needle aspiration biopsies,” *Anal. Bioanal. Chem.*, pp. 5007–5012, 2019.

[29] I. Piga *et al.*, “Feasibility Study for the MALDI-MSI Analysis of Thyroid Fine Needle Aspiration Biopsies: Evaluating the Morphological and Proteomic Stability Over Time,” *Proteomics - Clin. Appl.*, vol. 13, no. 1, pp. 1–9, 2019.

[30] J. M. Amann *et al.*, “Selective profiling of proteins in lung cancer cells from fine-needle aspirates by matrix-assisted laser desorption ionization time-of-flight mass spectrometry,” *Clin. Cancer Res.*, vol. 12, no. 17, pp. 5142–5150, 2006.

[31] J. Linder, “Recent Advances in Thin-Layer Cytology,” vol. 18, no. 1, 2000.

[32] A. Smith *et al.*, “Matrix-assisted laser desorption/ionisation mass spectrometry imaging in the study of gastric cancer: A mini review,” *Int. J. Mol. Sci.*, vol. 18, no. 12, 2017.

[33] I. Piga *et al.*, “Ultra-high resolution MALDI-FTICR-MSI analysis of intact proteins in mouse and human pancreas tissue,” *Int. J. Mass Spectrom.*, vol. 437, pp. 10–16, 2019.

[34] T. C. Rohner, D. Staab, and M. Stoeckli, “MALDI mass spectrometric imaging of biological tissue sections,” in *Mechanisms of Ageing and Development*, 2005, vol. 126, no. 1, pp. 177–185.

[35] K. J. Boggio, E. Obasuyi, K. Sugino, S. B. Nelson, N. Y. R. Agar, and J. N. Agar, “Recent advances in single-cell MALDI mass spectrometry imaging and potential clinical impact,” *Expert Review of Proteomics*, vol. 8, no. 5. pp. 591–604, Oct-2011.

[36] P. Ràfols *et al.*, “Signal preprocessing, multivariate analysis and software tools for MA(LDI)-TOF mass spectrometry imaging for biological applications.,” *Mass Spectrom. Rev.*, vol. 37, no. 3, pp. 281–306, 2018.

[37] T. Alexandrov, “MALDI imaging mass spectrometry: statistical data analysis and current computational challenges.,” *BMC Bioinformatics*, vol. 13 Suppl 1, no. Suppl 16, p. S11, 2012.

[38] C. G. Fraga, B. J. Prazen, and R. E. Synovec, “Objective data alignment and chemometric analysis of comprehensive two-dimensional separations with run-to-run peak shifting on both dimensions,” *Anal. Chem.*, vol. 73, no. 24, pp. 5833–5840, Dec. 2001.

[39] R. Tibshirani *et al.*, “Sample classification from protein mass spectrometry, by ‘peak probability contrasts,’” *BIOINFORMATICS*, vol. 20, no. 17, pp. 3034–3044, 2004.

[40] D. Trede, J. H. Kobarg, J. Oetjen, H. Thiele, P. Maass, and T. Alexandrov, “On the importance

of mathematical methods for analysis of MALDI-imaging mass spectrometry data.," *J. Integr. Bioinform.*, vol. 9, no. 1, p. 189, 2012.

3 Overview of Statistical Methods

Various statistical methods can be applied in the omics field, but two are the main approaches usually considered in the workflow of a omics' analysis. Firstly, unsupervised methods can be applied in order to explore the data structure. Data quality, such as outliers identification, is evaluated and potential clusterings and mutual relations are investigated. Then, supervised methods are used to construct either diagnostic, prognostic, or predictive models, based on the specific clinical question. The main aspect differentiating unsupervised from supervised methods is that the first can only be used for exploratory analysis because no prior information regarding the label of the data is available. Sometimes, when the unsupervised analysis fails, no further investigation with supervised analysis might be useful. In other situations in which the specific contexts of unsupervised and supervised analyses are so different and there is no relationship between the two, both the approaches are performed.

This chapter is divided into four sections. The first two present an overview of the different approaches used to handle high dimensional problems either with unsupervised and supervised analysis. Advantages and weaknesses of the different algorithms are discussed. None of these algorithms works best for every problem, because there are different factors, such as the size and structure of data, which play an important role on the choice of the approach to used. One of the most popular classifier in the omics field, the regularized regression model, is reviewed in the third section. It allows to create a linear regression model selecting the most informative features without losing information about individual features. In the last part of this chapter the typical process for the identification and evaluation of proteomics biomarkers is presented.

3.1 Unsupervised analysis

Unsupervised statistical analyses are useful to perform an initial exploration of the collected data, and can be divided into two major groups: dimensionality reduction (e.g. Variance thresholds, Correlation threshold, Principal Component Analyses, Neural networks) and clustering methods (e.g. Hierarchical Clustering Analyses and Partitioning methods). Class discovery uses structure inside the data to suggest interesting group's membership. If samples do not separate clearly, maybe also the classifier would not yield good results.

3.1.1 Dimensionality reduction

When dealing with hundreds to thousands number of features, it is useful to reduce the dimensionality of the space while preserving the information present in the entire dataset. Two main approaches are available for dimensionality reduction: feature selection and feature extraction.

3.1.1.1 Feature selection

Feature selection is used for filtering out irrelevant or redundant variables from the high dimensional dataset. The key difference with respect to feature extraction is that this last one creates new variables from the original ones as a combination of them, while feature selection keeps a subset of the original features.

Feature selection can be both supervised (e.g. Genetic Algorithms, Stepwise Selection, Univariate Analysis, Significance Analysis of Microarray) and unsupervised (e.g. Variance thresholds, Correlation threshold).

Furthermore, some supervised algorithms build a feature selection inside the model, i.e. Regularized Regression and Random Forest that are exhaustively explained later.

Variance thresholds:

Variance threshold is a feature selector that removes all low-variance features. The idea is really simple, feature that have not higher change between the observation do not add much information. It is recommended to use a lower threshold not to lose possible informative variables. As all the univariate analysis do not take into account any correlation between the features, the variance of individual features is evaluated. Because variance is dependent on scale, it is always recommended to normalize the features before applying variance thresholds. As an extreme use of this technique, if a threshold of zero is applied, only features with non-zero variance are kept, while all the features that have same value in all the samples are removed. Because the focus of this thesis is on proteomics analysis, a choice on this feature selector had to be done. In proteomics field it is not easy to have features with low variance due to the analytical variability of the sample (i.e. cytospin sample preparation, MALDI-MSI sample preparation, instrumental analysis).

Correlation threshold:

The same idea could be applied on correlation task. All the features that are highly correlated with the others can provide redundant information. All pairwise correlation between all the features had to be evaluated. Then, between two highly correlated features, those that have the largest mean absolute correlation with other features are discarded. Applying a lower threshold, informative variables could be lost. In the omics filed as well as in genetics, even if two features are highly correlated, they may have to be retained because are involved in the same genetic process and can not be separated for clinical interpretability .

3.1.1.2 Feature extraction

Aim of feature extraction is to retain all the information, reducing the number of variables by creating a new set of latent variables that are a combination of the original ones. As with feature selection, some algorithms already perform feature extraction inside the model. One of the most popular algorithms of this class are neural networks. As feature selection, feature extraction can be unsupervised (e.g. Principal Component Analysis, Neural networks) and supervised (e.g. Linear Discriminant Analysis and its variants, Support vector machine, K-nearest neighbour, Random forest).

Principal Component Analysis:

Principal Component Analysis (PCA) is one of the most popular techniques used in high dimensional data [1].

Dimensionality reduction in PCA is feasible by finding a new set of orthogonal latent variables (principal components) that are a linear combination of the originals and explain as much as possible variance of the independent variables. The first principal component (PC1) shows the direction of the highest variance, PC2 is orthogonal to the first and represent the direction of the maximum variance remained. The projection of the data into the new space PC1xPC2 yields maximum separation between data. The fraction of variance explained by a principal component is the ratio between the variance of that principal component and the total variance, which is the sum of variance of all the Principal Components. PCA is an unsupervised learning method; it works without taking into account the dependent variable. Before performing PCA, data have to be scaled and centered, because it is sensitive to the different scale of features, otherwise features that have the largest scale would dominate in the new latent variables, explaining the major amount of variance. A supervised version of the principal component analysis exists that works by estimating a sequence of Principal components that have maximal dependence on the response variable.

Neural networks:

Artificial Neural Networks (ANN) are machine learning models inspired by real biological neural networks which compose animal brains. They are characterized by a set of artificial neurons connected to one another in several ways, depending on the type of function that has to be learned by the network.

A Self-Organizing Map (SOM) is a type of ANN which applies the concept of competitive learning in order to train the neural network, instead of the error minimization approach used by other neural networks. Competitive learning enables the network to learn in an unsupervised way, since the input vectors are evaluated iteratively only against network neurons using an arbitrary distance function in order to find the so-called Best Matching Unit (BMU) neuron. Then its weights are updated allowing the network to learn a pattern based on the given training set [2]. SOMs are used for dimensionality reduction because they are essentially a map from a N-dimensional vector space to a 2-D topological space in which artificial neurons and their corresponding connections are arranged together.

Autoencoders are network models composed by two connected Feedforward Neural Network (a type of ANN), named encoder and decoder, respectively. The encoder reduces an input vector of dimension n to a vector of lower dimension m, and the decoder tries to map the reduced feature space to the original input of dimension n, but instead of minimizing the error between labels, it tries to minimize the error between the real input and the input reconstructed by the decoder. This approach is an unsupervised learning method because labels are not necessary to train the network, since input vectors are compared against a reconstructed version of themselves. Hence, the output of the encoder network is the representation of the original feature space encoded into a lower and user-defined dimension, and so identifying a dimensionality reduction approach using neural networks [3].

3.1.2 Clustering

Clustering is an unsupervised learning algorithm that searches for natural groupings of observations, called clusters. Methods are divided into two main approaches: partitioning and hierarchical techniques. Partitioning methods differ from hierarchical ones due to the fact of having to previously decide the number of clusters in which observations have to be divided.

3.1.2 .1 Hierarchical Clustering Analysis

HCA is used to group analogues observations into the same cluster according to the similarity among each other [4]. Divisive HCA uses a top-bottom approach, starting from a unique group, which is consecutively divided into different subgroups estimating the pairwise distance among data observations and generating a dendrogram, e.g. tree. Conversely, agglomerative HCA uses a bottom-up approach; single observations are grouped together according to their similarity into clusters. HCA can be performed after PCA to highlight the presence of different similar clusters that can be correlated with the outcome, using the selected principal components. In the clustering algorithms the use of different cluster techniques could highlight different behaviour of the sample. For example, in the hierarchical cluster trees, complete linkage is adequate to detect separation in groups of the samples, while single linkage was rather appropriate to identify outliers.

3.1.2 .2 Partitioning analysis

Partitioning methods can be divided into parametric and non-parametric models. Model based techniques are a broad family of algorithms designed for modelling an unknown distribution as a mixture of simpler ones, where each sub-groups of similar data follows a classical distribution [5]. They are more flexible with respect the non-parametric form because each cluster could have a different variance. Different algorithm based on non parametric models were proposed and the most widely used are described below.

K-mean:

Conversely, Heuristic partitioning methods are not based on formal models; an example was the k-means clustering method [6]. In k-means, data observations are partitioned into k different clusters, minimizing the distance intra cluster and maximizing the distance inter groups. The number of clusters, k, must be chosen a priori, as the metric to be used to calculate distance. Another disadvantage involves the structure beyond data. In K-means clusters are grouped around centroids, resulting in globular, perfectly separated, clusters with similar sizes. If the underlying structure in data are not globular, the algorithm could produce poor clusters.

Partitioning around medoids:

Partitioning around medoids (PAM) or k-medoids is a clustering algorithm that works similarly to the k-means, attempting to minimize the distance between data [7]. In contrast to k-means, in which center of a cluster is calculated as the average between the points in the cluster, PAM choses centers among input data points. As k-means the number of clusters into which the observations have to be partitioned, has to be chosen a priori. Compared to k-means it is more robust to noise and outliers, because k-means minimizes the sum of the squared Euclidean distance while k-medoids minimizes the sum of all the pairwise dissimilarities.

Affinity propagation:

Affinity propagation [8] is a relatively new clustering algorithm that unlike clustering algorithms such as k-means or k-medoids, does not require the number of clusters to be determined or estimated before running the algorithm. Affinity propagation takes as input a measure of similarity between each pair of data. Similarities between data are calculated and preference to group membership are voted.

3.2 Supervised analysis

By contrast, the class prediction is most closely associated with classification problems. Starting from samples with a priori knowledge of their class membership, the supervised method tries to allocate new observations to these classes. Several statistical tools had been developed that measure the strength of the (univariate) association between the individual features and the response variable and differ each other for the way to assess the weights given to the features. In the omics field (i.e. genomic, lipidomics, proteomics, metabolomics) two main problems could be encountered: data dimensionality higher than the observations and possible multicollinearity (high correlation levels among the variables).

When a large number of features needs to be included in a model, univariate and multivariable regression analyses are not recommended, because, although simpler and useful in ranking features according to their ability in prediction, they are prone to big mistakes. Univariate analysis is used to select variables that show good performance at separating samples in classes of interest, with the strong assumption of no interaction effects between features also in absence of information from individual variables. Furthermore, albeit an adjustment for multiple comparisons (type I error control) must be applied, the level of significance can become so little that it is impossible to find significant variables.

To overcome this problem, different statistical methods had been developed with the primary aim of reducing the number of features, filtering out irrelevant and redundant information. The strategy is to identify, in a large number of variables, those features that were differentially expressed in a pre-specified population so that they could be included in a class prediction. In such a way it is possible to reduce the number of variables (e.g. genes, proteins) needed in the future to classify the individual patient, according to the observed value of the biomarkers. For example, Tibshirani et al. developed the Significance Analysis of Microarray (SAM) [9] that uses repeated permutations of the data to determine if the expression of any features are significantly related to the response controlling for the False Discovery Rate.

In omics data, the number of detected variables usually exceeds the number of samples, even in a relatively large study with many biological samples, so overfitting is another potential pitfall. This reduction could be done in two different ways, selecting the most relevant features or by summarizing the multiple original variables constructing new latent variables.

3.2.1. Feature selection

Genetic algorithm:

Genetic algorithm (GA) is a search operation that reflects the process of natural selection. GA has two main use, find the best weights for a neural network or performed a supervised feature selection. In the second case GA seeks to find a collection of markers (called chromosome in GA) that separate cases and control, each chromosome is evaluated by a fitness function following these simple steps. Intensity variables had to be previously scaled to lie between 0 and 1, then sample is clustered according to the Euclidean distance.

GA is similar to the K-means, with the main difference that the second algorithm is an unsupervised method.

Few studies applied GA to mass spectrometry analysis, because of its computational complexity without an increment in classification accuracy when it is comparable to other popular approaches as classification tree, boosting and PAM algorithm [10].

Stepwise algorithm:

Stepwise algorithm is a supervised feature selection method based on a sequential process of selection [11]. Two different formulations of these methods exist: forward and backward. In forward stepwise search features were add one at a time at the model, if the new variable increased the accuracy of the classifier then the features are retained, otherwise, it is discarded. it is suggested to be used when a large set of predictor variables are to be managed. Backward stepwise is the same process but reversed, starting from a full model then one at a time each feature were discarded until performance reaches the optimal results in sensitivity and specificity. This method is used when a modest number of predictor variables are available and the focus is to eliminate a few of them. Stepwise regression is able to manage with a large number of potential biomarkers even if a high amount of potential predictor variable means a high quantity of different models to be tested in order to select the relevant variables. One of the main advantages of these methods is the possibility to look at the order in which variables were removed or added giving information about the different levels of importance of the predictor variables. In addition, it is not able to deal with correlated variables, in fact, if two predictor variables in the model are highly correlated, only one may be retained into the model. This is not a problem in the major clinical setting, conversely, in the omics field, it can be a problem if correlated variables involved in the same genetic process had both to be retained.

3.2.2. Feature extraction

The number of latent variables determines the complexity of the model, a high number of latent variables might lead to an overestimation of the effects of the variables, leading to perfect classification, while a low set of these new variables could lead to under-fitting data, due to the fact that the new variables do not retain enough information from the independent original variables. The most used latent variable approach for classification models in omics field are PLS-DA [12], Linear Discriminant Analysis (LDA), Quadratic Discriminant Analysis (QDA), diagonal linear discriminant analysis (DLDA), PCA followed by LDA [13]; in other works non-parametric classification approach was used, such as support vector machines (SVM), the nearest neighbour classification (KNN) and random forest classifiers. The aim is the development of a classification model able to discriminate between outcomes, following decision rules. In the supervised classification, the discrimination rule is built using a training sample, a set of patients with prior information about their group membership. Statistical methods are used to divide the feature space into regions that better separate samples in categories. These regions are divided by smooth curves that define decision boundaries. Non-parametric methods, such as SVM, are able to divided data points with non linear hyperplanes, while other parametric models as LDA separated samples linearly, which is not usually possible in high dimensional data, but are useful if the data distribution is known prior.

Support Vector Machines:

The SVM approach was developed by Vapnik [14], it maps validation data points into a high dimensional space where a maximal separating hyperplane is constructed to better maximize the minimal difference between observations belonging to different classes. SVM with linear kernel is similar to logistic regression, however, its strength is due to the fact that can be used with non-linear kernels to model non-linear decision boundaries. It is fairly robust respect overfitting.

K Nearest Neighbor:

The KNN method is first introduced by Fix and Hodges [13] to perform discriminant analysis when probability densities are unknown or difficult to determine. Each new observation is assigned to the class most common among its k nearest neighbors. Usually, Euclidean distance is used to identify the k neighbors. Larger k reduced the misclassification error but made boundaries rule between different class membership less distinct and more complex decreasing.

Classification tree:

Classification tree (CART model) [15] uses decision tree algorithm for classification or regression models. Each node represents a decision rule on input independent variable. Leaf nodes of the tree contain the independent variable, which determines prediction. Following rules through the decision tree, new observations are classified. Classification tree is robust to outliers, is scalable and could model non-linear decision boundaries thanks to its hierarchical structure.

Random Forest:

Random Forest combines predictions from different individual decision trees, in which thresholds of feature values determine whether the observation belongs to a class or to another [16].

Neural Network:

A Feedforward Neural Network (FNN) [17], which exploits backpropagation algorithm, is a type of ANN trained using a set of labelled input vectors, adjusting connections weights based on the distance between real labels and network predicted labels, thus this kind of model resides inside the category of supervised machine learning.

All these methods are the so called black-box: features were combined together in order to minimize the classification error without taking into account how there were groups and their real importance in the classification problem. Moreover, when another observation had to classify the model had to be run again.

Linear discriminant analysis and further extension:

Conversely, different parametric classifiers had been successful in handling high dimension dataset for classification problems. Different classifier derived by Linear Discriminant Analysis (LDA) specifically designed in multivariate analysis with a latent variable approach and depends on the assumption of data distribution. LDA assumes that data follow multivariate normal distributions with mean and covariance matrix estimated from the training sample.

Standard LDA procedure, similar to the regression analysis, attempted at expressing an outcome as a linear combination of other features or measurements, is closely similar to PCA and factor analysis because of look for linear combinations of features that better explained the data. In LDA covariance matrix is assumed to be the same for all classes. Then, the main objective of LDA is to find a projection matrix that maximizes the ratio of the determinant of the between-class covariance matrix respect to the determinant of the within-class covariance matrix (Fisher's criterion). In order to use LDA, we need to compute the inverse of the covariance matrix and so not only the estimation of k different means but also the estimation of k(k+1)/2 different variance, supposing to have k different variables. If the number of estimation is a fraction of the total number of samples, the models could be unstable due to the covariance matrix becomes singular. To solve this problem, Yu and Yang [18] have developed the Diagonal LDA algorithm (DLDA). Instead of using the entire covariance matrix, the DLDA used the main diagonal, diagonalized the two covariance matrix. Another extension is the QDA, in which the covariance matrix is not assumed to be equal in all the groups, so the decision boundary could be curved.

Finally, Partial least squares discriminant analysis (PLS-DA) became popular in omics field, where many predictor variables (frequently correlated) and relatively few samples are usual PLS-DA is a variant of Partial least squares regression PLS-R that could be used when the response variable Y is categorical. Under certain circumstances, PLS-DA provides the same results that of linear discriminant analysis (LDA) but is especially suited to deal with multicollinearity [12].

3.3 Penalized Regression Models

The main problem in dimensionality reduction through the construction of latent variables, as a combination of the original ones, is the loss of information about individual features. If the interest is to identify specific features, such as proteins to subsequently carry out a clinical investigation like genetic studies or identify biomarkers research in the serum samples, instead of only correctly classified new indeterminate samples, supervised regression models had to be performed. In the high dimensional setting we deal with a huge number of biological features and we are interested in identifying a subset of biomarkers which characterizes patients according to their label. From a statistical point of view, this consists in fitting a selected regression model. In regression models, the selection and elimination of irrelevant variables for the classification problem could be included in the workflow of the model itself, e.g. penalized regression models (LASSO).

The Ridge regression and the Least Absolute Shrinkage and Selection Operator (LASSO) are two mainly used penalized regression methods [19]. Both consist in fitting a model that includes all the predictors and the estimated coefficients are shrunken toward zero relative to the least square estimates. The shrinkage (or regularization) has the effect of reducing variance and, only for LASSO, can perform variable selection.

Given an omics dataset, we denote $\boldsymbol{X}$ the predictors matrix of dimension $n\ x\ p$, where n is the total number of dependent variables (observations/response) and p the number of independent variables (features). The estimation of $\boldsymbol{\beta}$ parameters is performed by maximizing the penalized log likelihood defined as:

$$\ell_{pen}(\boldsymbol{\beta}, \lambda; \boldsymbol{X}) = \ell(\boldsymbol{\beta}; \boldsymbol{X}) - p_{\lambda}(\boldsymbol{\beta})$$

Of note, when we maximize this function we are not only maximizing the log-likelihood:

$$\max_{\beta} \ell_{pen}(\boldsymbol{\beta}, \lambda; \boldsymbol{X}) = \max_{\beta}\{\ell(\boldsymbol{\beta}; \boldsymbol{X}) - p_\lambda(\boldsymbol{\beta})\}$$

The penalizing parameter, introduced in the models, induces bias but reduces the mean squared error; the penalty term is a factor that permits to balance the bias-variance trade-off, the balance between under- and over-fit. The higher the weight of the penalty term is, the closer to the origin are the $\boldsymbol{\beta}$ coefficients, due to the fact that the penalty term is non decreasing in the coefficients:

$$\forall(\boldsymbol{\beta}, \boldsymbol{\beta}^\star) : \ |\boldsymbol{\beta}| \leq |\boldsymbol{\beta}^\star|, \quad p_\lambda(\boldsymbol{\beta}) \leq p_\lambda(\boldsymbol{\beta}^\star)$$

Ridge regression:

The Ridge regression penalizes the size of regression coefficients using the squared L2 norm penalty factor, i.e adding the squared magnitude of the coefficient as penalty factor, thus reducing variability and improving the accuracy of linear regression models.

$$p_\lambda(\boldsymbol{\beta}) = \lambda \boldsymbol{\beta}^\top \boldsymbol{\beta} = \lambda \textstyle\sum_{j=1}^{p} \beta_j^2 \quad \text{with } |\boldsymbol{\beta}| = \left(|\beta_1|, \ldots, |\beta_p|\right)^\top$$

Ridge regression however does not perform variable selection.

Lasso regression:

Conversely, the standard LASSO method is a shrinkage and selection method that performs feature selection using an L1 norm penalty factor, i.e. adding the absolute value of magnitude of the coefficient as penalty factor, while constructing the predictive model (it is considered an embedded method).

$$p_\lambda(\boldsymbol{\beta}) = \lambda \mathbf{1}^\top |\boldsymbol{\beta}| = \lambda \textstyle\sum_{j=1}^{p} |\beta_j| \quad \text{with } |\boldsymbol{\beta}| = \left(|\beta_1|, \ldots, |\beta_p|\right)^\top$$

Elastic net regression:

The elastic net [20] is a regularized regression method that linearly combines the L1 and L2 penalties of the lasso and ridge methods.

$$\begin{aligned} p_\lambda(\boldsymbol{\beta}) &= \lambda_1 \mathbf{1}^\top |\boldsymbol{\beta}| + \lambda_2 \boldsymbol{\beta}^\top \boldsymbol{\beta} & \lambda_1, \lambda_2 \geq 0 \\ &= \lambda \left[\alpha \mathbf{1}^\top |\boldsymbol{\beta}| + \frac{1-\alpha}{2} \boldsymbol{\beta}^\top \boldsymbol{\beta}\right] & \alpha \in [0, 1] \end{aligned}$$

When the coefficient $\alpha = 0$, the Elastic net reduces to the Ridge regression, while $\alpha = 1$ yields to the Lasso regression. Balancing the two penalty terms, Elastic net can perform variable selection while holding accuracy in the prediction. Elastic net tends to retain or discard together groups of features that are highly correlated.

3.3.1 The choice of the shrinkage parameter through cross-validation

The results of a penalized regression can vary dramatically, depending on the value of the penalty factor: the higher it is, the more the partial likelihood is penalized.

Cross-validation is used to select the optimal penalty parameter. The idea is to find the value of λ that provides the best estimates of $\boldsymbol{\beta}$, which minimize the mean prediction error and maximise the cross validated penalized likelihood. Given a number of candidate values for λ, for each of them observations are randomly split in K different folds, and penalized likelihood is calculated in all the data without the k^{th}fold:

$$\ell_{-k}(\boldsymbol{\beta}) = \ell(\boldsymbol{\beta}) - \ell_{(-k)}(\boldsymbol{\beta}) = c - (y_k - \boldsymbol{X}_k^{\top}\boldsymbol{\beta})^2$$

where c is a constant value over the K folds due to the fact that the log-likelihood is calculated over all subjects.

The cross-validated log-likelihood is defined as

$$cvl_{\lambda} = \sum_{k} \ell_k\left(\widehat{\boldsymbol{\beta}}_{(-k)}^{\lambda}\right) = k * c - \sum_{k}\left[y_k - \boldsymbol{X}_k^{\top}\widehat{\boldsymbol{\beta}}_{(-k)}^{\lambda}\right]^2$$

where $k * c$ is the constant value summed across the K folds and $\widehat{\boldsymbol{\beta}}_{(-k)}^{\lambda}$ is obtained by maximizing the penalized log-likelihood with penalty λ when fold k is left out. Finally, the λ final value is $\lambda_{cvl} = argmax_{\lambda}\ cvl_{\lambda}$ [21].

3.3.2 Robustness of cross-validation

Given that the choice of the shrinkage parameter has an high effect on model results, the robustness of cross validation in lasso has been investigated during the years. Furthermore, the penalty factor could vary considerably due to the randomly assignment of observations to the k folds. The main approaches to evaluated the robustness of cross validation follows:

Percentile Lasso:

In the percentile-lasso [22] and Stability Selection [23] the penalized cross-validated process is iterated Mdifferent times, in each of which subjects are randomly assigned to the K folds. For each iteration the selected $\lambda_{cvl}^{[m]}$ are retained and the empirical distribution is computed over all iterations: $\left\{\lambda_{cvl}^{[m]}\right\}_{m=1,\dots,M}$.

Finally, the high-rank percentile of the empirical distribution is selected as the optimal value.

Stability Selection:

In the stability selection method the shrinkage parameter $\lambda_{cvl}^{[m]}$ is estimated for each of the Miterations, but the empirical distribution of the selection probability of each of the β coefficients is computed and informative biomarkers are selected as the ones that have an empirical selection probability that is higher than a pre-specified threshold.

3.3.3 More complex penalties models

More complex penalized models exist based on the specific aims of the feature selection and available information regarding the data.

Adaptive Lasso:

The adaptive lasso [24] differs from the standard lasso as the penalty factor λ is not equal for all the β coefficientsbut each regression coefficient is differently weighted. This situation can occur when prior information is available for some features and one does not want to give heavy weights to non-informative covariates but aims to retain the most important ones. This is achieved by variable-specific penalties $\lambda_j = \lambda\omega_j$:

$$p_{\lambda,\boldsymbol{\omega}}(\boldsymbol{\beta}) = = \lambda\boldsymbol{\omega}^{\top}|\boldsymbol{\beta}| = \lambda\sum_{j=1}^{p}\omega_j|\beta_j|$$

Usually, if no prior knowledge is available, weights are defined as the inverse of the regression coefficients $\omega_j = 1/|\tilde{\beta}_j|$ estimated on the same data in different ways, such as using a previously fitted full model or applying one of the penalized regression methods described above [25]. The problem of these methods is that the weight calculation process is performed on the same data successively used in the adaptive lasso, so we are bringing the solution in the direction we choosed, leading to overoptimistic results. On the other hand, we tend to find significant biomarkers even if there are not as we will never get a null model. For this reason it is perhaps better to perform an initial global test, testing the null hypothesis in which all the regression coefficients are equal to zero against the alternative hypothesis that al least one coefficient is different from zero. If the null hypothesis is rejected then one of the previous cited model can be used to calculate the weights of the adaptive lasso. Performing an initial ridge model, estimated regression coefficients are distorted towards zero without collapsing in zero as in the lasso model, and coefficients that are closest to zero will be weighted more. Conversely, using a lasso model, the solution of the adaptive lasso will be a subset of the simple lasso.

Group Lasso:

In the group lasso, the idea similar to the previous extended model, but with penalty factors applied differently for subgroups of covariates [26]. This model can be useful when different variables play a common role, i.e a group of proteins or genes in the same pathway.

$$p_{\lambda}(\boldsymbol{\beta}) = \lambda\sum_{g=1}^{G}\sqrt{p_g}\|\boldsymbol{\beta}_g\|_2 = \lambda\sum_{g=1}^{G}\sqrt{p_g\boldsymbol{\beta}_g{}^{\top}\boldsymbol{\beta}_g}$$

With $\boldsymbol{\beta}_g$ the vector of the coefficients in the group g.

The term $\|\boldsymbol{\beta}_g\|_2$represents the Euclidean distance of the sequence of group of coefficients from the origin. If a subgroup of covariates has an higher cardinality with respect to another group, the first group of biomarkers will be more penalized.

Sparse Group Lasso:

An extended version of the group lasso is the sparse group lasso [27]. Sometimes we are not necessarily interested in keeping all the biomarkers belonging to the same group in the final model. In addition to the group lasso, a simple lasso component could be added. In this way the penalization of the single markers increases.

$$p_\lambda(\boldsymbol{\beta}) = \alpha\lambda\|\boldsymbol{\beta}\| + (1-\alpha)\lambda\sum_{g=1}^{G}\sqrt{p_g}\|\beta_g\|_2$$

Fused Lasso:

In the fused lasso [28] the spatial component in the variable matrix is taken into account, so that two far away biomarkers are more penalized. This can be useful in genes sequences of mass spectra, in which the proximity of markers has to be considered, since two nearest genes possibly be involved in the same process have to be both retained in the model.

$$p_\lambda(\boldsymbol{\beta}) = \lambda_1\|\boldsymbol{\beta}\| + \lambda_2\sum_{j=2}^{p}|\beta_j - \beta_{j-1}|$$

3.4 Model building

A typical sequence of data analyses in proteomics biomarker research started checking for separability of the data by unsupervised cluster methods. Dealing with high dimensional setting, variable selection is another useful step. It could improve the performance of the classification model, dealing to an easy interpretation of the model by a biological point of view. Variable selection leads to over-optimistic and not generalizable results when it is performed using the same groups of samples that had been used to create the classification model. To avoid this, different approaches are available. It is usually preferred to have either an external set of data to use for a prior variable selection or to have prior information given by previous research (pilot study) or by known knowledge (additional information e.g. clinical knowledge). Another possibility is to use not supervised analysis on the same set of data on which the model had to be performed. Unsupervised analysis gives us information on features involved in the separation of observations without the use of prior information on these data, such as the class to which they belonged. Then this information could be used as prior information and then on the same training set classification models could be constructed. Last, as previously mentioned, supervised regression models that performed variable selection inside the model could be used an alternative when no external or prior knowledge are available.

Since it is always possible to find classifiers that accurately classify the data on which they were developed even if there is no relationship between expression of any of the genes and outcome, it is mandatory that the whole process underwent a rigorous validation procedure.

To validate the resulting models, a number of samples must be left out as a validation set. Using the constructed multivariate models, the class labels of the test set are predicted, so if the data to be predicted were used to train the classifier, then results look better than they should (problem known as "overfitting").

Thus, validation is a central and not-negligible step of classifier development to provide the transportability of the classifier to another population. What had been developed on the training cohort should be validated in a new cohort blinded to the previous results.

As previously described, cross-validation consists in partitioning the data in *k*-equal fold subsets. One of this subset is omitted from the development of the classifier and will be used as a test set. The remaining dataset is used to completely develop the classifier by fitting a penalized regression or other models. Then, patients of the test set will be classified. The procedure is repeated by including patients, previously classified as test patients, in the training set and switching patients of the training set to internal validation set. In this way all partitions work as training and test set in a *k*-loops cross-validation. In this way, optimal tuning parameter is selected, informative features are selected, their correspondence regression coefficients are estimated and the internal accuracy of the model is evaluated.

After cross validation, the prediction model has to be validates on a second independent dataset.

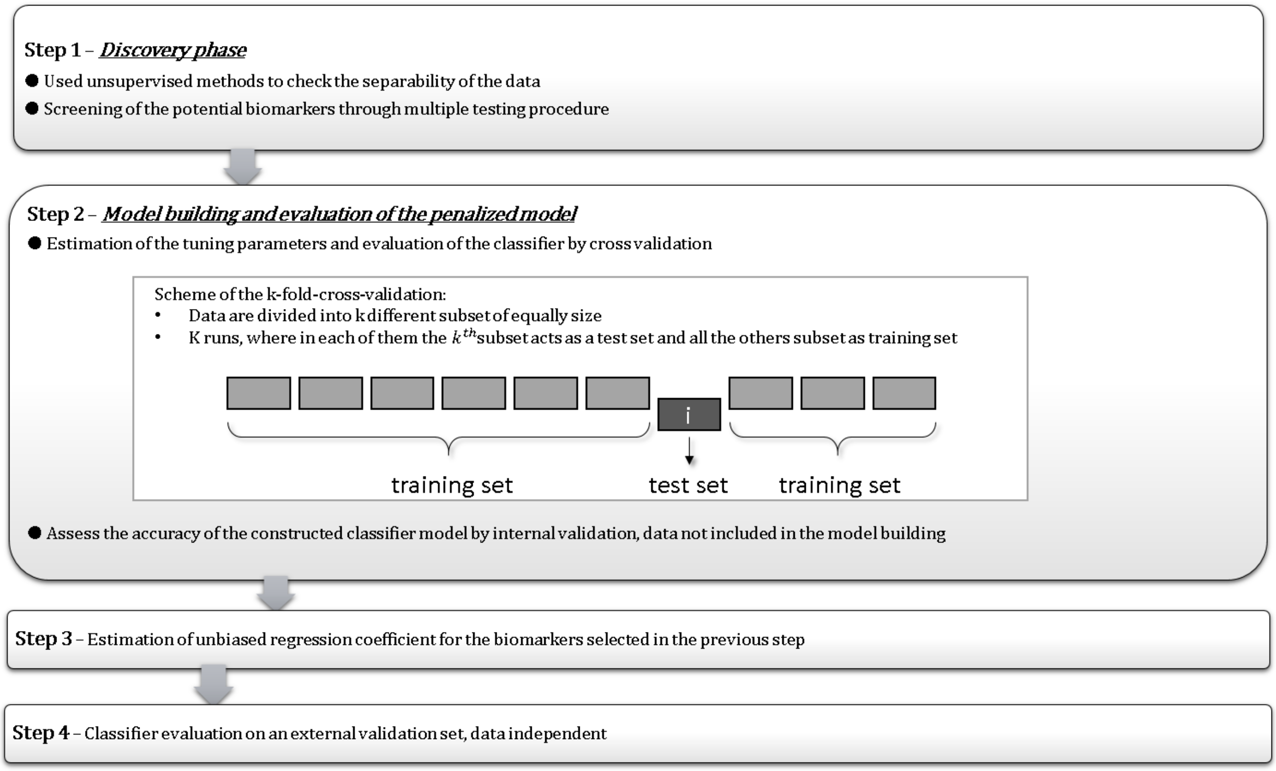
In this way, the performance of the classifier measured by this system truly reflects its accuracy in classifying patients who are not included in the development phase but are selected from an identical population.

The discrepancy between the predicted class and the actual class is evaluated with performance parameters of the model, such as sensitivity (True Positive Rate, TPR), specificity (True Negative Rate, TNR), Positive Predictive Value (PPV) and Negative Predictive Value (NPV).

Another step has to be performed to correctly identify a variable as a true biomarker and quantify its weight in the classification performance. Due to the fact that estimation of the regression coefficients are biased because of the shrinkage parameter, a standard regression with only the selected variables of a previous penalized regression had to be performed on external validation dataset to estimate the real weight of the variables.

Finally, the classifier had to be evaluated on an external validation cohort to assess performance and reliability of the model, in a different population from the one used to construct the classification model.

The schematic presentation of the typical procedure of data analysis is reported in the following figure.



Step 1 – ***Discovery phase***

● Used unsupervised methods to check the separability of the data

● Screening of the potential biomarkers through multiple testing procedure

Step 2 – ***Model building and evaluation of the penalized model***

● Estimation of the tuning parameters and evaluation of the classifier by cross validation

Scheme of the k-fold-cross-validation:

- Data are divided into k different subset of equally size
- K runs, where in each of them the k^{th}subset acts as a test set and all the others subset as training set

● Assess the accuracy of the constructed classifier model by internal validation, data not included in the model building

Step 3 – Estimation of unbiased regression coefficient for the biomarkers selected in the previous step

Step 4 – Classifier evaluation on an external validation set, data independent

Bibliography

[1] C. Syms, “Principal Components Analysis,” *Encycl. Ecol. Five-Volume Set*, pp. 2940–2949, 2008.

[2] T. Kohonen, “The Self-Organizing Map (Kohonen).pdf,” *Proceedings of the IEEE*, vol. 78, no. 9. pp. 1464–1480, 1990.

[3] “GitHub - janishar/mit-deep-learning-book-pdf: MIT Deep Learning Book in PDF format (complete and parts) by Ian Goodfellow, Yoshua Bengio and Aaron Courville.” [Online]. Available: https://github.com/janishar/mit-deep-learning-book-pdf. [Accessed: 27-Oct-2019].

[4] “Hierarchical Clustering,” in *Integrative Cluster Analysis in Bioinformatics*, Chichester, UK: John Wiley & Sons, Ltd, 2015, pp. 157–166.

[5] A. AdelekeR., I.-K. Hauwau, and O. Tinuke O., “Cluster Analysis of Data Points using Partitioning and Probabilistic Model-based Algorithms,” *Int. J. Appl. Inf. Syst.*, vol. 7, no. 7, pp. 21–26, Aug. 2014.

[6] Y. Li and H. Wu, “2012 International Conference on Solid State Devices and Materials Science A Clustering Method Based on K-Means Algorithm peer-review under responsibility of [name organizer],” *Phys. Procedia*, vol. 25, pp. 1104–1109, 2012.

[7] H. S. Park and C. H. Jun, “A simple and fast algorithm for K-medoids clustering,” *Expert Syst. Appl.*, vol. 36, no. 2 PART 2, pp. 3336–3341, 2009.

[8] D. Dueck, “AFFINITY PROPAGATION: CLUSTERING DATA BY PASSING MESSAGES,” 2009.

[9] V. G. Tusher, R. Tibshirani, and G. Chu, “Significance analysis of microarrays applied to the ionizing radiation response,” *Proc. Natl. Acad. Sci. U. S. A.*, vol. 98, no. 9, pp. 5116–5121, Apr. 2001.

[10] N. O. Jeffries, “Performance of a genetic algorithm for mass spectrometry proteomics,” *BMC Bioinformatics*, vol. 5, pp. 1–13, 2004.

[11] I. G. Chong and C. H. Jun, “Performance of some variable selection methods when multicollinearity is present,” *Chemom. Intell. Lab. Syst.*, vol. 78, no. 1, pp. 103–112, Jul. 2005.

[12] S. Wold, M. Sjöström, and L. Eriksson, “PLS-regression: A basic tool of chemometrics,” *Chemom. Intell. Lab. Syst.*, vol. 58, no. 2, pp. 109–130, 2001.

[13] Y. Tominaga, “Comparative study of class data analysis with PCA-LDA, SIMCA, PLS, ANNs, and k-NN,” *Chemom. Intell. Lab. Syst.*, vol. 49, no. 1, pp. 105–115, 1999.

[14] J. Valyon and G. Horváth, “A sparse least squares support vector machine classifier,” *IEEE Int. Conf. Neural Networks - Conf. Proc.*, vol. 1, pp. 543–548, 2004.

[15] D. Steinberg, “Chapter 10 CART: Classification and Regression Trees,” 2009.

[16] L. Breiman, “Random Forest Draft,” pp. 1–33, 2001.

[17] “Feedforward Neural Network Methodology - Terrence L. Fine - Google Libri.” [Online]. Available: https://books.google.it/books?hl=it&lr=&id=s-PlBwAAQBAJ&oi=fnd&pg=PR15&dq=Feedforward+Neural+Network&ots=abrFvtdb5q&sig=uYWSELcwnFg5WmTC5MZ2MVUrsQ4#v=onepage&q=Feedforward Neural Network&f=false. [Accessed: 27-Oct-2019].

[18] H. Yu and J. Yang, “A direct LDA algorithm for high-dimensional data — with application to face recognition,” *Pattern Recognit.*, vol. 34, no. 10, pp. 2067–2070, 2001.

[19] R. Tibshiranit, “Regression Shrinkage and Selection via the Lasso,” 1996.

[20] H. Zou and T. Hastie, “Erratum: Regularization and variable selection via the elastic net (Journal of the Royal Statistical Society. Series B: Statistical Methodology (2005) 67 (301-320)),” *J. R. Stat. Soc. Ser. B Stat. Methodol.*, vol. 67, no. 5, p. 768, 2005.

[21] J. C. Van Houwelingen and S. Le Cessie, “Predictive value of statistical models,” *Stat. Med.*, vol. 9, no. 11, pp. 1303–1325, 1990.

[22]S. Roberts and G. Nowak, “Stabilizing the lasso against cross-validation variability,” *Comput. Stat. Data Anal.*, vol. 70, pp. 198–211, 2014.
[23]N. Meinshausen and P. Bühlmann, “Stability selection,” *J. R. Stat. Soc. Ser. B Stat. Methodol.*, vol. 72, no. 4, pp. 417–473, Sep. 2010.
[24]H. Zou, “The adaptive lasso and its oracle properties,” *J. Am. Stat. Assoc.*, vol. 101, no. 476, pp. 1418–1429, 2006.
[25]S. van de Geer, P. Bühlmann, and S. Zhou, “The adaptive and the thresholded Lasso for potentially misspecified models (and a lower bound for the Lasso),” *Electron. J. Stat.*, vol. 5, no. 2010, pp. 688–749, 2011.
[26]M. Yuan and Y. Lin, “Model selection and estimation in regression with grouped variables,” *J. R. Stat. Soc. Ser. B Stat. Methodol.*, vol. 68, no. 1, pp. 49–67, 2006.
[27]N. Simon, J. Friedman, T. Hastie, and R. Tibshirani, “A sparse-group lasso,” *J. Comput. Graph. Stat.*, vol. 22, no. 2, pp. 231–245, 2013.
[28]R. Tibshirani, M. Saunders, S. Rosset, J. Zhu, and K. Knight, “Sparsity and smoothness via the fused lasso,” *J. R. Stat. Soc. Ser. B Stat. Methodol.*, vol. 67, no. 1, pp. 91–108, 2005.

4 Proteomic and Clinical studies

Preliminary to the clinical study, that was the main focus of the thesis, the protocol for the proteomic MALDI-MSI analysis was optimized to avoid degradation, alteration phenomena, contamination and artefacts formation. The methodological set-up of the proteomic protocol in a complicated field like that of the thyroid cytological specimens was a fundamental requirement for the conduction of the clinical study. Challenging technical aspects, such as i) the interference of haemoglobin due to the high vascularization of the thyroid organ and ii) the stability of the samples over time before the analysis from a morphological and proteomic point of view, were overcome through two studies whose design and results are reported in the first two sections of this chapter.

The clinical study for the detection of the potential cluster of proteomic signals with discriminant capability was originally planned to involve a large sample of thyroid nodules, however, due to the slow enrolment rate of malignant cases, section 3 contains only the results of a preliminary analysis. This interim evaluation involved data from 18 subjects with benign and malignant thyroid nodules and, an additional sample of 11 patients with different type of lesions (i.e. benign, indeterminate and malignant) was used for validation. Results are very promising and highlight the possibility to introduce MALDI-MSI as a complementary tool for the diagnostic characterization of thyroid lesions, but the final analysis is required to corroborate these findings.

4.1 The management of haemoglobin interference for the MALDI-MSI proteomics analysis of thyroid fine needle aspiration biopsies

4.1.1 Introduction

Although most thyroid nodules are diagnosed using a morphological approach, a significant challenge is related with the 20-30% of fine needle aspiration biopsy (FNAB) cases that are deemed to be indeterminate for malignancy (THY3 and THY4 according to the British system for reporting thyroid cytopathology) [7]. So far, patients with a THY3 diagnosis undergo diagnostic (and not therapeutic) total thyroidectomy, which has a significant impact on the lifestyle of the patient (in terms of potential related morbidity and the needing of lifelong hormone therapy) and places a hefty financial burden on the health care system. Moreover, the postoperative histological diagnosis of THY3 cases highlight that approximately 70% of the nodules were benign and the thyroidectomies unnecessary. Therefore, it is of paramount importance to assist pathologists in the diagnosis of the indeterminate lesions of thyroid FNABs. Matrix-assisted laser desorption/ionisation mass spectrometry imaging (MALDI-MSI) is a powerful tool in clinical proteomics allowing the investigation of the spatial distribution of biomarkers directly on tissues and cytological specimens, and the integration of molecular and morphological evidence [1–3]. Nowadays, only few studies have been published on MALDI-MSI for the identification of new possible proteins to support thyroid cancer diagnosis [4–6], all whilst using ex-vivo thyroid cytological samples. Nevertheless, significant blood contamination, generated by the abundant vasculature of thyroid lesions and by the cutaneous vasculature of the neck, is a frequent feature in thyroid FNABs [2]. In fact, the abundant presence of haemoglobin in FNABs is a cause of unsatisfactory rate in traditional cytology. Erythrocytes, present in large amounts in the suspension of cells, are also challenging for mass spectrometric analysis, as haemoglobin suppresses the ionisation of other protein signals. Amann et al. developed a sample preparation method for MALDI-MS analysis in order to reduce haemoglobin interference from simulated and clinical lung FNAB, suggesting the use of an erythrocyte lysis buffer [8]. They were able to obtain high quality MALDI-MS spectra of clinical FNABs, however the haemoglobin signal was still significantly observed, even after the erythrocyte lysis step [8]. Accordingly, we investigated the possibility to efficiently reduce the presence of haemoglobin and, consequently, increase the rate of analysable specimens in order to pave the way for future studies focusing on biomarkers discovery with MALDI-MSI. Moreover, this protocol may also be applicable to other specimens where significant contamination of haemoglobin is observed. For this purpose, we compare three protocols (the air-dried, the ethanol-fixed conventional smears and the liquid based preparation (LBP)) that are routinely used in clinical practice for the cytological diagnosis of thyroid FNABs in the context of MALDI-MSI proteomics analysis [9–11] (Supplementary Figure 1 in appendix A).

4.1.2 Material and Methods

4.1.2.1 Sample collection and preparation

The study was approved by the Ethical Committee of the San Gerardo Hospital (cod. AIRC MFAG- 18/11/2016). Ex-vivo cytological samples of 9 patients who underwent thyroidectomy at the Department of Surgery of San Gerardo Hospital, Monza, Italy, were collected within 30 minutes following the surgical procedure. Sampling was performed using a 25-Gauge needle. Three independent samples were collected from each surgical specimen and treated with different sample preparation methods: Protocol A) conventional air-dried smear (n=9), Protocol B) EtOH immediately fixed smear (n=7) and Protocol C) ThinPrep LBP (n=9). In protocol A, specimens were transferred and smeared from the syringe directly onto indium tin oxide (ITO) conductive slides (Bruker Daltonics, Bremen, Germany), then air-dried for approximately 30 minutes at room temperature and finally washed with 70%, 90% and 95% EtOH solutions for 30 seconds each. In fact, it is common practice for protein MSI analysis to perform washing steps, in order to remove salt and lipid contamination that can unfavourably affect the quality of MALDI-MSI data [12,13]. In protocol B, the air-drying step was excluded whilst in protocol C, the cytological samples were immediately transferred into a falcon tube filled with the ThinPrep® CytoLyt (Hologic, Marlbourough, MA, USA) methanol-based buffered solution, prepared following the manufacturer instruction of the ThinPrep® 2000 System (Hologic, Marlbourough, MA, USA) and transferred as a monolayer of cells onto ITO glass slides. Then, all slides were dried under vacuum for 15 minutes and stored at -80°C until the day of the analysis. A second group of samples (real in-vivo FNAB), taken from 19 patients who underwent FNAB, were prepared using protocol A (n=7) and protocol C (n=12). Before MALDI-MSI analysis, cytological specimens were equilibrated to room temperature, dried under vacuum for 30 minutes and the MALDI-matrix sinapinic acid (10 mg/ml in 60:40 acetonitrile:water w/0.2% trifluoroacetic acid) was uniformly deposited, with an optimised method, using the iMatrixSpray (Tardo Gmbh, Subingen, Switzerland) automated spraying system.

