Angela da Costa Carapito

Volatilome for the detection of bladder cancer in urine using gas chromatography-mass spectrometry

Dissertation thesis for the master's degree in Analytical, Clinical and Forensic Toxicology from the Faculty of Pharmacy, University of Porto.

Work elaborated under supervision of Doutora Maria Paula Guedes de Pinho and co-supervision of Doutora Joana Isabel Monteiro Pinto and Professora Doutora Maria de Lourdes P.A.S. Bastos.

December 2020

This work was financed by FEDER (COMPETE 2020–POCI) and Portuguese funds (FCT) in the framework of the project POCI-01-0145-FEDER-030388-PTDC/SAU-SER/30388/2017.

Abstract

Introduction: Cancer represents one of the biggest causes of death worldwide, being bladder cancer (BCa) the 10th most common cancer worldwide, with higher incidence and mortality rates specially in men. While many diagnostic tests are available and currently practiced today, such as urine cytology, cystoscopy, and imaging tests, some of those methods are invasive, expensive and present low sensitivity to detect cancer in its early stages. Considering these struggles, metabolomic analysis based on the detection of volatile organic compounds (VOCs) has been presented as an important approach in many investigation studies of new diagnostic methods about cancer in the last years.

Aim: The aims of this work were to investigate the potential of VOCs, in general, and volatile carbonyl compounds (VCCs) for BCa detection in urine by headspace solid-phase microextraction/gas chromatography-mass spectrometry (HS-SPME/GC-MS).

Methods: A total of 120 patients were included in this study, 60 BCa patients, being 17 women and 43 men; and 60 cancer-free individuals (controls). The extraction of VOCs and VCCs from urine samples of BCa and cancer-free individuals was performed by HS-SPME. Statistical analysis included both multivariate, like principal component analysis (PCA) and partial least square discriminant analysis (PLS-DA) and univariate tests. Area under the curve (AUC), sensitivity, specificity and accuracy were calculated through receiver operating characteristics (ROC).

Results: The volatile profile of urine (VOCs and VCCs) was able to discriminate BCa patients from controls. Statistically significant alterations were observed in the levels of 30 metabolites, demonstrating an AUC of 0.924, 85% sensitivity, 82% specificity and 83% accuracy for BCa detection. The main classes of the discriminant compounds were aldehydes, alkanes, ketones and aromatic hydrocarbons. Considering the discrimination of different BCa stages, namely stages II, III and IV compared with stage 0a, statistically significant alterations were also identified, demonstrating an AUC of 0.830, 50% sensitivity, 82% specificity and 80% accuracy.

Discussion/Conclusions: This work was able to demonstrate the important role of metabolomic analysis for the detection of BCa in urine samples. Some of the identified compounds were also found on previous works presented on the literature which proves that a panel of biomarkers for BCa detection can be established in the future as a diagnostic tool. Several challenges still exist regarding the clarification of metabolic pathways associated with the release of volatile compounds in cancer patients and

controls. More studies using urine samples of BCa and controls are necessary to elucidate VOCs metabolism as well as its potential applicability in portable medical devices. In that way, it will be possible to ensure an earlier diagnosis, with a more adapted treatment and less costs.

Keywords: Bladder cancer (BCa); volatile organic compounds (VOCs); volatile carbonyl compounds (VCCs); metabolomics; headspace solid-phase microextraction coupled with gas chromatography-mass spectrometry (HS-SPME/GC-MS).

Resumo

Introdução: O cancro representa uma das maiores causas de morte no mundo, sendo o cancro da bexiga (BCa) o 10º cancro mais comum, com elevada taxa de incidência e mortalidade especialmente em homens. Embora existam vários testes de diagnóstico disponíveis e atualmente praticados, como por exemplo a citologia urinária, cistoscopia e testes imagiológicos, alguns destes métodos são invasivos, dispendiosos e apresentam baixa sensibilidade para detetar cancro nos seus estadios mais precoces. Considerando estas dificuldades, as análises metabolómicas baseadas na deteção de compostos orgânicos voláteis (VOCs) têm sido apresentadas nos últimos anos, como uma importante abordagem em vários estudos de investigação de novos métodos de diagnóstico associados ao cancro.

Objetivo: Os objetivos deste trabalho consistiram em investigar o potencial de VOCs, em geral, e de compostos carbonílicos voláteis (VCCs) para a deteção de BCa na urina por cromatografia gasosa acoplada a espetrometria de massa através de microextração em fase sólida por *headspace* (HS-SPME/GC-MS).

Métodos: Um total de 120 pacientes foram incluídos neste estudo, 60 pacientes com BCa, sendo 17 mulheres e 43 homens; e 60 pacientes sem cancro. A extração de VOCs e VCCs das amostras de urina dos pacientes com BCa e dos pacientes sem cancro foi realizada através de HS-SPME. Os testes estatísticos incluíram ambas as análises multivariadas tais como as análises de componentes principais (PCA) e as análises discriminantes pelo método de mínimos