4.1.2.2 MALDI-MSI analysis and staining procedure

MALDI-TOF-MSI was performed using an ultrafleXtreme MALDI-TOF/TOF (Bruker Daltonik GmbH) in positive-ion linear mode, using 300 laser shots per spot, with a laser focus setting of 3 medium (diameter of 50 μm). Protein Calibration Standard I (Bruker Daltonics), that contains a mixture of standard proteins within the mass range of 5730 to 16950 Da, was used for external calibration (Mass accuracy ±30ppm). Spectra were recorded within the m/z 3000-20000 range. Data acquisition and visualisation was performed using the Bruker software packages (flexControl 3.4, flexImaging 4.1). After the analysis, the MALDI-matrix was removed with 70% EtOH and the slides were stained with haematoxylin and eosin (H&E), digitally scanned using an a ScanScope CS digital scanner (Aperio, Park Center Dr., Vista, CA, USA) and images were co-registered to the MSI-datasets in flexImaging.

4.1.2.3 Data analysis

Data pre-processing (MALDIquant package) and statistical analysis were performed using the open-source R software v.3.4.3. [14]. The individual average spectra were processed by performing baseline subtraction (SNIP method, iteration 100), smoothing algorithm (Moving

Average method, half window width 2), normalisation (Total Ion Current, TIC), alignment and peak picking (S/N ≥ 6). Peaks that appeared in at least 5% of all individual average spectra were used for the statistical analysis. The open-source software mMass v.5.5 (http://www.mmass.org) was used to confirm mass spectra alignment. The non parametric Kruskall-Wallis test (two-sided, α=0.05) was used to compare the different protocols in terms of 3 specific signal intensities (α and β Haemoglobin and Histone H4) and a post-hoc Dunn test, with Benjamini & Hochberg adjustment, was applied for pairwise comparisons.

4.1.3 Results and Discussion

The average spectra, obtained after TIC normalisation, of ex-vivo cytological samples collected from the same patient and treated with the three protocols A, B and C were compared, for descriptive purposes, in Figure 1a. The normalised intensities [A.U.] of the α and β Haemoglobin chains were clearly decreased in Protocol C, with respect to protocols A and B, with a concomitant increase of other signals in the m/z range 3000-15000, for example such as that corresponding with Histone H4 at m/z 11306 [8, 15], which was selected as an exemplary signal due to its presence in all samples (a panel of other statistically significant signals is provided in Supplementary Figure 2 in appendix A). The overall comparison of the signal intensities among the three protocols was statistically significant for αHaemoglobin (p=0.005), βHaemoglobin (p=0.02), and Histone H4 (p=0.00008), with paired post-hoc comparisons underlying significant differences between protocol A and B vs C for αHaemoglobin, and Histone H4, and between protocols B vs C for βHaemoglobin (Figure 1b, 1c and 1d). Mass error calculated on the average mass was αHaemoglobin -188 ppm; βHaemoglobin, -64 ppm; Histone H4, -89 ppm). One of the most remarkable features of MALDI-MSI is the possibility to co-register the in situ proteomic information with the histological image and to focus the data analysis solely on specific regions of interest (e.g. malignant or benign thyrocytes), excluding areas which could confound the results, in order to obtain their specific proteomic fingerprint [6].For this purpose, a region of interest (ROI) of 25 pixels was carefully selected by the pathologist containing mostly thyrocytes. Inter-patient variability of their specific spectra profiles was investigated analysing specimens from two malignant and two benign thyroid lesions (randomly chosen among all the samples).

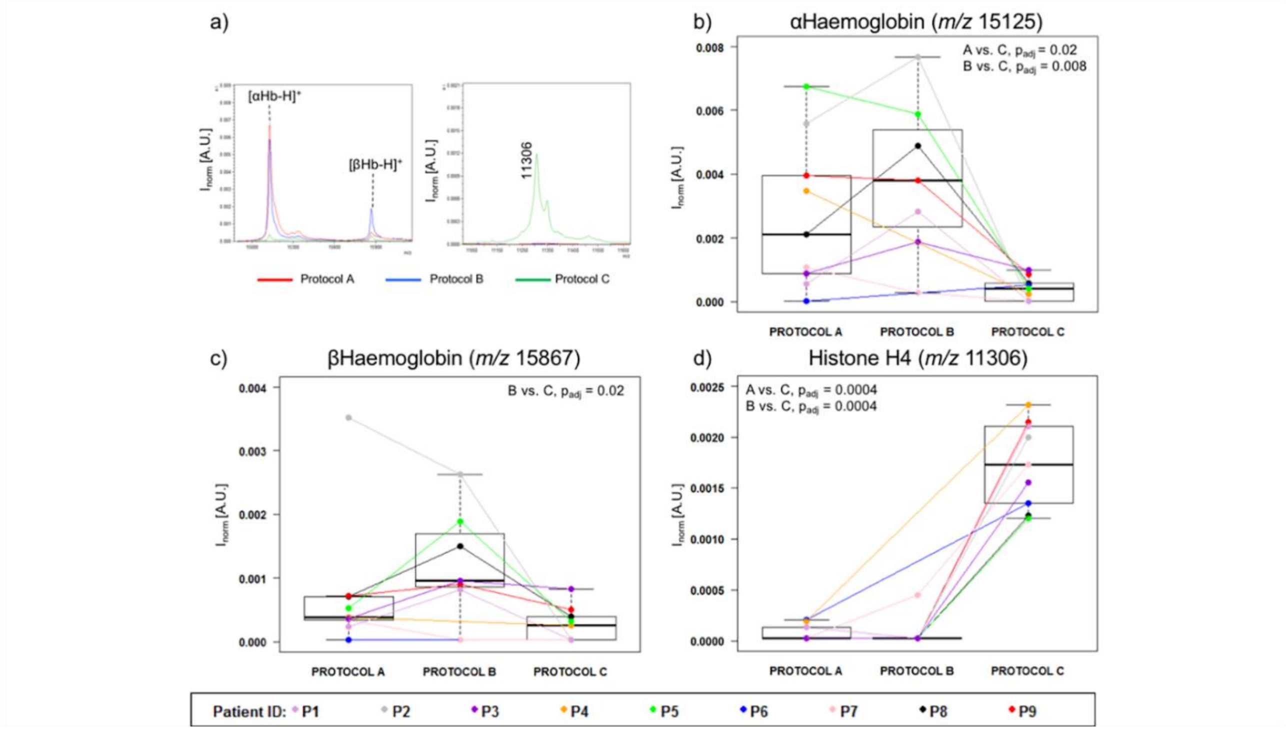


Figure 1: Comparison of the three independent sample preparation protocols using thyroid ex-vivo cytological samples from the same patients: a) spectra overlap of protocol A (red), protocol B (blue) and protocol C (green) from a single patient, zoomed in the regions m/z 15000-16000 and m/z 11000-11650; b) boxplots and individual values of the normalized intensity [A.U.] of αHaemoglobin, c) βHaemoglobin and d) Histone H4. The box contains data that fall between the first and third quartiles, the horizontal line indicates the median, and the brackets delineate 1.5 times the interquartile range (with data outside this range defining outliers).

A high degree of heterogeneity was evident when protocols A and B were used (Figure 2) due to the presence, or absence, of haemoglobin and its suppressing effect on any other protein signals. On the contrary, two comparable profiles, in which the haemoglobin interference was no longer a limiting factor, were obtained with protocol C (Figure 2). The number of total signals observed for each protocols in the samples a, b, c and d, as well as the number of common signals (peaks with a S/N≥6 present in both spectra) between THY5 samples (a vs. b) and THY2 samples (c vs. d) are shown in Supplementary Table 1 in appendix A. It is evident that when using protocols A and B the number of common signals is very low, in a range from 0 to 5 common signals, whereas with protocol C the number of common peaks increases noticeably. Moreover, Histone H4 and Haemoglobin signal intensities, taken from the ROIs, among the three protocols (Supplementary Figure 3 in appendix A) show the same trends observed in the average spectra, endorsing that protocol C clearly reduce ion suppression. The prospect to correlate molecular and cytological images and to select specific regions of interest, such as thyrocytes clusters, makes the MALDI-MSI proteomic approach highly valuable and suitable to enter the diagnostic routine in order to support the pathologist in the diagnosis of cytological samples.

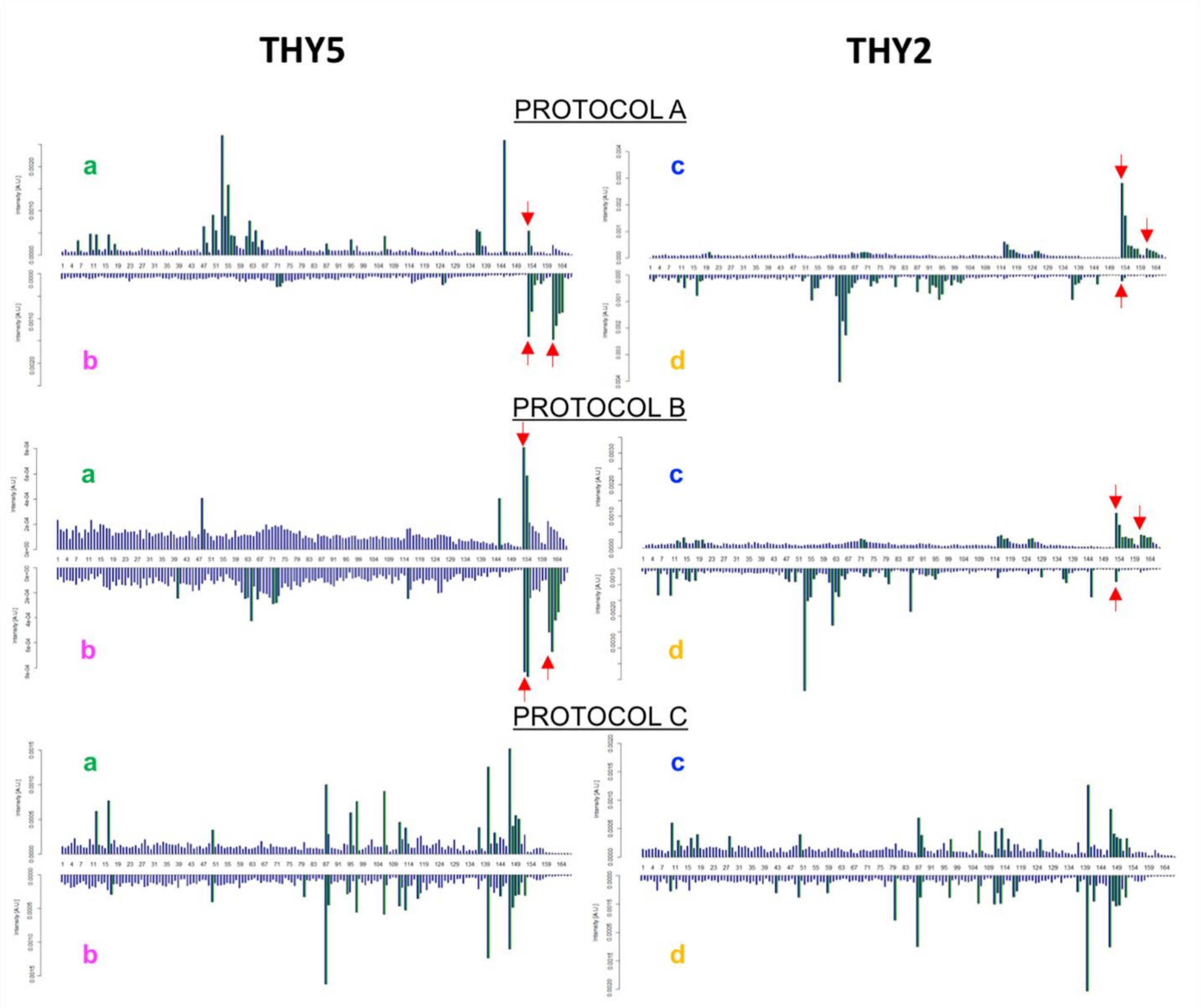


Figure 2. Histograms obtained from thyrocyte regions of interest taken from ex-vivo cytological samples of two THY5 patients (a, b) and two THY2 patients (c, d) prepared with protocols A, B and C. A total number of 167 peaks were detected in at least 5% of all the samples with S/N≥6. In the histogram of each single patient, the bars half coloured in green represent the peaks with S/N≥6 whereas the blue bars represent the peaks with S/N<6. The y-axes are the normalised intensity [A.U.] of the signals. The red arrows indicate the bars who correspond to αHb and βHb

The presence of haemoglobin signals per se is not the only challenging aspect; in fact, improper sample handling and collection could cause red blood cell haemolysis even in areas of sample where they are not present. During sample preparation with protocols A and B, the cytological sample is smeared onto the slide and this approach could lead to the rupture of erythrocyte cells membranes, thus the haemoglobin is released all over the sample. With protocol C, red blood cells haemolysis was markedly reduced, due to the minimal amount of red blood cells that remain in the sample following the treatment. In fact, the mechanical filtration step simplifies the complexity of the cytological sample by minimising not only the amount of red blood cells but also of debris and inflammatory cell [16]. As it stands, LBP in thyroid FNAB is an acceptable alternative to traditional smears [17,18]. Pathologists have now learnt to recognise specific diagnostic criteria related with thyroid LBP, such as nuclear modifications (the size is smaller, the nuclear-to-cytoplasmic ratio is bigger, nucleoli are prominent, grooves more obvious and pseudoinclusions are less evident). The cytoplasms are also less abundant in LBP. These changes are most probably due to a lack of the smearing effect that can be a potential cause of dry or degenerative artefacts in conventional smears [19]. Moreover, with protocol C, cells are distributed in a monolayer and it was much easier to select a ROI of only one type of cellular component. On the contrary, when conventional smears were used, the overlap of several cellular aggregates containing different cell types was commonly observed. The protocol C, that was able to reduce Hb contamination, was then evaluated with real in-vivo FNABs specimens, in comparison to protocol A, used in previous thyroid ex vivo studies [4–6]. The same trend, observed for protocol A and C of thyroid ex-vivo cytological samples, was also observed for thyroid FNABs (Supplementary Figure 4c in appendix A). Haemoglobin was detected in 7 out of the 7 conventional air-dried smears (protocol A), whereas Histone H4 was not detected in any of the samples. On the contrary, FNABs specimens treated with protocol C showed low amount of haemoglobin and an enhancement of any other proteins signals in the m/z 3000-16000 range. We also noticed that, when thyroid FNABs specimens were treated with protocol A, the rate of unusable samples was higher than that for ex-vivo cytological samples, since the vasculature of the neck skin is an additional source of haemoglobin contamination. The co-registration of the H&E stained image of the ThinPrep monolayer sample with the molecular image enabled us to underline the localisation of the signals present in our spectra with specific compartments (Supplementary Figure 4d in appendix A). As shown in Figure S4d, the signal at m/z 5063 was localised in the stroma and the one at m/z 11306 in the thyrocytes.

4.1.4 Conclusions

In conclusion, we highlight the possible role of LBP for the MALDI-MSI analysis of thyroid cytological samples. In particular, the LBP workflow based upon Protocol C allowed us to manage the haemoglobin interference, obtaining high-quality spectra to be used for a more reliable in situ profile comparison. The application of this protocol to in-vivo cytological samples should enable the discovery of protein biomarkers that can potentially assist cytopathologists in the diagnosis of thyroid nodules by integrating morphological information with proteomics data.

Bibliography

[1] E.H. Seeley, R.M. Caprioli, Molecular imaging of proteins in tissues by mass spectrometry. PNAS. 2008; 105:18126–18131. doi:10.1073/pnas.0801374105.

[2] N. Mosele, A. Smith, M. Galli, F. Pagni, F. Magni, MALDI-MSI Analysis of Cytological Smears: The Study of Thyroid Cancer, in: L.M. Cole (Ed.), Imaging Mass Spectrometry, Springer New York, New York, NY, 2017: pp. 37–47. doi:10.1007/978-1-4939-7051-3_5.

[3] K. Schwamborn, R.C. Krieg, S. Uhlig, H. Ikenberg, H. Wellmann, MALDI imaging as a specific diagnostic tool for routine cervical cytology specimens. Int J Mol Med. 2011; 27. doi:10.3892/ijmm.2010.587.

[4] F. Pagni, G. De Sio, M. Garancini, M. Scardilli, C. Chinello, A.J. Smith, F. Bono, D. Leni, F. Magni, Proteomics in thyroid cytopathology: Relevance of MALDI-imaging in distinguishing malignant from benign lesions. PROTEOMICS. 2016; 16:1775–1784. doi:10.1002/pmic.201500448.

[5] V. Mainini, F. Pagni, M. Garancini, V. Giardini, G. De Sio, C. Cusi, C. Arosio, G. Roversi, C. Chinello, P. Caria, R. Vanni, F. Magni, An Alternative Approach in Endocrine Pathology Research: MALDI-IMS in Papillary Thyroid Carcinoma. Endocr Pathol. 2013; 24:250–253. doi:10.1007/s12022-013-9273-8.

[6] F. Pagni, V. Mainini, M. Garancini, F. Bono, A. Vanzati, V. Giardini, M. Scardilli, P. Goffredo, A.J. Smith, M. Galli, G. De Sio, F. Magni, Proteomics for the diagnosis of thyroid lesions: preliminary report, Cytopathology. 2015; 26:318–324. doi:10.1111/cyt.12166.

[7] M. Melany, S. Chen, Thyroid Cancer: Ultrasound imaging and fine-needle aspiration biopsy. Endocrinol Metab Clin North Am. 2017; 46:691–711. doi:10.1016/j.ecl.2017.04.011.

[8] J.M. Amann, P. Chaurand, A. Gonzalez, J. Mombley, P. Massion, D. Carbone, R. Caprioli, Selective Profiling of Proteins in Lung Cancer Cells from Fine-Needle Aspirates by Matrix-Assisted Laser Desorption Ionization Time-of-Flight Mass Spectrometry. Clin Canc Res. 2006; 12:5142–5150. doi:10.1158/1078-0432.CCR-06-0264.

[9] B.E. Howitt, S. Chang, M. Eszlinger, R. Paschke, M.G. Drage, J.F. Krane, J.A. Barletta, Fine-Needle Aspiration Diagnoses of Noninvasive Follicular Variant of Papillary Thyroid Carcinoma. Am J Clin Pathol. 2015; 144:850–857. doi:10.1309/AJCPEIE12POICULI.

[10] J. Linder, Recent advances in thin-layer cytology. Diagn Cytopathol. 1998, 18:24–32. doi:10.1002/(SICI)1097-0339(199801)18:1<24::AID-DC5>3.0.CO;2-U.

[11] G. Ardito, E.D. Rossi, L. Revelli, F. Moschella, E. Giustozzi, G. Fadda, M.C. Marzola, D. Rubello, The role of fine-needle aspiration performed with liquid-based cytology in the surgical management of thyroid lesions. In Vivo. 2010; 24:333–337.

[12] A. Smith, I. Piga, M. Galli, M Stella, V. Denti, M. Del Puppo, F. Magni, Matrix-Assisted Laser Desorption/Ionisation Mass Spectrometry Imaging in the Study of Gastric Cancer: A Mini Review. Int J Mol Sci. 2017 Dec; 18(12): 2588. doi:10.3390/ijms18122588.

[13] I. Piga, B. Heijs, S. Nicolardi, L. Giusti, L. Marselli, P. Marchetti, M.R. Mazzoni, A. Lucacchini, L.A. McDonnell, Ultra-high resolution MALDI-FTICR-MSI analysis of intact proteins in mouse and human pancreas tissue. Int J Mass Spectrom. 2019. 437:10-16. doi: 10.1016/j.ijms.2017.11.001.

[14] S. Gibb, K. Strimmer, MALDIquant: a versatile R package for the analysis of mass spectrometry data. Bioinformatics. 2012; 28:2270–2271. doi:10.1093/bioinformatics/bts447.

[15] L.M. Zhang, M.A. Freitas, J. Wickham, M.R. Parthun, M.I. Klisovic, G. Marcucci, J.C. Byrd, Differential expression of histone post-translational modifications in acute myeloid and chronic lymphocytic leukemia determined by high-pressure liquid chromatography and mass spectrometry. J Am Soc Mass Spectrom. 2004; 15(1):77-86. doi: 10.1016/j.jasms.2003.10.001.

[16] C.W. Michael, B. Hunter, Interpretation of fine-needle aspirates processed by the ThinPrep technique: Cytologic artifacts and diagnostic pitfalls. Diagn Cytopathol. 2000; 23:6–13. doi:10.1002/1097-0339(200007)23:1<6::AID-DC2>3.0.CO;2-F.

[17] E.D. Rossi, G.F. Zannoni, S. Moncelsi, E. Stigliano, G. Santeusanio, C.P. Lombardi, A. Pontecorvi, G. Fadda, Application of Liquid-Based Cytology to Fine-Needle Aspiration Biopsies of the Thyroid Gland. Front Endocrinol. 2012; 3:57. doi:10.3389/fendo.2012.00057.

[18] E. Keyhani, S.A. Sharghi, R. Amini, S.A. Sharghi, M. Karimlou, F.A. Moghaddam, B. Larijani, Liquid base cytology in evaluation of thyroid nodules. J Diabetes Metab Disord. 2014; 13:82. doi:10.1186/s40200-014-0082-5.

[19] Y. Chong, S.-J. Ji, C.S. Kang, E.J. Lee, Can liquid-based preparation substitute for conventional smear in thyroid fine-needle aspiration? A systematic review based on meta-analysis. Endocr Connect. 2017; 6:817–829. doi:10.1530/EC-17-0165.

4.2 Feasibility study for the MALDI-MSI analysis of thyroid fine needle aspiration biopsies: evaluating the morphological and proteomic stability over time

4.2.1 Introduction

In clinical application and molecular pathology, matrix-assisted laser desorption/ionization mass spectrometry imaging (MALDI-MSI) is an emerging technology which enables the spatial distribution of biomolecules within tissue to be combined with the traditional morphological information [1]. Hence, for diagnostic, or prognostic purposes, along with predicting response to therapeutic treatment, biological specimens, such as tissues and biopsies, must be properly collected and handled in order to assure the preservation of the morphological structure and the proteomic profile, avoiding degradation or alteration phenomena, contamination and artefacts formation [2,3]. So far, several attempts have been aimed at preventing sample degradation [4]. Two of the most commonly used approaches for sample preservation and stabilization in pathology and proteomics are chemical fixation (e.g. formalin followed by paraffin wax embedding) [5,6] and snap-freezing [7]. Another approach developed to preserve proteins from degradation is the use of an additive-free procedure based on heat-fixation of the tissue [8]. The integrity of the complexity of cellular morphology is usually maintained, but changes in the fine structure has been observed [9]. In recent years, MALDI-MSI has been used in combination with conventional air-dried cytological smears with the ability to generate specific protein profiles for malignant, benign and different subtypes of thyroid lesions using fine needle aspiration biopsies (FNABs) specimens [10–12]. Centralized MSI analysis, typically carried out in multicenter studies, are challenging because the cytological smears have to be carefully shipped, avoiding sample degradation. Nowadays, the liquid based preparation (LBP) [13] is the gold standard for cervical preparation, which is usually performed by alcohol fixation of the conventional smear [14]. Moreover, the LBP is also more widely used among cytopathologists for evaluating thyroid FNABs specimens [15]. Regarding cytological samples, the use of methanol-based, buffered preservative mediums, such as CytoLyt and PreservCyt solutions during transportation became increasingly diffuse in the cytopathological practice [16]. These solutions avoid protein precipitation, lyse red blood cells, reduce the amount of mucus and preserve the morphology of the cytological samples during transportation and slide preparation [16,17]. Based on manufacturer’s instructions, the morphological stability of the cytological sample is guaranteed until six weeks when the sample is stored between 30 °C and 4 °C in the PreservCyt solution, whereas in CytoLyt up to 8 days at room temperature. Numerous studies [18–21] have investigated the stability of human genomic and human papilloma virus (HPV) DNA and RNA from cervical cytological samples in PreservCyt medium, underlining their stability for extended periods. Cuschieri et al. proved that HPV RNA within clinical cervical samples were stable in PreservCyt up to 14 days at room temperature [19]. Moreover, Tarkowski et al. have demonstrated that RNA was suitable for successful molecular assays, such as RT-PCR, even after one year of storage [20]. However, there are no studies evaluating how thyroid FNABs sample preparation in CytoLyt and PreservCyt solutions influence their stability over the time of storage for proteomic studies. Accordingly, assessing the stability of cytological specimens is of paramount importance for the collection of samples based on a robust and simple protocol to be implemented in clinical pathology units. In the present study, we investigated the morphological and proteomic stability over time, evaluating the intra-day and inter-day variability of the data generated by MALDI-MSI, analysing thyroid FNABs prepared with the proposed protocol. In addition, we explored the morphological and proteomic stability of the thyroid FNABs after 7 days, 14 days and 2 months of storage at 4 °C in the preservative solutions.

4.2.2 Material and Methods

4.2.2.1 Chemicals and Reagents

ThinPrep® CytoLyt and ThinPrep® PreservCyt solutions (methanol-based buffered preservative solutions) were purchased from Hologic (Marlbourough, MA, USA). Sinapinic acid matrix was purchased from Bruker Daltonics (Bremen, Germany). All the other chemicals were purchased from Sigma-Aldrich (Milan, Italy).

4.2.2.2 Sample collection and preparation

Fine needle aspirations were taken from (n=19) patients who underwent ultrasound-guided procedures at the Department of Radiology, San Gerardo Hospital, Monza, Italy. Cytological samples were diagnosed as THY1c/not diagnostic-cystic (n=3), THY2/benign (n=9), THY3/indeterminate (n=2), THY4/suspicious for malignancy (n=1) and THY5/malignant (n=4) according to the British system for reporting thyroid cytopathology. THY4 and THY5 samples were confirmed to be malignant by post-operative histopathology. Supplementary Table 1 in appendix B summarizes the demographic and clinical characteristics of the enrolled subjects. Samples were collected with a 25-Gauge (G) needle and immediately transferred into a falcon tube filled with CytoLyt solution. The Ethics Committee of the hospital San Gerardo Hospital, Monza, Italy, approved the study (AIRC MFAG - 2016). Cytological samples deposited into CytoLyt solution were centrifuged at 800 g for 10 minutes at room temperature, using a Centrifuge 5804 R equipped with an S-4-72 rotor (Eppendorf, Hamburg, Germany), the supernatant was discarded, the pellet was re-suspended in 200 µL of PreservCyt solution and transferred in an eppendorf tube. Subsequently, samples were centrifuged at 800 g for 10 minutes at room temperature using a Centrifuge 5424 R equipped with a FA-45-24-11 rotor (Eppendorf, Hamburg, Germany); the supernatant was discarded and the pellet resuspended in a final volume of 100 µL of PreservCyt solution. When the amount of cellular material was adequate, the sample was equally divided in order to obtain multiple replicates. Finally, samples were transferred onto indium tin oxide (ITO) conductive slides (Bruker Daltonics, Bremen, Germany) by centrifugation (800 g for 15 minutes, at room temperature) using a Hettich® ROTOFIX 32A centrifuge equipped with a Swing-Out rotor 1624, carriers 1660 and slide carriers for two chambers 1670 (Hettich Lab Technology, Tuttlingen, Germany). The cytospin preparations were prepared using cyto chambers with a diameter of 6.2 mm in order to obtain a monolayer of cells. A maximum of eight cytospin spots were positioned onto one ITO-conductive slide. ITO-slides with cytospin samples were dried under vacuum for 30 minutes. Finally, consecutive washing steps of 30 seconds each, with increased concentration of ethanol (70%, 90% and 95%), were performed in order to remove salt contamination. The slides were then dried under vacuum for 15 minutes and stored at -80°C until the day of the analysis.

4.2.2.3 MALDI-MSI Sample preparation

Cytospin cytological specimens were brought to room temperature and dried under vacuum for 30 minutes. MALDI matrix (10 mg/ml sinapinic acid in 60:40 acetonitrile:water w/0.2% trifluoroacetic acid) was uniformly deposited, with an optimized method, using the iMatrixSpray (Tardo Gmbh, Subingen, Switzerland) automated spraying system.

4.2.2.4 Experimental design

FNABs specimens were split into several samples for the intra-day (n=7 patients) and inter-day repeatability (n=5 patients) evaluations. In addition, the sample stability in PreservCyt solution after 7 days (n=6 patients), 14 days (n=6 patients) and 2 months (n=2 patients), and the sample stability in CytoLyt solution after 7 days (n=2 patients) were studied. The malignant samples (n=5) were prepared at t0 in PreservCyt solution. The general workflow of the study is represented in Figure 1 and the experimental design is summarized in Supplementary Table 2 in appendix B. Samples were stored in PreservCyt and CytoLyt solutions at 4 °C until the day of the cytospin deposition onto the ITO-slides.

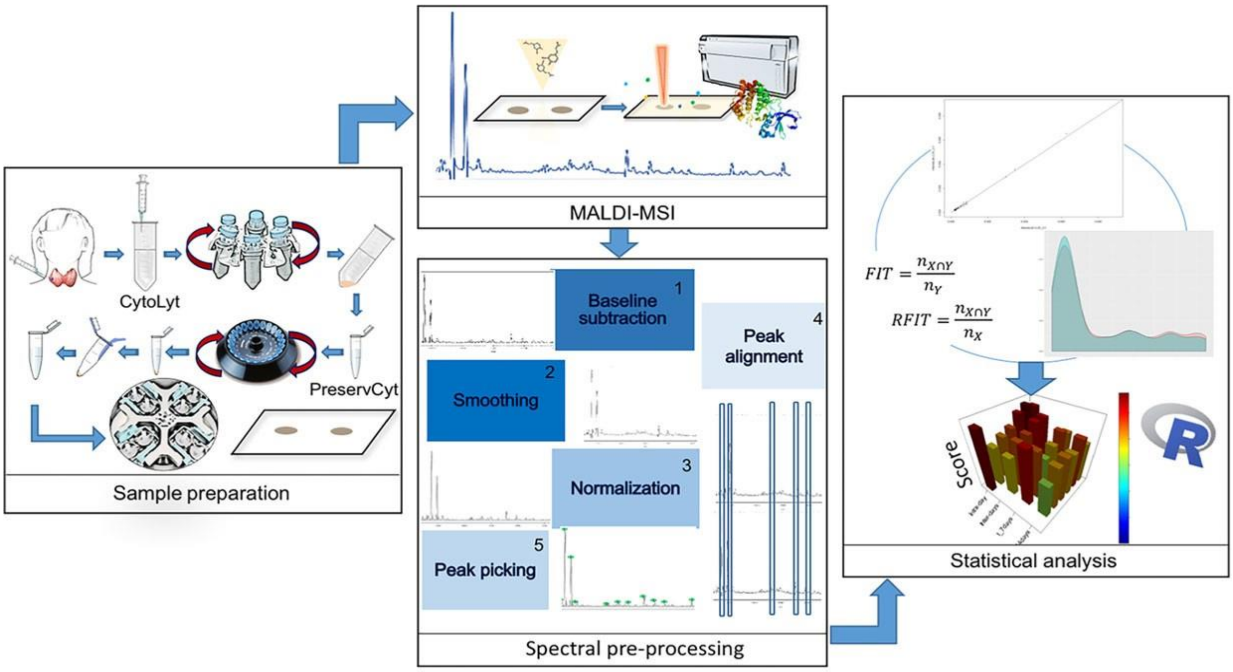


Figure 4. General Workflow: i) Cytological sample preparation, ii) MALDI-MSI sample preparation and analysis, iii) spectral pre-processing and iv) statistical analysis

4.2.2.5 MALDI-MSI analysis

MALDI time of flight (TOF) MSI was performed using an ultrafleXtreme MALDI-TOF/TOF (Bruker Daltonik GmbH) in positive-ion linear mode, using 300 laser shots per spot, with a laser focus setting of 3 medium (diameter of 50 µm) and a raster width of 50 x 50 µm. A mixture of standard proteins within the *m/z* range of 5,730 to 16,950 (Protein Calibration Standard I, Bruker Daltonics) was used for external calibration. Spectra were recorded in the *m/z* range 3,000-20,000. Data acquisition and visualization were performed using the flex software package by Bruker Daltonics (flexControl 3.4, flexImaging 4.1). After MALDI-MSI analysis, the MALDI-matrix was removed by increasing concentration of EtOH (70% and 95%) and the cytological specimens were stained with hematoxylin and eosin (H&E). High resolution cytological images were recorded using a ScanScope CS digital scanner (Aperio, Park Center Dr., Vista, CA, USA).

4.2.2.6 Data analysis

Overall average mass spectra from the MALDI-TOF-MSI datasets were exported in CSV format and loaded in the open-source R software v.3.4.3 to perform the pre-processing operations that were carried out using the MALDIquant R package [22]. Baseline subtraction with SNIP method and iteration 100; moving average smoothing used half window size of 2; total ion current (TIC) normalization, divided spectrum intensities by the sum of all the intensities values of the spectrum itself; spectra alignment and peak picking with a S/N of 6, were performed. Spectra alignment was verified using the open-source software mMass v5.5 (http://www.mmass.org) [23,24].

The similarity of the mass spectra was evaluated by using two score systems, accounting for the frequency of common peaks and the matching of signal intensity.

Assume $X = (x_i)_{i=1,\dots,N_X}$and $Y = (y_i)_{i=1,\dots,N_Y}$are the sequences of the intensities in the reference (X) and the query spectra (Y), where N_X and N_Y are the cardinality of the *m/z* values. The first score system (S3) ranges from 0–3 and is the sum of three components (i.e. fit, retrofit and spearman's correlation) which contributes 1 at most [25]. The fit is defined as the ratio of the common peaks in the two spectra and the N_Y peaks detected in the query spectrum:

$$FIT = \frac{n_{X\cap Y}}{n_Y},$$

while, the retrofit is defined as the ratio of common peaks in the two spectra and the N_X peaks in the reference spectra:

$$RFIT = \frac{n_{X\cap Y}}{n_X},$$

where $n_{X\cap Y}$ is the number of shared mass peaks. The Spearman's Correlation is a measure of association between the ranks of the intensities of the common peaks $n_{X\cap Y}$:

$$\rho_s = \frac{\sum_j^{n_{X\cap Y}}[(r_j - \bar{r}) * (s_j - \bar{s})]}{\sqrt{\sum_j^{n_{X\cap Y}}(r_j - \bar{r})^2} * \sqrt{\sum_j^{n_{X\cap Y}}(s_j - \bar{s})^2}}$$

where r_j and s_j are the ranks of x_j and y_j ($j = 1, \dots, n_{X\cap Y}$), while $\bar{r}$ and $\bar{s}$ are their mean values.

The second score system (S4) ranges from 0–4 and extends S3 to include a fourth feature that measures the overlap (OV), which takes into account the whole shape of the two spectra. This latter index measures the overlapping area between the empirical distributions of two sequences of intensities on ranked *m/z*:

$$OV = \hat{F}^X_{n_{X\cup Y}} \cap \hat{F}^Y_{n_{X\cup Y}}$$

where $\hat{F}^X_{n_{X\cup Y}}$ and $\hat{F}^Y_{n_{X\cup Y}}$are the empirical distribution function, and $n_{X\cup Y}$ the *m/z* values either in the X or the Y spectra [26]. For calculations, we used the overlapping R package.

Mimicking an equivalence trial, in order to establish whether no meaningful difference exists between the proteomic profiles in time (i.e. t_0 vs. t_{7days} and t_0 vs. t_{14days}), the 95% confidence intervals (CI) of the observed mean similarity indices should be inside a pre-specified interval of equivalence [27]. Since no recognized reference exist, we have conservatively used the CI calculated on the inter-day comparison.

To further assess spectra similarity in time, Principal Component Analysis (PCA) and hierarchical clustering analysis were also performed. PCA was carried out with the prcomp function in the stats R package. Data were scaled and centered before the analysis due to PCA being sensitive to different scales of features. Hierarchical clustering analysis was performed with the function hclust in the stats R package, using complete linkage method to show similar clusters on the selected principal components that explained as much as possible variance of the original independent variables.

4.2.3 Results and Discussion

Tissue specimens need to be properly handled to ensure not only the integrity of the morphological structure but also to avoid protein degradation [2]. Several protocols have been developed for this purpose, such as chemical fixation followed by paraffin wax embedding and snap-freezing [4]. Heat-fixation of the tissue is also used, but changes in the fine structures of the cells has been observed [9]. The stabilization of cytological samples, based on CytoLyt and PreservCyt solutions, is employed in the clinical laboratory to preserve cervical cytological samples for the analysis of RNA and DNA after long-term storage [14,20].

Here, we propose a sample preparation protocol for the analysis of thyroid FNA by MALDI-MSI that combines stabilization in preservative solutions followed by cytospin deposition. In particular, we investigate the experimental repeatability of the proteomics analysis of cytological samples and their stability in time in both preservative solutions, focusing on mass spectra similarities.

4.2.3.1 Cytospin sample preparation: morphology evaluation

The washing steps in preservative solutions followed by cyto-deposition were able to guarantee high cellular adequacy (Supplementary Figure 1 in appendix B). Cell morphology was generally satisfactory in different specimens from the same patient, independently from the time of storage in the preservative media (Supplementary Figure 1 in appendix B). With regard to the conventional smear [11], our protocol has the advantage of being more efficient, reducing the sample-to-sample variability and enabling up to 8 spots to be placed onto a single ITO-slide. Moreover, the cytospin spot size is approximately 6 mm in diameter, while the conventional smear size is extremely variable and in the order of centimetres (approximately 2-3 cm). Therefore, the time of MALDI-MSI analysis of one cytospin-spot is drastically reduced compared to the conventional smear.

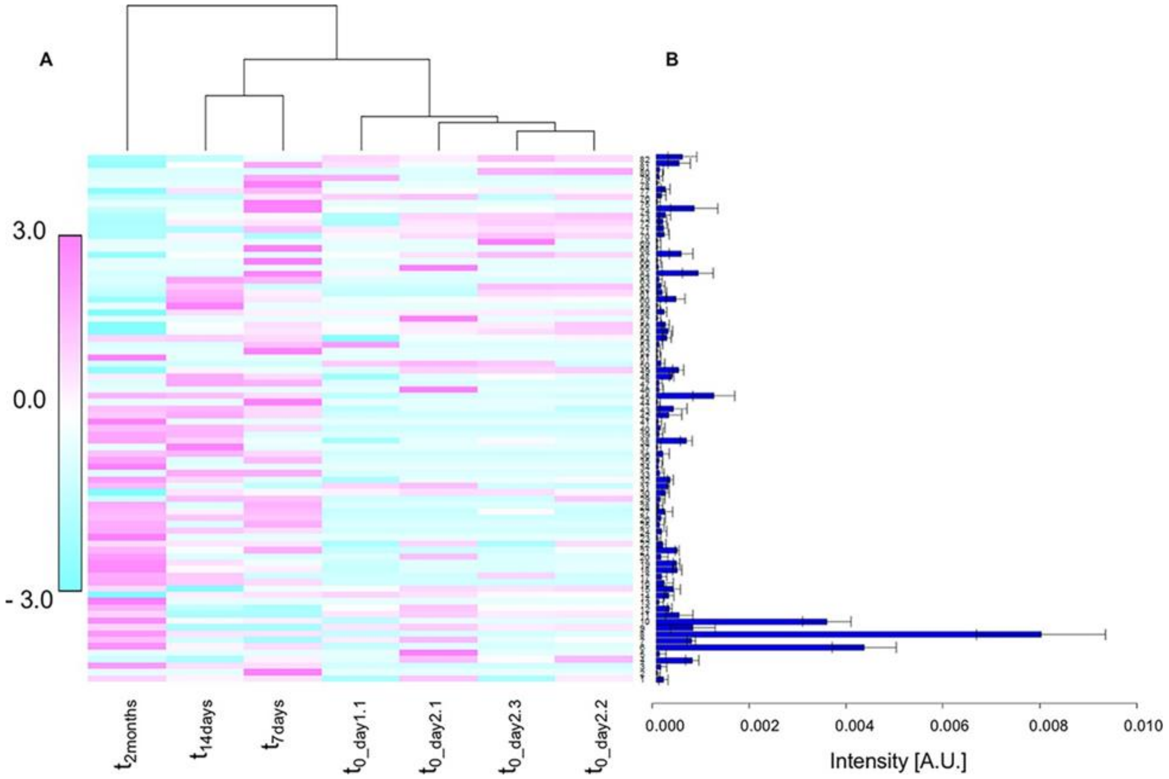


Figure 2. Normalized intensities [AU] of the peaks (M/Z) detected in seven replicates of patient P316 (t0 day1.1, t0 day2.1, t0 day2.2, t0 day2.3, t 7days, t 14days…….t 2months). A) The heatmap shows the signal intensities of each of the peaks that were rescaled to have a mean of 0 and SD of 1; B) the histogram shows the mean intensities of each of the peaks detected at different times (error bars represent the 95% confidence interval).

4.2.3.2 Mass spectra similarity: qualitative comparison

The variability of the signal intensities among replicates prepared at different times in one patient is reported in Figure 2 and, similarly, a comparison of spectra obtained from all the replicates is shown in Supplementary Figure 2A (ID: P316) as well as spectra comparison of multiple sample prepared at t_0 (Supplementary Figure 2B in appendix B). The heatmap in Figure 2A shows how the normalized signal intensities of the peaks in ranked *m/z* change among the replicates whose mean intensity values (and 95% CI) are represented in the distribution in Figure 2B. Furthermore, the dendrogram in the upper part of Figure 2A shows that the t_0 intra-day replicates are highly similar even if a slight separation is seen in samples evaluated in t_{0_day1} versus t_{0_day2}, most probably reflecting an additional source of analytical variability. The distance between t_{7days}/t_{14days} and t_0 is higher than the one observed between t_{0_day1} and t_{0_day2}, since these samples were stored in the *PreservCyt* solution for 7 days and 14 days before deposition onto the ITO-slides. Nevertheless, the two spectra, t_{7days} and t_{14days}, were clustered together, also reflecting the same sources of analytical variability (i.e. cytospin sample preparation, MALDI-MSI sample preparation, instrumental analysis) given that they were similar to the spectra at t_0 (Supplementary Figure 2A in appendix B). The spectrum of the replicate at $t_{2months}$ is very marginal, suggesting that longer storage time in the *PreservCyt* does not preserve the specimens (Figure 2A) and, looking at the spectra in Supplementary Figure 2A, it is evident how peak intensities and the spectra as a whole change with respect to the other replicates. On the contrary, the variability of the peak intensities among replicates of the same patient until 14 days seems to not depend on the increased storage time in the *PreservCyt* solution, but more likely reflects the analytical variability due to the cytospin-deposition onto the ITO-slides in different days. These results are in good agreement with those previously reported for cervical cytological samples that were stable until 14 days [19]. Similar results are seen in the short (t_{0_day1} and t_{0_day2}) and in the long term (from t_0 to $t_{2months}$) in Supplementary Figure 3 (P262) and Supplementary Figure 4 (P319) in appendix B, respectively.

The unsupervised learning method PCA was employed to further investigate and visualize mass spectra similarity among all those obtained from all the patient replicates (Supplementary Figure 5A in appendix B). From the PCA score chart, it is evident that replicate samples from the same patient are grouped together and that the similarity of spectra were preserved independently from the time of storage in *PreservCyt* solution. Moreover, the hierarchical clustering analysis (Supplementary Figure 5A in appendix B) highlights how malignant samples are clustered together and separated from benign samples. When further replicates have been prepared at t_0 (intra-day and inter-day), one sample (for each patient) was randomly chosen as reference among all the intra/inter-day replicates. To evaluate the intra/inter-day variability at t_0, two query spectra were randomly chosen from the remaining intra-day and inter-day spectra, respectively. Finally, the spectrum of each replicate prepared at different time of storage in *PreservCyt* solution were compared with the reference spectrum. Figure 3 illustrates all the paired comparisons between replicates in one patient (ID: P316), using the reference spectrum $t_{0\_day2.2}$. The graphs in A and B are obtained considering only the common peaks between spectra. In graphs A, the normalized intensities of the peaks in the reference spectra (*x*-axis) are plotted against the normalized intensity of the peaks of the query spectra (*y*-axis). When the intra-day and inter-day replicates are compared with the reference spectrum, common peaks have very similar intensities, since the points lie mainly near the bisector. In graphs B, the common peaks are ranked with respect to their increasing intensities in the reference spectra (*x*-axis) and plotted versus the normalized intensities (*y*-axis). The way in which the intensity of the common peaks increased are very similar in the query and reference spectra, when the intra/inter-day comparisons are considered. Dissimilarities in the highest signals intensities, in the comparisons with t_{7days}, t_{14days} and $t_{2months}$, were observed. However, as previously stated, these differences are not entirely surprising. The analytical variation of the peak intensities is a well-known problem in MALDI protein analysis. It was reported that the mean CV in the peak intensities for protein profiling varies among studies from 4% to 26% [28,29]. This variability is strictly related to the MALDI-sample preparation, which involved matrix deposition (crystals heterogeneity), and the desorption/ionization processes [30]. However, it also depends on the heterogeneity of the tissues consecutive sections or, like in our experiments, on the intra-sample heterogeneity of cytological replicates. For this reason, the graphs presented in C provide a better representation since they considered the whole spectra and take into account overlap of the complete spectra (see section 2.6). The percentage of the overlap of the comparisons varies from 98-93% for the intra/inter-day, to 51% for the one at $t_{2months}$. The shape of the spectra density is conserved for the comparisons until t_{14days}, where the overlap is 80%. Noticeably, the minimum OV index is 74% when all patients are considered.

4.2.3.3 Mass spectra similarity: quantitative scores for stability evaluation

To better investigate the stability of the samples over time, we quantified the degree of mass spectra similarity with a score (S3) that we derived from a previous study [25] We also considered a new score (S4) that equally weight the number of signals and peak intensities, differently from S3, that placed more emphasis on the number of signals. In order to evaluate intra-day and inter-day repeatability, both S3 and S4 were calculated for all the possible comparisons in each patient (e.g. for P316, intra-day: t0_day2.1 vs. t0_day2.2, t0_day2.1 vs. t0_day2.3, t0_day2.2 vs. t0_day2.3; inter-day: t0_day2.1 vs. t0_day1.1, t0_day2.2 vs. t0_day1.1, t0_day2.3 vs. t0_day1.1). The distributions of intra/inter-day of S3 and S4 values in the box plots of Supplementary Figure 6 (in appendix B) overlap, showing slight heterogeneity. This is also underlined by the CV values reported in Supplementary Table 3 in appendix B, which ranged from 7.37 to 12.43. Moreover, the CVs calculated in all the replicates of each subject reached values below 12.31% (Supplementary Table 5 in appendix B). When one query spectrum at t0_intra-day, t0_inter-day, t7days and t14 days was compared with one reference spectrum randomly selected among the t0 replicates, no remarkable differences among the scores of all paired combinations were observed (Figure 4). The 95% CI of each paired comparison, using both S3 and S4 scores, are reported in Table 1, and show that the one calculated for the comparison of t0 vs. t7days is almost completely contained in the 95% CI of the inter-day comparison. This suggests that the two evaluations can be considered equivalent, but the same conclusion does not hold for the t_0 vs. t_{14days} comparison. However, the inter-day

CV was surprisingly low, compared to both literature and our experience using MALDI-MSI [28,29], with values of 12.03% and 10.54% for S_3 and S_4, respectively.

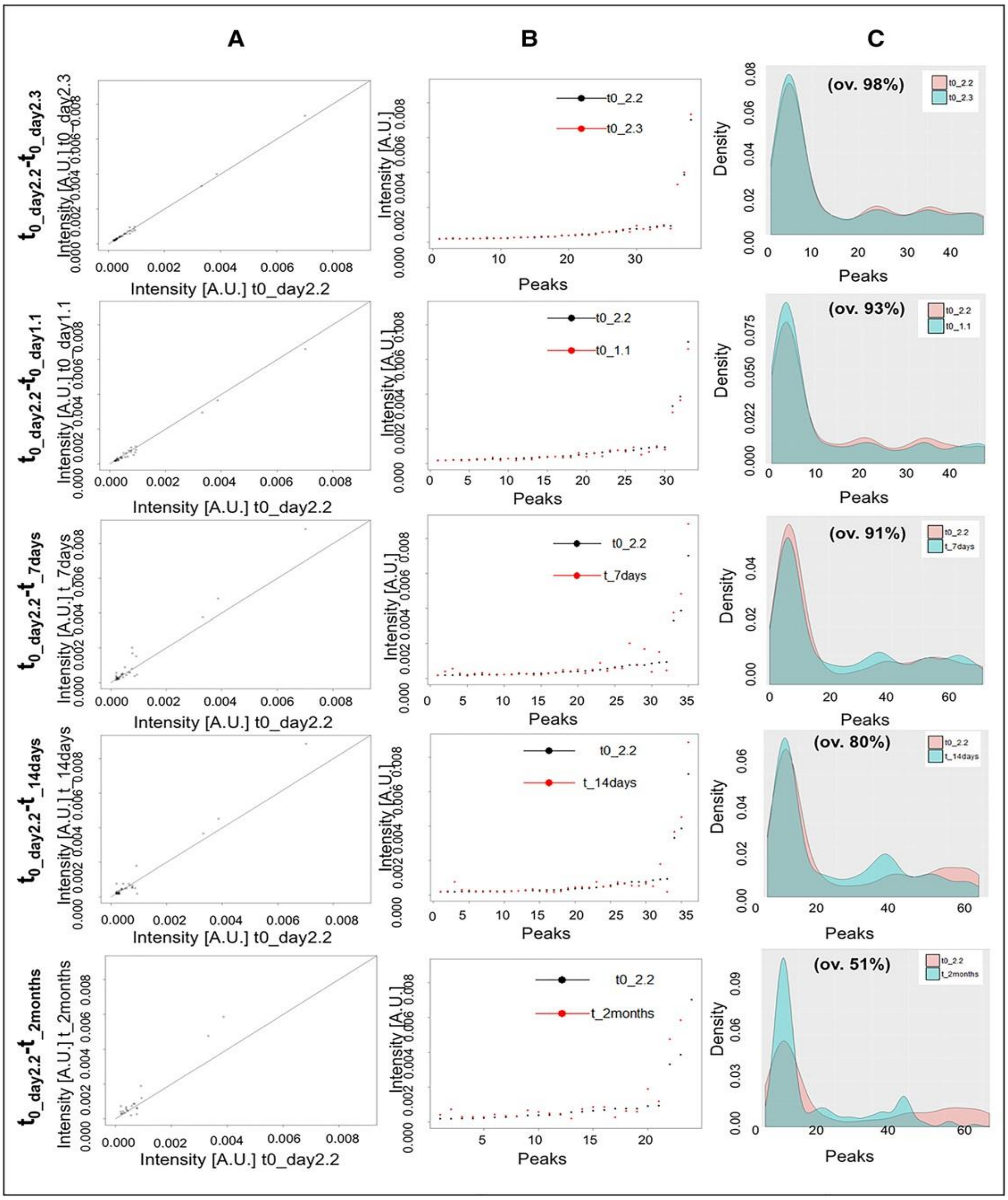


Figure 3. P316 comparisons between the reference spectrum $t_{0\ day2.2}$ and the query spectra $t_{\ day2.3}$, $t_{0\ day1.1}$, t_{7days}, t_{14days}, $t_{2months}$ prepared at different times of storage, are shown. A) Graphs show the normalized intensity [AU] of the peaks of the reference spectrum (x-axis) plotted against the normalized intensity [AU] of the peaks of the query spectra (y-axis). Only common peaks are considered. The black line is the bisector. B) Graphs show the common peaks ranked respect to their increasing intensities in the reference spectrum (x-axis) and plotted versus the normalized intensities [AU] (y-axis). The peaks intensities of the reference spectrum are represented by black dots, whereas the peaks intensities of the query spectra are represented by red dots. C) Graphs show the overlapping area of the shape of the spectra density of the reference (pink area) and the query (blue area).

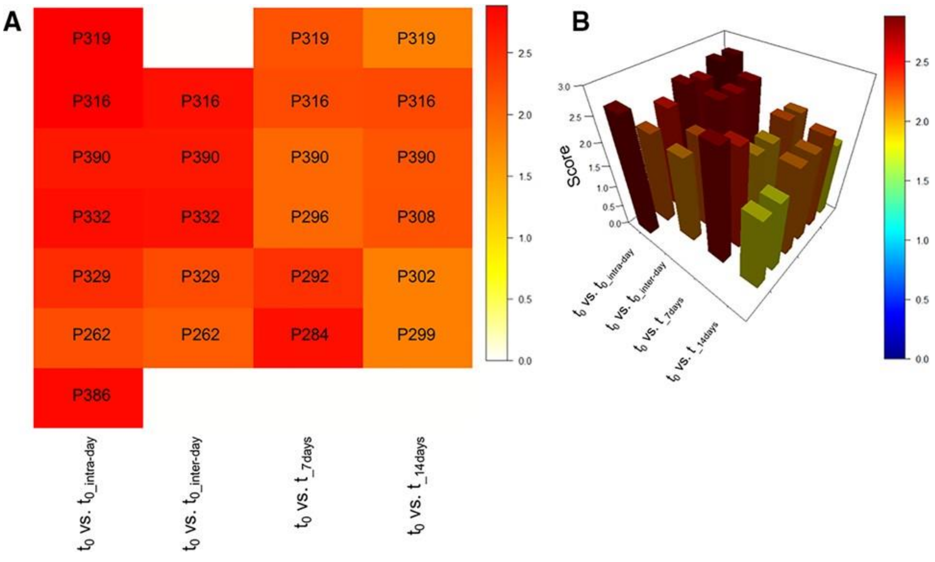


Figure 4. Heatmap (A) and 3-D histograms (B) of the S_3 score obtained from all the spectra comparison between a randomized reference spectrum t_0 and the query spectra $t_{0\_intra\text{-}day}$, $t_{0\_inter\text{-}day}$, t_{7days}, t_{14days}

Based on this consideration, we recalculated the 95% CI under the assumption of a 20% CV value, which is more representative of the routinely MALDI-TOF variability, and of 30%, in case of the worst acceptable hypothetical analytical session (Supplementary Table 4 in appendix B). When a CV of 20% and 30% were used to calculate the 95% CI of the comparison t_0 vs $t_{inter\text{-}day}$, a high intersection and a complete overlap where also observed for the t_0 vs. t_{14days} 95% CI. These results suggest no substantial deviations from t_0 when the cytological samples are stored in *PreservCyt* until 14 days. Both S_3 and S_4 scores were split in their components (fit and retrofit, spearman's correlation and overlap) to investigate their relative contribution (Supplementary Figure 7 in appendix B). Results showed that spearman's correlation was the one that most unfavorably affected the evaluation of mass spectra similarities (Supplementary Table 5 in appendix B), and this might depend on variations among peak intensities observed in MALDI-MSI experiments. Finally, to further ascertain the sample stability until 14 days, all spectra referred to a different storage time for three patients (P316, P319, P390) were compared to the t_0 spectrum of all patients (example: P262 t_0 versus t_{0_day1}, t_{0_day2}, t_{7days}, t_{14days} of P316) in order to assess whether differences between patients remain unchanged independently from the sample storage. Regardless of which spectra was used as a reference (t_{0_day1}, t_{0_day2}, t_{7days} or t_{14days}), the relation between spectra of two patients was not influenced by the time of storage. Indeed, in all paired combinations, the S_4 had an average CV of 10.72% (range 2.73% - 28.66%), which is comparable to the value observed in inter-day variability (CV = 10.43%, Supplementary Table 3 in appendix B). We also have investigated the stability after 7 days of storage in the *CytoLyt* solution (P384, P386). The mass spectra similarity scores (S_4)

between spectra at t_0 and t_{7days} were 3.04 and 2.98, respectively, and were beyond the 74.5% of the total score S_4. Although these results refer solely to two patients, they are in line with what we previously observed for the sample stability in the *PreserveCyt* solution and suggest that the samples also remain stable in the *CytoLyt* solution for at least one week.

4.2.4 Conclusions

In this work, we have assessed the morphological and proteomic stability of thyroid FNABs in *PreservCyt* (until 14 days of storage) and *CytoLyt* (until 7 days of storage) solutions, with MALDI-MSI analysis. In addition, we introduced a new feature in the similarity score to equally take into account the number of signals (fit and retrofit) and their intensities (spearman's correlation and spectra overlap). The major limitation of this study was the low cellularity of thyroid FNABs which cannot always be split in multiple replicates in order to increase the sample size. However, this study represents a step forward towards the implementation of MALDI-MSI, combined with a trustworthy and robust sample preparation methodology, into the cytopathology routine, integrating the morphology with the proteomics data to improve the diagnosis of thyroid FNABs and pave the way for a further study aimed at the classification of benign, malignant and indeterminate FNABs. Towards this second aim, we are now enrolling more than 500 patients of THY2-THY3 and THY4/5.

Likewise, the findings of this study could be useful and straightforwardly extended to other biological liquid based specimens.

Bibliography

[1] Aichler, M., Walch, A., MALDI Imaging mass spectrometry: current frontiers and perspectives in pathology research and practice. Lab. Invest. 2015, 95, 422–431.

[2] Bennike, T.B., Kastaniegaard, K., Padurariu, S., Gaihede, M., et al., Comparing the proteome of snap frozen, RNAlater preserved, and formalin-fixed paraffin-embedded human tissue samples. EuPA Open Proteomics 2016, 10, 9–18.

[3] Chandana, R., Mythri, R.B., Mahadevan, A., Shankar, S.K., Srinivas Bharath, M.M., Biochemical analysis of protein stability in human brain collected at different post-mortem intervals. Indian J. Med. Res. 2009, 129, 189–199.

[4] Unhale, S.A., Skubitz, A.P.N., Solomon, R., Hubel, A., Stabilization of Tissue Specimens for Pathological Examination and Biomedical Research. Biopreservation Biobanking 2012, 10, 493–500.

[5] Steiner, C., Ducret, A., Tille, J.-C., Thomas, M., et al., Applications of mass spectrometry for quantitative protein analysis in formalin-fixed paraffin-embedded tissues. PROTEOMICS 2014, 14, 441–451.

[6] Gustafsson, O.J.R., Arentz, G., Hoffmann, P., Proteomic developments in the analysis of formalin-fixed tissue. Biochim. Biophys. Acta BBA - Proteins Proteomics 2015, 1854, 559–580.

[7] Ericsson, C., Franzén, B., Nistér, M., Frozen tissue biobanks. Tissue handling, cryopreservation, extraction, and use for proteomic analysis. Acta Oncol. 2006, 45, 643–661.

[8] Svensson, M., Borén, M., Sköld, K., Fälth, M., et al., Heat Stabilization of the Tissue Proteome: A New Technology for Improved Proteomics. J. Proteome Res. 2009, 8, 974–981.

[9] Cazares, L.H., Van Tongeren, S.A., Costantino, J., Kenny, T., et al., Heat fixation inactivates viral and bacterial pathogens and is compatible with downstream MALDI mass spectrometry tissue imaging. BMC Microbiol. 2015, 15.

[10] Mainini, V., Pagni, F., Garancini, M., Giardini, V., et al., An Alternative Approach in Endocrine Pathology Research: MALDI-IMS in Papillary Thyroid Carcinoma. Endocr. Pathol. 2013, 24, 250–253.

[11] Pagni, F., De Sio, G., Garancini, M., Scardilli, M., et al., Proteomics in thyroid cytopathology: Relevance of MALDI-imaging in distinguishing malignant from benign lesions. PROTEOMICS 2016, 16, 1775–1784.

[12] Pagni, F., Mainini, V., Garancini, M., Bono, F., et al., Proteomics for the diagnosis of thyroid lesions: preliminary report. Cytopathology 2015, 26, 318–324.

[13] Tripathy, K., Misra, A., Ghosh, J., Efficacy of liquid-based cytology versus conventional smears in FNA samples. J. Cytol. 2015, 32, 17.

[14] Karnon, J., Peters, J., Platt, J., Chilcott, J., et al., Liquid-based cytology in cervical screening: an updated rapid and systematic review and economic analysis. Health Technol. Assess. Winch. Engl. 2004, 8, iii, 1–78.

[15] Rossi, E.D., Zannoni, G.F., Moncelsi, S., Stigliano, E., et al., Application of Liquid-Based Cytology to Fine-Needle Aspiration Biopsies of the Thyroid Gland. Front. Endocrinol. 2012, 3.

[16] Bonnier, F., Traynor, D., Kearney, P., Clarke, C., et al., Processing ThinPrep cervical cytological samples for Raman spectroscopic analysis. Anal Methods 2014, 6, 7831–7841.

[17] Norimatsu, Y., Ohsaki, H., Masuno, H., Kagawa, A., et al., Efficacy of CytoLyt® Hemolytic Action on ThinPrep® LBC Using Cultured Osteosarcoma Cell Line LM8. Acta Cytol. 2014, 58, 76–82.

[18] Agreda, P.M., Beitman, G.H., Gutierrez, E.C., Harris, J.M., et al., Long-Term Stability of Human Genomic and Human Papillomavirus DNA Stored in BD SurePath and Hologic PreservCyt Liquid-Based Cytology Media. J. Clin. Microbiol. 2013, 51, 2702–2706.

[19] Cuschieri, K.S., Beattie, G., Hassan, S., Robertson, K., Cubie, H., Assessment of human papillomavirus mRNA detection over time in cervical specimens collected in liquid based cytology medium. J. Virol. Methods 2005, 124, 211–215.

[20] Tarkowski, T., A., Rajeevan, M.S., Lee, D.R., Unger, E.R., Improved detection of viral RNA isolated from liquid-based cytology samples. *Molecular Diagnostic* 2001, 6, 125–130.

[21] Castle, P.E., Solomon, D., Hildesheim, A., Herrero, R., et al., Stability of archived liquid-based cervical cytologic specimens. *Cancer* 2003, 99, 89–96.

[22] Gibb, S., Strimmer, K., MALDIquant: a versatile R package for the analysis of mass spectrometry data. *Bioinformatics* 2012, 28, 2270–2271.

[23] Strohalm, M., Hassman, M., Košata, B., Kodíček, M., mMass data miner: an open source alternative for mass spectrometric data analysis. *Rapid Commun. Mass Spectrom.* 2008, 22, 905–908.

[24] Strohalm, M., Kavan, D., Novák, P., Volný, M., Havlíček, V., mMass3: A Cross-Platform Software Environment for Precise Analysis of Mass Spectrometric Data. *Anal. Chem.* 2010, 82, 4648–4651.

[25] Hollemeyer, K., Altmeyer, W., Heinzle, E., Pitra, C., Species identification of Oetzi's clothing with matrix-assisted laser desorption/ionization time-of-flight mass spectrometry based on peptide pattern similarities of hair digests. *Rapid Commun. Mass Spectrom.* 2008, 22, 2751–2767.

[26] Schmid, F., Schmidt, A., Nonparametric estimation of the coefficient of overlapping—theory and empirical application. *Comput. Stat. Data Anal.* 2006, 50, 1583–1596.

[27] Lesaffre, E., Superiority, equivalence, and non-inferiority trials. *Bull. NYU Hosp. Jt. Dis.* 2008, 66, 150–154.

[28] Albrethsen, J., Reproducibility in Protein Profiling by MALDI-TOF Mass Spectrometry. *Clin. Chem.* 2007, 53, 852–858.

[29] Piga, I., Heijs, B., Nicolardi, S., Giusti, L., et al., Ultra-high resolution MALDI-FTICR-MSI analysis of intact proteins in mouse and human pancreas tissue. *Int. J. Mass Spectrom.* 2017.

[30] Gessel, M.M., Norris, J.L., Caprioli, R.M., MALDI imaging mass spectrometry: Spatial molecular analysis to enable a new age of discovery. *J. Proteomics* 2014, 107, 71–82.

4.3 MALDI-MSI as a Complementary Diagnostic Tool in Cytopathology: A Pilot Study for the Characterization of Thyroid Nodules

4.3.1 Introduction

The application of innovative technologies, such as Matrix-Assisted Laser Desorption/Ionization (MALDI) Mass Spectrometry Imaging (MSI), on cytological thyroid specimens is feasible and robust protocols are now available, enabling the molecular signature of different lesions to be characterized [1-3]. After the pioneering phase, challenging technical aspects of this approach, such as the interference of haemoglobin and the stability of the samples, were overcome [4,5]. Furthermore, recent technical improvements related to the increased lateral resolution that can be achieved by MALDI-TOF-MS instrumentation enable the detection of small cell subpopulations based on their different protein profiles (i.e. profiles of single cell-types), even within regions that are indistinguishable at the microscopic level, highlighting how molecular imaging can be combined with traditional pathology to generate protein signatures and build classification models [7-9]. Moreover, we have reported that MALDI-MSI is able to distinguish benign and malignant cases in different cytological thyroid specimens [1-3]. Moving forwards from the first results obtained using *ex-vivo* cytological smears taken from surgical procedures, the present study applies MALDI-MSI on *real* Fine Needle Aspirates (FNAs). Even if thyroid FNAs are safe, cost effective and efficient diagnostic tools, a significant rate of 20-30% of cases is still indeterminate for malignancy [10]. Ancillary tests like immunohistochemistry and genetics may improve the diagnostic performances but, theoretically, MALDI-MSI could represent an alternative option too [1-3]. To the state of the art, MALDI-MSI was restricted to translational research and the reproducibility across multiple centres was the largest remaining obstacle in moving it towards clinical routine. However, promising results came from microbiology, where MALDI-MSI based classifiers applied the technology in real time in the diagnostic setting. Recently published studies showed the usefulness, advantages, and applicability of MALDI MSI in different fields of pathology (diagnosis, prognosis and treatment response) [10]. The preliminary findings of our trial are encouraging especially for the methodological improvement of the protocol and the feasibility of the technique in a particularly complicated field like thyroid cytological specimens. A statistical model, able to manage the big data that were generated by this high-throughput proteomics approach, was applied for the characterization of thyroid lesions. Our results suggest an association between pathological thyroid features and proteomic information from the FNAs, representing the basis for proteomic signatures that are predictive of disease status.

4.3.2 Material and Methods

The study was carried out in accordance with the relevant guidelines and regulations; the protocol was approved by the ASST Monza Ethical Board (Associazione Italiana Ricerca sul Cancro Associazione Italiana Ricerca sul Cancro-AIRC-MFAG 2016 Id. 18445, HSG Ethical Board Committee approval October 2016, 27102016). Appropriate informed consent was obtained from all patients included in the study. The present study considers a subset of the consecutive series of subjects who underwent ultrasound (US)-guided FNAs in Monza and were prospectively enrolled in an AIRC-granted clinical study that was powered to discover new markers for the diagnosis of thyroid nodules.

4.3.2.1 Pathology

US-guided FNAs were performed using a 25-gauge needle at the Department of Radiology, San Gerardo Hospital. One or two passes per nodule were executed and needle washing from every pass was sent for proteomics MALDI-MSI analysis [6]. In blind, pathologists evaluated the corresponding Pap-stained smears for traditional morphological diagnosis and were classified according to the 5-tiered Italian SIAPEC system for reporting thyroid cytopathology [11]. We certified benign-Thy2 cases by performing a US examination of patients 12-months after the first US-guided FNA confirming absence of new echographic malignant features, absence of significant increasing of nodule size, absence of nodes metastasis, and no incidence of new suspicious nodules. For malignant cases, histological diagnoses were progressively collected after thyroidectomy to certify the nature of the nodules. The training set included 9 subjects with a confirmed benign diagnosis at the pathologist's morphological examination (hyperplastic nodules/Thy2) and 9 patients that were classified as malignant papillary thyroid carcinoma (PTCs/Thy5). An additional 11 patients were involved in the validation set and their cytological classes included: Thy2 (n=4), Thy3 (n=1), Thy4 (n=1), Thy5 (n=4) and 1 PTC-metastatic lymph node. Table 1 summarizes the relevant clinical-pathological characteristics for all the cases in the study.

TRAINING SET					
Study lesion code	**Age (years)**	**Sex**	**Nodule size (mm)**	**FNA**	**Classification at follow-up or histology**
🟩 262	81	F	30	THY2	Hyperplastic
🟩 268	81	F	10	THY2	Hyperplastic
🟩 302	63	F	15	THY2	Hyperplastic
🟩 308	32	F	10	THY2	Hyperplastic
🟩 384	71	F	20	THY2	Hyperplastic
🟩 475	39	F	25	THY2	Hyperplastic
🟩 565	69	M	22	THY2	Hyperplastic
🟩 1046	56	F	18	THY2	Hyperplastic
🟩 1122	76	F	11	THY2	Hyperplastic
🟥 213	48	F	15	THY5	PTC
🟥 250	87	F	20	THY5	PTC
🟥 436	69	M	14	THY5	PTC
🟥 440	45	F	23	THY5	PTC-FV
🟥 442	40	F	15	THY5	PTC
🟥 992	46	F	13	THY5	PTC-FV
🟥 995	61	F	50	THY5	PTC-FV
🟥 1012	69	M	18	THY5	PTC-FV
🟥 1076	38	F	14	THY5	PTC
VALIDATION SET					
🟩 1081	79	F	35	THY2	Hyperplastic
🟩 1083	49	F	15	THY2	Hyperplastic
🟩 1123	36	F	36	THY2	Hyperplastic
🟥 1156	53	F	11	THY2	Hyperplastic
🟥 1149	30	F	15	THY5	PTC
🟥 1084	60	M	11	THY5	PTC-FV
🟥 1126	54	M	20	THY5	PTC
🟥 1187 *	24	F	25	THY5	PTC
🟦 1082	49	F	35	THY3	Hyperplastic
🟦 1202	36	M	20	THY4	PTC- FV
🟥 1188 *	24	F	25	Metastasis	Lymph node

Legend: M=male, F=female, PTC=Papillary Thyroid Carcinoma, FV=Follicular Variant

*The two lesions are from the same patient

Table 1. Clinical information of the lesions and the patients included in the study. Green corresponds to Thy2 hyperplastic nodules; blue corresponds to nodules with an indeterminate for malignancy or suspicious cytological diagnosis; in red malignant Thy5 cases are listed.

4.3.2.2 In situ proteomics: MALDI-MSI

Needle washing from thyroid FNAs were collected into a *CytoLyt* solution (20% buffered methanol-based solution, ThinPrep™ 2000 system, CYTYC Corporation, Hologic), samples were prepared as previously described and finally transferred as a cytospin spot onto ITO glass slides [4,5; 12-14]. Then, all slides were washed with increased concentration of ethanol (70%, 90% and 95%) for 30 s each, dried under vacuum for 15 min and stored at − 80 °C until the day of the analysis (mean 24-48 hours after the time of biopsy). Before MALDI-MSI analysis, cytological specimens were equilibrated to room temperature, dried under vacuum for 30 min and the MALDI-matrix sinapinic acid (10 mg/ml in 60:40 acetonitrile:water w/0.2% trifluoroacetic acid) was uniformly deposited, with an optimised method, using the iMatrixSpray (Tardo GmbH, Subingen, Switzerland) automated spraying system. MALDI-TOF-MSI was performed using an ultrafleXtreme MALDI-TOF/TOF (Bruker Daltonik GmbH) in positive-ion linear mode, using 300 laser shots per spot, with a laser focus setting of 3 medium (diameter of 50 μm) and a pixel size of 50 x 50 μm. Protein Calibration Standard I (Bruker Daltonics), that contains a mixture of standard proteins within the mass range of 5730 to 16,950 Da, was used for external calibration (mass accuracy ± 30 ppm). Spectra were recorded within the *m/z* 3000–20,000 range. Data acquisition and visualisation were performed using the Bruker software packages (flexControl 3.4, flexImaging 5.0). After the analysis, the MALDI matrix was removed with 70% EtOH and the slides were stained with haematoxylin and eosin (H&E), digitally scanned using a ScanScope CS digital scanner (Aperio, Park Center Dr., Vista, CA, USA) and images were co-registered to the MSI datasets in flexImaging for the integration of proteomic and morphological data. Regions of interest (ROIs) containing pathological areas will be comprehensively annotated. Satisfactory specimens should include at least 6 groups of 10 thyrocytes, as SIAPEC guidelines [15].

4.3.2.3 Statistical analysis

Quartiles, ranges, mean and standard deviation (sd) were calculated for descriptive purposes. The analysis on proteomic data in the training set was performed on ROIs that included only epithelial cells, while for each patient in the validation set, three different approaches were tested: the average spectra generated from the MALDI-MSI analysis, the spectra from each ROIs selected by the pathologist and all the single spectra of the imzML MALDI-MSI analysis (pixel by pixel*)*. The spectra were processed by performing baseline subtraction (median method), smoothing (moving average method, half window width 2.5), normalization (total ion current, TIC), peak alignment, and peak picking ($S/N \geq 6$). Pre-processing was performed separately between training and validation set, in order to not influence the data of the validation. The open-source software mMass v.5.5 (http://www.mmass.org) was used to confirm mass spectra alignment. Only peaks with an absolute intensity of more or equal than 0.0003, after TIC normalization, were retained. Intra- and inter-patient filters were applied on the detected features in the training set: i) only the features (*m/z*) detected in at least 25% of the ROIs within the same patient were considered and ii) the features (*m/z*) that were common to at least 25% of the Thy2 and to 25% of the Thy5 were included in the model and considered to be those most representative of benign and malignant lesions, respectively. For the two groups in the training set (benign vs malignant lesions), a logistic regression with a Lasso regularization method was performed [16-18]. To select the Lasso penalising parameter, and to assess the predictive accuracy within the training set, cross-validation was performed. The validation was done in blind from the patient's histological diagnosis and considering only the features selected by the Lasso model to quantify the probability of malignancy. Data pre-processing (MALDIquant package) and statistical analyses (glmnet package) were performed using the open-source R software v.3.5.0.

4.3.3 Results

The cohort of 28 patients included in this study had an average age of 54 years old (sd=17) and 23 (79%) were females. The average nodule diameter was 20 mm (sd=9). In the group of patients used in the training phase, the selected ROIs varied in terms of the number of clusters and cells that composed the placards. In the Thy2 cases, an average number of 10 ROIs (range=5-22, median=9) was recorded by the pathologist, while, in the Thy5 cases, a mean of 8 ROIs (range=4-19, median=6) was selected. To compensate for this variability, equivalent groups of ROIs were generated for each patient: 5 groups of ROIs for Thy2 cases and 4 ROIs for Thy5 cases, each comprising from 1 to 7 ROIs. These were then used to calculate the average spectra for the statistical analysis. ROIs from Thy2 lesions had an average number of 9 pixels (range=3-39, median=7) while in the Thy5 had an average of 31 (range=3-162, median=13). Therefore, 45 mean spectra were generated for the benign and 36 for the malignant lesions and used for the statistical analysis of the training data. After pre-processing and the two intra- and inter-patient filters, 69 features were found to be the most representative of Thy2 and Thy5 lesions, 20 of these were selected from the statistical model as the most discriminant to correctly distinguish samples and quantify their probability of being malignant lesions (Supplementary Table 1 in appendix C). Then, the capability of the features included in the model to discriminate benign from malignant lesions was also tested on each single pixel present in the analysed specimens. This was performed using the same groups of patients included in the training phase (Supplementary Figure 2a, 2b in appendix C). A complete overlap of the cytological diagnosis and MALDI-MSI results was observed. In particular, specimens of the benign group were observed to be very homogeneous (uniformly distributed green colour, Figure 1a), indicating that all the protein profiles were similar. In the validation phase on 11 additional lesions (10 patients), three different approaches were applied based on: i) average of spectra of the ROIs, ii) overall average spectra of the entire FNA, irrespective of the morphological selection of the ROIs, and iii) pixel by pixel analysis (Supplementary Figure 1 in appendix C).

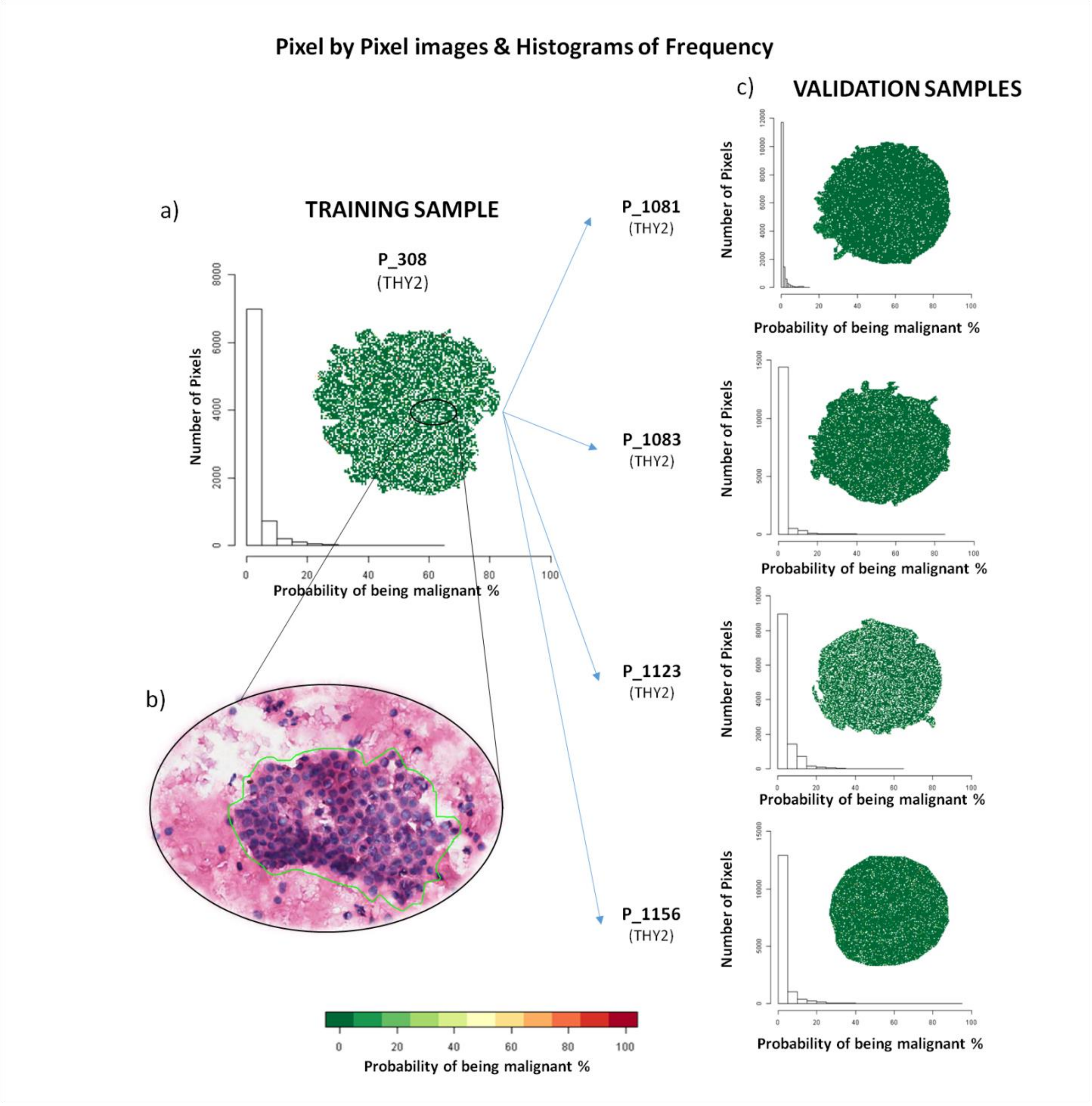
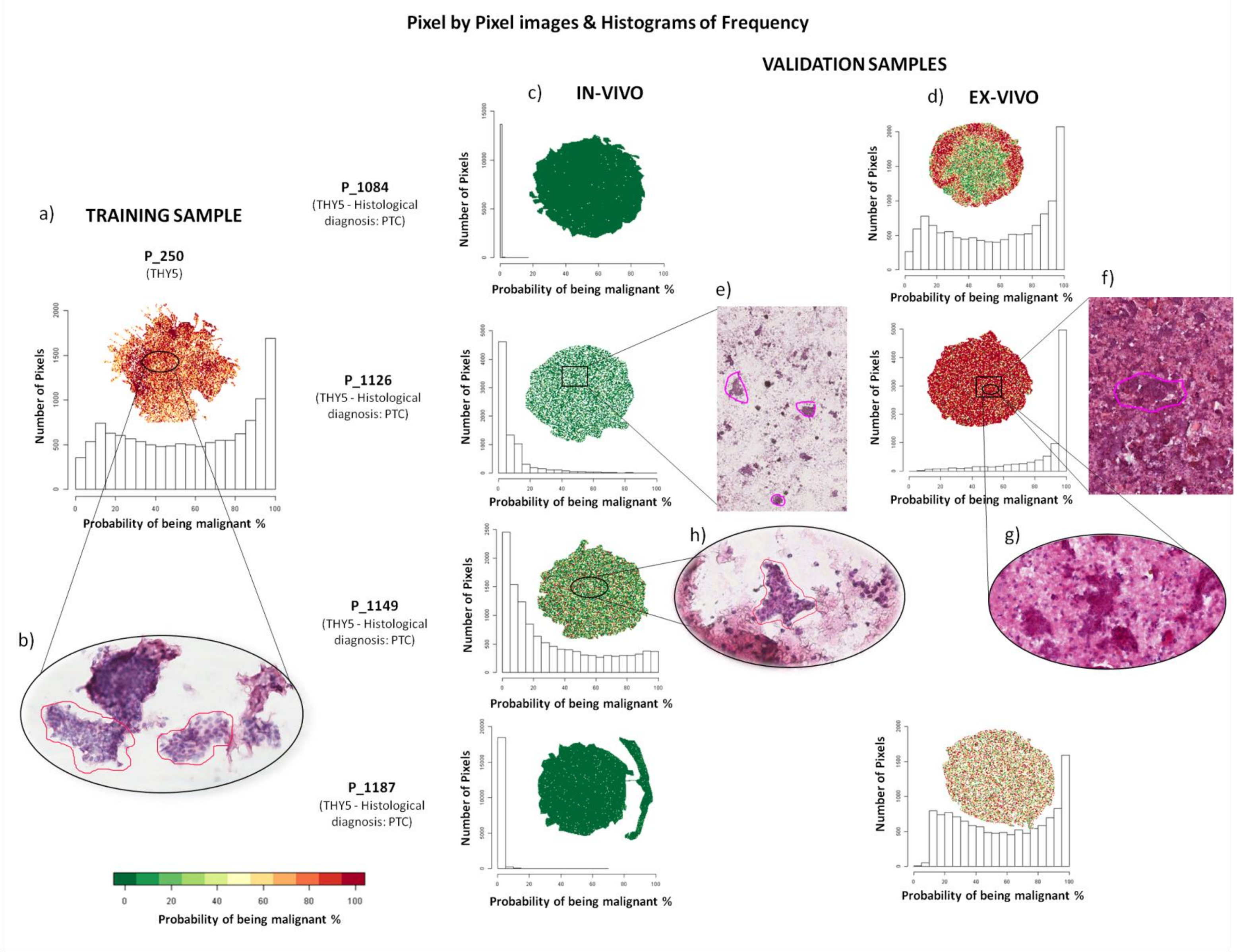


Figure 1. Examples of pixel by pixel images and distributions of the probabilities of being malign in the training and validation set of benign Thy2 nodules. a) imzML MALDI-MSI data of the Thy2 P_308 in the training sample; b) Haematoxylin and Eosin (H&E) staining of P_308; c) Validation of Thy2 samples using imzML MALDI-MSI data.

Figure 2. Examples of pixel by pixel images and distributions of the probabilities of being malign in the training and validation set of malignant Thy5 nodules. a) imzML MALDI-MSI data of the Thy5 P_250 in the training sample; b) H&E staining image of P_250; Validation of c) in-vivo Thy5 samples and d) ex-vivo Thy5 samples using imzML MALDI-MSI data; e) low cellularity in the H&E staining image of the P_1126 in-vivo sample; f) high cellularity in the H&E staining image of P_1126 ex-vivo sample and g) a zoom-in of thyrocyte clusters; h) H&E staining image of high quality cluster of thyrocytes cells of P_1149 in-vivo.of Thy2 samples using imzML MALDI-MSI data.



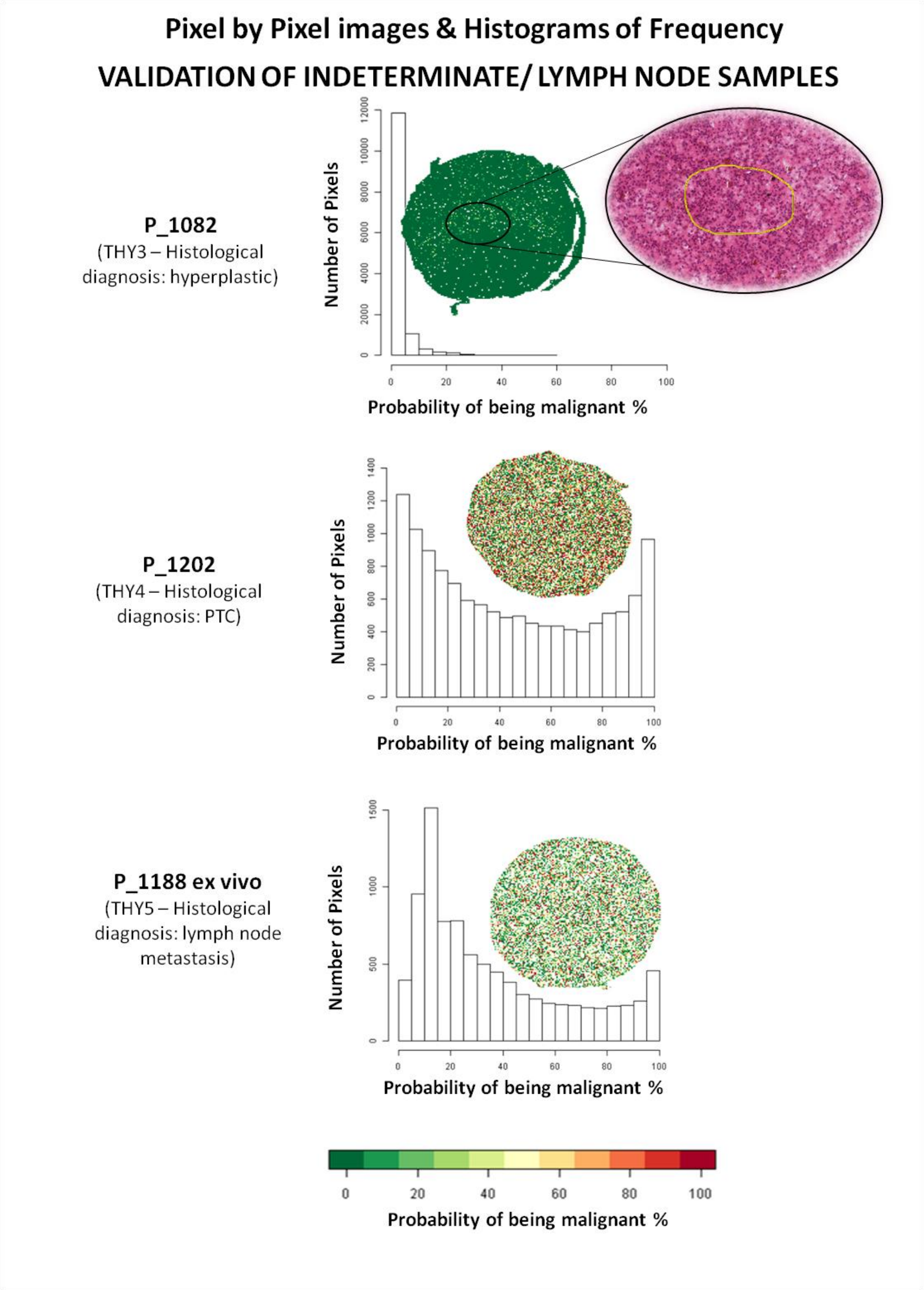


Figure 3. Validation set of indeterminate for malignancy (Thy3), suspicious (Thy4) cases and metastatic lymph node. Pixel by pixel images and distribution of the probabilities of being malignant for each pixel in the MALDI-MSI analysis.

The average number of ROIs for the specimens used in the validation phase was 12 (2-25, median 11), with a mean number of pixel for each ROI of 15 (1-208, median 5). The model correctly classified all the benign cases (four Thy2, as shown in Figure 1c, and one morphological Thy3, as shown in Figure 3). In the malignant scenario, three Thy5 cases were particularly challenging due to the paucity of cells (Figure 2e: P_1126) or to a heterogeneous background of benign/malignant cells (Figure 2c: P_1084, cytological image not shown) or colloid-rich, cystic variant PTC (Figure 2c: P_1187, cytological image not shown). As a consequence, the proteomic analysis did not identify diagnostic signals of alert at the first screening classifying these samples as benign (Figure 2c). Patients Thy5 P_1149 and Thy4 P_1202, both adequate specimens, were correctly classified based on the distribution of the probabilities to be malignant using both ROIs and pixel by pixel data (Figure 2c, 2h and 3; Supplementary Table 2b, 2c and 3c in appendix C)). Then, an additional experiment was planned to support the hypothesis to justify the incorrect classification using ex-vivo specimens. Samples from the same nodules (taken ex-vivo after thyroidectomy, as previously described [19]) were now correctly classified as malignant by the model, due to a greater amount of neoplastic clusters that did not limit the analysis (Figure 2d, 2f, 2g). Analysis of an in-vivo specimen of a metastatic lymph node (P_1188) resulted in a correct classification as malignant based on ROIs but as benign in the pixel by pixel classification (P_1188). Specimen collected ex-vivo from this lymph node was correctly classified both based on ROIs and pixel by pixel model (Figure 3). Finally, the comparison of the three methodological approaches employed for the validation set highlights improved discriminant power in both the pixel by pixel and ROI-analysis with respect to when the average spectra of the whole sample was employed (Supplementary Table 2 and 3 in appendix C)). This result underlines the particular strengths of MALDI-MSI that could be exploited to support, as complementary tool, the fundamental diagnostic role of the pathologist.

4.3.4 Discussion

4.3.4.1 Proteomics for the diagnosis of thyroid carcinoma

The development of new diagnostic tools to support cytopathologists in the diagnostic triage of indeterminate for malignancy thyroid nodules can be approached from the alternative perspective offered by proteomics [20,21]. Previous reports enlightened the possibility to apply imaging methods such as MALDI-MSI to cytological specimens to combine the analytical power of traditional morphology and molecular signatures [22]. Preliminary experiments were done using ex-vivo specimens taken from surgical samples [19], while in the present study true needle washing specimens were used. The feasibility of the MALDI-MSI approach to spatially localize proteins in a cancer cells area is enlighten in Supplementary Figure 3 in appendix C. This represents an intriguing and important methodological step, leading to the recovery of left-over material from the FNAs that can be recovered by washing the needle and stabilizing the cells for 2 weeks [5]. This procedure allows specimens to be collected from centers that don't have a diagnostic unit with proteomics facilities and then shipped to the referee lab within the following ten days. In the near future, the more systematic enrollment of patients from multiple centres could ensure the generation of diagnostic libraries containing molecular signatures, which include different malignant and rare histotypes for research purposes.

4.3.4.2 Big data and biostatistics: a requirement for the introduction of proteomics in clinics

Indeed, the application of proteomics as a routine option for the characterization of challenging cases also requires the development of an enlarged network given that the validation of protocols, biostatistic models and putative analytical features is related with the inter-laboratory reproducibility, standardization of workflows and diagnostic strengthening of the methods. In particular, with the advent of molecular techniques like next generation sequencing (NGS) and proteomics approaches (MALDI-MSI), biostatistics models and bioinformatics that can manage big data are necessary for improving the confidence of pathologists [23,24]. Statistical models of cancer at the genomic, proteomic and transcriptomic levels have proven effective in developing diagnostic and prognostic molecular signatures, as well as in identifying pathogenetic pathways [25]. High-throughput experimental tools allow for the simultaneous measurement of thousands of biomolecules, integrating heterogeneous data into quantitative predictive models to significantly improve cytological diagnoses. Molecular diagnostic workflows can be divided into those that employ unbiased statistical inference and those that also incorporate *a priori* constraints of specific biological interactions from data [26]. In the present study, a diagnostic model was trained using clear-cut benign or malignant cases to identify specific discriminant features to be tested in the validation phase. Three different approaches were used: the analysis of groups of ROIs that the pathologists selected using morphological criteria, a pixel by pixel approach, and examination of the average spectrum of the whole sample.

4.3.4.3 Training phase: features selection for benign and malignant thyroid FNAs discrimination

The histograms in Figure 1 and 2 show how the probability of being malignant could be associated can be effectively represented with curves and the samples from FNAs should not pass the diagnostic proteomic triage whenever a signal of alert was pointed out. After the application of filters, biostatisticians designed a combination of features that was able to correctly distinguish all the training cases, in blind, when they were re-tested. The highest probability to be malignant of 7% (overall mean of the 3rd quartiles = 2.89%, sd = 2.03%) for the Thy2 and the minimum of 28% (overall mean of the 3rd quartiles = 81.81%, sd = 22.66%) for Thy5 was observed in the training phase (Supplementary Table 3a, 3b in appendix C).

4.3.4.4 Validation phase of the selected features and pixel-by-pixel classification of thyroid FNAs

Results obtained in the pixel by pixel validation phase showed that all benign lesions, including the Thy3 (later confirmed as benign after surgical resection), have a 3rd quartile value of the probability of being malign below 7%. The malignant lesions had a 3rd quartile above 28% with the exception of specimens with scarce cellularity or a heterogeneous background. These specimens stressed the model due to particularly challenging nodules that were representative of the diagnostic situation characterizing routine thyroid pathology. Samples with issues in terms of quantitative adequacy, haemorrhagic slides, colloid-rich or very heterogeneous FNAs with interspersed macrophages and lymphocytes are all good examples of challenging specimens. In benign lesions, a minimum amount of cells was sufficient to confirm the nature of the hyperplastic goiter and no signal of alert was recorded. In the malignant group, three FNAs from histologically proven PTC (Thy5) were not correctly assigned (Figure 2c) due to the quality of the samples taken from the patient. Two different situations were highlighted: samples with paucity of malignant thyrocytes or with high inflammatory or colloidal background. In fact, when the analyses were repeated with samples taken ex-vivo from the thyroid of same patients after surgical removal, they were easily diagnosed as malignant by our diagnostic tool due to the increased quality of the specimens with a greater amount of neoplastic clusters (Figure 2d). An in-vivo specimen of a metastatic lymph node was also misclassified as benign only in the pixel by pixel classification (P_1188). A possible explanation for this failure could be due to the low number of thyrocytes present in the sample. As a consequence, the correct classification was obtained when using the ROIs, where the background was less impacted by the quality of the spectra, but this confounded the model in the pixel by pixel classification. However, the specimen that was collected ex-vivo was correctly classified using either the ROIs or the pixel by pixel model (P_1188: Figure 3 and Supplementary Table 2c1 in appendix C). This suggests that, once the pathologist certified the presence of a satisfying quantity of neoplastic cells in the washing material, the model also correctly triaged malignant PTC cells in a sample taken from a metastatic lymph node.

4.3.5 Conclusions

Notwithstanding the consideration that the diagnostic validity of the model needs to be verified in the large cohort of patients that is currently under enrollment, the present study introduces an original methodological approach to build a proteomic diagnostic tool in thyroid cytopathology by taking advantage of MALDI-MSI technology. The next step will be to systematically test the workflow and to putatively identify the most significant features employed by the classification model. The direct consequences of successful results could be the use of MALDI-MSI proteomics as a complementary approach for the characterization of indeterminate for malignancy thyroid nodules. Despite the technical challenges of this study, the application of proteomics and imaging may help to elucidate key biomolecular events and pathways in oncogenic processes [27,28]. Collectively, this represents an important paradigm for both the fundamental characterization of cancer systems and the discovery of molecular targets for diagnostic application.

Bibliography

[1] Mosele N, Smith A, Galli M, Pagni F, Magni F. MALDI-MSI Analysis of Cytological Smears: The Study of Thyroid Cancer. Methods Mol Biol. 2017;1618:37-47.

[2] Mainini V, Pagni F, Garancini M, Giardini V, De Sio G, Cusi C et al. An alternative approach in endocrine pathology research: MALDI-IMS in papillary thyroid carcinoma. Endocr Pathol. 2013;24:250–3.

[3] Pagni F, Mainini V, Garancini M, Bono F, Vanzati A, Giardini V et al. Proteomics for the diagnosis of thyroid lesions: preliminary report. Cytopathology. 2015;26:318–24.

[4] Piga I, Capitoli G, Denti V, Tettamanti S, Smith A, Stella M et al. The management of haemoglobin interference for the MALDI-MSI proteomics analysis of thyroid fine needle aspiration biopsies. Anal Bioanal Chem. 2019;411(20):5007-5012

[5] Piga I, Capitoli G, Tettamanti S, Denti V, Smith A, Chinello C et al. Feasibility Study for the MALDI-MSI Analysis of Thyroid Fine Needle Aspiration Biopsies: Evaluating the Morphological and Proteomic Stability Over Time. Proteomics Clin Appl. 2019 Jan;13(1):e1700170

[6] Ucal Y, Tokat F, Duren M, Ince U, Ozpinar A. Peptide Profile Differences of Noninvasive Follicular Thyroid Neoplasm with Papillary-Like Nuclear Features, Encapsulated Follicular Variant, and Classical Papillary Thyroid Carcinoma: An Application of Matrix-Assisted Laser Desorption/Ionization Mass Spectrometry Imaging. Thyroid. 2019 Jun 20. doi: 10.1089/thy.2018.0392

[7] Smith A, L'Imperio V, Denti V, Mazza M, Ivanova M, Stella M et al. High Spatial Resolution MALDI-MS Imaging in the Study of Membranous Nephropathy. Proteomics Clin Appl. 2019 Jan;13(1):e1800016.

[8] Smith A, Galli M, Piga I, Denti V, Stella M, Chinello C et al. Molecular signatures of medullary thyroid carcinoma by matrix-assisted laser desorption/ionisation mass spectrometry imaging. J Proteomics. 2019;191:114-123.

[9] Judd AM, Gutierrez DB, Moore JL, Patterson NH, Yang J, Romer CE et al A recommended and verified procedure for in situ tryptic digestion of formalin-fixed paraffin-embedded tissues for analysis by matrix-assisted laser desorption/ionization imaging mass spectrometry. J Mass Spectrom. 2019 Jun 28. doi: 10.1002/jms.4384

[10] PM, Heeren RMA, Porta T, Balluff B. Mass spectrometry imaging for clinical research - latest developments, applications, and current limitations. Analyst. 2017;142(15):2690-2712

[11] Bongiovanni M, Crippa S, Baloch Z, Piana S, Spitale A, Pagni F et al. Comparison of 5-tiered and 6-tiered diagnostic systems for the reporting of thyroid cytopathology: a multi-institutional study. Cancer Cytopathol. 2012;120(2):117-25.

[12] Amann JM, Chaurand P, Gonzalez A, et al. Selective profiling of proteins in lung cancer cells from fine-needle aspirates by matrix-assisted laser desorption ionization time-of-flight mass spectrometry.Clin Canc Res. 2006;12:5142–50

[13] Howitt BE, Chang S, Eszlinger M, Paschke R, Drage MG, Krane JF et al. Fine-needle aspiration diagnoses of noninvasive follicular variant of papillary thyroid carcinoma. Am J Clin Pathol. 2015;144:850–7

[14] Linder J. Recent advances in thin-layer cytology. Diagn Cytopathol. 1998;18:24–32.

[15] Sparano C, Parenti G, Cilotti A, Bencini L, Calistri M, Mannucci E et al. Clinical impact of the new SIAPEC-IAP classification on the indeterminate category of thyroid nodules. J Endocrinol Invest. 2019;42(1):1-6

[16] Eberlin LS, Margulis K, Planell-Mendez, Zare RN, Tibshirani R, Longacre TA et al. Pancreatic Cancer Surgical Resection Margins: Molecular Assessment by Mass Spectrometry Imaging. PLoS Med 2016; 13(8): e1002108.

[17] Margulis K., Chiou A.S., Aasi S.Z., Tibshirani R.J., Tang J.Y., Zare R.N. Distinguishing malignant from benign microscopic skin lesions using desorption electrospray ionization mass spectrometry imaging Proceedings of the National Academy of Sciences of the United States of America, 2018; 115 (25) , pp. 6347-6352.

[18] Galli M, Zoppis I, De Sio G, Chinello C, Pagni F, Magni F et al. A Support Vector Machine Classification of Thyroid Bioptic Specimens Using MALDI-MSI Data. Adv Bioinformatics. 2016;2016:3791214

[19] Pagni F, De Sio G, Garancini M, Scardilli M, Chinello C, Smith AJ et al. Proteomics in thyroid cytopathology: Relevance of MALDI-imaging in distinguishing malignant from benign lesions. Proteomics. 2016;16(11-12):1775-84.

[20] Wojakowska A, Cole LM, Chekan M, Bednarczyk K, Maksymiak M, Oczko-Wojciechowska M et al. Discrimination of papillary thyroid cancer from non-cancerous thyroid tissue based on lipid profiling by mass spectrometry imaging. Endokrynol Pol. 2018;69(1):2-8

[21] Pagni F, L'Imperio V, Bono F, Garancini M, Roversi G, De Sio G et al. Proteome analysis in thyroid pathology. Expert Rev Proteomics. 2015;12(4):375-90

[22] Schwamborn K, Krieg RC, Uhlig S, Ikenberg H, Wellmann A. MALDI imaging as a specific diagnostic tool for routine cervical cytology specimens. Int J Mol Med. 2011;27(3):417-21

[23] Sivakumar K, Nithya NS, Revathy O. Phenotype Algorithm based Big Data Analytics for Cancer Diagnose. J Med Syst. 2019;43(8):264

[24] Nathoo FS, Kong L, Zhu H. A Review of Statistical Methods in Imaging Genetics. Can J Stat. 2019; 47(1):108-131

[25] Stemmer A, Galili T, Kozlovski T Zeevi Y, Marcus-Kalish M, Benjamini Y et al.Current and Potential Approaches for Defining Disease Signatures: a Systematic Review. J Mol Neurosci. 2019;67(4):550-558

[26] Frey LJ. Data integration strategies for predictive analytics in precision medicine. Per Med. 2018;15(6):543-551.

[27] Ganly I, McFadden DG. Short Review: Genomic Alterations in Hürthle Cell Carcinoma. Thyroid. 2019;29(4):471-479

[28] Mayson SE, Haugen BR. Molecular Diagnostic Evaluation of Thyroid Nodules. Endocrinol Metab Clin North Am. 2019;48(1):85-97.

5 An insight into the indices of similarity between mass spectra

The quantitative evaluation of mass spectra similarity has been often employed to investigate reproducibility and repeatability of analytical methods. Moreover, these mass spectra similarity approaches have been occasionally used for classification purposes comparing query samples to reference spectral libraries. Several mass spectra similarity indexes had been developed, but the statistical evaluation of their reliability for mass spectra comparison was never assessed.

Different similarity measures present in literature are reviewed and compared in this chapter, and a new score system based on overlap is also proposed. The statistical performances of these measures have been compared through simulated mass spectra that mimic those from real proteomic data observed in thyroid cancer. Simulation protocols and results are reported and the limits and benefits of the different approaches are also discussed.

5.1 Introduction

In the clinical application of mass spectrometry, similarity scores have been often employed to evaluate not only the agreement of an unknown sample with respect to a reference compound in spectral libraries [1], but also to evaluate analytical methods reproducibility [2] and repeatability [3]. Different mass spectra similarity scores have been used for different purposes over the years. Several papers evaluated the performances of one or two scores per time using real mass spectrometry data. Only few studies investigated mass spectra similarity scores using mass spectra generated from replicates of real reference library spectra [4] or limited simulated mass spectra [5].

Two scores were initially proposed for the comparison of mass spectrometry profiles: the Probabilitity-Based Matching (PBM) of McLafferty et al. [1] and the similarity measure of Hertz et al. [6]. Subsequently, Stein and Scott reviewed and compared the performances of five different algorithms, namely: PBM, dot-product (cosine correlation), Hertz similarity measure, Euclidean and absolute value distance. These five algorithms were used to compare a test spectrum against reference spectra using a mass spectra database. Results indicated that the PBM and the Hertz similarity measure, the two algorithms constructed specifically for mass spectra comparisons, performed worst, while the dot-product function had good performances. Finally, they described a new optimal composite algorithm (i.e. the Stein and Scott measure), which achieved the best performances. This index was obtained from the cosine correlation score by weighting spectra for intensity and *m/z* value, and by adding a term based on ratio of peak pair intensity. Moreover, Wan et al. highlighted that the cosine index is able to differentiate between very similar mass spectra, where the Similarity Index (SI) fails [5].

Koo et al. investigated for the first time the compound identification accuracy of different mass spectra similarity measure through simulations, generating spectra from existing reference library. These simulations showed that Partial and Semipartial correlation indices had the best performance in accuracy, but the worst in computation time, compared to cosine correlation, Stein and Scott measure, and Discrete Fourier Transform (DFT) composite measure [7].

Additional similarity scores that have been used to investigate mass spectra similarity are: fold difference, intensity match, Spearman's correlation, Person's correlation, Euclidean distance, Manhattan distance, fit, retrofit [1]. Most of these similarity scores were applied in electrospray ionization mass spectrometry analysis and no studies investigated their performances. We proposed a new score never used in the evaluation of the similarity of protein spectra, called overlap [3]. All the measures we reviewed were included in the simulation study that we made. The purpose was to investigate for the first time the performances of different mass spectral similarity scores applied to linear matrix assisted laser desorption ionization -time of flight-mass spectrometry (MALDI-TOF-MS) data.

5.2 Review of the existing measures

In the description of the different similarity scores, the PBM, the Hertz, the Partial and Semipartial correlation indices were not taken into account. The first two were excluded due to their low performance, which is widely documented in the literature, while Partial and Semipartial correlation indices were not considered due to their field of application that is different from our context.

Consider two given spectra $X = (x_i)_{i=1,...,N_X}$ and $Y = (y_i)_{i=1,...,N_Y}$, where the generic i^{th} peak represents a mass-to-charge value (m/z), while N_X and N_Y are the total number of m/z values in X and Y spectra.

1)Cosine correlation:

The dot-product term [5], also known as the cosine correlation index, is used to obtain the cosine angle between the direction in space of the query and reference sequences of intensity signals. It is defined as follows:

$$S_C = \frac{X \circ Y}{\|X\| \cdot \|Y\|}$$

where $X \circ Y = \sum_{i=1}^{N} x_i y_i$, $\|X\| = \sqrt{\sum_{i=1}^{N} x_i^2}$ and N was the total number of m/z values used in the comparison. Note that S_C ranges between -1 and 1, and it is always non-negative if X and Y are non-negative intensities.

The dot-product index varies between 0 when the spectra are completely different and even when do not contain common ions and 1 when spectra are identical. Even if spectra are pre-processed, in particular normalized to the sum of peak intensities, the cosine similarity index, as well as others measures (e.g. fold difference) that will be introduced later, is independent from normalization.

2)Stein and Scott index:

An improved optimized dot-product cosine correlation system, called Stein and Scott Composite similarity [1], weights spectra for intensity and *m/z* value and adds a term based on the ratio of peak pair intensity.

Firstly, the ratio of peak pair S_R is as follows:

$$S_R(X,Y) = \frac{1}{N_{X\cap Y}} \sum_{i}^{X\cap Y} \left(\frac{y_i}{y_{i-1}} \cdot \frac{x_{i-1}}{x_i}\right)^n$$

where $n = -1$ or 1 if the term in parentheses was less than or greater than unity, respectively. The value $N_{X\cap Y}$ is the number of non-zero peaks in both the reference and the query spectra, i.e the number of shared mass peaks.

The Stein and Scott Composite similarity is calculated by:

$$S_{SS}(X,Y) = \frac{N_X \cdot S_C(X,Y) + N_{X\cap Y} \cdot S_R(X,Y)}{N_X + N_{X\cap Y}}$$

where N_X is the number of non-zero peak intensities existing in the query spectra. In the literature this similarity score is constructed only for weighted intensities, so its original formulation is:

$$S_{SS}(X,Y) = \frac{N_X \cdot S_{WC}(X,Y) + N_{X\cap Y} \cdot S_{WR}(X,Y)}{N_X + N_{X\cap Y}}$$

where S_{WC} is the weighted cosine correlation formula in which X and Y are considered as weighted spectra W_X and W_Y, $S_{WC}(X,Y) = S_C(W_X, W_Y) = \frac{W_X \circ W_Y}{\|W_X\| \cdot \|W_Y\|} = \frac{\sum_{i=1}^{N} W_{x,i} \cdot W_{y,i}}{\sqrt{\sum_{i=1}^{N} W_{x,i}{}^2} \cdot \sqrt{\sum_{i=1}^{N} W_{y,i}{}^2}}$,

and $S_{WR}(X,Y) = S_R(W_X, W_Y) = \frac{1}{N_{X\cap Y}} \sum_{i}^{X\cap Y} \left(\frac{W_{y,i}}{W_{y,i-1}} \cdot \frac{W_{x,i-1}}{W_{x,i}}\right)^n$, where $W_{x,i}$ and $W_{y,i}$ are non-zero weighted intensities having common m/z value defined as $W = [Peak\ Intensity]^a [Mass]^b$, a

and b are the weight factors for peak intensity and m/z values, respectively. For our purpose, these formulae are considered without weighted intensity as previously reported.

3)Similarity Index:

Similarity Index (SI) method [8], is defined as:

$$S_{SI} = \sqrt{\frac{\sum_{i=1}^{N}\left\{\frac{x_i - y_i}{x_i + y_i} \times 100\right\}^2}{N}}$$

4)Discrete Fourier composite index:

The Discrete Fourier transform (DFT) index converts an original spectral signal $Z = (z_1, \ldots, z_n)$ into a new signal $Z^F = ({z_1}^F, \ldots, {z_n}^F)$ as follows [9]:

$${z_k}^F = \sum_{d=1}^{n} z_d \exp\left(-\frac{2\pi i}{n} kd\right), k = 1, \ldots, n$$

where the notation i in this case is the imaginary unit and not the i^{th}peak, called k. By using the Euler's formula that defines $\exp(i\phi) = \cos\phi + i\sin\phi$, the original equation becomes:

$${z_k}^F = \sum_{d=1}^{n} z_d cos\left(-\frac{2\pi}{n} kd\right) + i\sum_{d=1}^{n} z_d sin\left(-\frac{2\pi}{n} kd\right), k = 1, \ldots, n$$

We have a new transformed signal, whose real part $Z^{FR} = ({z_1}^{FR}, \ldots, {z_n}^{FR})$ is defined as follow:

$${z_k}^{FR} = Re({z_k}^F) = \sum_{d=1}^{n} z \cdot cos\left(-\frac{2\pi}{n} kd\right)$$

where the function $Re(\cdot)$ is the real part of the imaginary number.

The DFT with real composite similarity is defined as follow:

$$S_{DFT}(X,Y) = \frac{N_X \cdot S_C(X,Y) + N_{X\cap Y} \cdot S_C(X^{FR}, Y^{FR})}{N_X + N_{X\cap Y}}$$

5)Pearson's correlation:

The correlation between two sequences of intensities is defined in standard terms as the covariance of the two sequences divided by the product of the standard deviations:

$$S_P = Corr(X,Y) = \frac{Cov(X,Y)}{\sqrt{Var(X)}\sqrt{Var(Y)}}$$

where $Cov(X,Y)$ was the covariance between X and Y and $Var(X)$ was the variance of X.

6)Fold difference:

It is the ratio of protein amount between two spectra, and it is used to detect differences of various proteins concentration in complex mixtures [10]. It is defined as

$$S_{FD} = \frac{\sum_{i=1}^{N} \frac{x_i}{y_i}}{N}$$

7)Spearman's correlation:

Spearman's correlation is the non-parametric version of the Pearson's correlation coefficient $S_P = Corr(X,Y)$. It is defined as a measure of association between the ranks of the intensities of the common peaks $n_{X\cap Y}$:

$$\rho_s = \frac{\sum_j^{n_{X\cap Y}}\left[\left(r_j - \bar{r}\right) * \left(s_j - \bar{s}\right)\right]}{\sqrt{\sum_j^{n_{X\cap Y}}\left(r_j - \bar{r}\right)^2} * \sqrt{\sum_j^{n_{X\cap Y}}\left(s_j - \bar{s}\right)^2}}$$

where r_j and s_j were the ranks of x_j and y_j $(j = 1, \dots, n_{X\cap Y})$, while $\bar{r}$ and $\bar{s}$ were their median values.

8-9)Fit and retrofit:

These two measures are only based on the cardinality of the *m/z* values in the two spectra, and the number of shared mass peaks [11]. The fit is defined as the ratio of the common peaks in the two spectra and the n_Y peaks detected in the query spectrum:

$$FIT = \frac{n_{X\cap Y}}{n_Y}$$

while the retrofit is defined as the ratio of common peaks in the two spectra and the n_X peaks in the reference spectra:

$$RFIT = \frac{n_{X\cap Y}}{n_X}$$

10)Overlap:

The overlap (OV) takes into account the whole shape of two spectra [12]. This latter index measured the overlapping area between the empirical distributions of two sequences of intensities on ranked m/z:

$$OV = \hat{F}_{n_{X\cup Y}}^{X} \cap \hat{F}_{n_{X\cup Y}}^{Y}$$

where $\hat{F}_{n_{X\cup Y}}^{X}$ and $\hat{F}_{n_{X\cup Y}}^{Y}$ were the empirical distribution function, and $n_{X\cup Y}$ were the m/z values either in the X or the Y spectra.

11)Intensity match:

Intensity match is an improvement of the correlation system, in which the perfect correspondence of peaks abundance between spectra is investigated. This score focus on the agreement between two sequences of signals within a 30% range on intensity variability.

12)Lin's concordance correlation:

The Lin's concordance correlation coefficient measures the agreement between two sequences of variables [13][14]. The Lin's CCC is defined as:

$$S_{LIN} = \frac{2rs_Xs_Y}{(\bar{X}-\bar{Y})^2+(s_X)^2+(s_Y)^2}$$

This index is equivalent to 1 minus the ratio of expected orthogonal squared distance from the line $Y = X$ and the expected orthogonal squared distance from the line $Y = X$ assuming independence. Like a correlation coefficient -1 ≤ S_{LIN}≤ 1: values near +1 indicate strong concordance between $X$and Y, values near -1 indicate strong discordance and values near zero indicate no concordance.

13)Kendall's correlation:

It is used to measure the ordinal association between two measured quantities.

$$S_K = \frac{(number\ of\ concordant\ pairs) - (number\ of\ discordant\ pairs)}{\frac{n(n-1)}{2}}$$

14-15)Euclidean and Manhattan distances:

They calculated the distance between the intensity of two sequences of data.

$$dist_{euclidean} = \sqrt{\sum_{i=1}^{N}(x_i - y_i)^2} \quad \text{and} \quad dist_{manhattan} = \sqrt{\sum_{i=1}^{N}|(x_i - y_i)|}$$

A summary of the properties and characteristics of the reviewed similarity measures are reported below.

ID	Score	Domain	Value of perfect agreement	Characteristics
1	Cosine correlation	[0;1]	1	Angle between direction of intensities
2	Stein and Scott index	[1;+∞]	1	Ratio of intensities
3	Similarity index	[0;100]	0	Difference of intensities
4	Discrete Fourier Transformation	[0;1]	1	Composition of waves sinusoid
5	Person's correlation	[-1;1]	1	Linear relation of intensities
6	Fold difference	[1;+∞]	1	Ratio of intensities
7	Spearman's correlation	[-1;1]	1	Ranks of the intensities
8	Fit	[0;1]	1	Number of signals
9	Retrofit	[0;1]	1	Number of signals
10	Overlap	[0;1]	1	Density function
11	Intensity match	[0;1]	1	Linear relation of intensities
12	LIN's concordance correlation	[-1;1]	1	Concordance of intensities
13	Kendall's correlation	[-1;1]	1	Ranks of the intensities
14	Euclidean distance	[0;+∞]	0	Difference of intensities
15	Manhattan distance	[0;+∞]	0	Difference of intensities

5.3 Simulation study

In order to evaluate the performances of the reviewed indices a comprehensive simulation study was carried out mimicking the proteomic profiles observed in thyroid cancer. Two main simulation protocols were set-up to consider two completely different shapes of the spectra, contrasting a situation of picks uniformly distributed vs picks not uniformly distributed. In each protocol, various scenarios of mass profiles, with masses ranging from 3,000 to 15,000 m/z, were investigated, with different numbers of truly relevant peaks, percentage of shared common peaks, different variability on peaks intensity and localization along the *m/z* axis.

All the simulated spectra were pre-processed following the same strategy: baseline subtraction (median method), smoothing (moving average method, half window width 2.5), normalization (total ion current, TIC), peak alignment, and peak picking (signal-to-noise ratio, S/N ≥ 6). Only peaks with S/N ≥ 6 and an abundance greater or equal to 0.0003 were considered as relevant peaks. The choice of the cut-off was justified by the magnitude of the intensities that we identified as relevant in our experience in thyroid cancer.

Spectral comparisons were made between couples of simulated mass profiles for intra- and inter-group comparisons.

5.3.1 Simulation protocol 1

The first simulation protocol have explored a uniform distribution of the peaks along the *m/z* axis. The spectra generation followed this strategy:

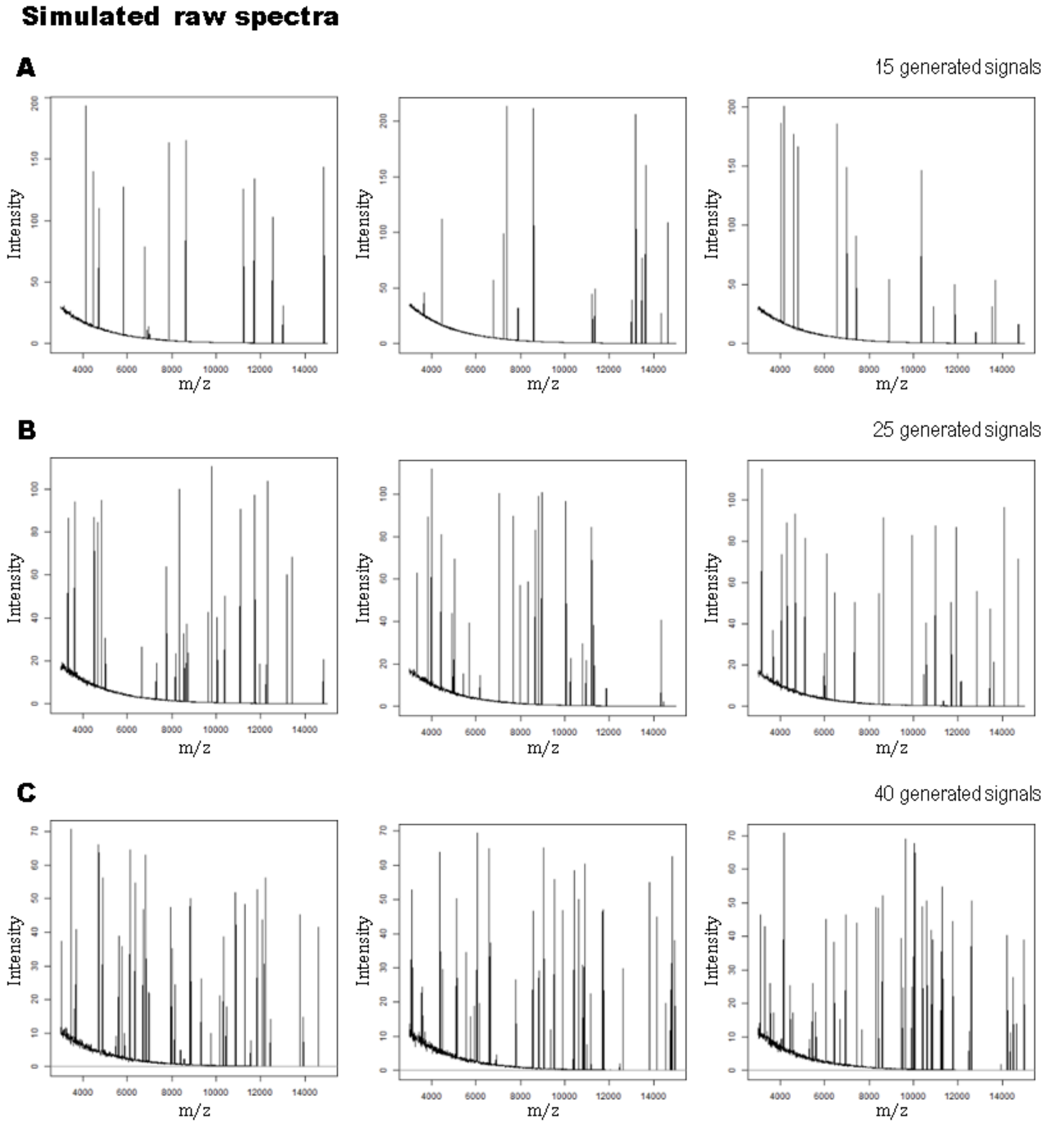
1. Two reference spectra, one for each group, were generated with the same pre-defined number of informative peaks, but with different percentages of common *m/z* values. The remaining *m/z* values (allowing to reach the pre-defined number of true peaks) were randomly generated by a uniform distribution along the *m/z* axis.
2. Given the localization of *m/z* values along the axis, the corresponding peaks abundance was simulated from a uniform distribution with an absolute intensity ranging from 3.5 to 200. The abundance of *m/z* common peaks were forced to overlap, with small random variations from a normal distribution. Non-common peaks were free to have different intensity values. An exponential distribution was used in generating the baseline spectra, while the background random noise followed a gaussian distribution with mean equal to the baseline value at each *m/z* point and standard deviation 10^(-5).
3. For each reference spectrum, 99 replications were generated by small random changes in peak intensities with respect to the reference ones to obtain 100 replicates for each of the two groups.

4. Each configuration was replicated 20 times changing the *m/z* localisation along the axis.

Overall, we investigated 30 scenarios obtained by combining 3 different values for the number of informative peaks, 5 values for the percentage of common peaks and 2 values for intensity variability of common signals, as reported in the table below:

n° of relevant peaks:	40, 25, 10
% of common signals:	90, 75, 50, 25, 10
% of intensity variability:	30, 5

Examples of the raw spectra (without pre-processing) generated with 15 (A), 25 (B) and 40 (C) relevant peaks between a mass range of 3000-15000 m/z are reported in the graph below.



5.3.2 Simulation protocol 2

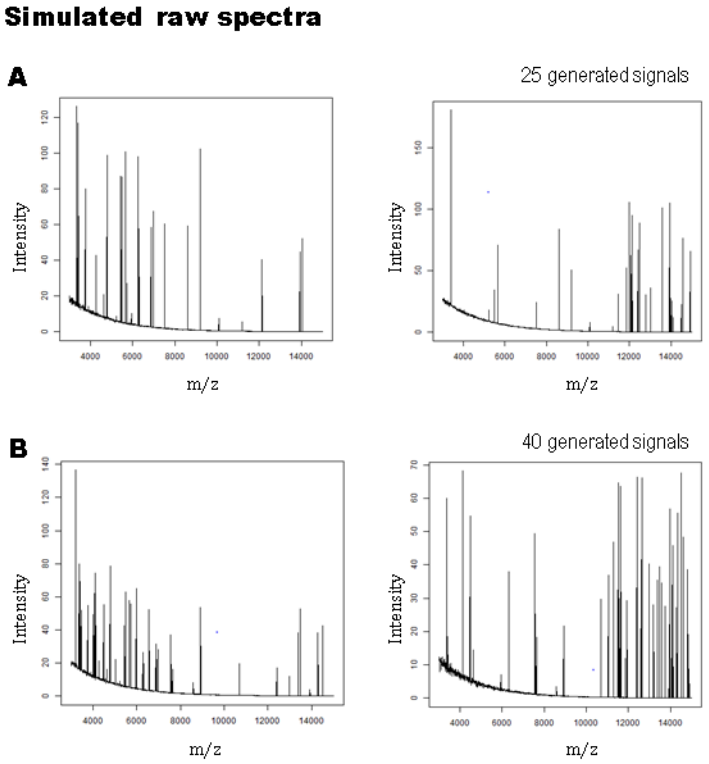
The second simulation protocol have explored a non uniform distribution of the peaks along the *m/z* axis. Specifically, the *m/z* axis was divided into three regions (3000-7000; 7000-11000; 11000-15000) and the spectra generation followed this strategy:

1. The first reference spectrum was generated in order to have in the first region twice or five times the percentage of peaks compared to each of the other two regions.
2. The second reference spectrum was generated symmetrically with respect to the first one, with the same percentage of peaks that fell into the third region, while the remaining percentage of signals were equally distributed in the first two regions.
3. No common peaks was generated between the two different reference spectra. In order to overcome this problem, peaks generated in the second region were the same between the two reference spectra and with the same number of *m/z*. Moreover, the percentage of signals in the third region of the first reference spectrum was chosen among the percentage of signals in the third region of the second reference spectrum, due to its highest amount of signals in that region. Similarly, we have done the same for the first region of the second reference spectrum, which were chosen among the *m/z* value generated for the first reference spectrum in the first region. In the same way, the informative peaks in the first region of the second spectrum were chosen among the informative peaks generated for the first spectrum in the first region.
4. For each reference spectrum, 99 replicates that differ from the reference spectra only for peak intensities were generated.
5. Each configuration was replicated 20 times changing the *m/z* localisation along the axis.

Overall, we investigated 4 scenarios obtained by combining 2 different values of the number of informative peaks, 2 values for the percentage of relevant peaks in each of the three regions and 1 values for intensity variability of common signals, as reported in the table below:

n° of relevant peaks	40, 25
% of relevant peaks in each region:	
- first group	(70, 15, 15) - (50, 25,25)
- second group	(15, 15, 70) - (25, 25, 50)
% of intensity variability	5

Examples of the raw spectra (without pre-processing) generated with 25 (A) and 40 (B) relevant peaks between a mass range of 3000-15000 m/z and with a density of 50% of the relevant peaks in the richest regions, are reported in the graph below.



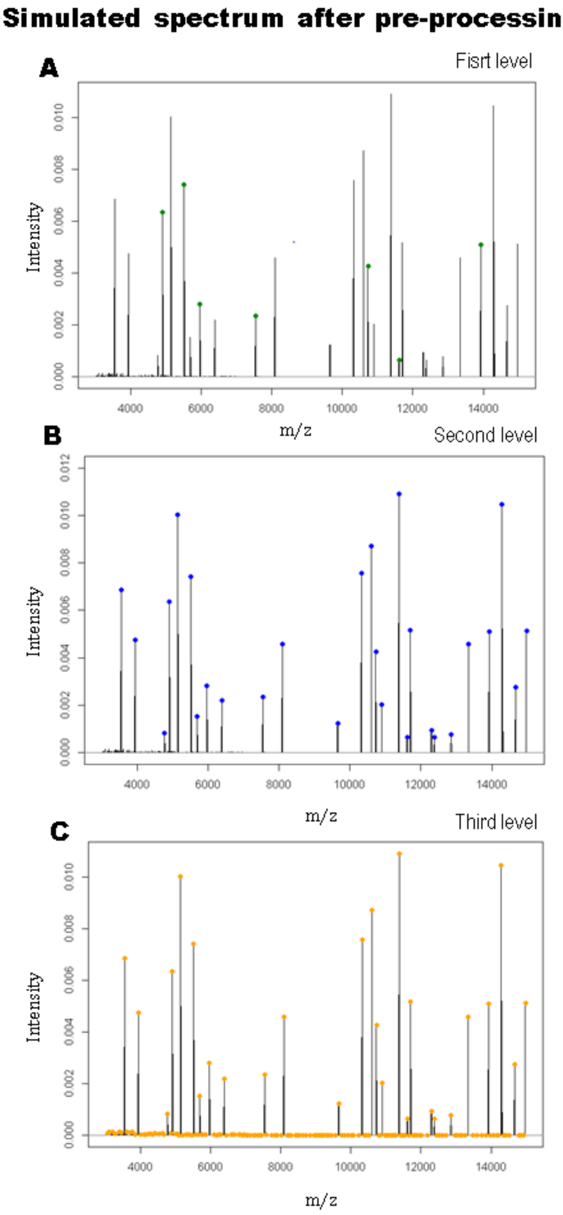
5.3.3 Details on the analyses

The 200 spectra generated for each scenario led to a 5000 paired spectra comparisons in each of the two groups evaluating similarity between replicates, while 5050 inter-group comparisons were performed. Each configuration was replicated 20 times changing the *m/z* localisation along the axis, for a total of 5000x20x2 intra-group comparisons and 5050x20 inter-group comparisons.

A total of 15 scores were considered and each score was evaluated at three different level of detected signals:

1. First level: only common peaks with an absolute intensity greater or equal than the cut-off of 0.0003 (after TIC normalization) were retained for paired comparisons.
2. Second level: all the *m/z* values with an absolute intensity greater or equal than the cut-off either in the first or the second spectrum are taken into consideration for the comparison.
3. Third level: the whole spectra were retained and all the detected signals with S/N ≥ 6, were taken into account.

An example of a spectrum (after pre-processing) generated with 25 relevant peaks in a mass range of 3000-15000 is reported in the graph below. Colored points show the detected signal involved in the three levels evaluation: (A) only signals greater than the threshold and in common with another hypothetical spectrum (25% of common peaks) (green points) were retained for the analysis; (B) all the signals greater than the threshold (blue points) were retained for the analysis; (C) all the signals of the spectrum (orange points) were retained for the analysis.



Replicates of the same sample were required to evaluate the reliability of the different similarity measures, since no recognized standard of reference exists. The 95% confidence interval (CI) of the median of the scores obtained for each paired comparison between two different replicates represented the gold standard. In order to assess the performance of different scores in evaluating spectra similarity, the CIs of the median of each paired inter-group comparisons were calculated and compared to the intra-sample CIs [6].

5. 4 Results

5.4.1 Protocol 1

First level (Figure 1,2,3,10,11,12 in appendix D):

At the first level, only common peaks with intensity greater than the threshold were considered in the pairwise comparison. Since we looked at common peaks with the same intensity values, a similar behavior among the different scenarios is expected. In each score an increment in variability of the distribution of the estimated values was observed when the percentage of common peaks in the inter-sample analysis decreased. When comparing the same score, this behavior is attenuated when the number of informative peaks increased (15, 25, 40).

Pearson, Spearman, Kendall and LIN correlation reached negative estimated values and had high heterogeneity. These scores work on intensity agreement: the less the number of signals to be correlated (i.e. for the scenario of 15 relevant generated *m/z* and 10% of common signals, correlation is calculated on only 2 intensity values), the higher was the bias in the results even though little variability in intensity (5%). The Stein and Scott index, Euclidean and Manhattan distance showed a tendency to underestimate similarity when the percentage of common peaks decreased. The degree of underestimation cannot be evaluated because these measure had not a bounded domain, therefore only the comparison with the intra-sample analysis can be discussed. In Euclidean and Manhattan distances the differences in the estimated median values were of low magnitude: the 95% CIs for each scenario in the inter-sample analysis was completely contained in the 95% CI of the intra-sample comparison. Conversely, Stein and Scott index showed a decreasing trend of the inter-samples analyses with respect to intra-sample analysis.

With an increment in peaks intensity variability (30%) these results were more marked, except for Cosine, DFT and Overlap that showed expected estimate values.

Second level (Figure 4,5,6,13,14,15 in appendix D):

At the second level analysis all the *m/z* values greater than the threshold of 0.0003 were retained in the analysis. If a peak was greater than the threshold in only one spectrum (non-common peaks), the intensity value at the same *m/z* was retained for the second spectrum, irrespectively from its value. The focus of the analysis was to highlight the ability of different scores to detect dissimilarities between spectra at different percentage of common peaks.

Pearson, Spearman and Kendall correlation indices showed median estimated values around 0.5 in the inter-sample comparison in the presence of 75% of common peaks and when 15 peaks were generated. When the number of relevant peaks increased (i.e. 25, 40), the estimated median value was around 0.5 in the scenario with the 90% of common peaks, even if it was expected to be (asymptotically) near 1. When the intensity variability increases at 30% also the estimated median value related to 90% of common peaks scenario decreased to a value lower than 0.5, irrespective to the number of relevant peaks. This suggested that non-common peaks heavily affected Pearson, Spearman and Kendall correlation, although, boxplots decreased according to the decrease of the percentage of common peaks. Intensity match and SI gave more reliable results. Indeed, the first score worked with a greater tolerance when comparing signal intensities, while the second took into account the percentage of dissimilarity in peaks intensities. An increment of intensity variability (30%) lead to underestimation in similarity for Intensity match (e.g. intra-sample analysis showed an estimated median value around 0.45), while SI lead to the same results showed in the scenarios with 5% variability. Conversely, scores that considered the whole shape of the spectra, like the overlap, lead to an overestimation of mass spectra similarity when the percentage of common peaks decreased. On the other hand, with the increment in intensity variability the estimated median value slightly decreased. This phenomenon derives from the fact that there is always an overlap between two density functions due to the uniform distribution of *m/z* value along the entire mass range. The overestimation of mass spectra similarity decreased when the number of relevant signals decreased.

Results showed that Fold difference and Stein and Scott index are sensible to intensity value of non-common signals. Since all the *m/z* values of the two spectra were taken into account, the ratio between the intensity value of the background noise and informative peaks was calculated, leading to infinite value. A possible solution to this problem, as suggested in the literature, was to substitute noise values with the threshold value. In the Fold difference score this brought to reasonable results, the trend on boxplots of "fold difference" and "fold difference literature" was the same. On the other hand, Stein and Scott index using the same solution was not able to take difference between the different scenarios, leading to the same result for all the different percentage of common peaks.

Third level (Figure 7,8,9,16,17,18 in appendix D):

No relevant differences were found between the results at second and third level of retained peaks in Cosine correlation, DFT, Overlap, Fold difference and Euclidean and Manhattan distance. Pearson and LIN correlations led to better performances. Conversely, Spearman and Kendall correlations failed both in the intra and in the inter-sample analysis for all the three sets of generated signals (i.e. 15, 25, 40), showing similarity measures lower than zero. Intensity match had estimated median values lower than 0.08 when 15 relevant signals were generated, and 0.20 when 15 relevant signals were generated. Similarity index underestimated both in the intra and in the inter-sample analyses. Indeed, in all the scenarios the estimated median values were higher than 90. Stein and Scott index showed no trend with the decrease of the percentage of common peaks, with all the boxplots lying on the same range of values.

5.4.2 Protocol 2

First level (Figure 19,20,25,26 in appendix D):

The four different types of correlations (i.e. Pearson, Spearman, Kendall and LIN correlation) underestimated the median expected value for the inter-sample analysis, with median values around 0.6-0.7 for a total number of 40 informative peaks generated. This effect increased with the decrease of the generated peaks (25), with estimated similarity values around 0.5-0.6. The worst performance was reached by the Intensity match that showed results in the inter-sample analysis lower than 0.2. The Cosine, DFT, Fold difference and overlap indices reached the best results, with similarity responses in the inter-sample analysis superimposable to the ones of the intra-sample analysis: median values greater than 0.8 for the overlap score, around 0.9 for the cosine score, and 0.94 for the DFT. A completely overlap in the 95% CIs of the intra and inter-analysis for the Fold difference score was observed. The Euclidean and Manhattan distance, Similarity index and Stein and Scott index, showed discrepancies in their estimated median values between intra and inter-sample analysis, leading to completely non overlapping boxplots.

Second level (Figure 21,22,27,28 in appendix D):

When a percentage of 50% of generated peaks fell in the first or third regions and the remaining 50% was equally distributed in the other two regions, an overlap of 70% was expected. Cosine correlation, DFT, Overlap reached this response in the inter-sample analysis, with better results obtained for the DFT and overlap, and an minimal underestimation for the cosine (median results around 0.6) and overestimation for the DFT (median results around 0.8). With the increment of the number of generated peaks (40), these estimated median values decreased.

When the 70% of the peaks fell in the first or third region and the remaining 30% was equally distributed in the other two regions, the expected overlap was of 35%. In the case of 25 generated peaks, DFT had a median value greater than 0.6, Cosine around 0.4 and Overlap score around 0.35. When the number of generated peaks increased to 40, DFT showed results slightly lower than 0.6 and were around 0.35 for the cosine. The Overlap index showed the same behavior as before. Pearson, Spearman, Kendall, LIN correlation and Intensity match showed median estimated values around 0.2 in the inter-sample analysis, and the first four measures reached negative values increasing the number of peaks in the richest region.

The Euclidean and Manhattan distance showed an increment in the differences between the median estimated values of the intra and inter analysis with the decrease of the number of relevant peaks. This is reasonable, because the lower the number of signals, the greater should be the influence of the difference between peak intensities. In the same way, the greater the number of peaks within the richest regions on the m/z axis that resulted in fewer common peaks between the two spectra, the greater the spectra dissimilarity.

Third level (Figure 23,24,29,30 in appendix D):

The results in analysis at the third level were mainly comparable to those obtained at the second level.

5.5 Discussion

5.5.1 Protocol 1

The relevant factors to discuss about the results of the simulations from Protocol 1 are: the number of relevant signals, the percentage of common peaks, the localization of *m/z* signals and their intensity variability.

Localization of m/z signals:

Looking at the intra- and the inter-sample comparison between spectra with the 90% of common signals, an overall increase of variability was reached according to the decrease of the number of informative peaks generated. A small number of peaks migth bring different localization scenarios of *m/z* values, leading to similar or completely different shape of the spectra (Figure in paragraph 5.3.1).

Number of informative signals:

Another problem was the effect of the background noise. The average number of relevant generated signals among spectra was pre-defined, with small variability given by the presence of noise that might get out from the spectrum other signals. Scores were more sensitive to noise when the number of relevant peaks decreased. In this case, an increase in the number of informative peaks reduced the effect of this problem, because the non-real signals were mediated by the effect of the increasing number of relevant signal.

Intensity variability:

A serious problem in mass spectra similarity measures is the variability in signal intensities caused by analytical variability (i.e. sample preparation, instrumental analysis, spectra misalignment). As suggested in the literature various method to align and scaling spectra to compensate for spectra differences (pre-processing workflow) had to be performed to discard anomalous peaks. The problem could not be totally eliminated, but only controlled. The less are the detected signals, the greater the influence of this problem. A little variability in peak intensities could cause a loss in the reliability of the match, when the score is only based on peaks intensities. These evidences suggested that the correlation measures should be used only when the comparison is at the first level because at the second level they might lead to non-informative results. The percentage of common peaks to use had to be consistent, since the risk is to obtain biased results. Other measures that take into account intensity variability, such as the Intensity match, might have better performance than the correlation indices at either levels. SI, which works similarly to the Intensity match, could be a valid similarity measure, with opposite interpretation with respect to the Intensity match. When comparing two spectra, the differences between the spectra had to be greater than the SI that was calculated in a repeatability study (intra-sample). Euclidean and Manhattan distances, Stein and Scott index, and fold difference always need a reference value to be compared, that is obtained in the repeatability study that is not always feasible in the clinical practice. As previously reported in the result section, Fold difference and Stein and Scott index are sensible to the intensity value of non-common signals. A possible solution, suggested in the literature by Wan et al. was to substitute noise value with the threshold value. In this way, false signals are introduced, leading to false results and a higher risk of changing the data.

In conclusion, overlap and DFT led to good results also when intensity variability increases, but showed a flat estimation when the percentage of common peaks decreased. The Overlap score is preferable due to the higher computational time of DFT score, and to the fact that DFT depends on parameters that had to be set and that could lead to different results (no investigation were done on this aspect). Overlap and DFT failed when the Cosine score was able to differentiate between very similar spectra. The Cosine correlation showed to have the best performance in all the scenarios.

Lastly, some of the proposed measure are constructed to work on the whole mass spectrum (e.g. overlap, cosine correlation), rather than on the restricted spectrum obtained considering only the only most abundant peaks detected (e.g. Correlation scores). In the first case the whole shape of the spectra was considered and low abundance peaks that could provide information were retained. While, in the second case the spatial distribution of the ions along mass-to-charge (*m/z*) axis was lost, but certainly the background noise was discarded.

5.5.2 Protocol 2

The second protocol of simulations led to slightly different conclusions. Correlation's scores showed the best performances, but always with an underestimation of the expected results. Again, this suggests that the correlation scores are heavily affected by intensity values, since they looked for completely correspondence in intensities. Moreover, in proteomics analysis, the background noise that perturbs the intensity values brings to non reliable results mostly for the correlation scores that work on ranked data (i.e. Spearman and Kendall correlation). This negative effect increased when the analyses were performed at the third level for all the types of scenarios taken into consideration in this work. Overlap showed the best results as compared to DFT and Cosine, differently to the results found in Protocol 1.

In conclusion:

- A perfect score usable in all the situations does not exist.
- Euclidean distance, Manhattan distance, Fold difference, Similarity index and Stein and Scott index can only be considered when an intra-sample analysis is available as gold standard to make the interpretation of inter-sample comparisons possible. Only the ratio between the estimated median values of the scores in the inter-sample analysis compared to the ones obtained in the intra-sample analysis can be evaluated. Moreover, we suggest to use Fold difference and Stein and Scott index only for the comparison of common peaks.
- Pearson's correlation, Spearman's correlation, Kendall's correlation and LIN's concordance correlation led to the worst results since they search for a complete correspondence in peaks intensities between two spectra. The suggestion is to use them only in the comparison of common peaks, to highlight if common peaks had also common intensities. Intensity match can be a useful substitute of Correlation due to the fact that it takes into account a greater intensity variability that is common in mass spectrometry analysis.
- The best performers are the Cosine and DFT when the *m/z* values are uniformly distributed along the *m/z* axis (simulation protocol 1), and Overlap when the scores are not uniformly distributed along the m/z axis (simulation protocol 2).

- The global results of these simulations provide the rationale for the construction of composite scores able to take into account different aspects. Fit and Retrofit, which compared two spectra based only on the number of common peaks, without bias generated by looking at peaks intensity, had to be retained. Cosine correlation and overlap could be two additional scores to include in the composite score system since they efficiently consider signal intensities. Furthermore, the Overlap index include information also on the whole shape of the spectra.

Bibliography

[1] S. E. Stein and D. R. Scott, “Optimization and testing of mass spectral library search algorithms for compound identification,” *J. Am. Soc. Mass Spectrom.*, vol. 5, no. 9, pp. 859–866, 1994.

[2] F. L. Bazsó, O. Ozohanics, G. Schlosser, K. Ludányi, K. Vékey, and L. Drahos, “Quantitative Comparison of Tandem Mass Spectra Obtained on Various Instruments,” *J. Am. Soc. Mass Spectrom.*, vol. 27, no. 8, pp. 1357–1365, 2016.

[3] I. Piga *et al.*, “Feasibility Study for the MALDI-MSI Analysis of Thyroid Fine Needle Aspiration Biopsies: Evaluating the Morphological and Proteomic Stability Over Time,” *Proteomics - Clin. Appl.*, vol. 13, no. 1, pp. 1–9, 2019.

[4] I. Koo, S. Kim, and X. Zhang, “Comparative analysis of mass spectral matching-based compound identification in gas chromatography-mass spectrometry,” *J. Chromatogr. A*, vol. 1298, pp. 132–138, Jul. 2013.

[5] K. X. Wan, I. Vidavsky, and M. L. Gross, “Comparing similar spectra: From similarity index to spectral contrast angle,” *J. Am. Soc. Mass Spectrom.*, vol. 13, no. 1, pp. 85–88, 2002.

[6] C. E. Costello, H. S. Hertz, T. Sakai, and K. Biemann, “Routine use of a flexible gas chromatograph-mass spectrometer-computer system to identify drugs and their metabolites in body fluids of overdose victims.,” *Clin. Chem.*, vol. 20, no. 2, pp. 255–265, 1974.

[7] I. Koo, X. Zhang, and S. Kim, “Wavelet- and fourier-transform-based spectrum similarity approaches to compound identification in gas chromatography/mass spectrometry,” *Anal. Chem.*, vol. 83, no. 14, pp. 5631–5638, 2011.

[8] L. Drahos and K. Vékey, “Quantification of isomeric differences in mass spectra,” in *Rapid Communications in Mass Spectrometry*, 1996, vol. 10, no. 10, pp. 1309–1315.

[9] P. D. Welch, *The Fast Fourier Transform and Its Applications by E. Oran Brigham*, vol. 12, no. 1. 1988.

[10] Q. Li, “Assigning Significance in Label-Free Quantitative Proteomics to Include Single-Peptide-Hit Proteins with Low Replicates,” *Int. J. Proteomics*, vol. 2010, pp. 1–15, 2010.

[11] K. Hollemeyer, W. Altmeyer, E. Heinzle, and C. Pitra, “Species identification of Oetzi’s clothing with matrix-assisted laser desorption/ionization time-of-flight mass spectrometry based on peptide pattern similarities of hair digests.,” *Rapid Commun. Mass Spectrom.*, vol. 22, no. 18, pp. 2751–67, Sep. 2008.

[12] F. Schmid and A. Schmidt, “Nonparametric estimation of the coefficient of overlapping - Theory and empirical application,” *Comput. Stat. Data Anal.*, vol. 50, no. 6, pp. 1583–1596, Mar. 2006.

[13] L. I.-K. Lin, “A Concordance Correlation Coefficient to Evaluate Reproducibility,” *Biometrics*, vol. 45, no. 1, p. 255, Mar. 1989.

[14] C. A. E. Nickerson, “A Note On ‘A Concordance Correlation Coefficient to Evaluate Reproducibility,’” *Biometrics*, vol. 53, no. 4, p. 1503, Dec. 1997.

6 Discussion

MALDI-MSI represents an ideal tool to explore the spatial distribution of proteins directly in-situ, integrating molecular and cytomorphological information and enabling the discovery of potential diagnostic markers in thyroid cytopathology. Given the amount of data generated from MALDI-MSI analysis, it is of paramount importance to use proper statistical methods in order to find discriminant features for thyroid nodules classification. However, many technical challenges had to be solved in order to reach this goal. The first one was the interference of haemoglobin. In fact, red blood cells present in the fine needle aspiration biopsy (FNAB) specimens caused ion suppression of other proteins during the MALDI-MSI analysis due to large amounts of haemoglobin. We planned a study comparing three protocols that used ex-vivo cytological samples collected from fresh thyroid nodules of 9 patients who underwent total thyroidectomy: (A) conventional air-dried smears; (B) cytological smears immediately fixed in ethanol; (C) ThinPrep liquid base preparation (LBP). Protocol C and A were also evaluated using real FNABs. The study highlighted the possibility to manage the haemoglobin interference when LBP was used as sample preparation protocol, obtaining high-quality MALDI-MS spectra that could be used for a more reliable comparison of in situ protein profiles.

The sample preparation protocol was then further improved and the second technical challenge regarded the morphological and proteomic stability of the samples in the preservative solutions. Mass spectra similarity was investigated on intra-day, inter-day replicates and on samples stored at 4°C and prepared at different time points. Results showed no degradation of the cellular morphology and good stability of the protein profiles when the specimen was placed for up to 14 days in PreservCyt solution.

Assessing the similarity of mass spectra is a major topic in mass spectrometry data comparison. A review of the most used scores for the evaluation of mass spectra proteomic profiles similarity was performed and a new index of Overlap was proposed here. A simulation study was implemented, investigating different scenarios, in order to identify the best similarity measures to compare proteomic profiles. In particular, it was observed that the best similarity measure could be reached by combining the scores with the best performances into a unique composite score: fit, retrofit, overlap and cosine correlation.

The optimization of the proteomic protocol paved the way for the clinical question of this thesis: the classification of benign vs malignant for the qualification of the indeterminate TYR 3 FNABs. Unfortunately, this was done only on a subset of the target sample size due to a low rate of enrolment of patients with malignant lesions. However, the statistical model was based on the analysis of a considerable number of Region of Interests (ROIs), according to the morphological triage performed by the pathologist. For the two groups in the training set (benign vs malignant lesions), a logistic regression with a Lasso regularization method was performed and twenty features were selected from the statistical model as the most discriminant to correctly distinguish samples and quantify their probability of being malignant lesions. Finally, the model was validated on a different group of patients using the overall average spectra of all the analysis, the spectra from each ROI and a pixel by pixel approach using all the single spectra of the MALDI-MSI analysis. Successful results were obtained, with the correct classification of different types of thyroid lesions being achieved. Notwithstanding the consideration that the diagnostic validity of the model needs to be verified in the large cohort of patients that is currently under enrolment.

The plan for the future is to perform the final analysis of the MALDI spectra on a training set that involves 80 clear-cut diagnosis of THY2 and 25 THY5, numbers quite far from the ones originally planned (i.e. 160 vs 80). This is due to the convergence of two conditions: i) the current rate of THY5 has declined compared to the past and ii) some samples have been discarded since the challenging needle washing material from FNABs is sometimes too scarce to make possible the MALDI analysis. However, it should be noted that the final analysis will be performed on data derived from the ROIs, so that each subject will contribute with multiple spots of information.

Furthermore, since for certain types of thyroid lesions, as the follicular lesions and Noninvasive Follicular Thyroid Neoplasm With Papillary-Like Nuclear Features (NIFTP), the distinction between benign and malignant nodules is possible only after histology, because the cytological report is almost always THY3. We will introduce into the training set some THY3 cytological specimens with these diagnoses in order to train the model to recognize also these uncertain cases for which the traditional morphological diagnosis is not possible.

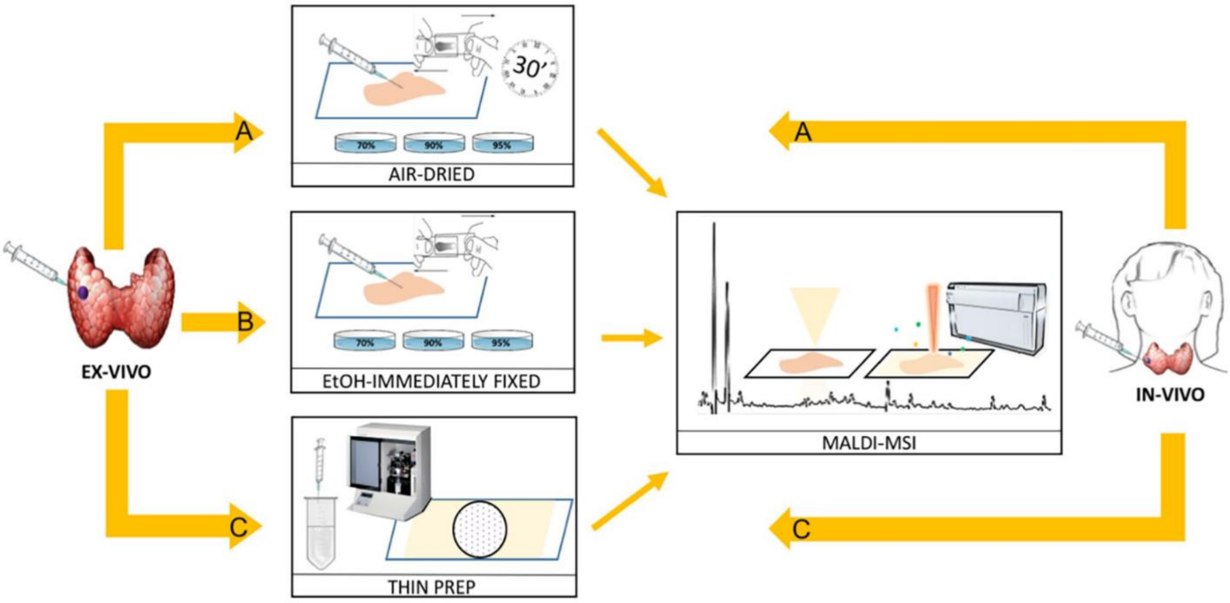
Moreover, since multiple sources of information are available for the same subject, we will consider a second analysis that integrates different omics data. Proteomics, genetics and clinical/pathological data will be combined using the Integrative Lasso with Penalty Factors, method (IPF-LASSO). This method is based on the regression model with L1 LASSO penalization in which each omics group will be differently penalized assigning one penalty factor to each modality determined by cross-validation. Finally, we foresee also a methodological development that extends the aforementioned Lasso models to handle the classification of the heterogeneous group of benign and malignant thyroid lesions.

In conclusion, the direct consequences of these successful results reported in this thesis could be the use of MALDI-MSI proteomics as a complementary approach for the characterization of indeterminate for malignancy thyroid nodules also in particularly challenging situations, as in case of “needle washing” material from FNABs. Despite the technical challenges of this study, the application of proper proteomics, imaging and statistical approaches may help to elucidate key bio molecular events and pathways in oncogenic processes. Collectively, this represents an important paradigm for both the fundamental characterization of cancer systems and the discovery of molecular targets for diagnostic application.

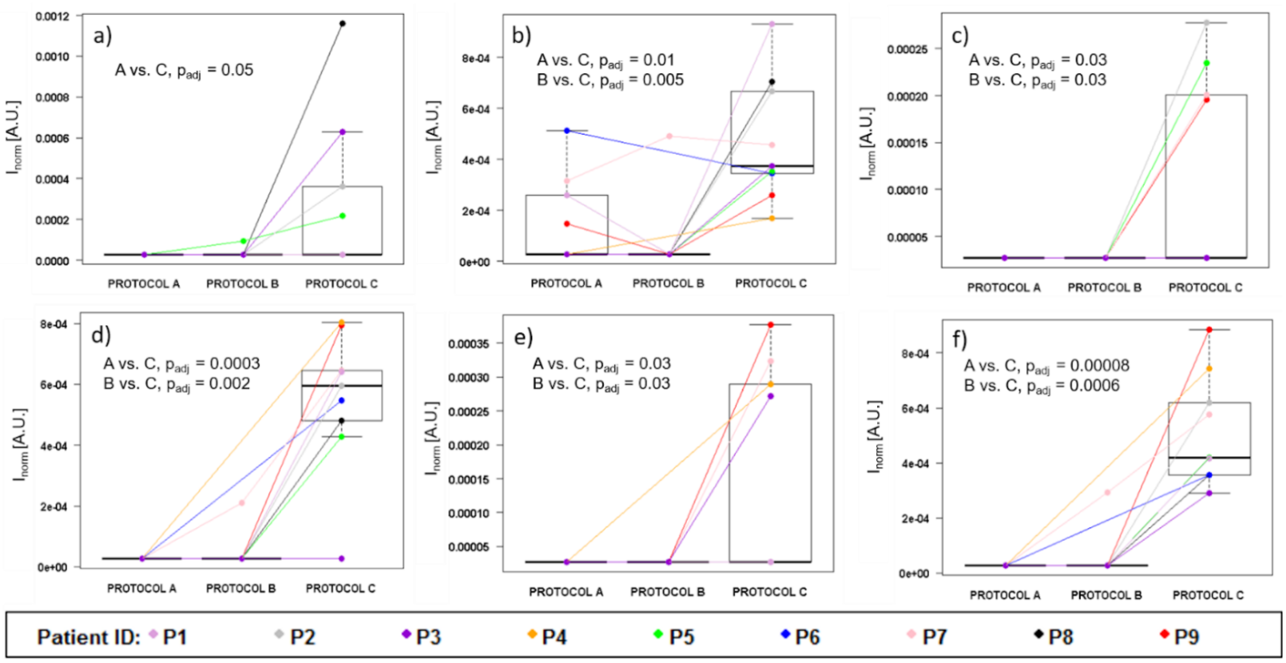
APPENDIX A

Supplementary materials of:

The management of haemoglobin interference for the MALDI-MSI proteomics analysis of thyroid fine needle aspiration biopsies



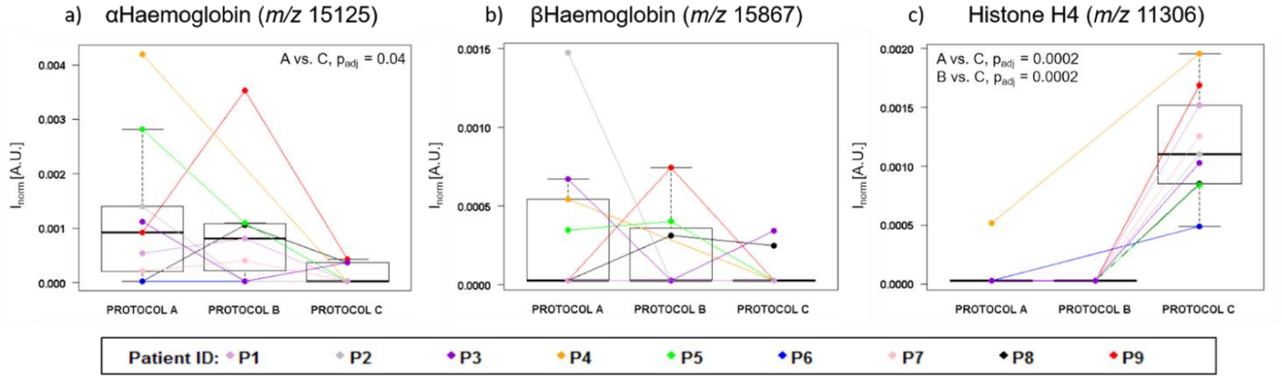
Supplementary Figure 1. General Workflow for the sample preparation of ex-vivo and in-vivo FNAB by MALDI.



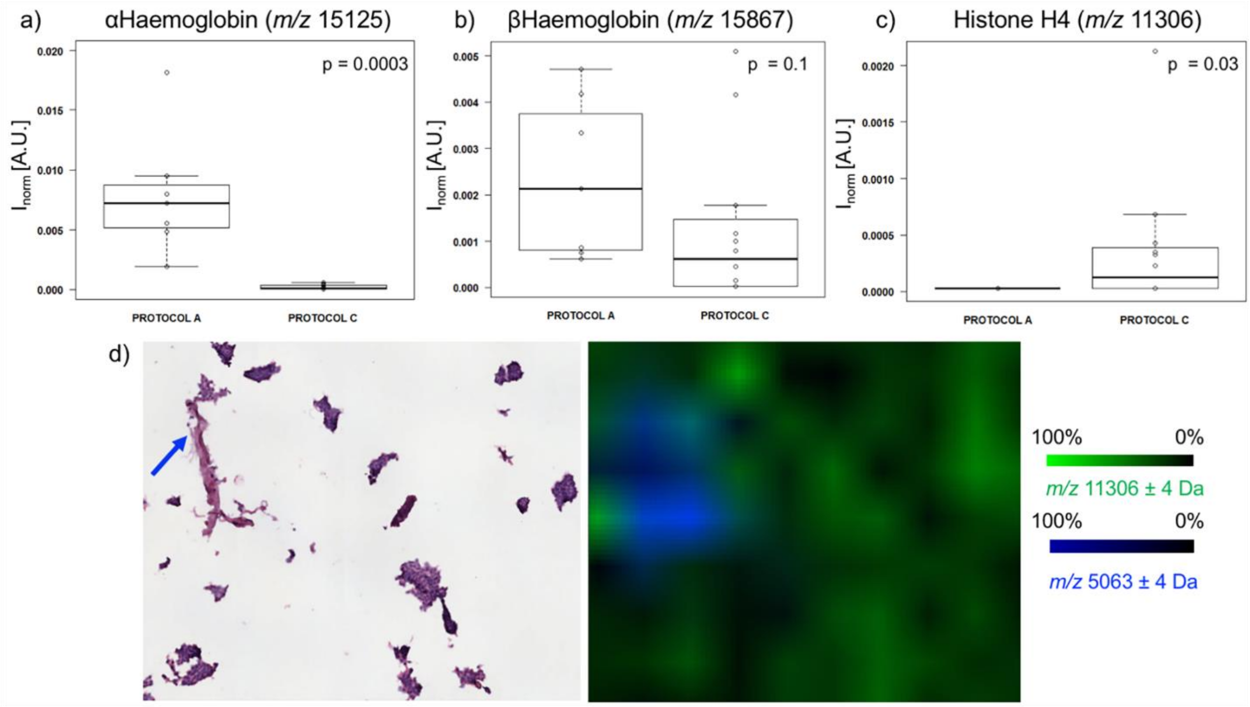
Supplementary Figure 2. Comparison of the three independent sample preparation protocols using thyroid ex-vivo cytological samples from the same patients and boxplots and individual values of the normalized intensity [A.U.] of six signals (a,b,c,d,e,f). The box contains data that fall between the first and third quartiles, the horizontal line indicates the median, and the brackets delineate 1.5 times the interquartile range (with data outside this range defining outliers).

1) Sample	N° of peaks		
	Protocol A	Protocol B	Protocol C
a	26	4	16
b	11	13	18
c	22	22	20
d	35	33	22
2) Samples comparision	N° of common peaks		
	Protocol A	Protocol B	Protocol C
a vs. b	1	2	11
c vs. d	2	5	14

Supplementary Table 1. Summary of the peak histogram presented in Figure 2: 1) total number of peaks in samples a, b, c, d with Protocols A, B and C, respectively; 2) number of common peaks when comparing a with b and c with d using the three protocols



Supplementary Figure 3. Comparison of the three independent sample preparation protocols using ROIs of thyroid ex-vivo cytological samples from the same patients: boxplots and individual values of the normalized intensity [A.U.] of a) αHaemoglobin, b) βHaemoglobin and c) Histone H4. The box contains data that fall between the first and third quartiles, the horizontal line indicates the median, and the brackets delineate 1.5 times the interquartile range (with data outside this range defining outliers).

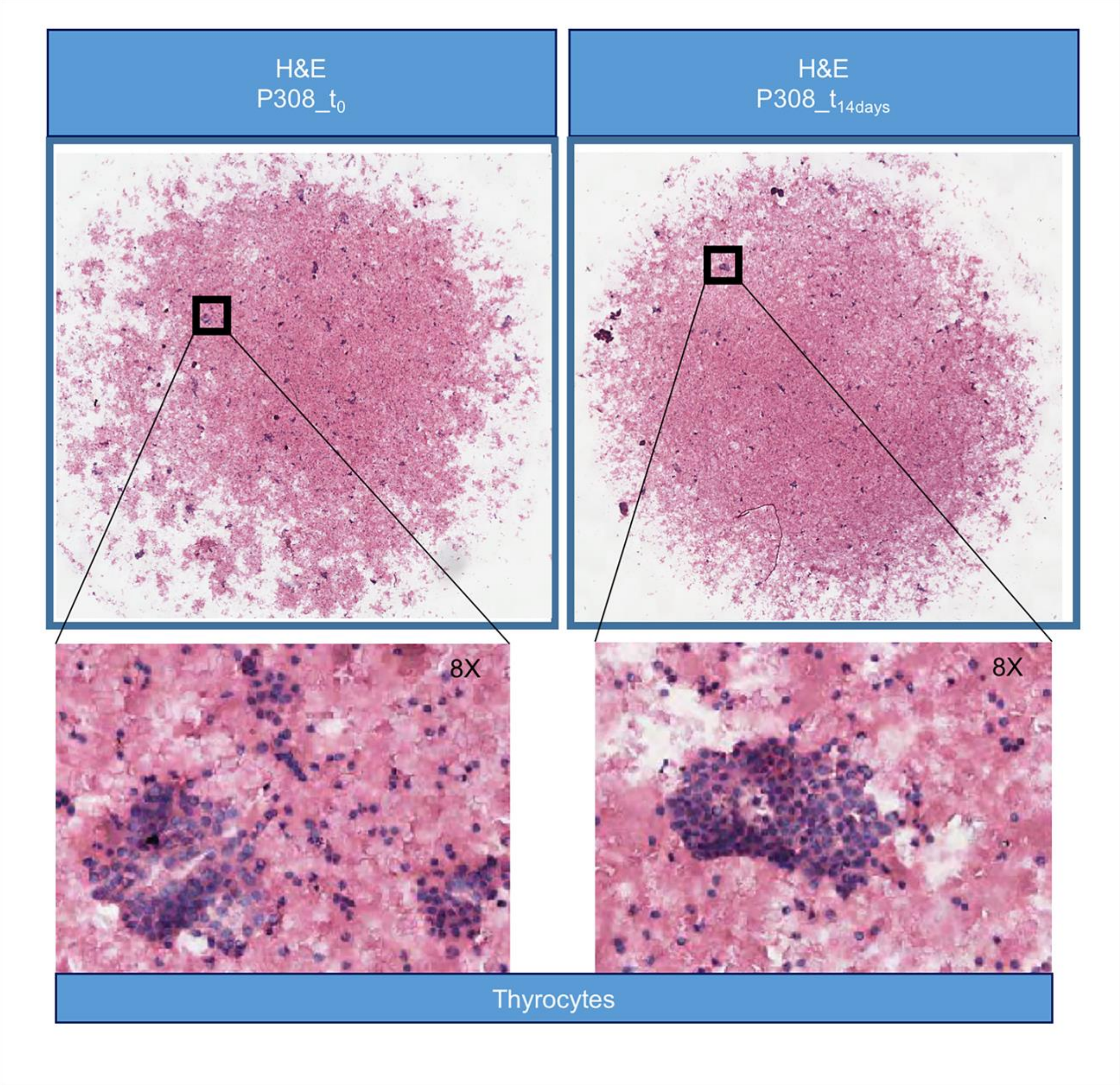


Supplementary Figure 4. Boxplots and individual values of the normalised intensity [A,U.] of a) αHaemoglobin, b) βHaemoglobin and c) Histone H4, comparing 7 air-dried and 12 thyroid FNABs samples. d) H&E staining image and MALDI molecular image and localization of the signal at m/z 5063 in the stroma region (blue) and the signal at m/z 11306 in the thyrocyte cells (green).

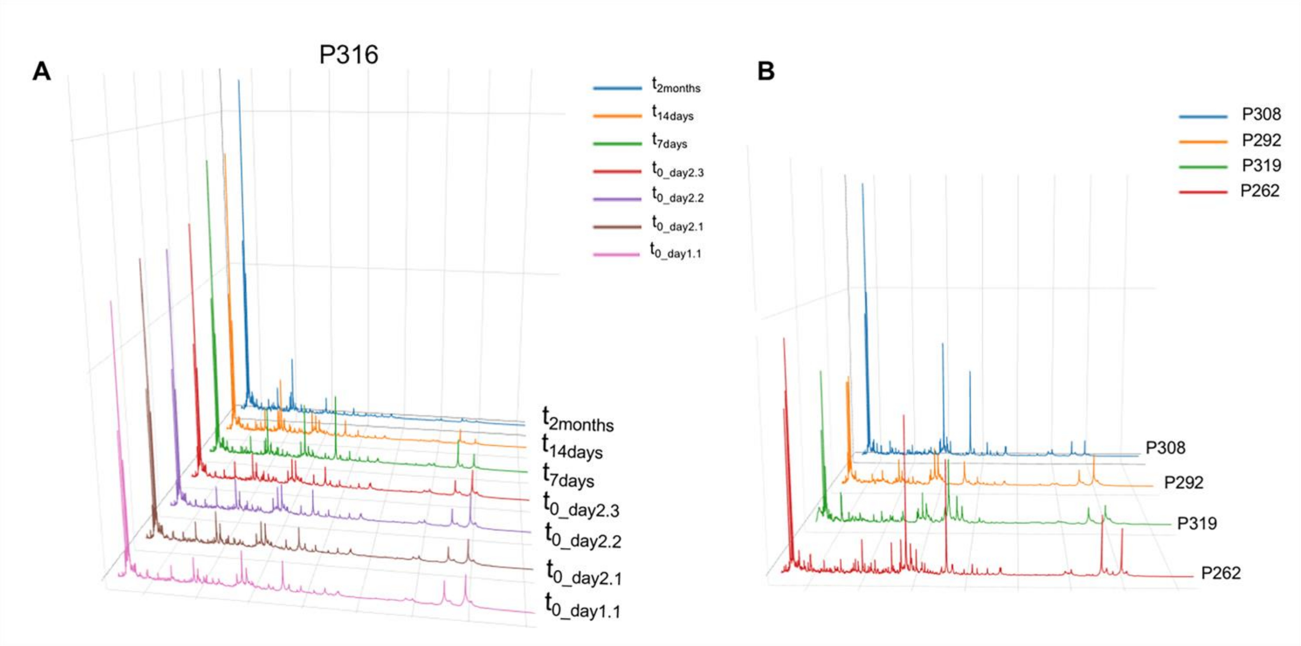
APPENDIX B

Supplementary materials of:

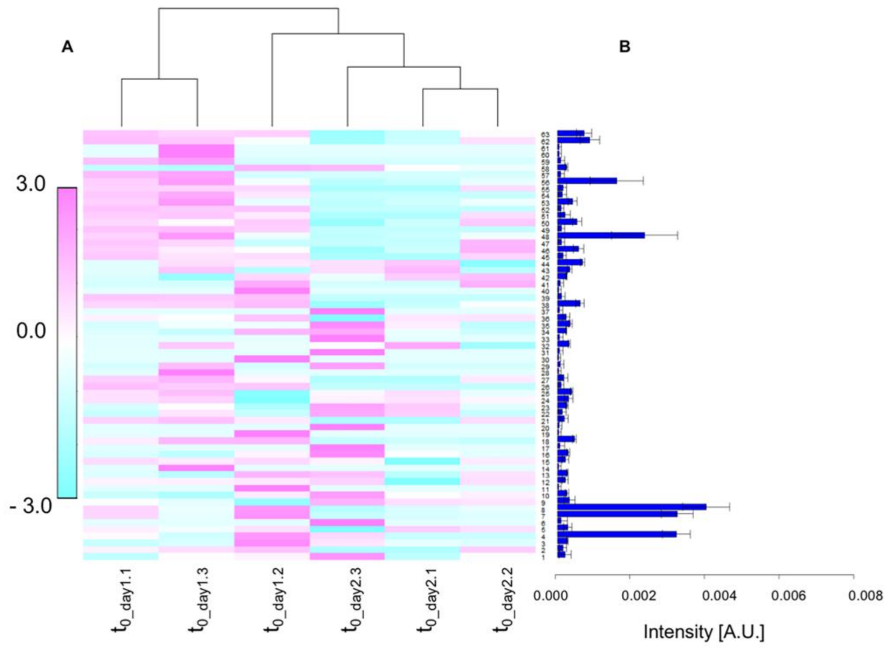
Feasibility study for the MALDI-MSI analysis of thyroid fine needle aspiration biopsies. evaluating the morphological and proteomic stability over time



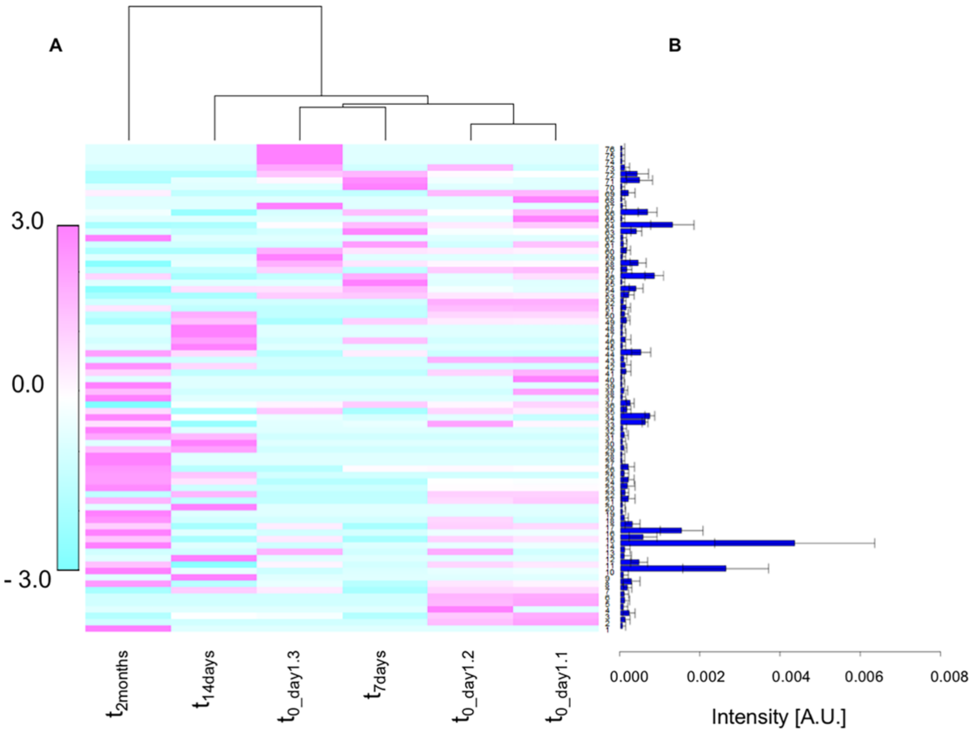
Supplementary Figure 1. H&E staining of the cytospin-based sample preparation of P308 prepared at t_0 and after 14 days of storage in *PreservCyt* solution. An 8X zoom on a cluster of thyrocytes is shown for both cytospin samples.



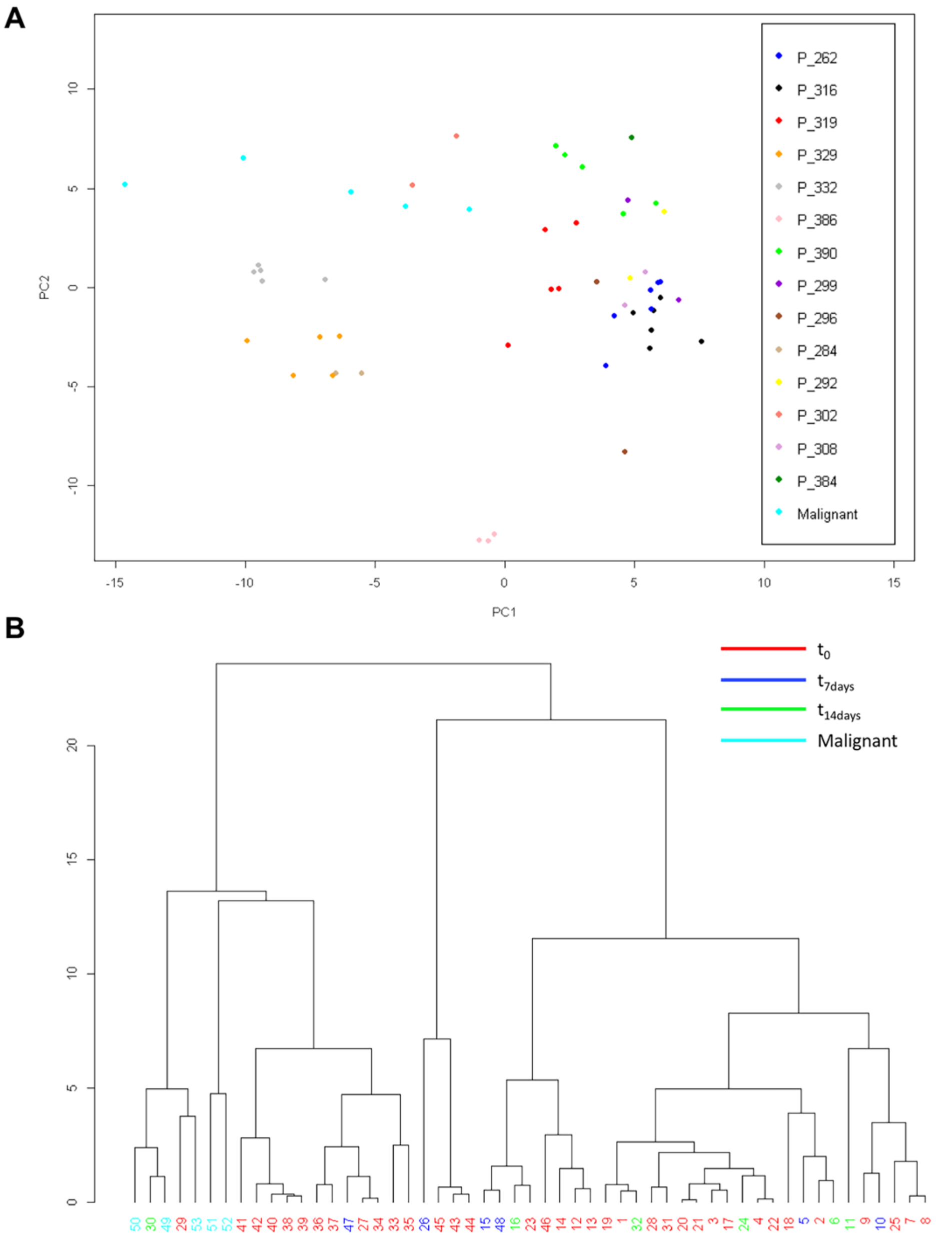
Supplementary Figure 2. (A) Spectra comparison of all replicates of P316; (B) Spectra comparison of the replicates at t_0 in *PreservCyt* of P262, P319, P292, P308.



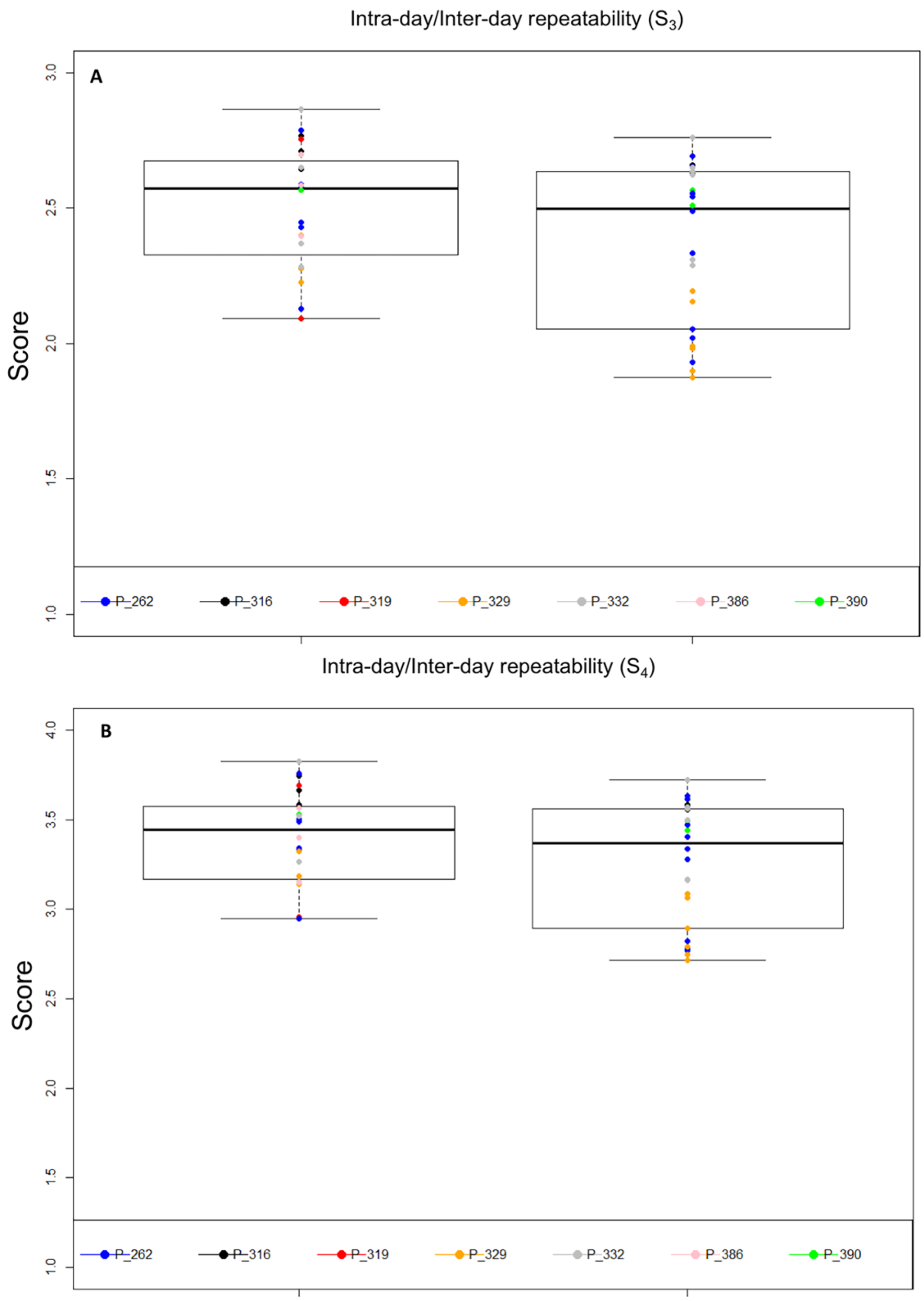
Supplementary Figure 3. Normalized intensities [A.U.] of the peaks (*m/z*) detected in seven replicates of patient P262 ($t_{0\_day1.1}$, $t_{0\_day1.2}$, $t_{0\_day1.3}$, $t_{0\_day2.1}$, $t_{0\_day2.2}$, $t_{0\_day2.3}$,). (A) The heatmap shows the signal intensities of each peaks that were rescaled to have mean 0 and standard deviation 1; (B) the histogram shows the mean intensities of each peaks detected at different times (error bars represent the 95% confidence interval).



Supplementary Figure 4. Normalized intensities [A.U.] of the peaks (*m/z*) detected in seven replicates of patient P319 ($t_{0\_day2.1}$, $t_{0\_day2.2}$, $t_{0\_day2.3}$, $t_{\_7days}$, $t_{\_14days}$, $t_{\_2months}$). (A) The heatmap shows the signal intensities of each peaks that were rescaled to have mean 0 and standard deviation 1; (B) the histogram shows the mean intensities of each peaks detected at different times (error bars represent the 95% confidence interval).



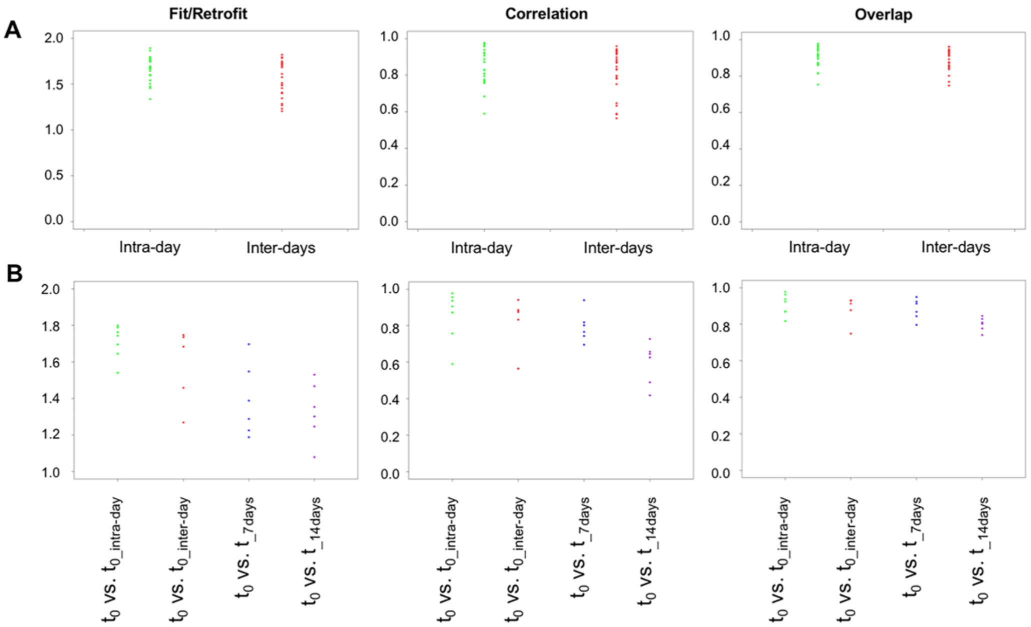
Supplementary Figure 5. (A) Principal Component Analysis of the spectra of all the replicates at t_0 and after storage at 4°C in *PreservCyt* after 7 and 14 days, the dots in light blue are the malignant samples; (B) Hierarchical Clustering Analysis of the spectra of all the replicates at t_0 (red) and after storage at 4°C in *PreservCyt* after 7 (blue) and 14 (green) days, the five malignant samples are colored in light blue.



Intra-day/Inter-day repeatability (S_3)

Intra-day/Inter-day repeatability (S_4)

Supplementary Figure 6. Box-plots, for the intra-day and inter-day repeatability, of the (A) S_3 and (B) S_4 scores. Color dots represent the score paired comparisons for each patient.



Supplementary Figure 7. Scatter plots for the components of S_3 and S_4 (fit and retrofit sum together, spearman's correlation and overlap). (A) Graphs show the three components of all the possible intra-patient combinations for intra-day and inter-day scores. (B) Graphs show the three components for all the spectra comparisons between the randomized reference spectrum t_0 and the query spectra $t_{0\_intra-day}$, $t_{0\_inter-day}$, t_{7days}, t_{14days}.

Patient ID	Cytological classification	Sex	Age (years)
P213	THY5	F	47
P250	THY5	F	87
P262	THY2	F	80
P284	THY2	M	67
P292	THY1	F	78
P295	THY4	F	34
P296	THY3	M	75
P299	THY1	F	54
P302	THY2	F	63
P308	THY2	F	32
P316	THY2	F	70
P319	THY2	F	50
P329	THY2	F	46
P332	THY3	F	43
P384	THY2	F	71
P386	THY2	F	76
P390	THY1	F	51
P442	THY5	F	39
P453	THY5	F	66

Supplementary Table 1. Clinical and demographic characteristics of the patients included in the study.

Patient ID	t_0	t_{7days}	t_{14days}	$t_{2months}$
P262	3x t_{0_day1} 3x t_{0_day2}	--	--	--
P284	1	1	--	--
P292	1	1	--	--
P296	1	1	--	--
P299	1	--	1	--
P302	1	--	1	--
P308	1	--	1	--
P316	1x t_{0_day1} 3x t_{0_day2}	1	1	1
P319	3x t_{0_day1}	1	1	1
P329	2x t_{0_day1} 3x t_{0_day2}	--	--	--
P332	2x t_{0_1day} 3x t_{0_2day}	--	--	--
P384	1	1*	--	--
P386	3x t_{0_day1}	1*	--	--
P390	2x t_{0_day1} 1x t_{0_day2}	1	1	--
P213	1	--	--	--
P250	1	--	--	--
P295	1	--	--	--
P442	1	--	--	--
P453	1	--	--	--

Supplementary Table 2. Experimental design. The number of replicates for each patient at different time of preparation (t_0, t_{7days}, t_{14days}, $t_{2months}$) are reported in each column. The asterisk is referred to the replicates of P384 and P386 stored in *CytoLyt* solution for 7 days.

Index	Evaluation	N	Min	Max	Median	Mean	SD	CV%
S_3	**Intra day**	24	2.09	2.87	2.57	2.51	0.27	8.64
	Inter day	26	1.88	2.76	2.50	2.37	0.30	12.43
S_4	**Intra day**	24	2.95	3.83	3.45	3.41	0.25	7.37
	Inter day	26	2.71	3.72	3.37	3.26	0.34	10.43

Supplementary Table 3. S_3 and S_4 scores of the intra-day and inter-day paired comparisons. The total number of observations (N), minimum (Min), maximum (Max), mean, median, standard deviation (SD) and the coefficient of variation (CV) are calculated.

Index	Comparison				
	t_0 vs. $t_{0\_\text{inter-day}}$	**t_0 vs. $t_{0\_\text{inter-day}}$**	**t_0 vs. $t_{0\_\text{inter-day}}$**	**t_0 vs. $t_{7\text{ days}}$**	**t_0 vs. $t_{14\text{ days}}$**
S_3	**(CV 12.03%)**	**(CV_H 20%)**	**(CV_H 30%)**		
	2.15-2.65	1.98-2.82	1.77-3.03	1.96-2.40	1.75-2.09
S_4	**(CV 10.54%)**	**(CV_H 20%)**	**(CV_H 30%)**		
	2.97-3.58	2.70-3.85	2.42-4.14	2.80-3.33	2.53-2.92

Supplementary Table 4. In the first column are reported the 95% confidence intervals (CIs) for the inter-day comparison. The 95% CI of t_0 vs. $t_{7\text{days}}$ and t_0 vs. $t_{14\text{days}}$ are compared to the 95% CIs of t_0 vs. $t_{0\_\text{inter-day}}$ calculated with an hypothetical coefficient of variation (CV_H) of 20% and 30%.

	INTRA-DAY REPEATABILITY											
	P_386		P_332		P_329		P_319		P_316		P_262	
	Score up to 4	Score up to 3	Score up to 4	Score up to 3	Score up to 4	Score up to 3	Score up to 4	Score up to 3	Score up to 4	Score up to 3	Score up to 4	Score up to 3
	3,361	2,519	3,166	2,277	2,815	1,906	3,658	2,715	3,601	2,666	2,617	1,887
	3,766	2,824	3,525	2,651	3,049	2,120	3,100	2,187	3,666	2,707	3,502	2,655
	3,575	2,715	3,064	2,191	3,548	2,598	3,410	2,499	3,853	2,885	3,069	2,200
			3,809	2,846	3,311	2,392					3,604	2,635
											3,349	2,427
											3,448	2,537
Mean	3,567246333	2,686210267	3,39112325	2,49133295	3,18102925	2,254026025	3,389328733	2,466722033	3,7067308	2,7527335	3,265044617	2,390245067
SD	0,202570382	0,154581357	0,341699829	0,30926313	0,317535412	0,303617069	0,279352945	0,26539358	0,130931115	0,116110552	0,365964095	0,297447304
CV%	6%	6%	10%	12%	10%	13%	8%	11%	4%	4%	11%	12%

	INTER-DAY REPEATABILITY									
	P_390		P_332		P_329		P_316		P_262	
	Score up to 4	Score up to 3	Score up to 4	Score up to 3	Score up to 4	Score up to 3	Score up to 4	Score up to 3	Score up to 4	Score up to 3
	3,442097	2,508865	3,571816	2,634958	3,086376	2,194301	3,584171	2,658442	3,633471	2,692026
	3,496023	2,567602	3,160842	2,288489	2,895106	1,982432	3,556254	2,625619	3,473446	2,555364
			3,562582	2,65154	3,06519	2,153512	3,579264	2,636539	3,613018	2,653482
			3,721197	2,760162	2,792265	1,990703			2,822544	2,053908
			3,167958	2,310005	2,712639	1,875			2,769682	2,021273
			3,499176	2,622838	2,747612	1,899434			2,781811	1,931197
									3,403377	2,543131
									3,335527	2,489715
									3,279436	2,333721
Mean	3,46906	2,5382335	3,447261833	2,544665333	2,883198	2,015897	3,573229667	2,6402	3,234701333	2,363757444
SD	0,03813144	0,041533331	0,230909069	0,196375419	0,161423006	0,131100049	0,014904684	0,016714949	0,351956061	0,291068465
CV%	1%	2%	7%	8%	6%	7%	0%	1%	11%	12%

	INTRA-DAY REPEATABILITY	
	Score up to 3	Score up to 4
Total Mean	10%	8%
Total SD	4%	3%

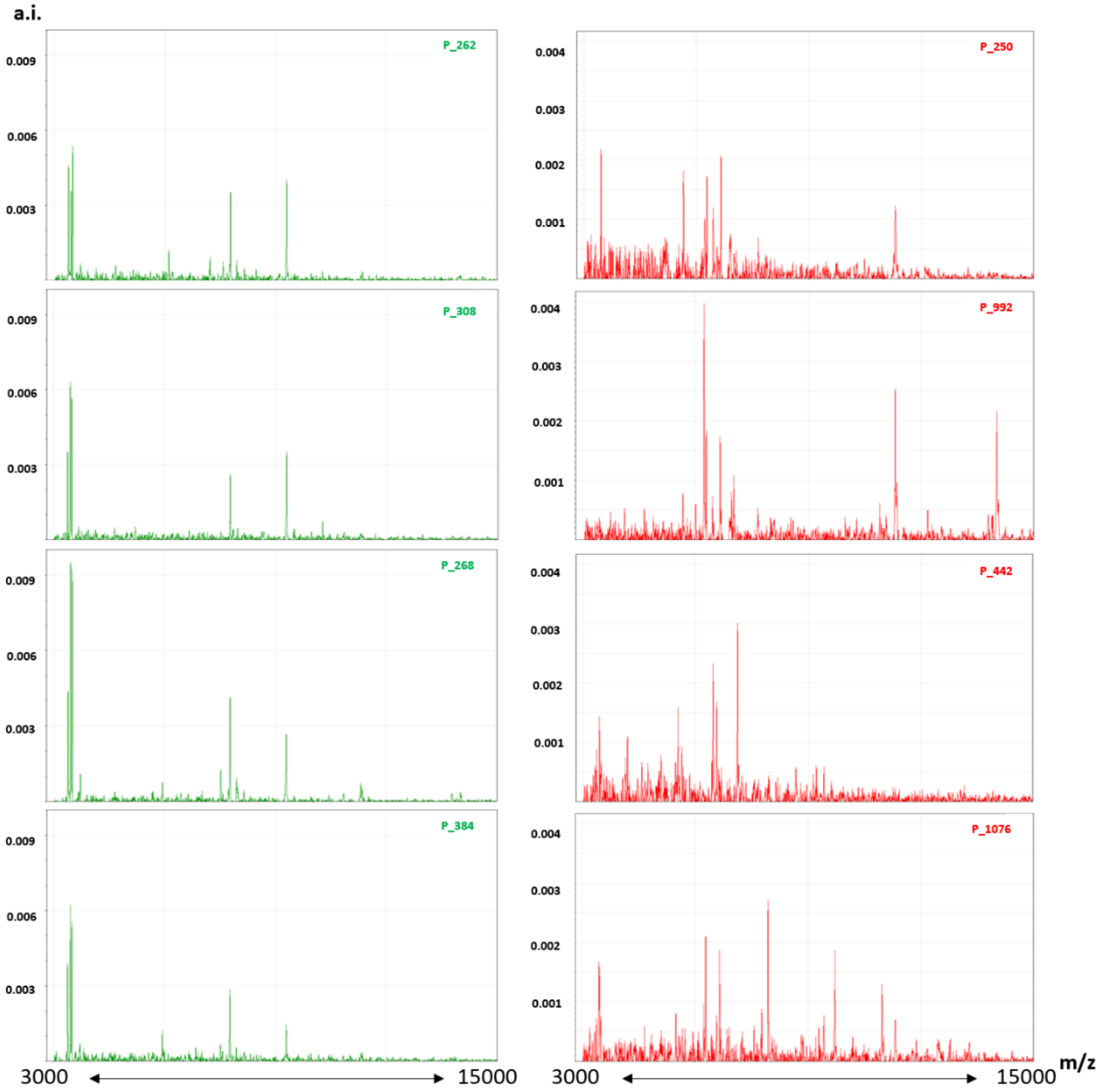
	INTER-DAY REPEATABILITY	
	Score up to 3	Score up to 4
Total Mean	7%	6%
Total SD	5%	4%

Supplementary Table 5. S_3 and S_4 scores of the intra-day and inter-day paired comparisons of $t0_{intra\text{-}day}$ and $t0_{inter\text{-}day}$. The mean, standard deviation (SD) and the coefficient of variation (CV) are calculated.

APPENDIX C

Supplementary materials of:

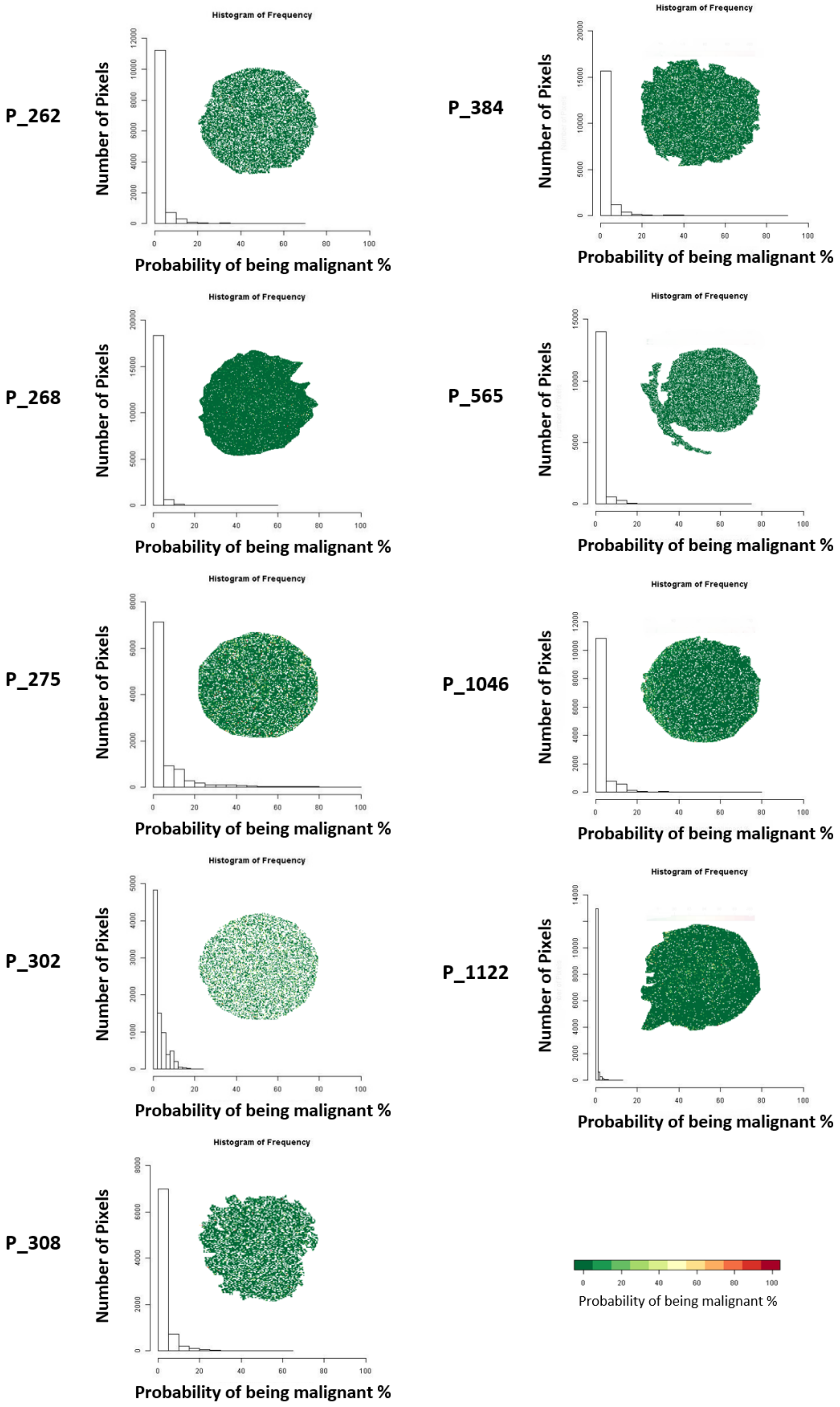
MALDI-MSI as a Complementary Diagnostic Tool in Cytopathology: A Pilot Study for the Characterization of Thyroid Nodules



Supplemetary figure 1. Example of single spectra from 4 benign (green) and 4 malignant (red) patients of the training set.

Pixel by Pixel images & Histograms of Frequency

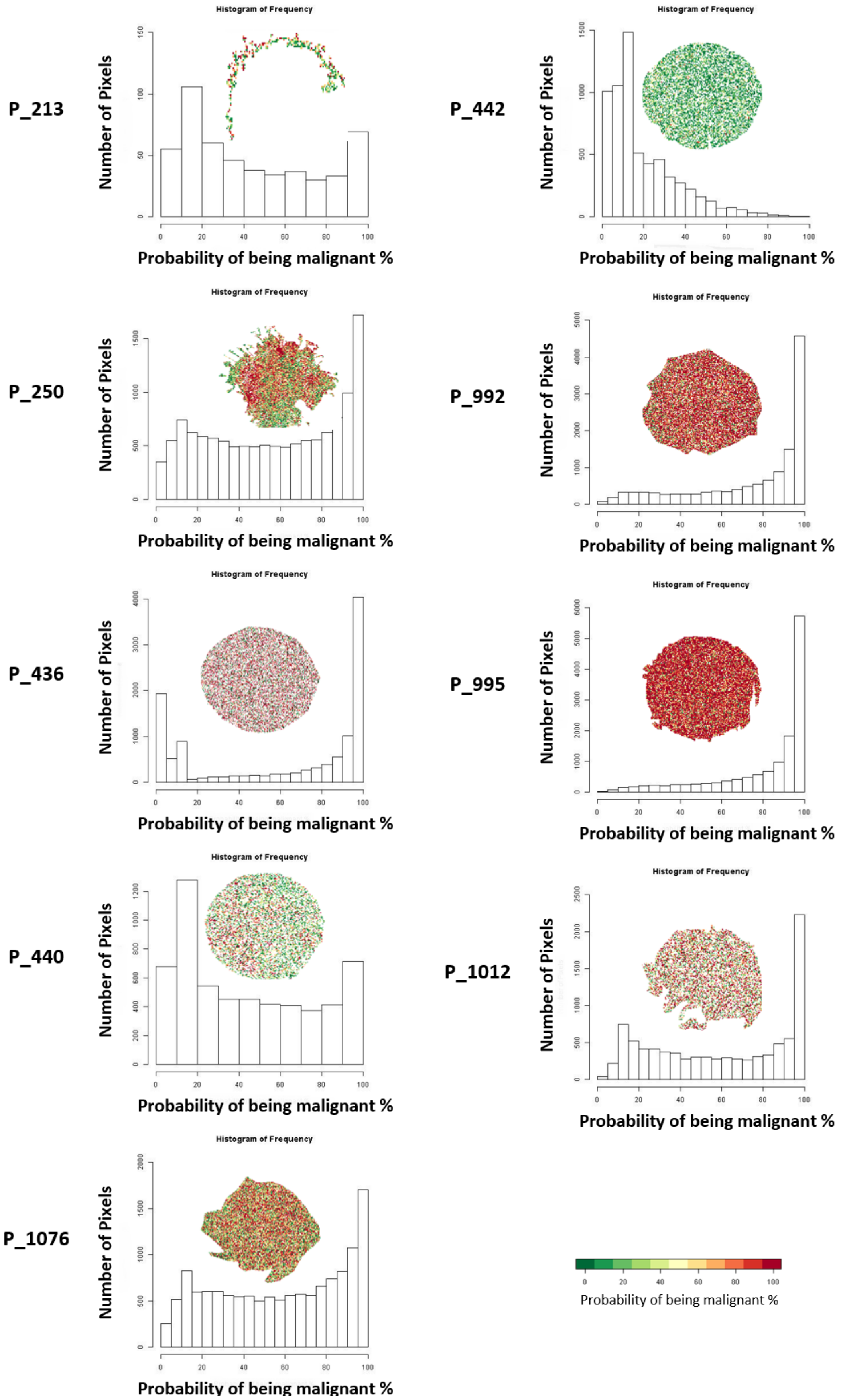
THY2 - TRAINING SAMPLE



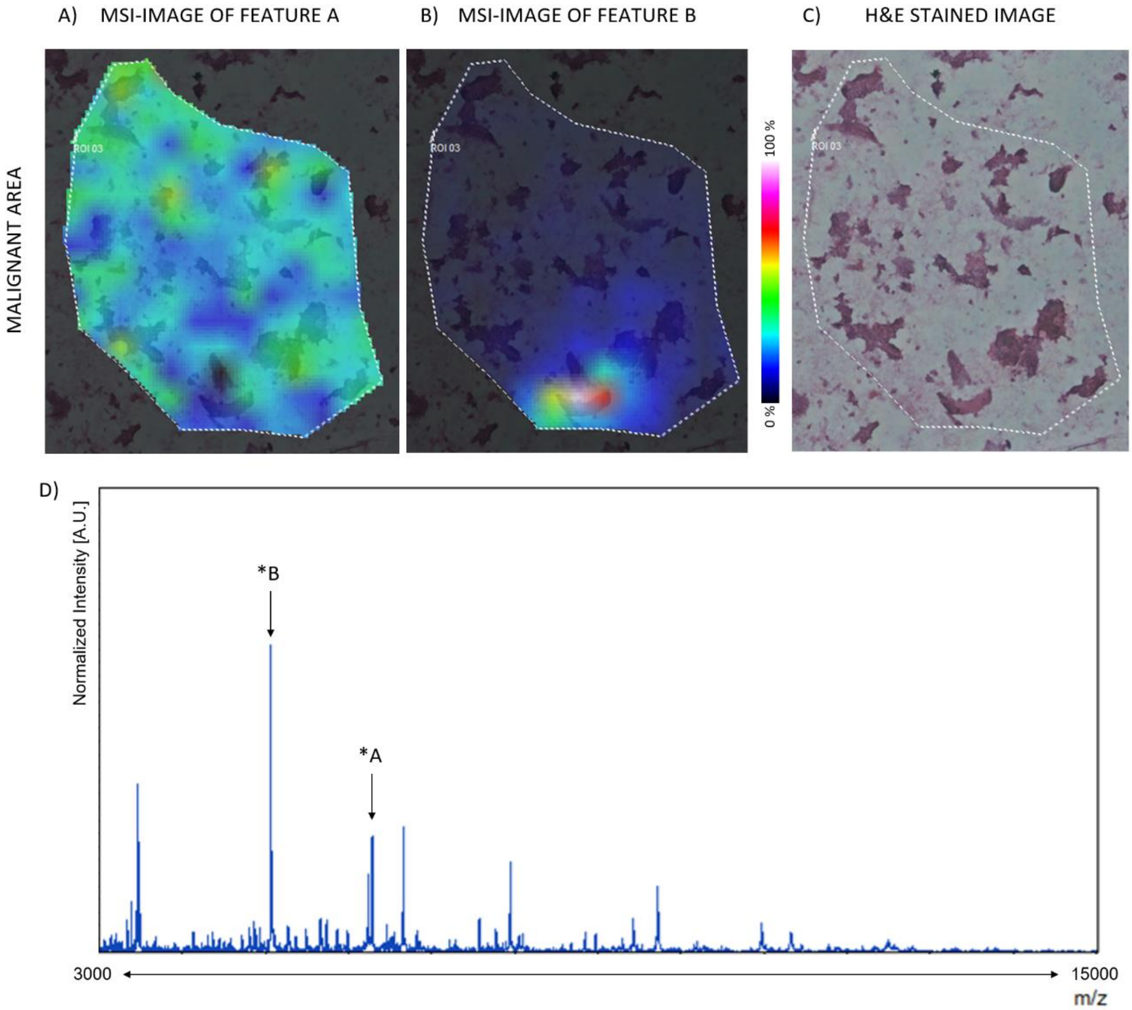
Supplementary Figure 2a. Pixel by pixel images and distribution of the probabilities of being malignant for each pixel in the training set of Thy2 nodules.

Pixel by Pixel images & Histograms of Frequency

THY5 - TRAINING SAMPLE



Supplementary Figure 2b. Pixel by pixel images and distribution of the probabilities of being malignant for each pixel in the training set of Thy5 nodules.



Supplementary Figure 3. MALDI-MSI molecular images of an area of an area of a malignant specimen and spatial localization of two m/z features in the (A) cancer cell clusters (feature A); and (B) stromal area (feature B); (C) Haematoxylin and eosin stained image and (D) Total ion count normalized average spectrum.

	Patient ID	Percentage of features per ROI per sample																			
		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
TIR 5	213	0%	100%	25%	0%	75%	0%	100%	100%	0%	100%	0%	0%	0%	0%	0%	100%	0%	100%	25%	100%
	250	0%	100%	0%	0%	100%	0%	100%	75%	50%	100%	100%	0%	50%	100%	25%	100%	0%	100%	25%	100%
	436	0%	100%	100%	0%	100%	50%	100%	0%	100%	100%	75%	50%	50%	0%	25%	100%	0%	100%	75%	100%
	440	0%	100%	25%	50%	100%	25%	100%	100%	100%	100%	0%	75%	50%	0%	0%	0%	25%	0%	0%	100%
	442	100%	0%	0%	0%	0%	25%	0%	0%	0%	0%	0%	0%	0%	100%	0%	0%	0%	0%	0%	25%
	992	100%	0%	0%	0%	100%	50%	25%	25%	75%	0%	75%	25%	75%	50%	0%	25%	0%	0%	50%	100%
	995	75%	25%	0%	25%	50%	25%	75%	100%	50%	25%	100%	0%	0%	0%	0%	50%	100%	100%	0%	100%
	1012	0%	0%	0%	0%	100%	0%	100%	50%	0%	100%	0%	100%	25%	0%	0%	100%	0%	0%	0%	100%
	1076	0%	100%	75%	0%	0%	0%	100%	100%	0%	100%	0%	0%	0%	0%	0%	75%	100%	0%	75%	50%
Overall																					
	mean	30,60%	58,30%	25,00%	8,30%	69,40%	19,40%	77,80%	61,10%	41,70%	69,40%	38,90%	27,80%	27,80%	27,80%	5,60%	61,10%	25,00%	44,40%	27,80%	86,10%
	sd	0,46	0,82	1	1,77	2,36	2,76	3,38	4	4,43	4,81	5,35	5,91	6,5	6,68	7,07	8	8,5	8,97	9,5	9,94

	Patient ID	Percentage of features per ROI per sample																			
		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
TIR 2	262	0%	100%	100%	20%	20%	100%	60%	0%	20%	60%	100%	0%	0%	100%	80%	0%	0%	0%	0%	40%
	268	0%	100%	100%	40%	0%	60%	0%	0%	0%	60%	0%	0%	0%	100%	80%	0%	0%	0%	0%	0%
	275	0%	0%	100%	100%	80%	0%	20%	0%	20%	0%	0%	60%	0%	0%	100%	0%	0%	20%	0%	0%
	302	0%	100%	100%	0%	0%	100%	0%	0%	0%	100%	20%	0%	0%	100%	60%	0%	0%	0%	0%	100%
	308	0%	100%	100%	100%	100%	20%	60%	0%	20%	80%	20%	60%	20%	60%	20%	20%	0%	0%	0%	60%
	384	0%	100%	100%	60%	40%	80%	0%	60%	0%	20%	20%	0%	0%	0%	80%	0%	20%	0%	0%	0%
	565	0%	100%	100%	20%	20%	80%	60%	0%	0%	60%	80%	0%	0%	100%	20%	0%	0%	0%	0%	100%
	1046	0%	100%	100%	40%	40%	100%	0%	0%	20%	40%	100%	0%	40%	100%	40%	20%	0%	0%	0%	20%
	1122	0%	100%	0%	100%	0%	60%	0%	0%	40%	0%	0%	0%	0%	100%	0%	0%	0%	0%	0%	0%
Overall																					
	mean	0,00%	88,90%	88,90%	53,30%	33,30%	66,70%	22,20%	6,70%	13,30%	46,70%	37,80%	13,30%	6,70%	73,30%	53,30%	4,40%	2,20%	2,20%	0,00%	35,60%
	sd	0	0,33	0,33	0,39	0,36	0,36	0,29	0,2	0,14	0,35	0,43	0,26	0,14	0,44	0,35	0,09	0,07	0,07	0	0,42

Supplementary Table 1. Percentage of features per ROI in each Thy2 and Thy5 sample of the training set.

A)

Cytological Diagnosis	Patient	OBJECTS	Probability	Histological Diagnosis
THY2	P_1081 in-vivo	Average Spectrum	0,03%	Hyperplasia
		ROI 01	0,96%	
		ROI 02	0,12%	
		ROI 03	0,21%	
		ROI 04	0,71%	
		ROI 05	0,03%	
		ROI 06	0,01%	
		ROI 07	0,14%	
		ROI 08	0,38%	
		ROI 09	1,60%	
		ROI 10	1,74%	
		ROI 11	0,01%	

Cytological Diagnosis	Patient	OBJECTS	Probability	Histological Diagnosis
THY2	P_1083 in-vivo	Average Spectrum	0,00%	Hyperplasia
		ROI 01	0,00%	
		ROI 02	0,87%	
		ROI 03	0,00%	
		ROI 04	0,00%	
		ROI 05	0,00%	
		ROI 06	0,00%	
		ROI 07	0,00%	
		ROI 08	0,02%	
		ROI 09	0,06%	
		ROI 10	0,00%	
		ROI 11	0,00%	
		ROI 12	0,00%	
		ROI 13	0,00%	
		ROI 14	0,00%	
		ROI 15	0,01%	
		ROI 16	0,00%	
		ROI 17	0,00%	
		ROI 18	0,00%	
		ROI 19	0,03%	
		ROI 20	0,01%	
		ROI 21	1,12%	
		ROI 22	0,04%	
		ROI 23	0,03%	
		ROI 24	0,00%	
		ROI 25	0,04%	

Cytological Diagnosis	Patient	OBJECTS	Probability	Histological Diagnosis
THY2	P_1123 in-vivo	Average Spectrum	0,00%	Hyperplasia
		ROI 01	0,01%	
		ROI 02	2,01%	
		ROI 03	5,61%	
		ROI 04	36,48%	
		ROI 05	1,29%	
		ROI 06	17,73%	
		ROI 07	3,42%	
		ROI 08	0,09%	
		ROI 09	1,68%	
		ROI 10	0,22%	
		ROI 11	0,27%	
		ROI 12	0,24%	
		ROI 13	0,02%	
		ROI 14	0,01%	
		ROI 15	0,05%	
		ROI 16	0,00%	

Cytological Diagnosis	Patient	OBJECTS	Probability	Histological Diagnosis
THY2	P_1156 in-vivo	Average Spectrum	0,00%	Hyperplasia
		ROI 01	0,02%	
		ROI 02	0,02%	
		ROI 03	0,07%	
		ROI 04	0,00%	
		ROI 05	0,57%	
		ROI 06	0,01%	
		ROI 07	0,01%	
		ROI 08	0,16%	
		ROI 09	0,00%	
		ROI 10	0,68%	
		ROI 11	0,27%	
		ROI 12	0,03%	
		ROI 13	0,10%	
		ROI 14	3,94%	
		ROI 15	8,20%	
		ROI 16	0,58%	
		ROI 17	0,25%	
		ROI 18	0,02%	
		ROI 19	0,01%	
		ROI 20	0,06%	
		ROI 21	0,00%	
		ROI 22	3,13%	

B)

Cytological Diagnosis	Patient	OBJECTS	Probability	Histological Diagnosis
THY5	P_1149 in-vivo	Average Spectrum	16,42%	PTC
		ROI 01	85,17%	
		ROI 02	98,32%	
		ROI 03	89,47%	
		ROI 04	20,91%	
		ROI 05	1,19%	
		ROI 06	10,03%	
		ROI 07	0,15%	
		ROI 08	0,09%	
		ROI 09	0,94%	
		ROI 10	72,44%	
		ROI 11	76,47%	
		ROI 12	27,48%	

Cytological Diagnosis	Patient	OBJECTS	Probability	Histological Diagnosis
THY5	P_1084 in-vivo	Average Spectrum	0,00%	PTC
		ROI 01	0,00%	
		ROI 02	0,00%	
		ROI 03	0,00%	
		ROI 04	0,00%	
		ROI 05	0,00%	
		ROI 06	0,01%	
		ROI 07	0,00%	
		ROI 08	0,03%	
		ROI 09	0,01%	
		ROI 10	0,02%	
		ROI 11	0,01%	
		ROI 12	0,04%	
		ROI 13	0,00%	
		ROI 14	0,01%	
		ROI 15	0,00%	

Cytological Diagnosis	Patient	OBJECTS	Probability	Histological Diagnosis
THY5	P_1126 in-vivo	Average Spectrum	0,00%	PTC
		ROI 01	0,23%	
		ROI 02	0,66%	
		ROI 03	27,93%	
		ROI 04	1,93%	
		ROI 05	0,08%	
		ROI 06	1,42%	
		ROI 07	1,33%	
		ROI 08	24,61%	
		ROI 09	31,11%	
		ROI 10	12,63%	
		ROI 11	2,27%	
		ROI 12	12,51%	
		ROI 13	31,16%	
		ROI 14	11,63%	
		ROI 15	1,50%	

Cytological Diagnosis	Patient	OBJECTS	Probability	Histological Diagnosis
THY5	P_1187 in-vivo	Average Spectrum	0,00%	PTC
		ROI 01	0,00%	
		ROI 02	0,00%	
		ROI 03	0,00%	
		ROI 04	0,00%	
		ROI 05	0,00%	
		ROI 06	0,00%	
		ROI 07	0,00%	
		ROI 08	0,02%	
		ROI 09	0,00%	
		ROI 10	0,00%	

B1)

Cytological Diagnosis	Patient	OBJECTS	Probability	Histological Diagnosis
THY5	P_1084 ex-vivo	Average Spectrum	99,41%	PTC
		ROI 01	89,26%	
		ROI 02	96,84%	
		ROI 03	93,18%	

Cytological Diagnosis	Patient	OBJECTS	Probability	Histological Diagnosis
THY5	P_1126 ex-vivo	Average Spectrum	100,00%	PTC
		ROI 01	100,00%	
		ROI 02	99,93%	
		ROI 03	100,00%	
		ROI 04	99,98%	
		ROI 05	100,00%	
		ROI 06	100,00%	
		ROI 07	100,00%	
		ROI 08	100,00%	
		ROI 09	100,00%	
		ROI 10	100,00%	
		ROI 11	100,00%	

Cytological Diagnosis	Patient	OBJECTS	Probability	Histological Diagnosis
THY5	P_1187 ex-vivo	Average Spectrum	99,98%	PTC
		ROI 01	100,00%	
		ROI 02	100,00%	
		ROI 03	74,22%	
		ROI 04	100,00%	
		ROI 05	100,00%	
		ROI 06	64,23%	
		ROI 07	98,93%	
		ROI 08	100,00%	

C)

Cytological Diagnosis	Patient	OBJECTS	Probability	Histological Diagnosis
THY3	P_1082	Average Spectrum	1,21%	Hyperplasia
		ROI 01	1,81%	
		ROI 02	0,62%	

Cytological Diagnosis	Patient	OBJECTS	Probability	Histological Diagnosis
THY4	P_1202	Average Spectrum	99,67%	PTC
		ROI 01	99,99%	
		ROI 02	98,67%	
		ROI 03	99,83%	
		ROI 04	93,21%	
		ROI 05	99,96%	
		ROI 06	99,99%	
		ROI 07	99,99%	

C1)

Cytological Diagnosis	Patient	OBJECTS	Probability	Histological Diagnosis
THY5 LYMPH	P_1188 in-vivo	Average Spectrum	83,11%	PTC
		ROI 01	48,75%	
		ROI 02	58,62%	
		ROI 03	41,86%	
		ROI 04	95,83%	
		ROI 05	51,06%	
		ROI 06	16,70%	
		ROI 07	54,56%	
		ROI 08	70,71%	
		ROI 09	56,18%	
		ROI 10	70,97%	

Cytological Diagnosis	Patient	OBJECTS	Probability	Histological Diagnosis
THY5 LYMPH	P_1188 ex-vivo	Average Spectrum	99,84%	PTC
		ROI 01	90,15%	
		ROI 02	58,92%	
		ROI 03	2,92%	
		ROI 04	20,86%	
		ROI 05	97,56%	
		ROI 06	86,03%	
		ROI 07	98,27%	
		ROI 08	99,44%	
		ROI 09	97,02%	
		ROI 10	53,85%	
		ROI 11	99,67%	
		ROI 12	83,40%	
		ROI 13	100,00%	
		ROI 14	67,52%	
		ROI 15	96,53%	
		ROI 16	100,00%	
		ROI 17	96,56%	
		ROI 18	34,18%	
		ROI 19	59,42%	

Supplementary Table 2. Probability of being malignant for each patient of the validation set (ROIs and average spectrum). A) Thy2 patients, B) Thy5 patients, B1) ex-vivo samples of Thy5 patients, C) Thy3 and Thy4 patient, C1) lymphnode sample (P_1188), in-vivo and ex-vivo.

A) THY2 Training

Study lesion code	Minimum	1st Quartile	Median	Mean	3rd Qurtile	Maximum
P_262	0,00	0,00	0,00	2,02	2,00	69,00
P_268	0,00	0,00	0,00	1,12	1,00	56,00
P_302	0,00	1,00	2,00	3,11	5,00	24,00
P_308	0,00	1,00	2,00	3,03	4,00	64,00
P_384	0,00	0,00	0,00	2,36	2,00	88,00
P_475	0,00	1,00	2,00	6,74	7,00	97,00
P_565	0,00	0,00	1,00	1,54	2,00	72,00
P_1046	0,00	0,00	0,00	2,42	2,00	77,00
P_1122	0,00	0,00	0,00	0,46	1,00	13,00
Overall						
mean	0,00	0,33	0,78	2,53	2,89	62,22
sd	0,00	0,50	0,97	1,80	2,03	27,75

B) THY5 Training

Study lesion code	Minimum	1st Quartile	Median	Mean	3rd Qurtile	Maximum
P_213	1,00	16,00	36,00	44,91	72,25	100,00
P_250	0,00	28,00	59,00	57,09	88,00	100,00
P_436	0,00	12,00	84,00	61,98	98,00	100,00
P_440	0,00	13,00	39,00	44,40	72,00	100,00
P_442	0,00	8,00	14,00	20,02	28,00	98,00
P_992	1,00	58,00	89,00	75,42	98,00	100,00
P_995	2,00	73,00	93,00	81,69	98,00	100,00
P_1012	1,00	29,00	64,00	61,11	95,00	100,00
P_1076	0,00	30,00	61,00	57,89	87,00	100,00
Overall						
mean	0,56	29,67	59,89	56,06	81,81	99,78
sd	0,73	22,16	26,59	18,22	22,66	0,67

Supplementary Table 3. Distribution of the probabilities to be malignant in the pixel by pixel analysis for Thy2 (a) and Thy5 (b) training set.

C) Validation

Study lesion code	Cytologic Diagnosis	Minimum	1st Quartile	Median	Mean	3rd Quartile	Maximum
P_1081	THY2	0,00	0,00	0,00	0,98	1,00	15,00
P_1082	THY3	0,00	1,00	2,00	2,93	4,00	57,00
P_1083	THY2	0,00	0,00	0,00	1,84	1,00	85,00
P_1084 in vivo	THY5	0,00	0,00	0,00	0,06	0,00	17,00
P_1084 ex vivo	THY5	1,00	28,00	63,00	58,88	90,00	100,00
P_1123	THY2	0,00	0,00	2,00	3,63	5,00	59,00
P_1126 in vivo	THY5	0,00	1,00	4,00	8,88	11,00	98,00
P_1126 ex vivo	THY5	3,00	81,00	97,00	85,98	100,00	100,00
P_1149	THY5	0,00	7,00	20,00	31,95	53,00	100,00
P_1156	THY2	0,00	0,00	1,00	3,02	3,00	93,00
P_1187 in vivo	THY5	0,00	0,00	0,00	0,40	0,00	66,00
P_1187 ex vivo	THY5	0,00	30,00	57,00	57,51	86,00	100,00
P_1188 in vivo	THY5	0,00	0,00	0,00	1,39	2,00	15,00
P_1188 ex vivo	THY5	1,00	13,00	27,00	37,00	56,00	100,00
P_1202	THY4	0,00	15,00	40,00	45,11	75,00	100,00

Supplementary Table 3c. Distribution of the probabilities to be malignant in the pixel by pixel analysis for the validation set.

APPENDIX D

First level analysis

15 generated signals

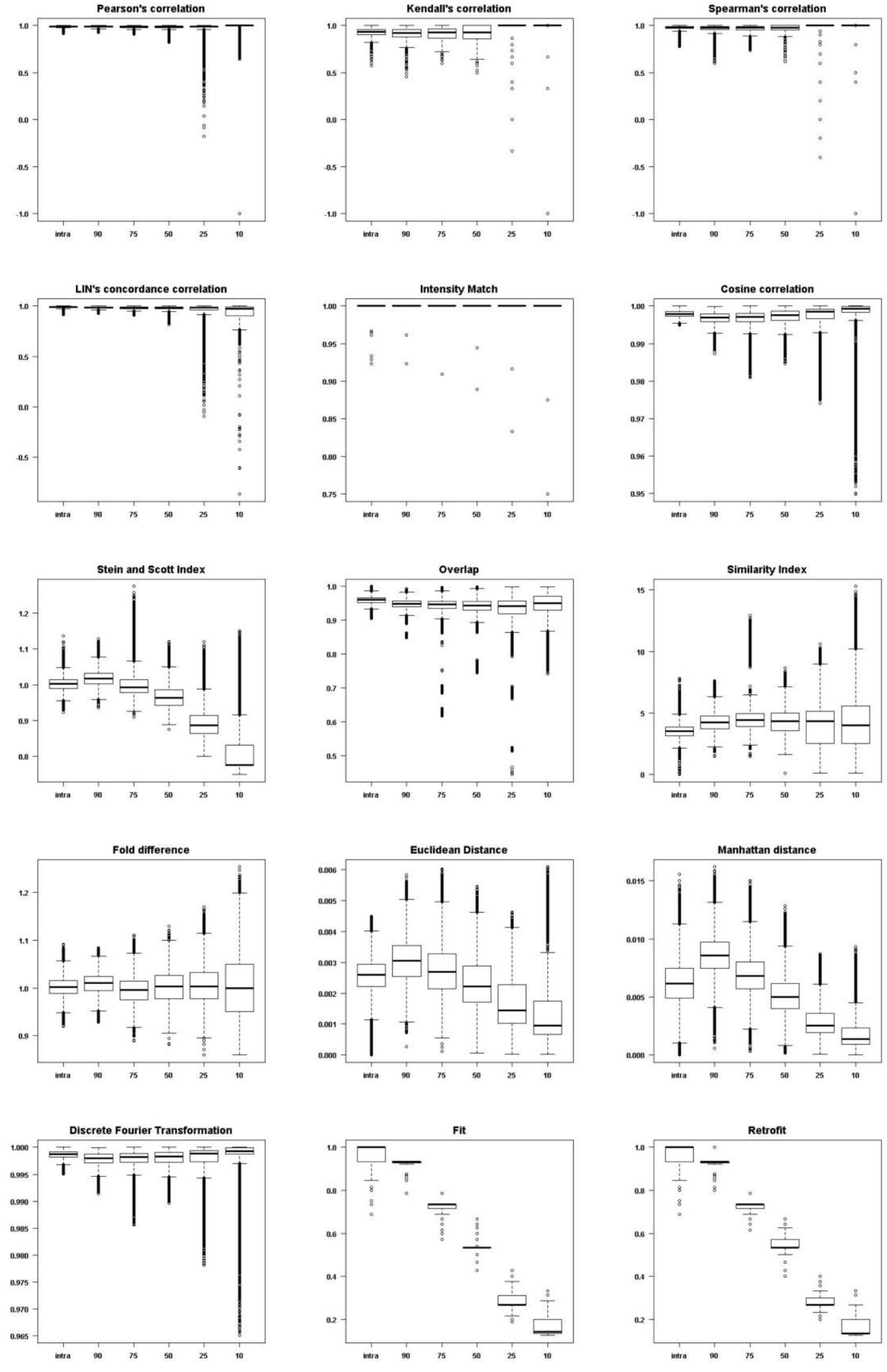


Figure 1: Boxplots of the different scores evalueted in pairwise comparison between spectra generated with 15 informative peaks and a 5% intensity variability. On the x-axis the first reported measure is referred to the intra-sample analysis, while the others are the percentage of common peaks used in the inter-sample analysis.

First level analysis

25 generated signals

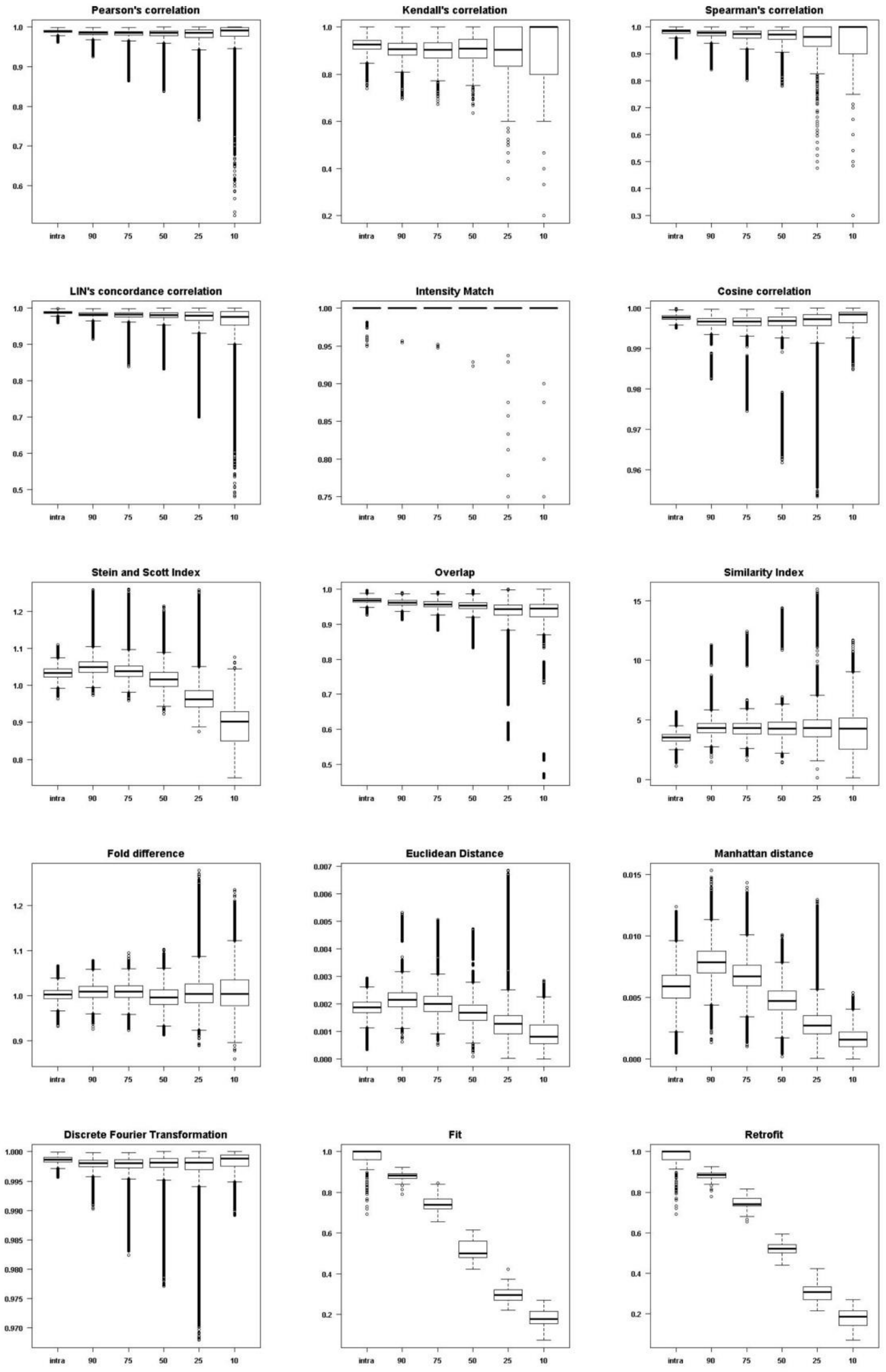


Figure 2: Boxplots of the different scores evalueted in pairwise comparison between spectra generated with 25 informative peaks and a 5% intensity variability. On the x-axis the first reported measure is referred to the intra-sample analysis, while the others are the percentage of common peaks used in the inter-sample analysis.

First level analysis

40 generated signals

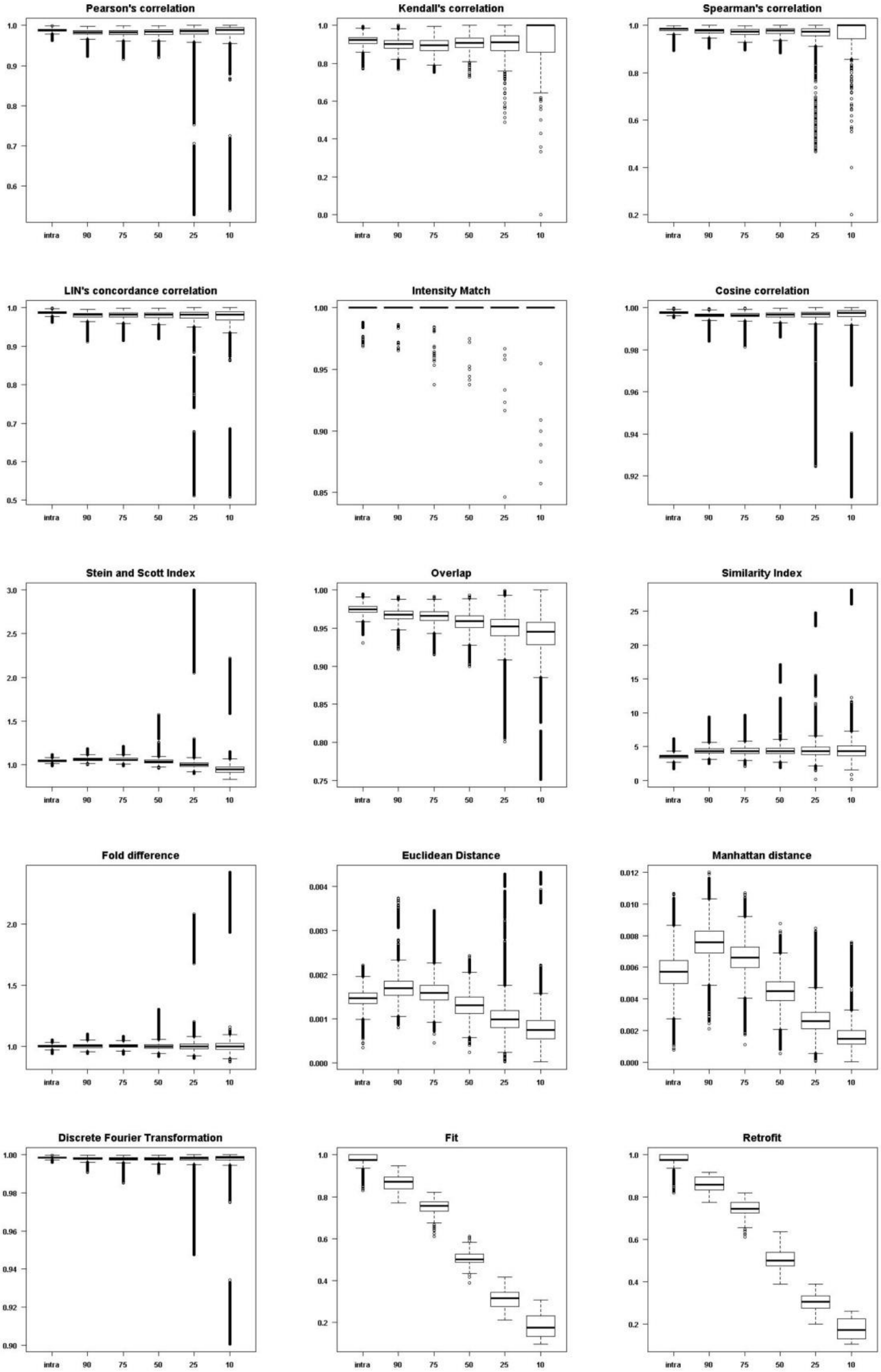
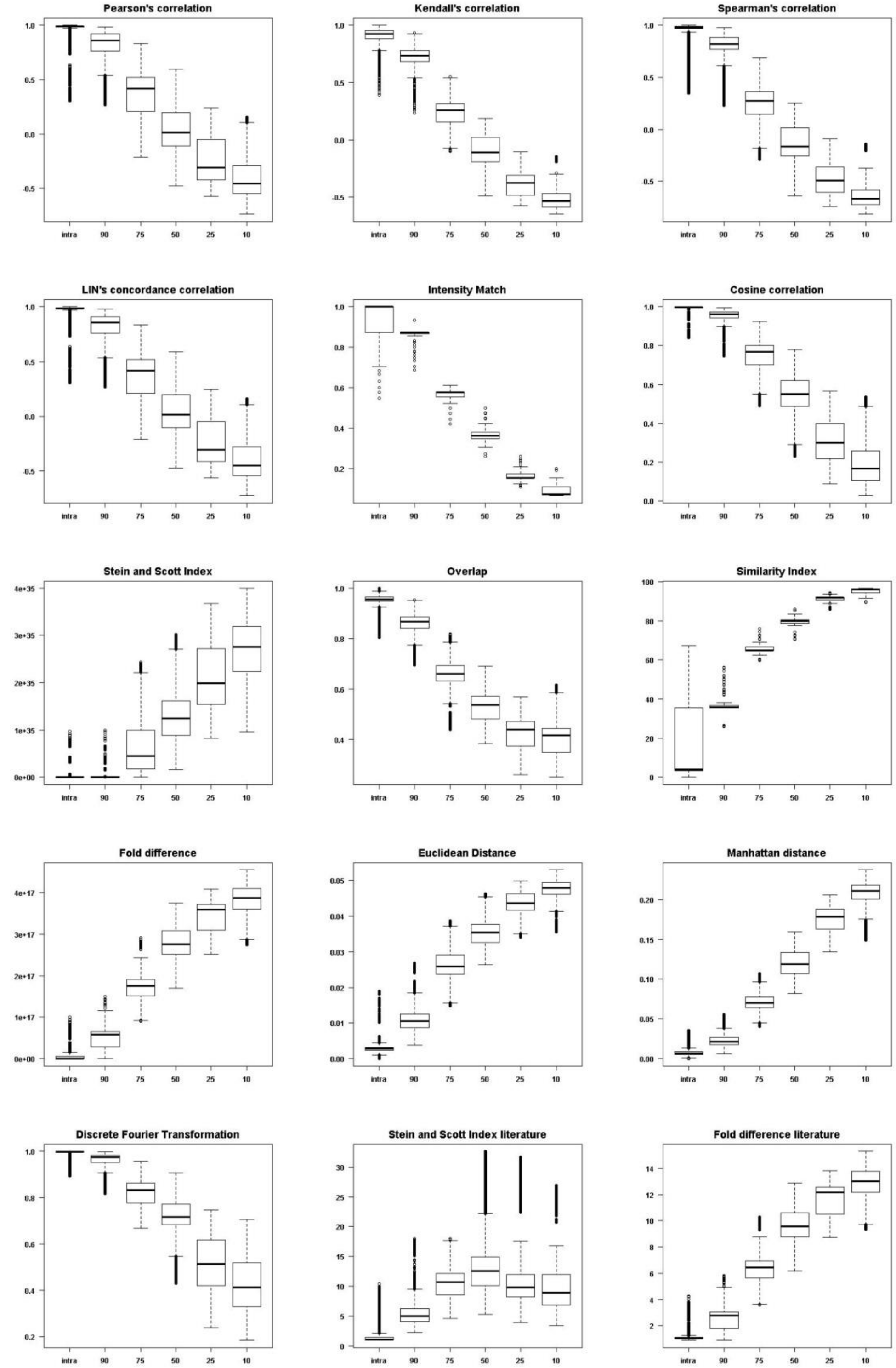


Figure 3: Boxplots of the different scores evalueted in pairwise comparison between spectra generated with 40 informative peaks and a 5% intensity variability. On the x-axis the first reported measure is referred to the intra-sample analysis, while the others are the percentage of common peaks used in the inter-sample analysis.

Second level analysis

15 generated signals


Pearson's correlation

Kendall's correlation

Spearman's correlation

LIN's concordance correlation

Intensity Match

Cosine correlation

Stein and Scott Index

Overlap

Similarity Index

Fold difference

Euclidean Distance

Manhattan distance

Discrete Fourier Transformation

Stein and Scott Index literature

Fold difference literature

Figure 4: Boxplots of the different scores evalueted in pairwise comparison between spectra generated with 15 informative peaks and a 5% intensity variability. On the x-axis the first reported measure is referred to the intra-sample analysis, while the others are the percentage of common peaks used in the inter-sample analysis.

Second level analysis

25 generated signals

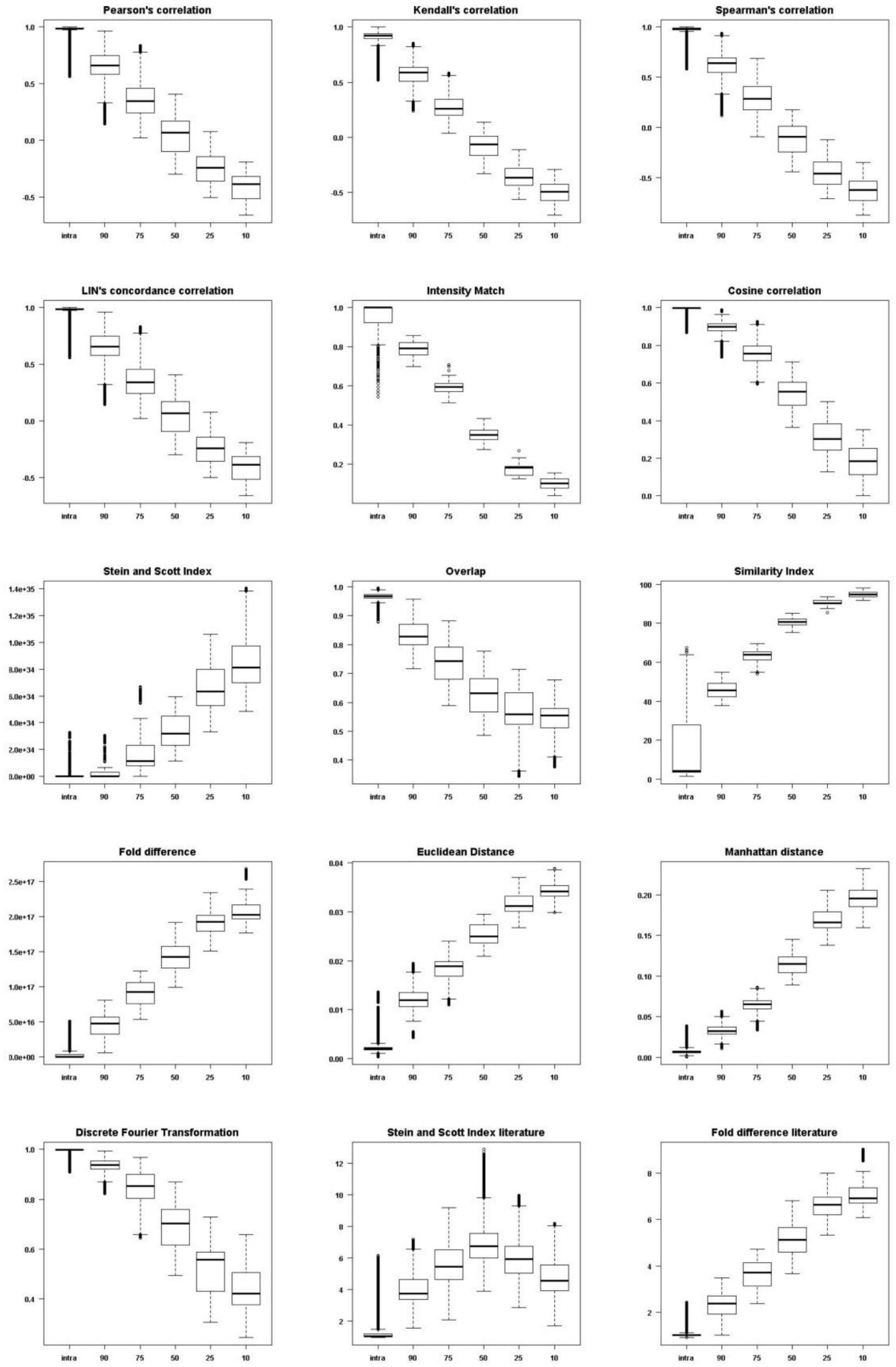


Figure 5: Boxplots of the different scores evalueted in pairwise comparison between spectra generated with 25 informative peaks and a 5% intensity variability. On the x-axis the first reported measure is referred to the intra-sample analysis, while the others are the percentage of common peaks used in the inter-sample analysis.

Second level analysis

40 generated signals

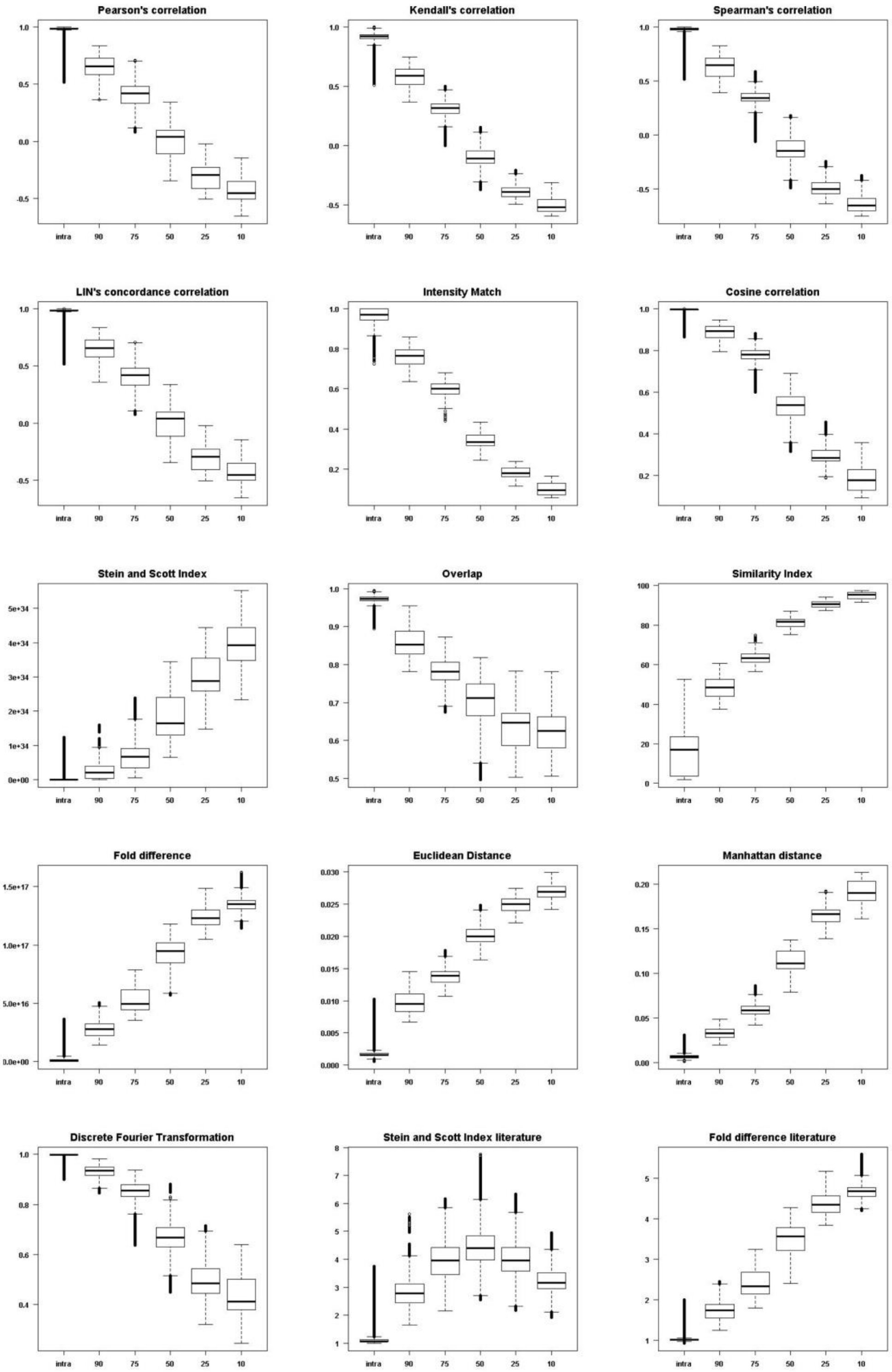


Figure 6: Boxplots of the different scores evalueted in pairwise comparison between spectra generated with 40 informative peaks and a 5% intensity variability. On the x-axis the first reported measure is referred to the intra-sample analysis, while the others are the percentage of common peaks used in the inter-sample analysis.

Third level analysis

15 generated signals

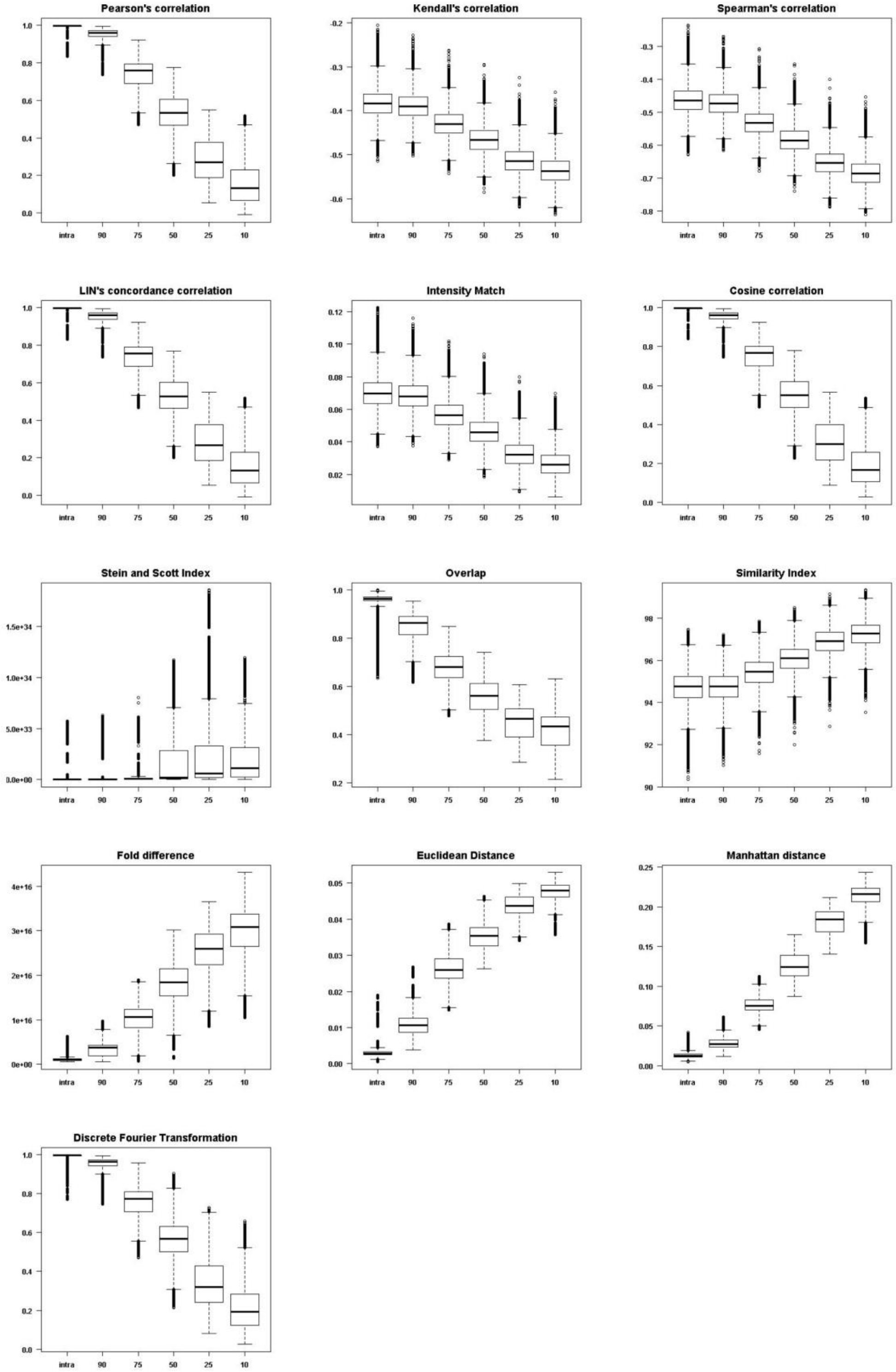


Figure 7: Boxplots of the different scores evalueted in pairwise comparison between spectra generated with 15 informative peaks and a 5% intensity variability. On the x-axis the first reported measure is referred to the intra-sample analysis, while the others are the percentage of common peaks used in the inter-sample analysis.

Third level analysis

25 generated signals

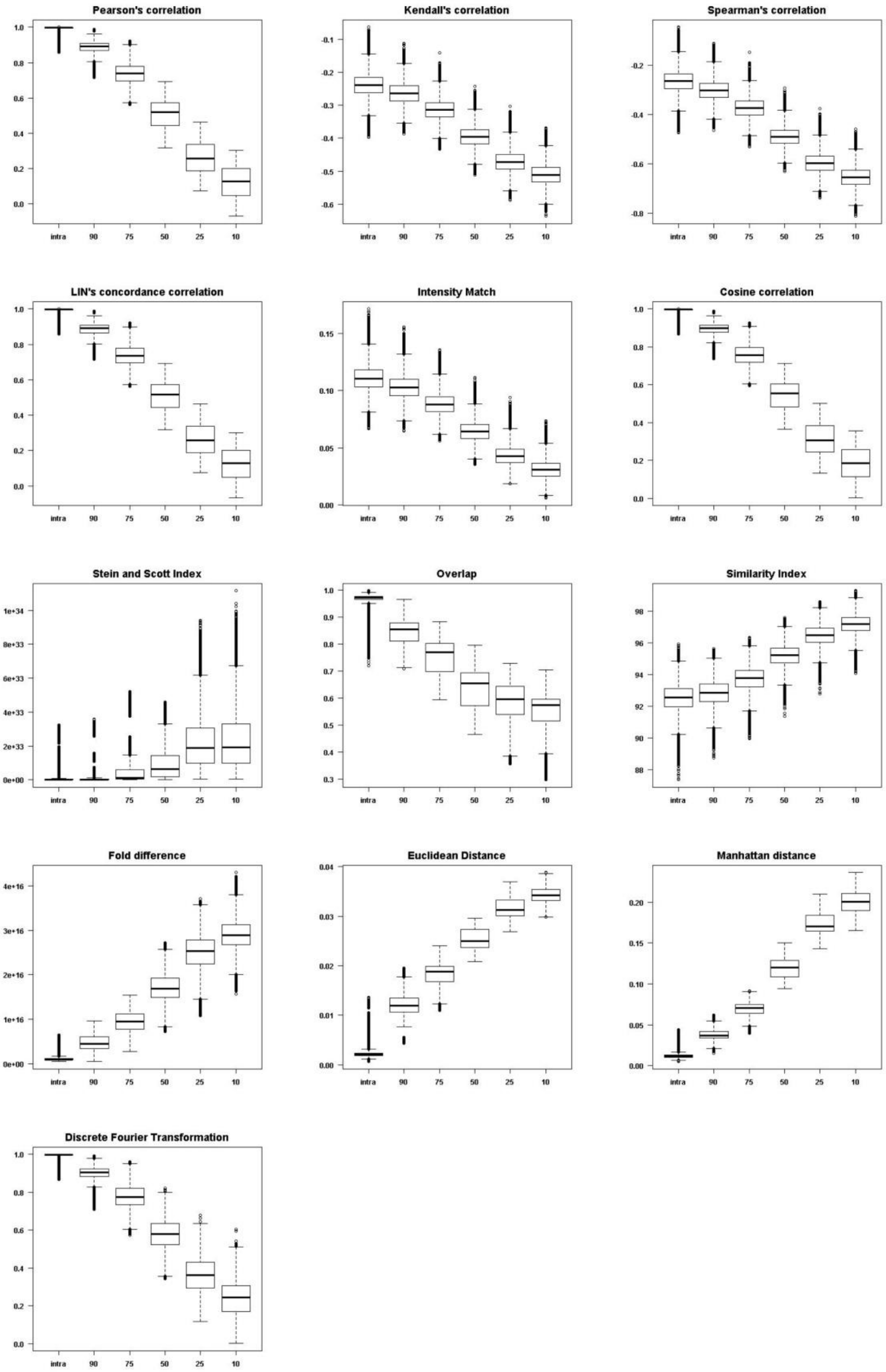


Figure 8: Boxplots of the different scores evalueted in pairwise comparison between spectra generated with 25 informative peaks and a 5% intensity variability. On the x-axis the first reported measure is referred to the intra-sample analysis, while the others are the percentage of common peaks used in the inter-sample analysis.

Third level analysis

40 generated signals

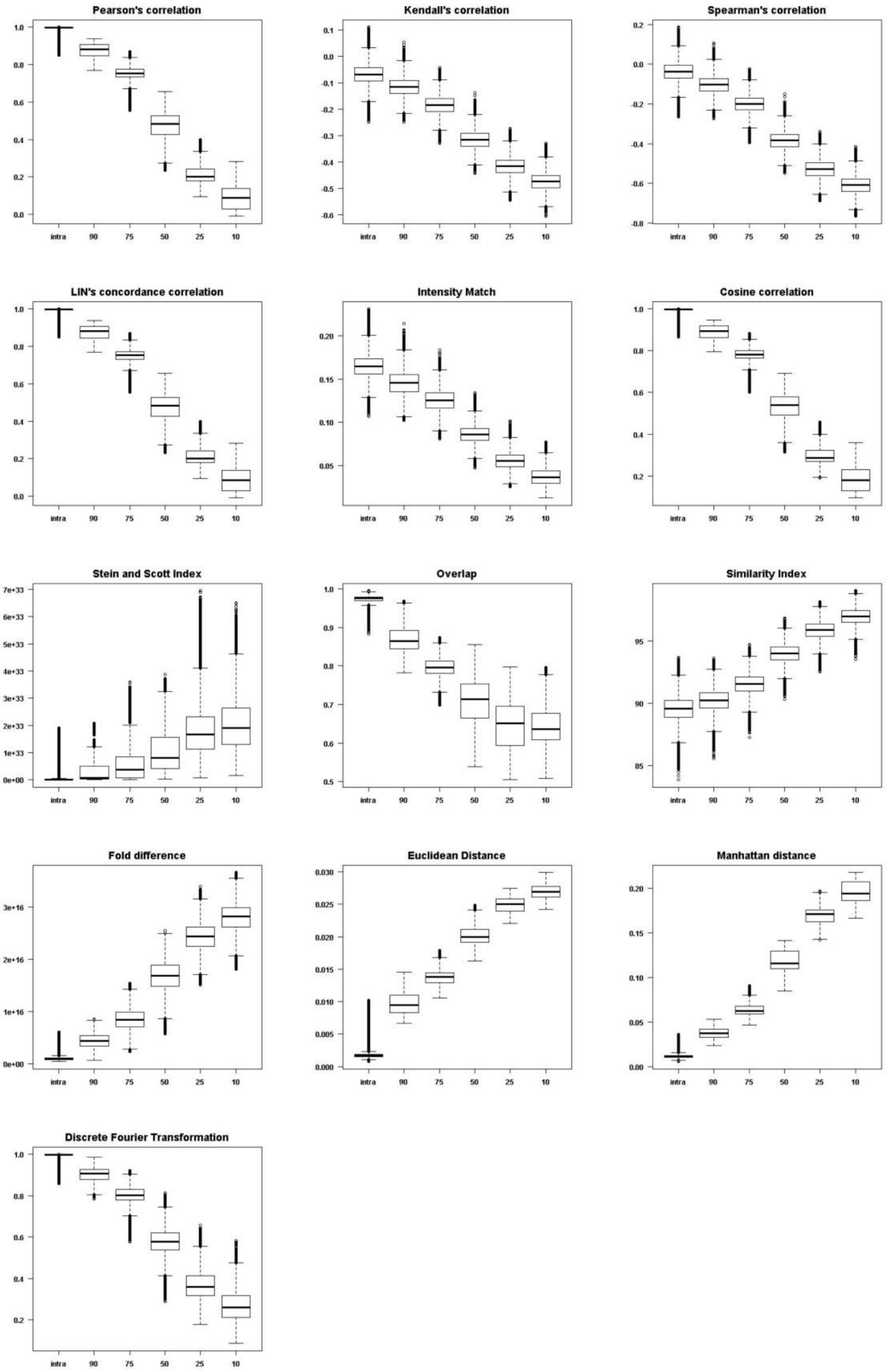


Figure 9: Boxplots of the different scores evalueted in pairwise comparison between spectra generated with 40 informative peaks and a 5% intensity variability. On the x-axis the first reported measure is referred to the intra-sample analysis, while the others are the percentage of common peaks used in the inter-sample analysis.

First level analysis

15 generated signals

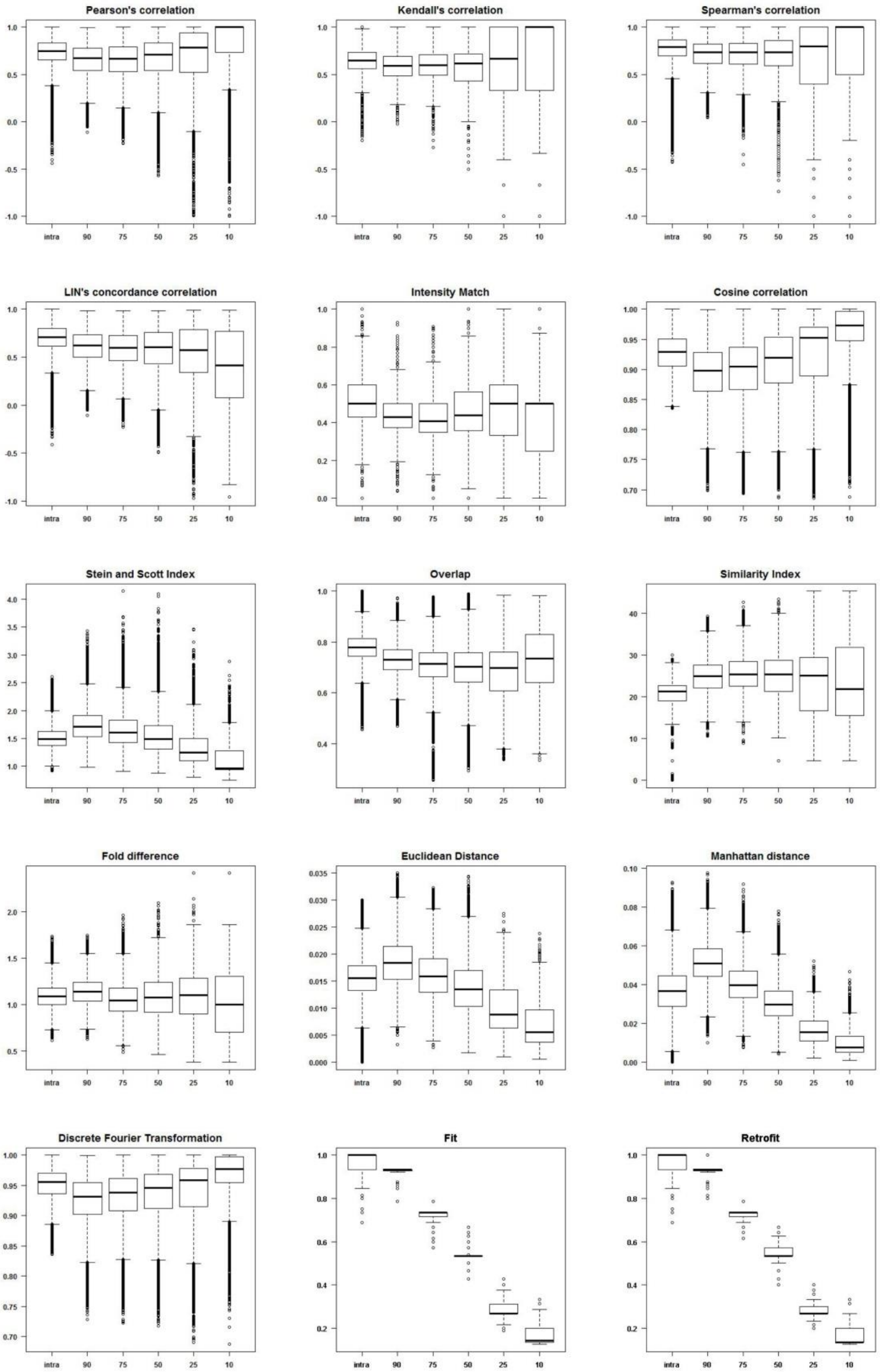


Figure 10: Boxplots of the different scores evalueted in pairwise comparison between spectra generated with 15 informative peaks and 30% intensity variability. On the x-axis the first reported measure is referred to the intra-sample analysis, while the others are the percentage of common peaks used in the inter-sample analysis.

First level analysis

25 generated signals

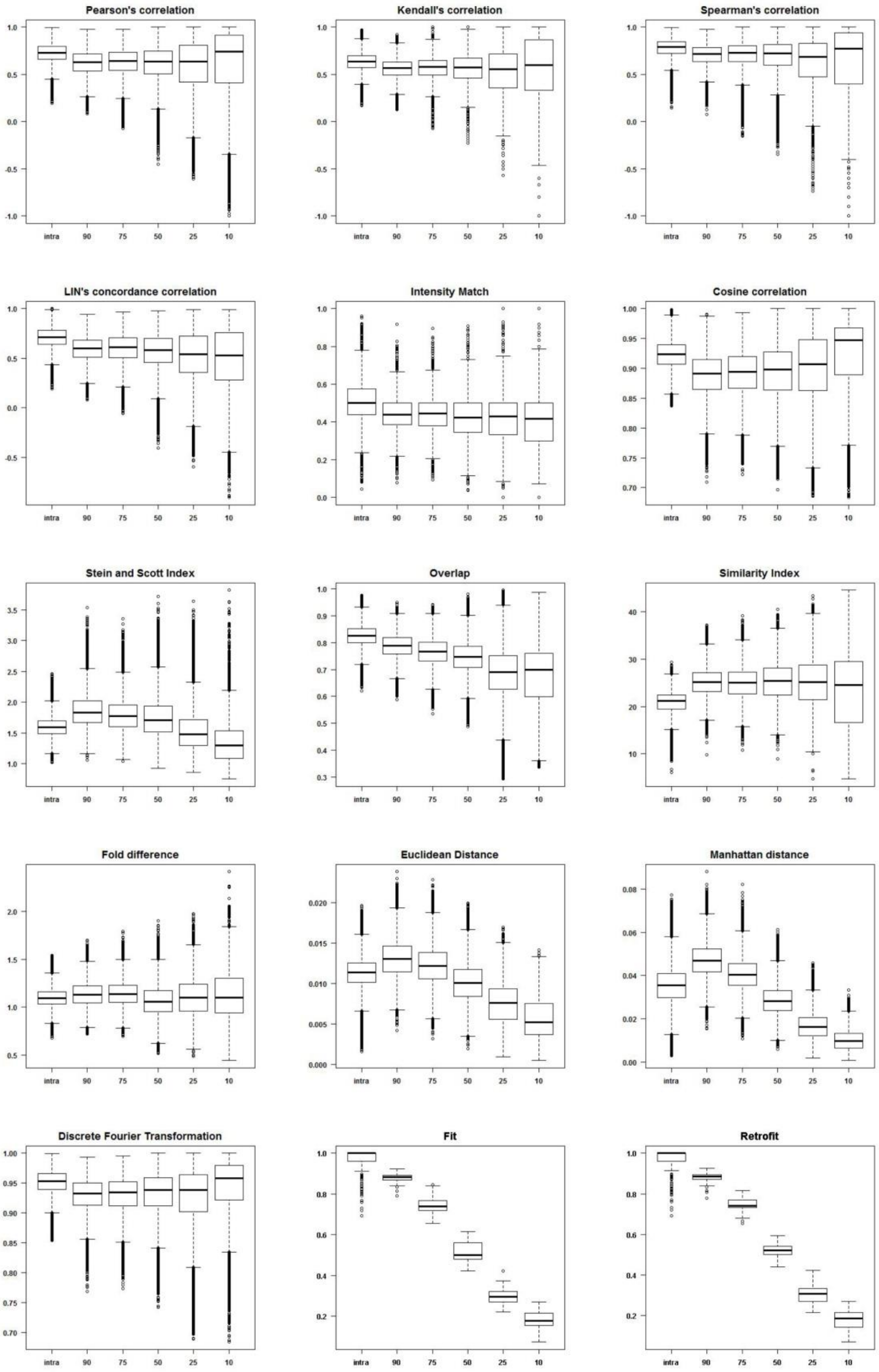


Figure 11: Boxplots of the different scores evalueted in pairwise comparison between spectra generated with 25 informative peaks and 30% intensity variability. On the x-axis the first reported measure is referred to the intra-sample analysis, while the others are the percentage of common peaks used in the inter-sample analysis.

First level analysis

40 generated signals

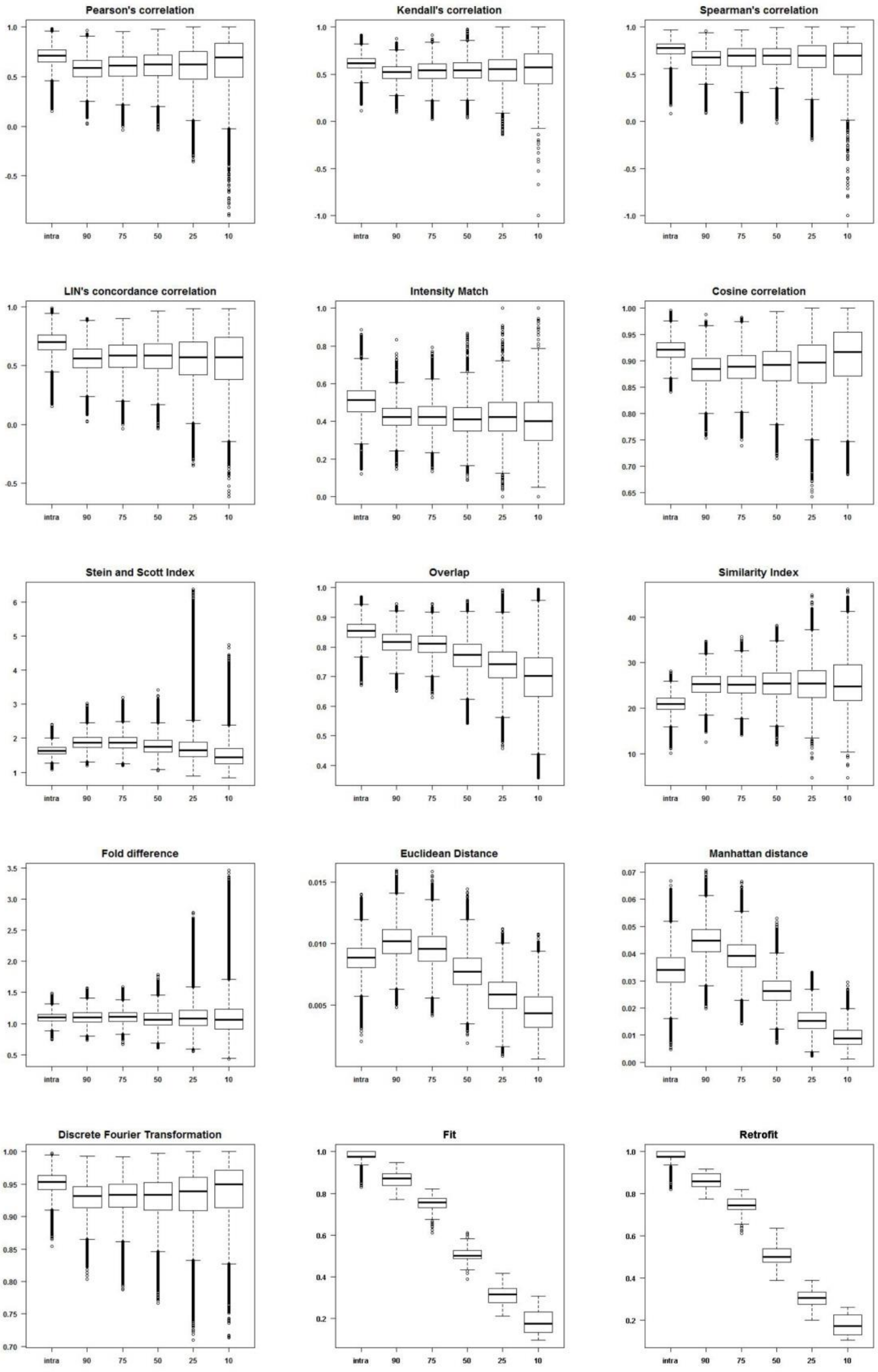


Figure 12: Boxplots of the different scores evalueted in pairwise comparison between spectra generated with 40 informative peaks and 30% intensity variability. On the x-axis the first reported measure is referred to the intra-sample analysis, while the others are the percentage of common peaks used in the inter-sample analysis.

Second level analysis

15 generated signals

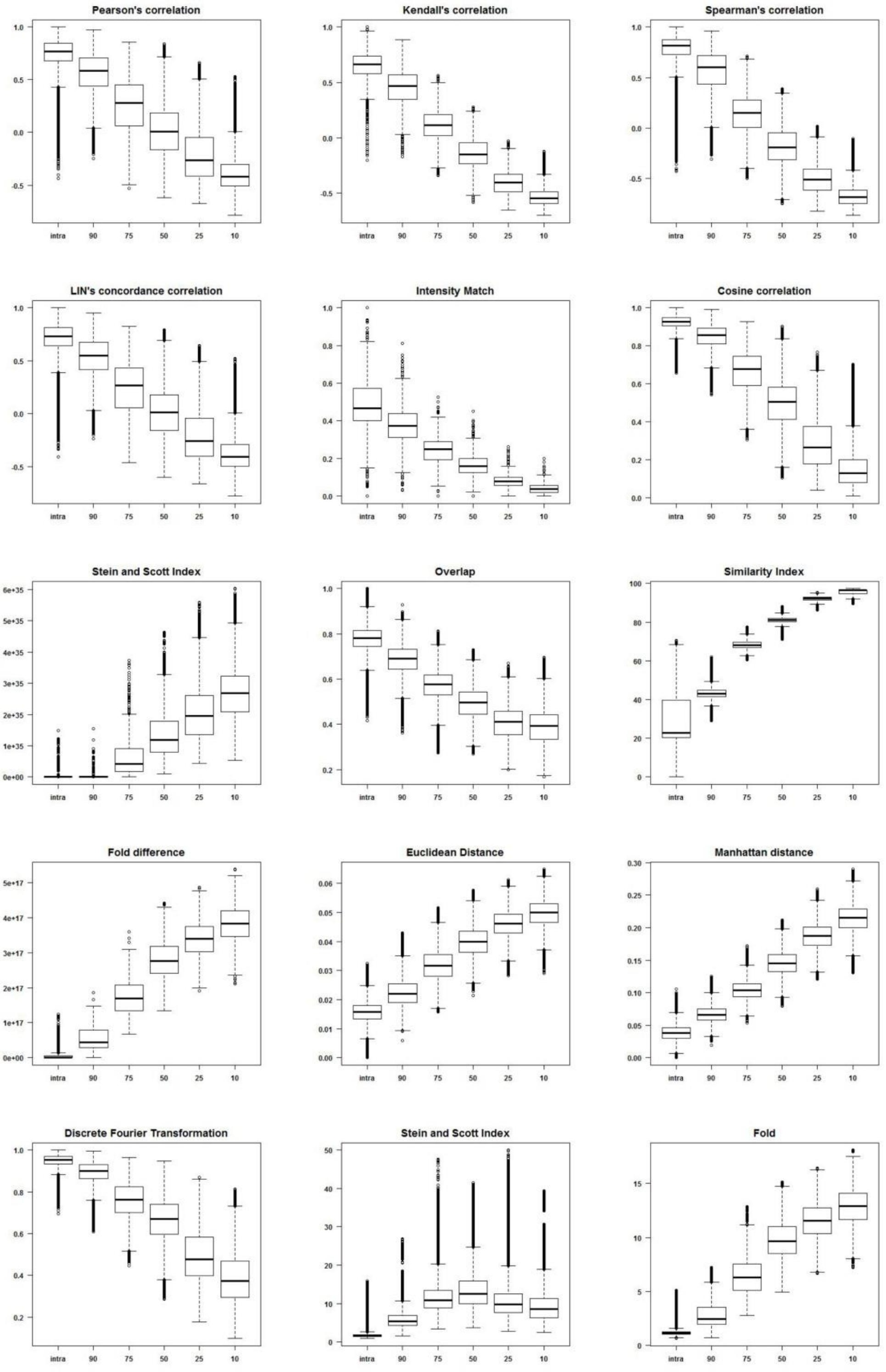


Figure 13: Boxplots of the different scores evalueted in pairwise comparison between spectra generated with 15 informative peaks and 30% intensity variability. On the x-axis the first reported measure is referred to the intra-sample analysis, while the others are the percentage of common peaks used in the inter-sample analysis.

Second level analysis

25 generated signals

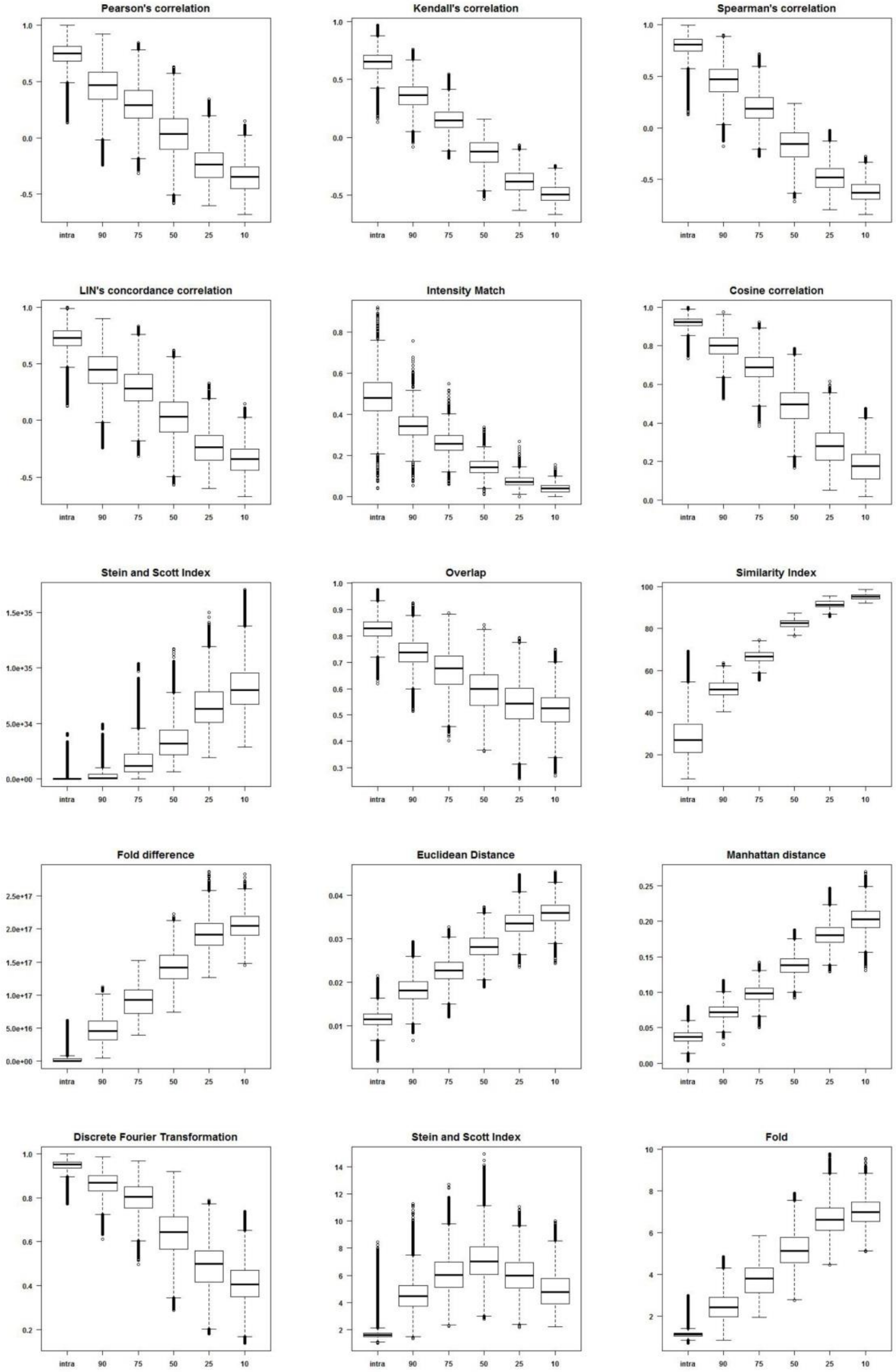


Figure 14: Boxplots of the different scores evalueted in pairwise comparison between spectra generated with 25 informative peaks and 30% intensity variability. On the x-axis the first reported measure is referred to the intra-sample analysis, while the others are the percentage of common peaks used in the inter-sample analysis.

Second level analysis

40 generated signals

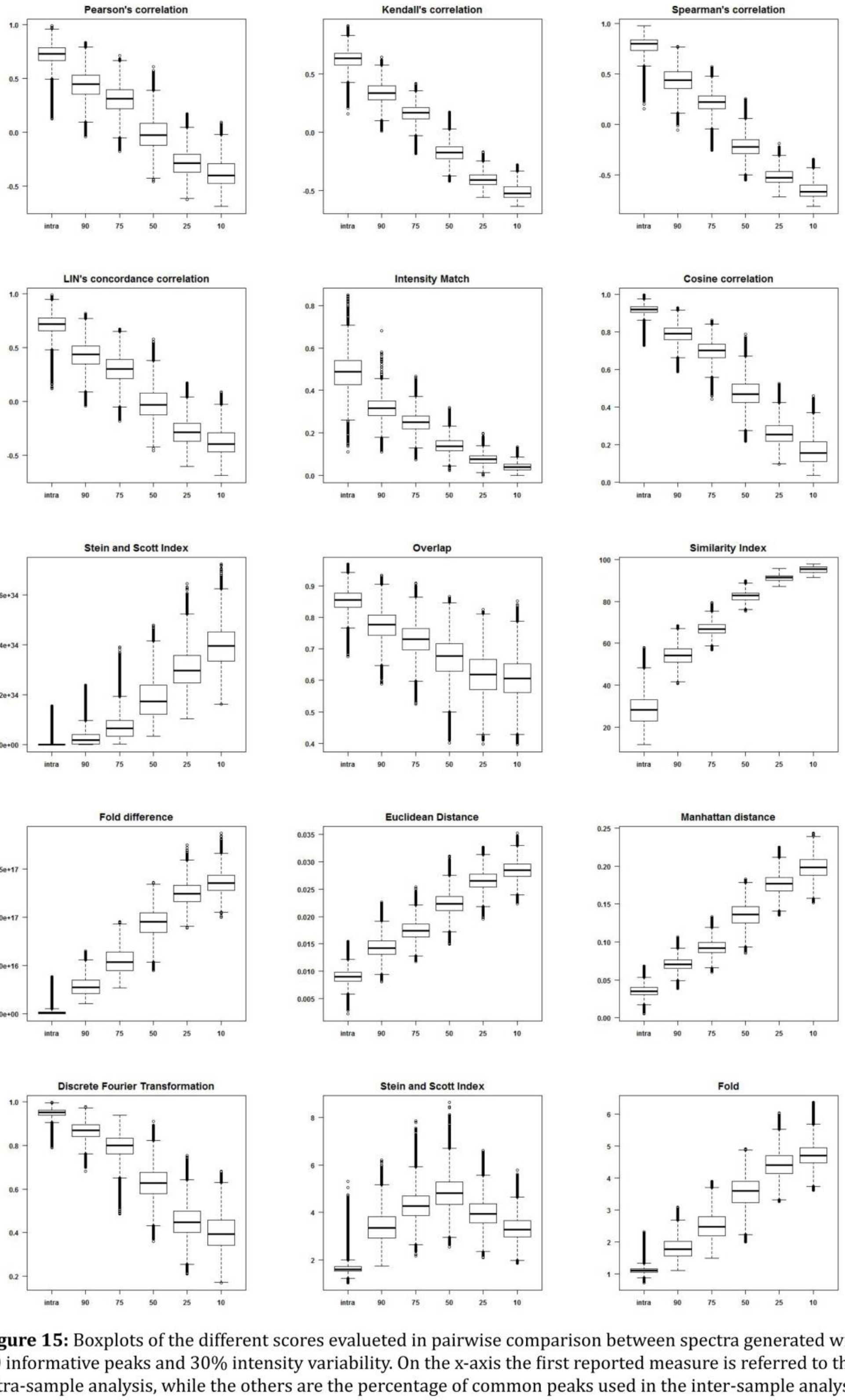
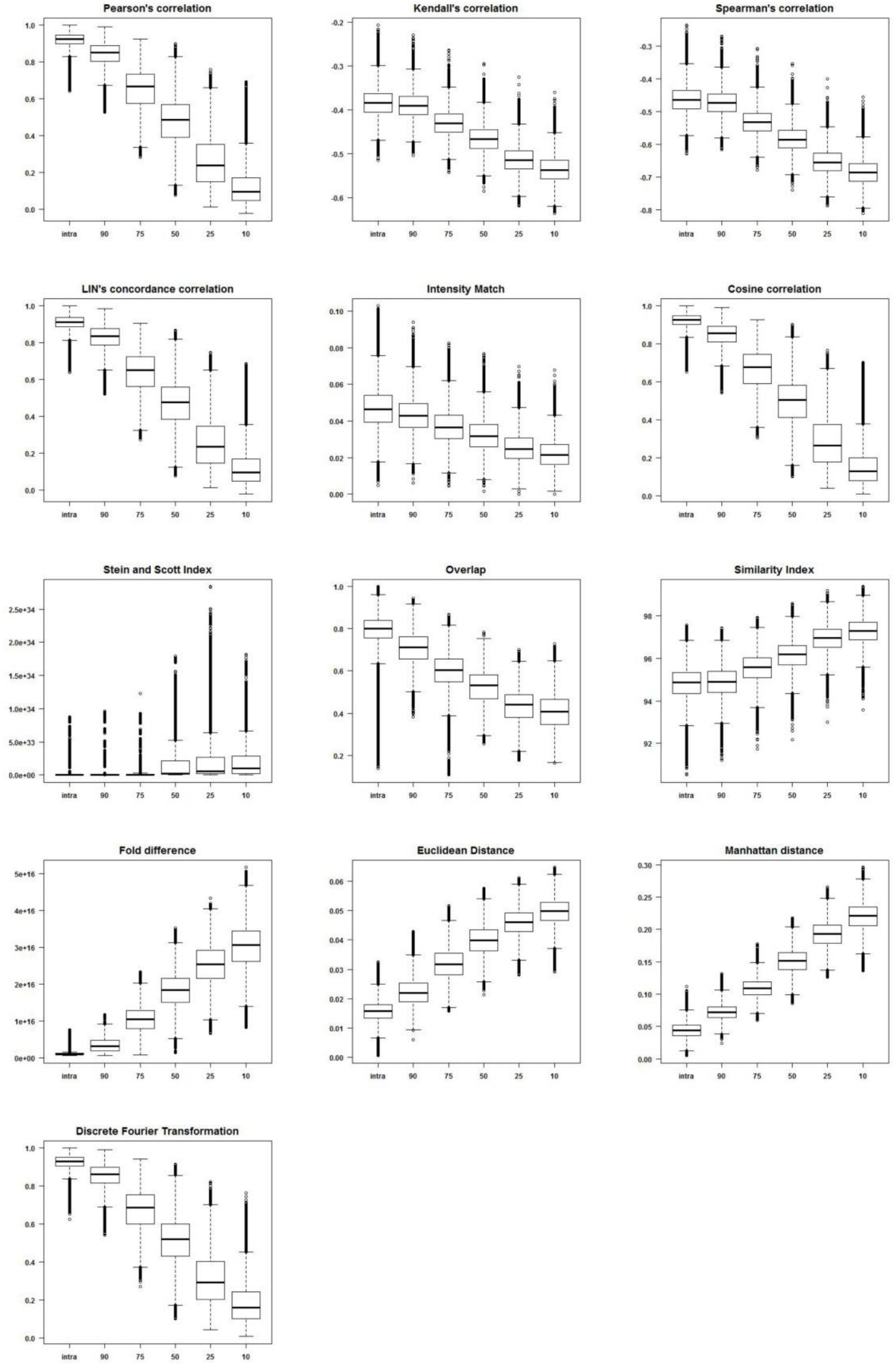


Figure 15: Boxplots of the different scores evalueted in pairwise comparison between spectra generated with 40 informative peaks and 30% intensity variability. On the x-axis the first reported measure is referred to the intra-sample analysis, while the others are the percentage of common peaks used in the inter-sample analysis.

Third level analysis

15 generated signals



Pearson's correlation

Kendall's correlation

Spearman's correlation

LIN's concordance correlation

Intensity Match

Cosine correlation

Stein and Scott Index

Overlap

Similarity Index

Fold difference

Euclidean Distance

Manhattan distance

Discrete Fourier Transformation

Figure 16: Boxplots of the different scores evalueted in pairwise comparison between spectra generated with 15 informative peaks and 30% intensity variability. On the x-axis the first reported measure is referred to the intra-sample analysis, while the others are the percentage of common peaks used in the inter-sample analysis.

Third level analysis

25 generated signals

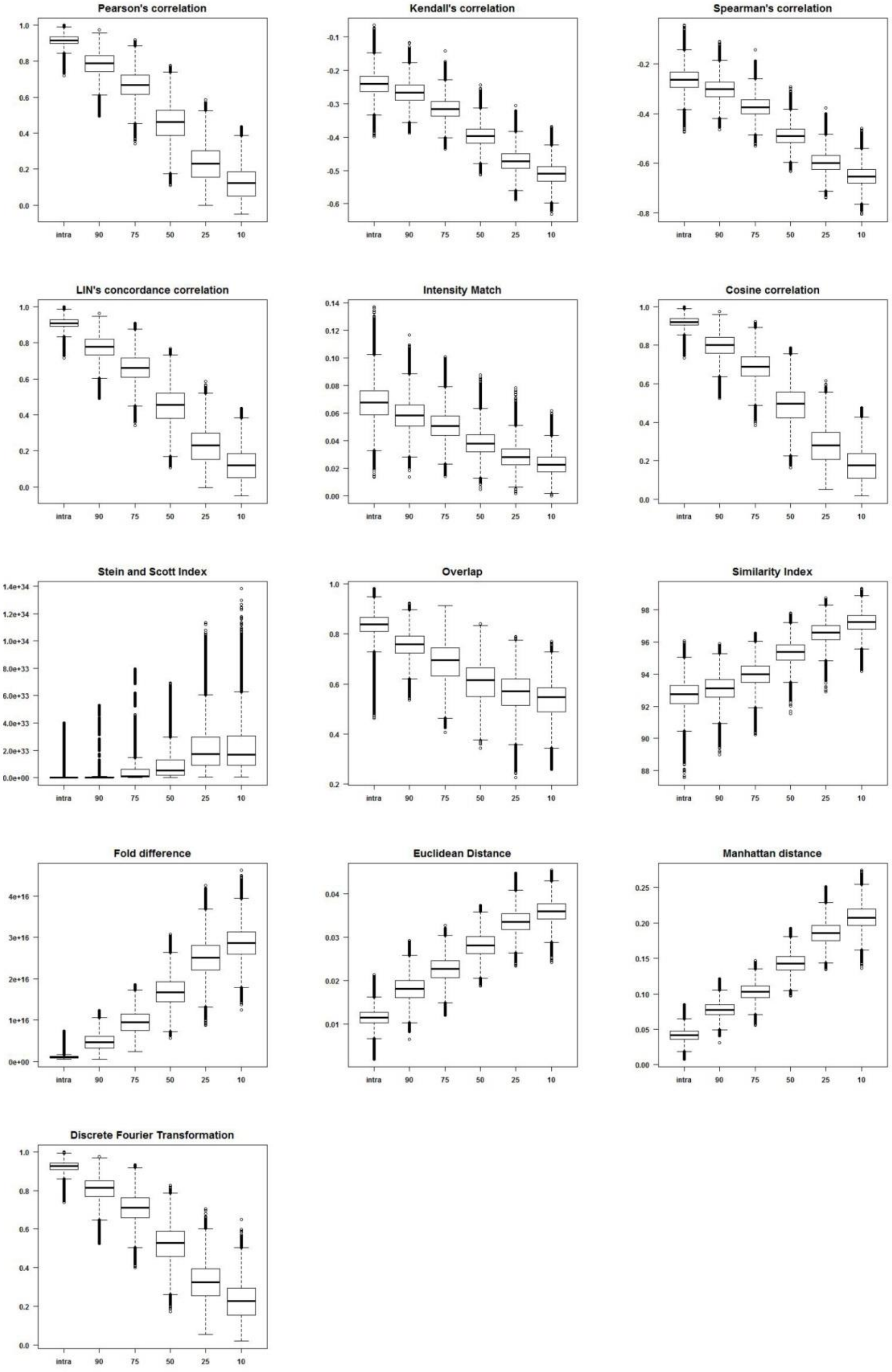


Figure 17: Boxplots of the different scores evalueted in pairwise comparison between spectra generated with 25 informative peaks and 30% intensity variability. On the x-axis the first reported measure is referred to the intra-sample analysis, while the others are the percentage of common peaks used in the inter-sample analysis.

Third level analysis

40 generated signals

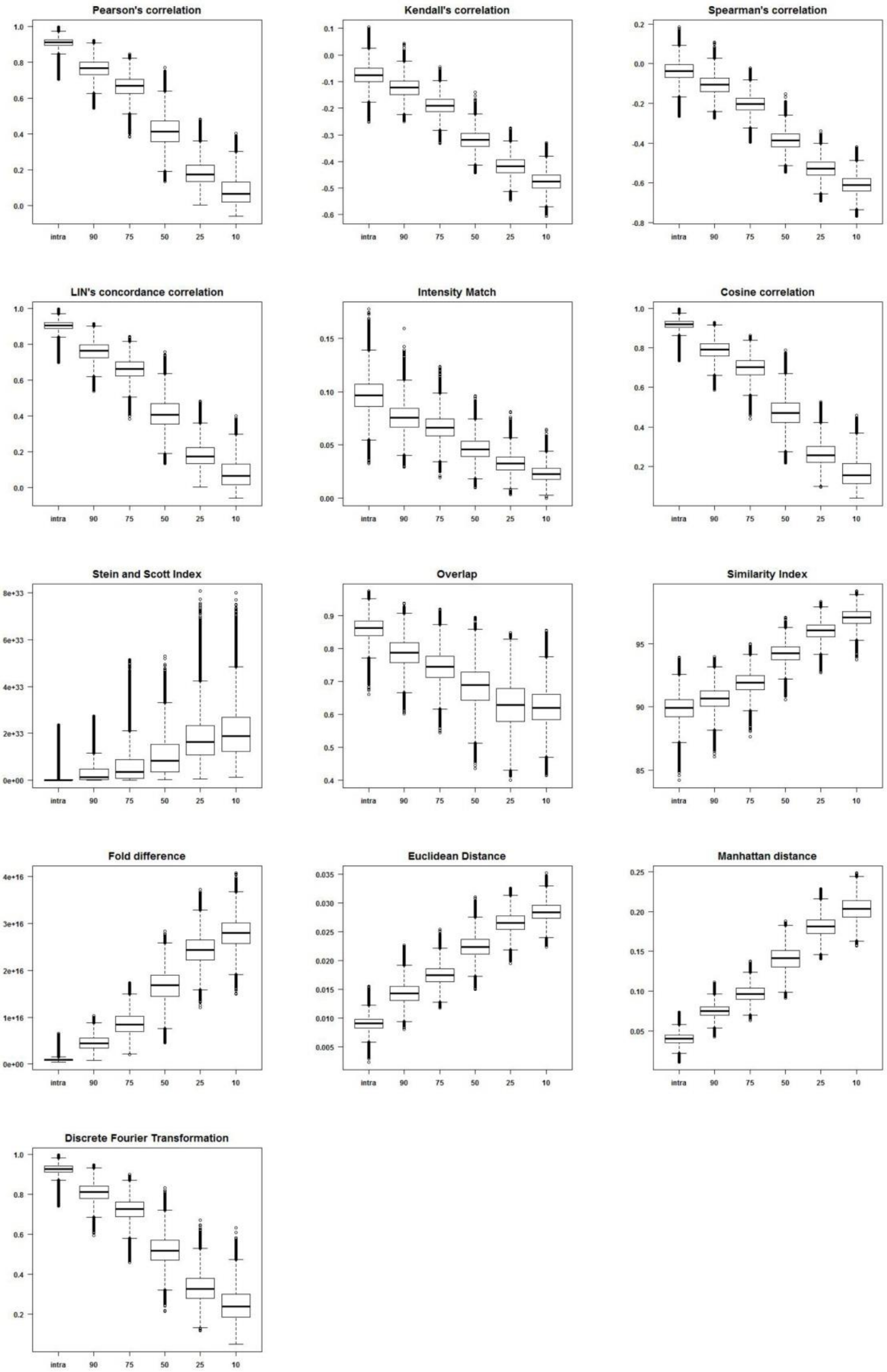


Figure 18: Boxplots of the different scores evalueted in pairwise comparison between spectra generated with 40 informative peaks and 30% intensity variability. On the x-axis the first reported measure is referred to the intra-sample analysis, while the others are the percentage of common peaks used in the inter-sample analysis.

First level analysis

25 generated signals

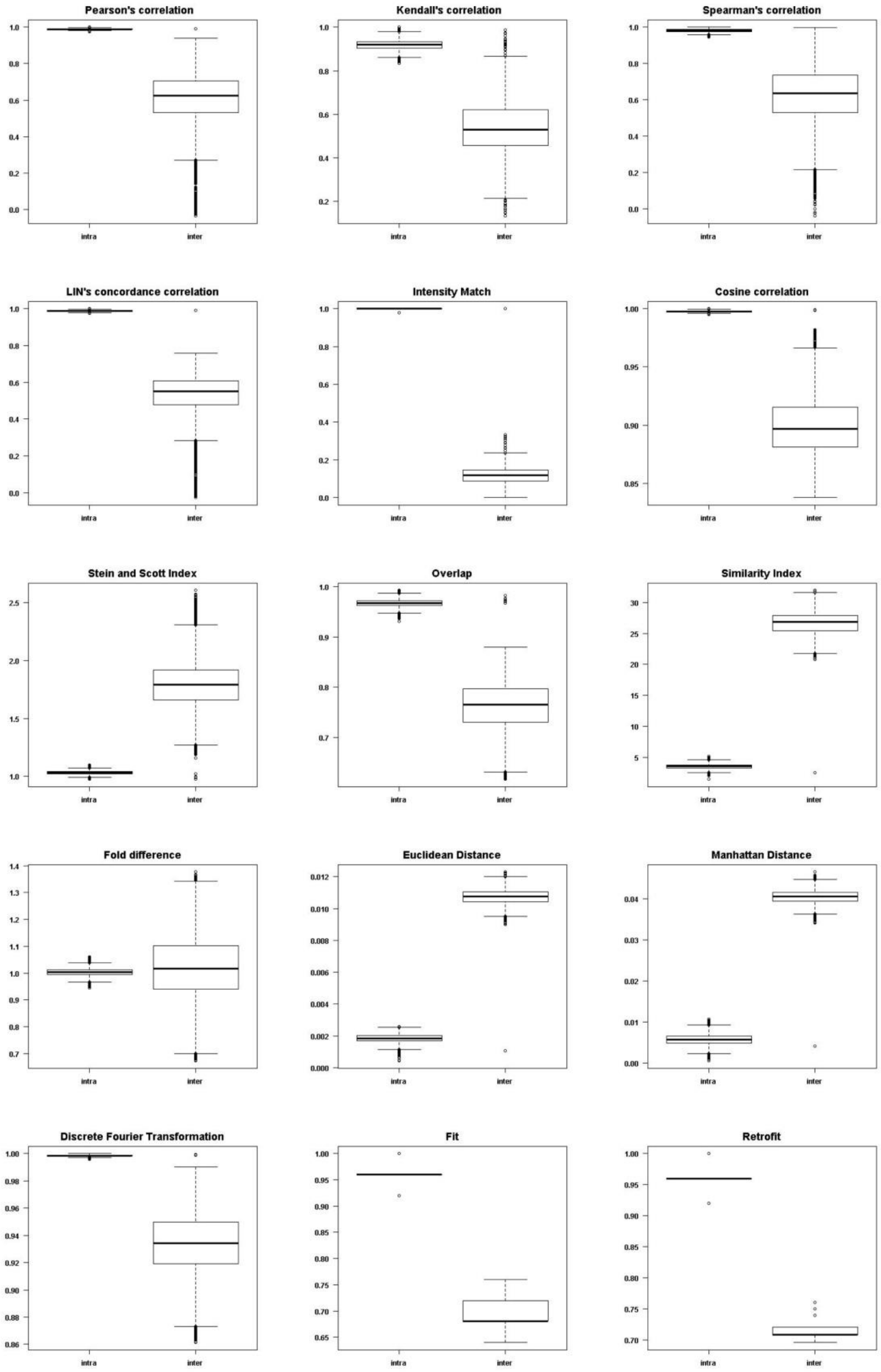


Figure 19: Boxplots of the different scores evalueted in pairwise comparison between spectra generated with 25 informative peaks. The 50% of peaks fall into the richest region. On the x-axis the first reported measure is referred to the intra-sample analysis, while the second is the inter-sample analysis.

First level analysis

40 generated signals

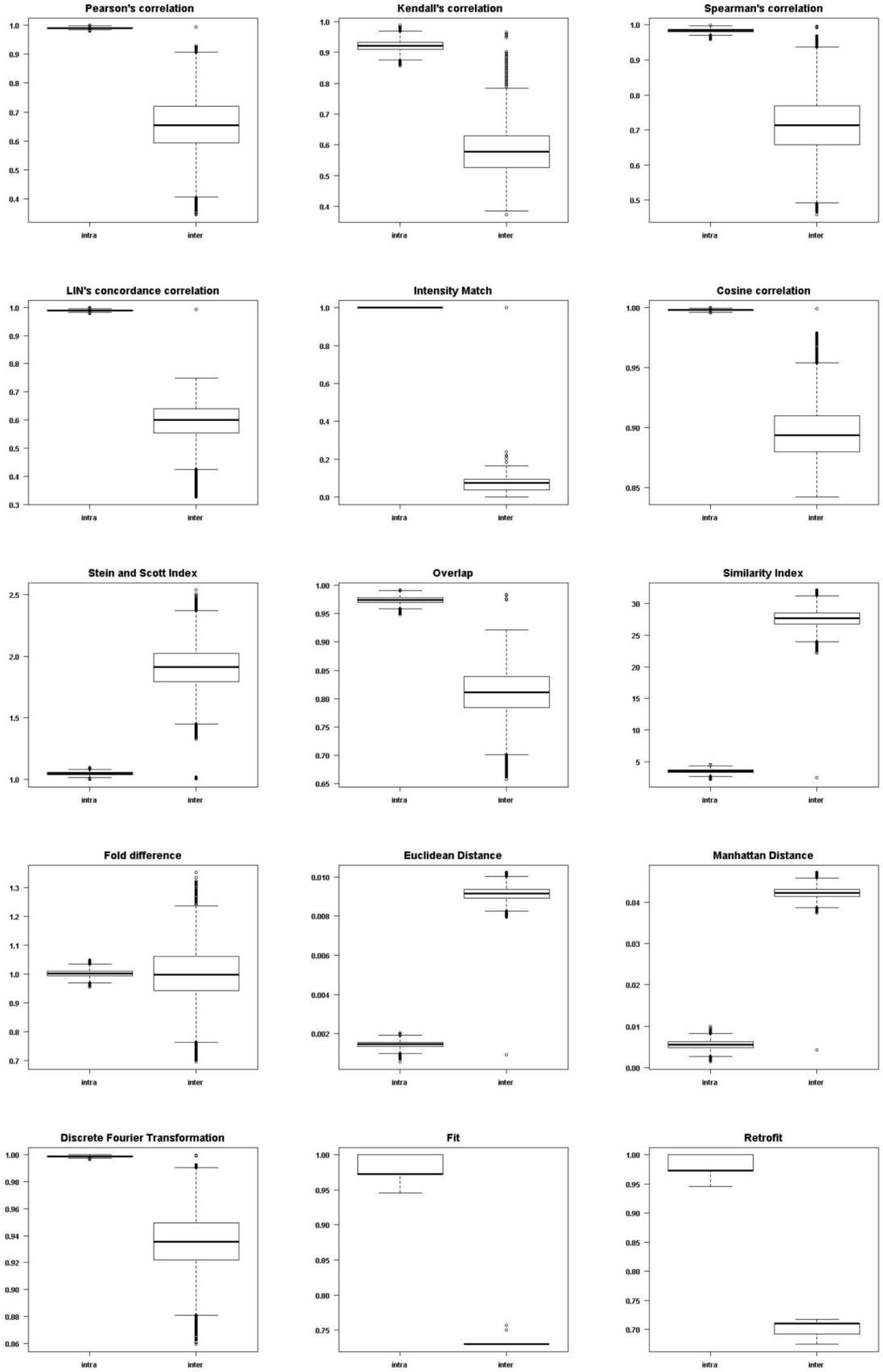


Figure 20: Boxplots of the different scores evalueted in pairwise comparison between spectra generated with 40 informative peaks. The 50% of peaks fall into the richest region. On the x-axis the first reported measure is referred to the intra-sample analysis, while the second is the inter-sample analysis.

Second level analysis

25 generated signals

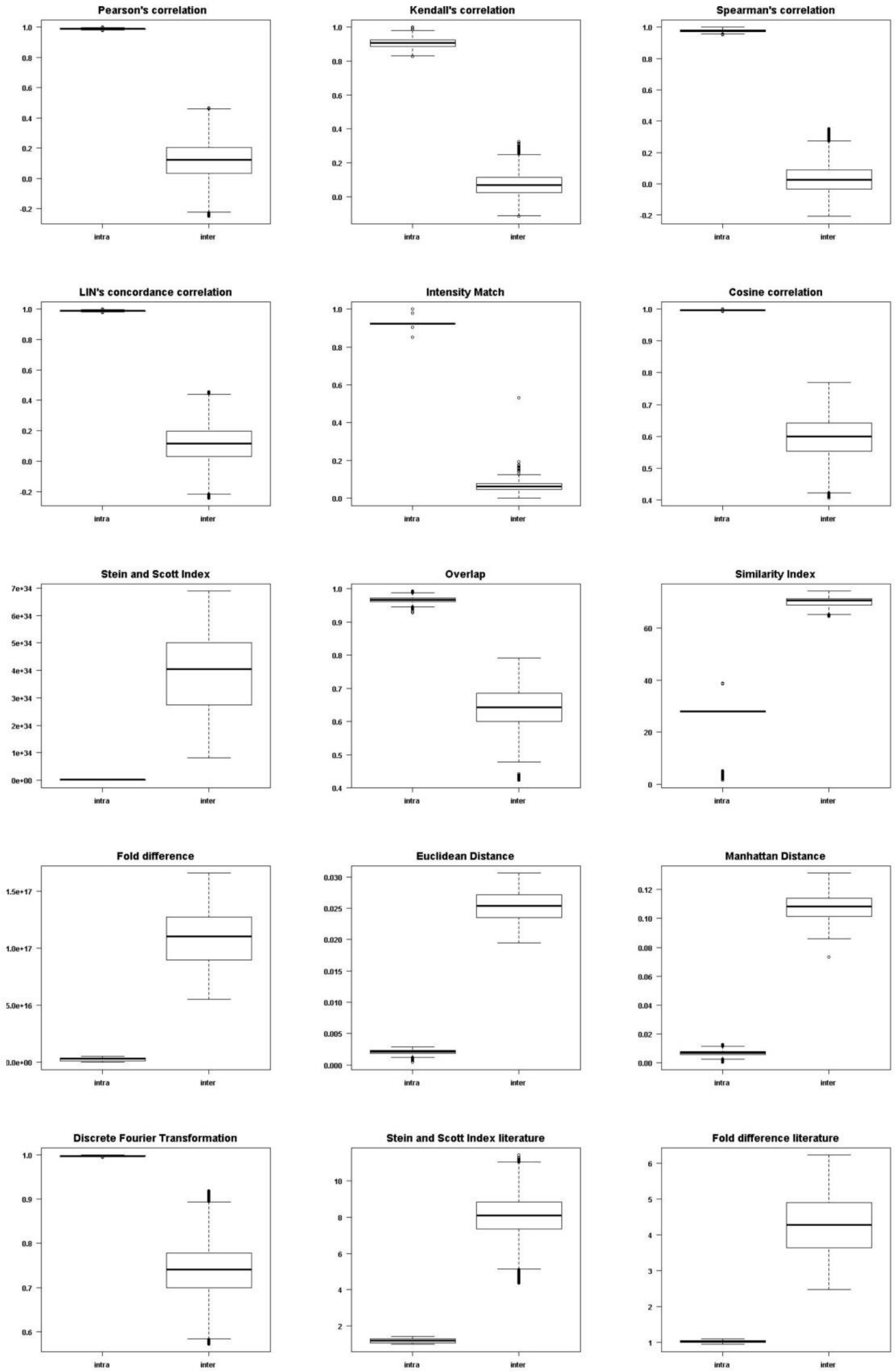


Figure 21: Boxplots of the different scores evalueted in pairwise comparison between spectra generated with 25 informative peaks. The 50% of peaks fall into the richest region. On the x-axis the first reported measure is referred to the intra-sample analysis, while the second is the inter-sample analysis.

Second level analysis

40 generated signals

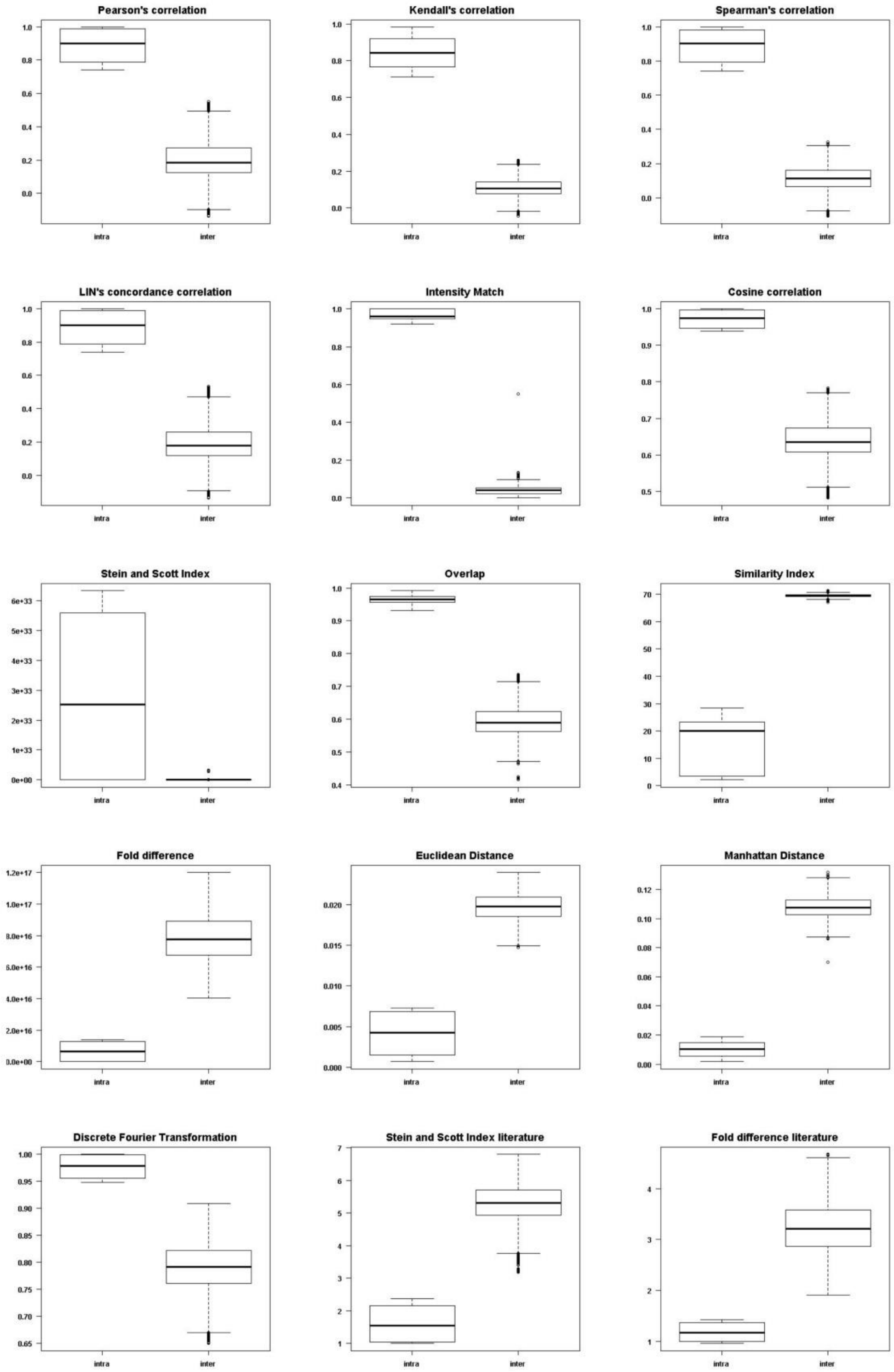


Figure 22: Boxplots of the different scores evalueted in pairwise comparison between spectra generated with 40 informative peaks. The 50% of peaks fall into the richest region. On the x-axis the first reported measure is referred to the intra-sample analysis, while the second is the inter-sample analysis.

Third level analysis

25 generated signals

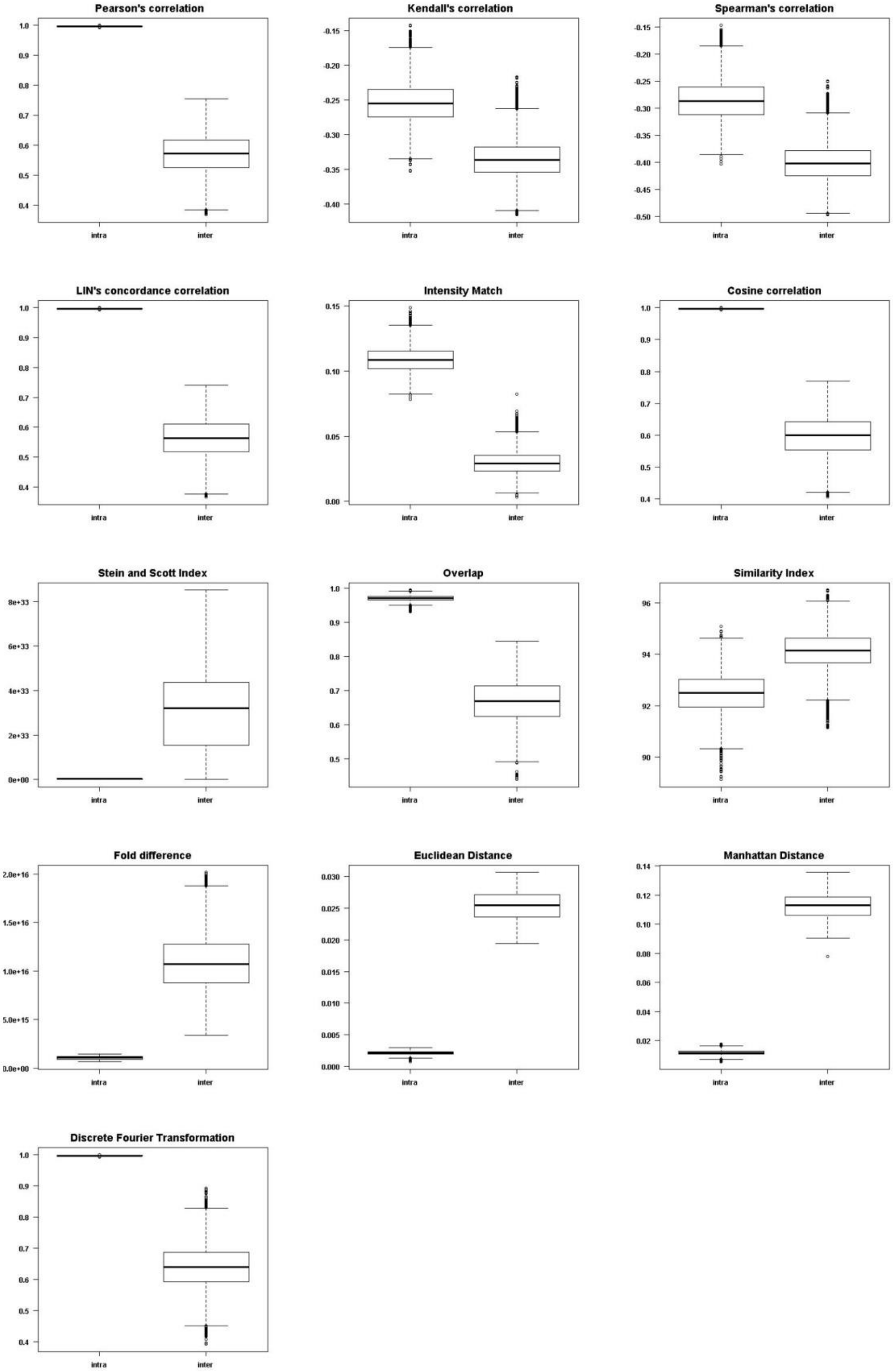


Figure 23: Boxplots of the different scores evalueted in pairwise comparison between spectra generated with 25 informative peaks. The 50% of peaks fall into the richest region. On the x-axis the first reported measure is referred to the intra-sample analysis, while the second is the inter-sample analysis.

Third level analysis

40 generated signals

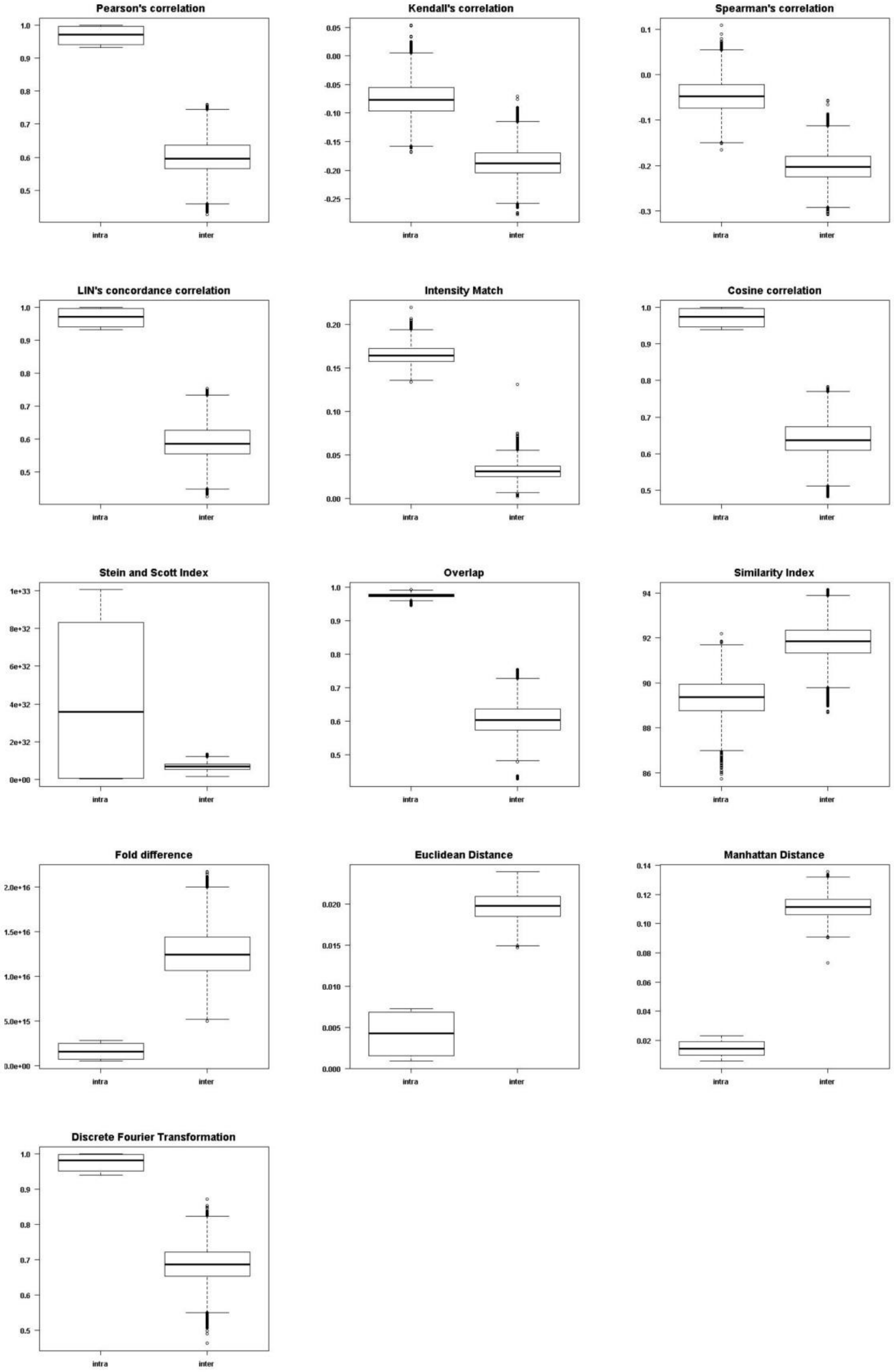


Figure 24: Boxplots of the different scores evalueted in pairwise comparison between spectra generated with 40 informative peaks. The 50% of peaks fall into the richest region. On the x-axis the first reported measure is referred to the intra-sample analysis, while the second is the inter-sample analysis.

First level analysis

25 generated signals

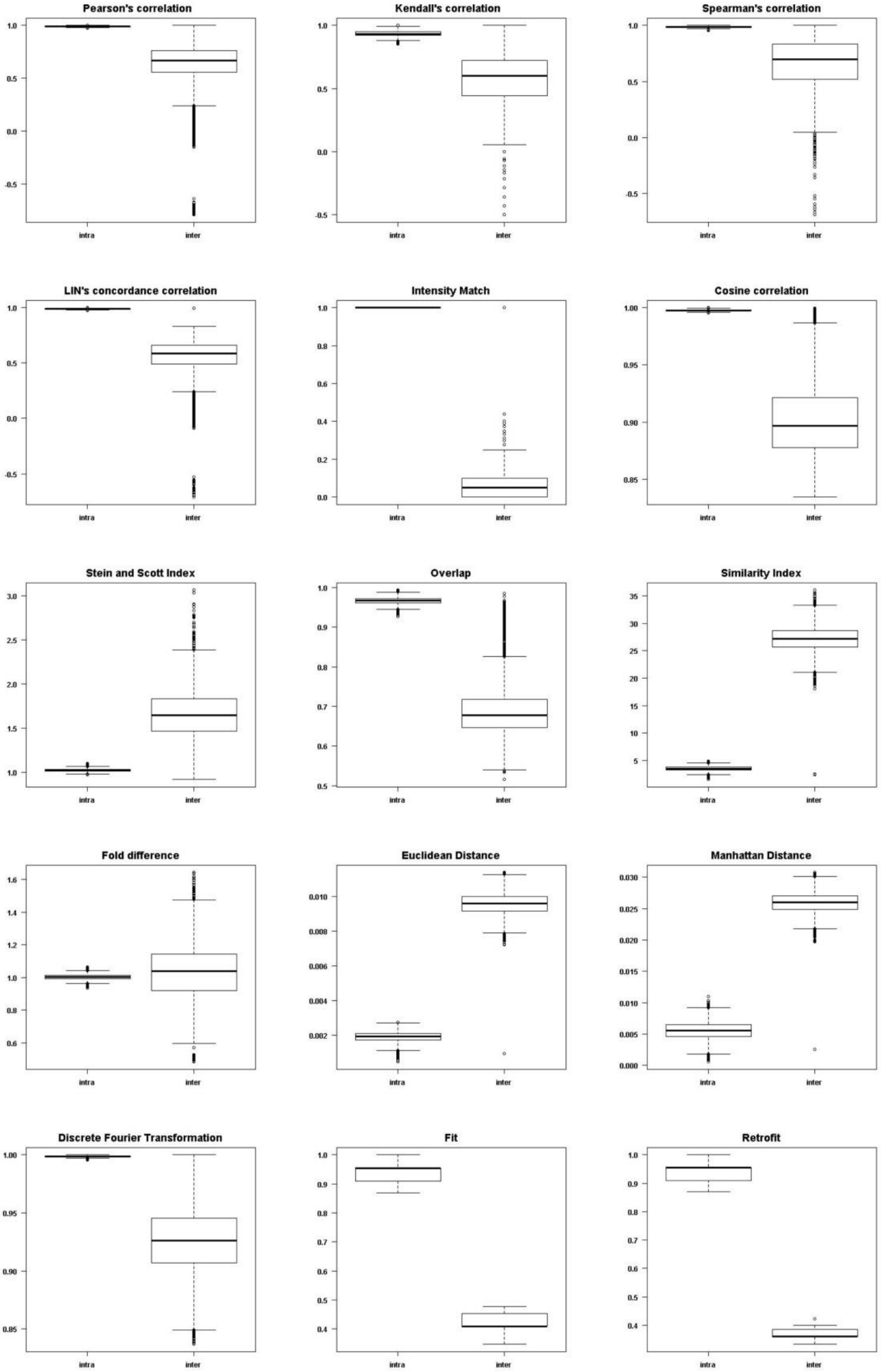


Figure 25: Boxplots of the different scores evalueted in pairwise comparison between spectra generated with 25 informative peaks. The 70% of peaks fall into the richest region. On the x-axis the first reported measure is referred to the intra-sample analysis, while the second is the inter-sample analysis.

First level analysis

40 generated signals

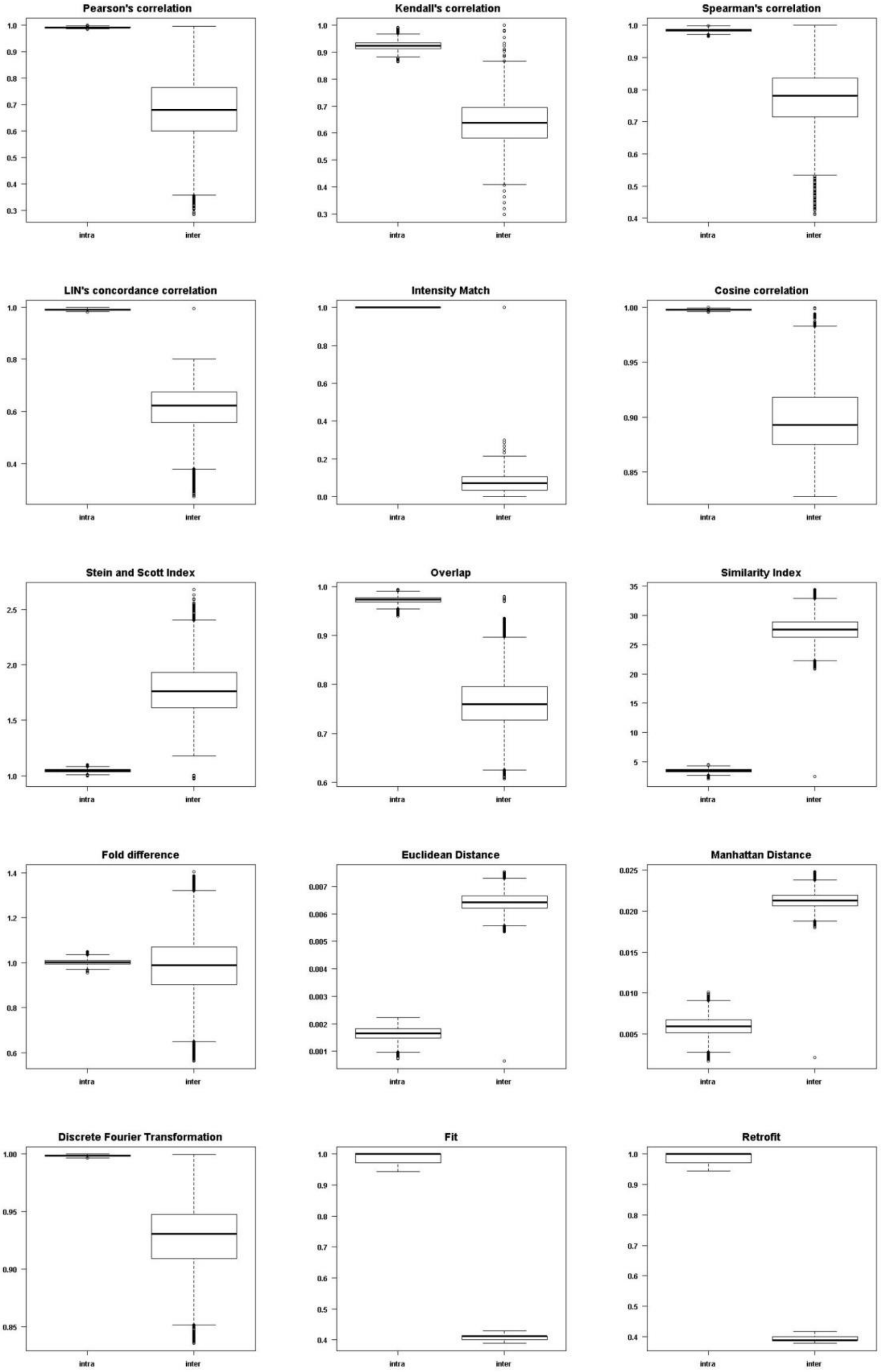


Figure 26: Boxplots of the different scores evalueted in pairwise comparison between spectra generated with 40 informative peaks. The 70% of peaks fall into the richest region. On the x-axis the first reported measure is referred to the intra-sample analysis, while the second is the inter-sample analysis.

Second level analysis

25 generated signals

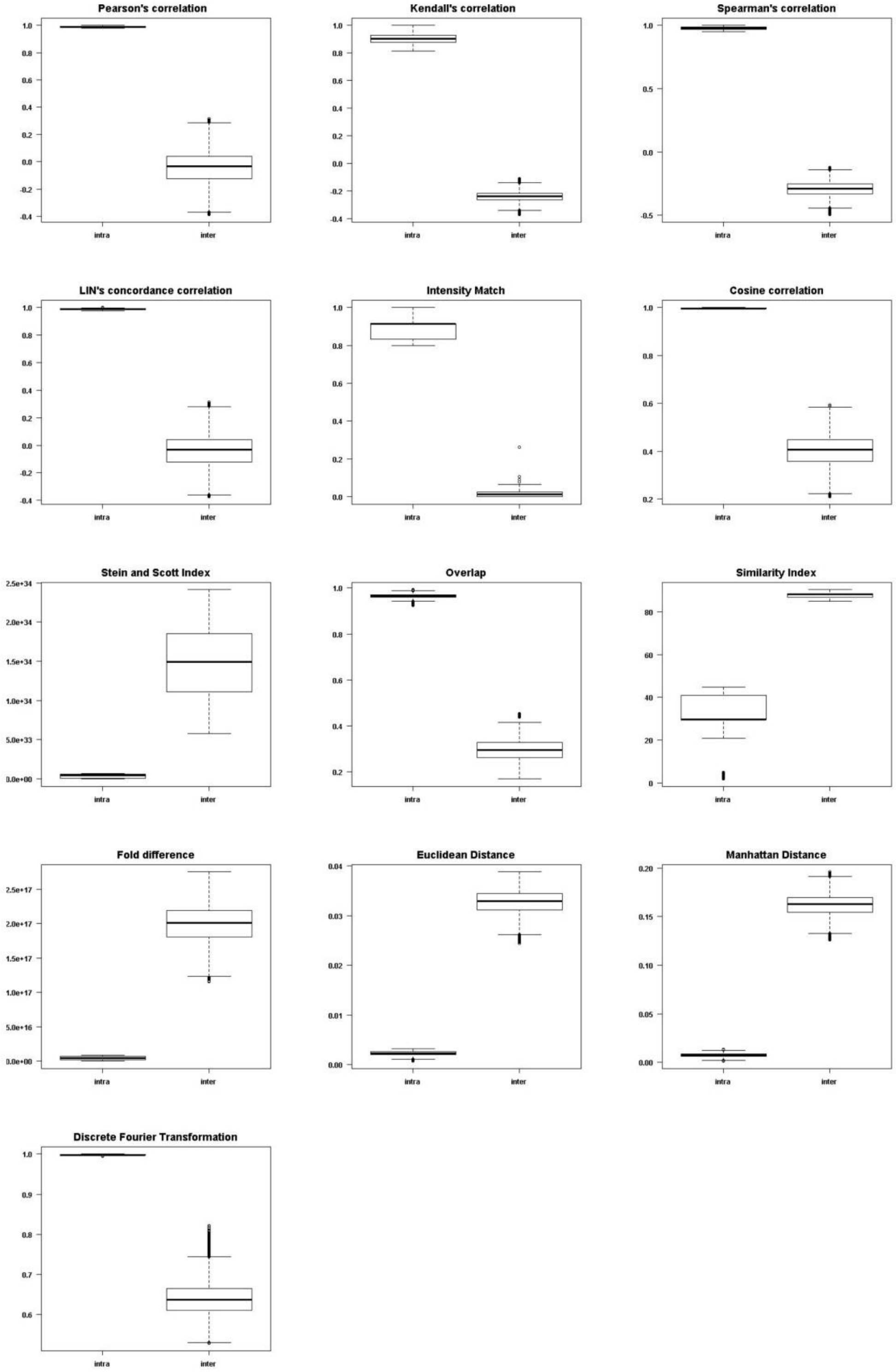


Figure 27: Boxplots of the different scores evalueted in pairwise comparison between spectra generated with 25 informative peaks. The 70% of peaks fall into the richest region. On the x-axis the first reported measure is referred to the intra-sample analysis, while the second is the inter-sample analysis.

Second level analysis

40 generated signals

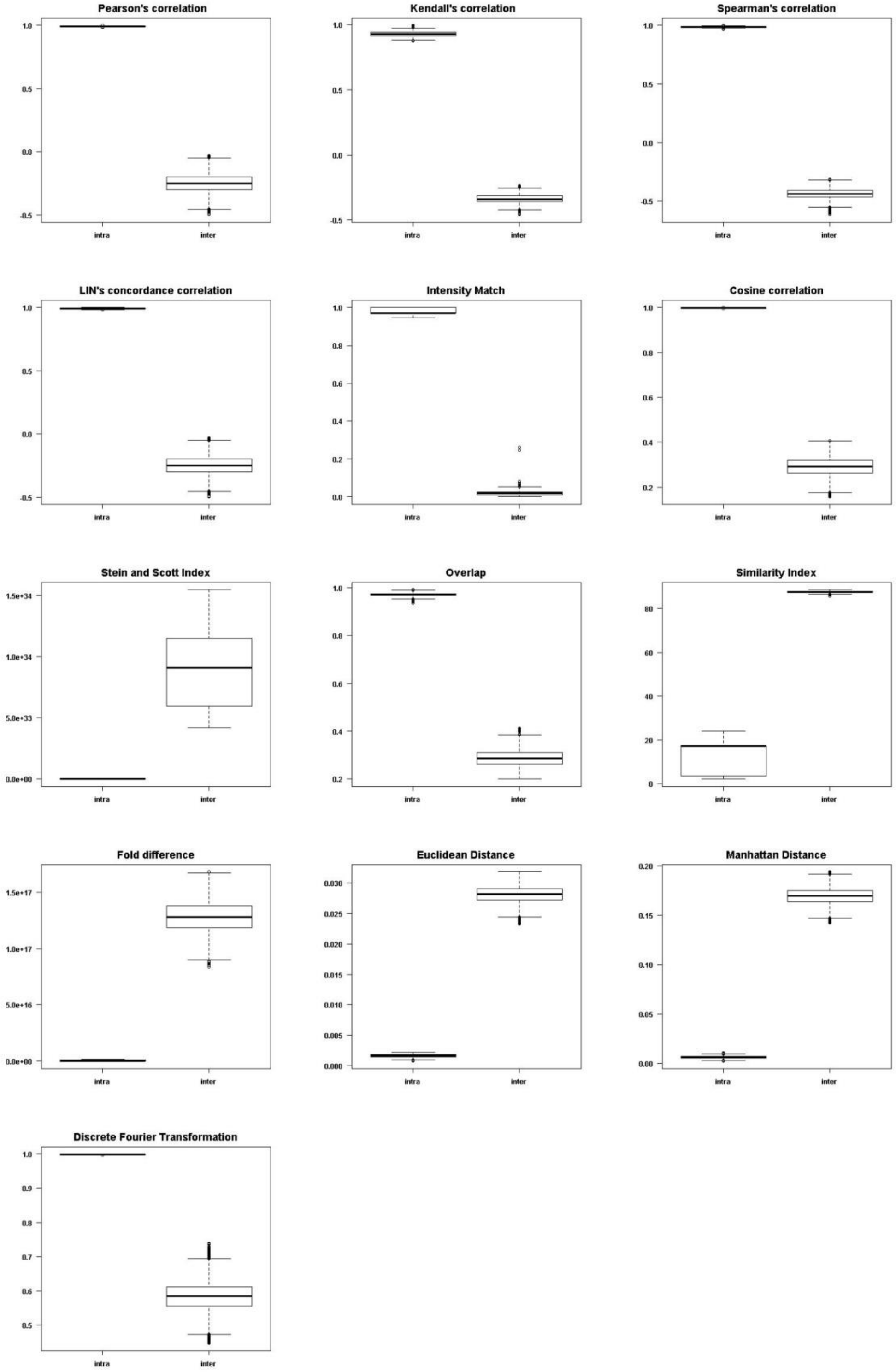


Figure 28: Boxplots of the different scores evalueted in pairwise comparison between spectra generated with 40 informative peaks. The 70% of peaks fall into the richest region. On the x-axis the first reported measure is referred to the intra-sample analysis, while the second is the inter-sample analysis.

Third level analysis

25 generated signals

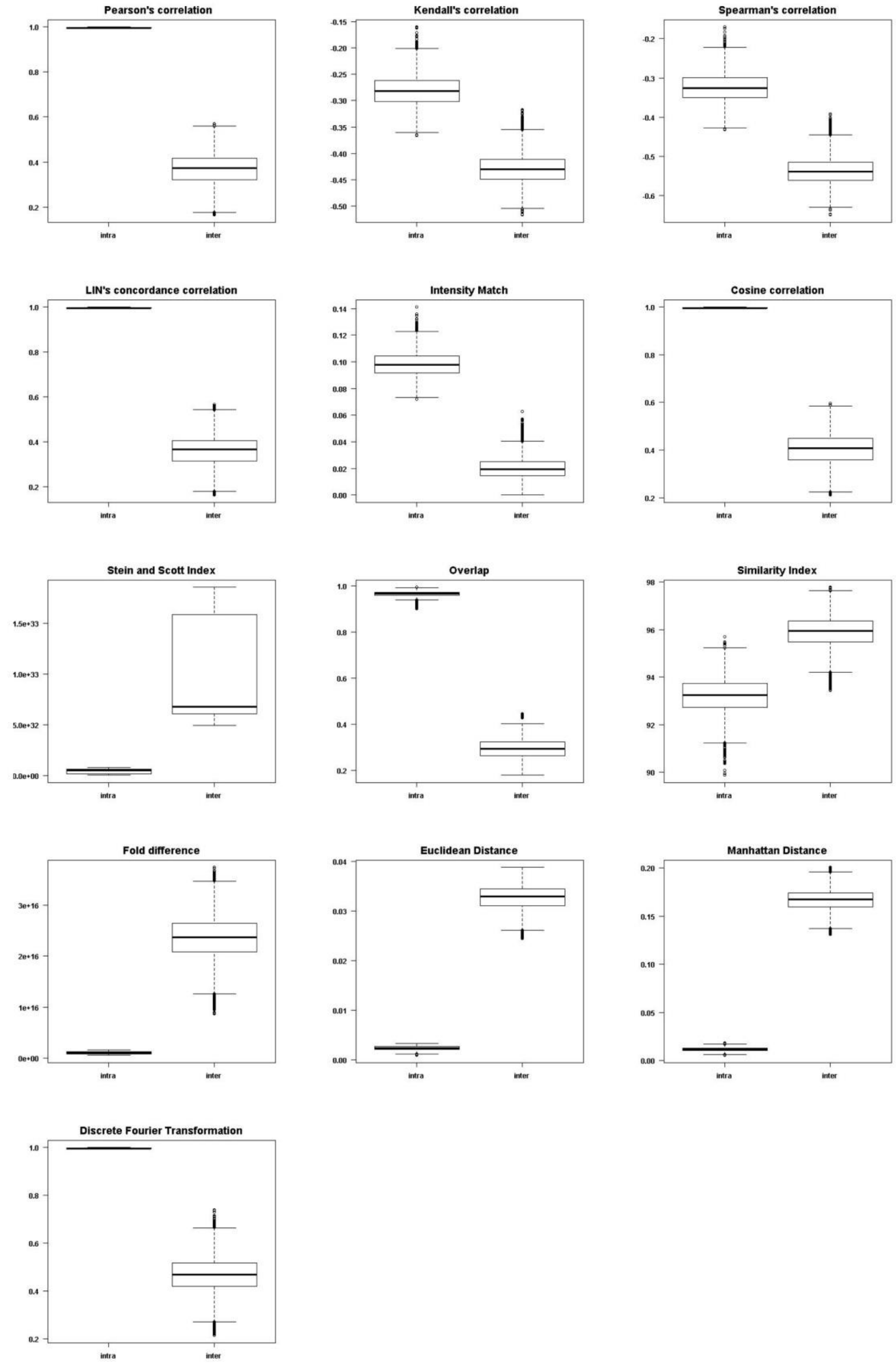


Figure 29: Boxplots of the different scores evalueted in pairwise comparison between spectra generated with 25 informative peaks. The 70% of peaks fall into the richest region. On the x-axis the first reported measure is referred to the intra-sample analysis, while the second is the inter-sample analysis.

Third level analysis

40 generated signals

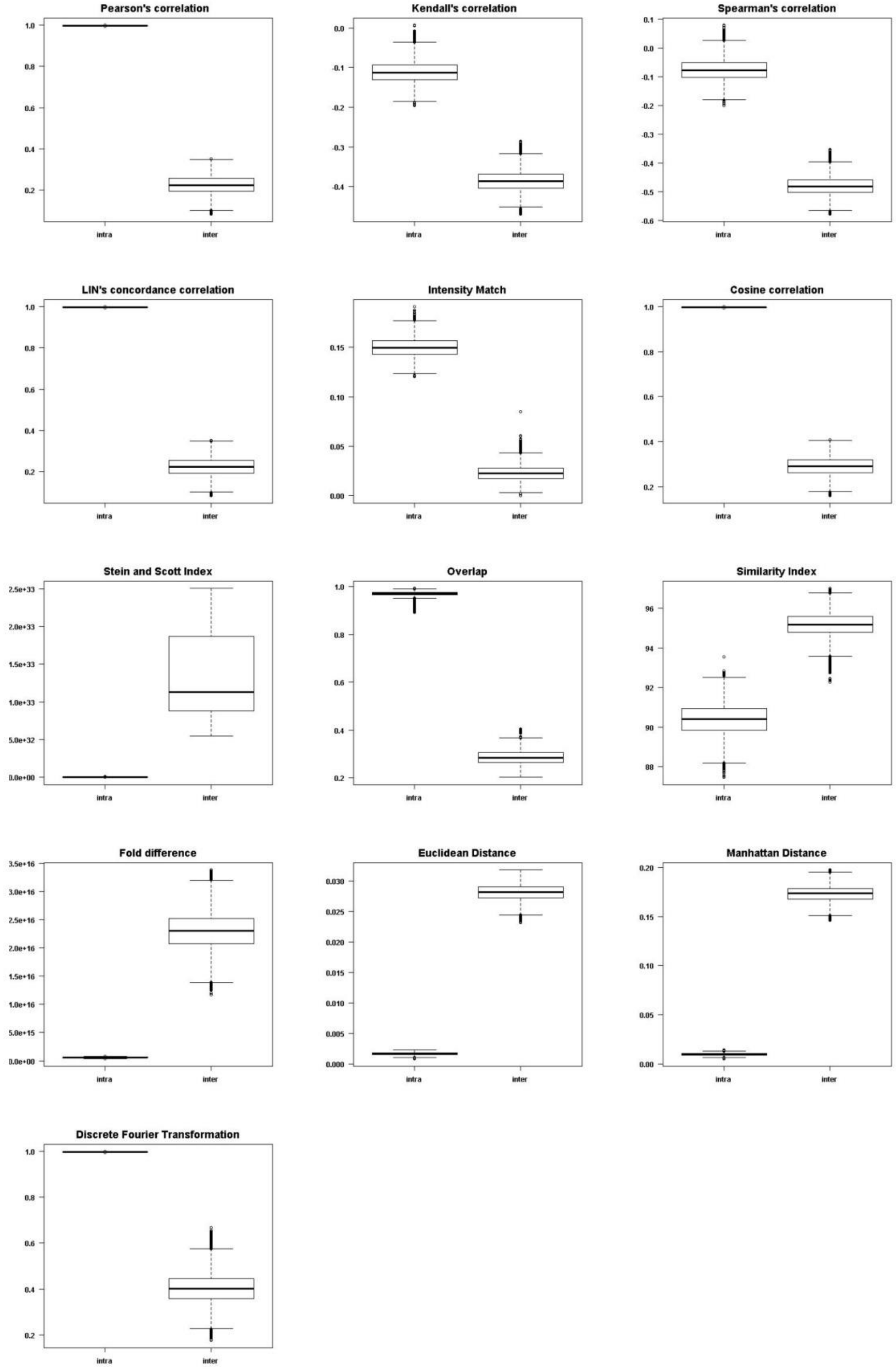


Figure 30: Boxplots of the different scores evalueted in pairwise comparison between spectra generated with 40 informative peaks. The 70% of peaks fall into the richest region. On the x-axis the first reported measure is referred to the intra-sample analysis, while the second is the inter-sample analysis.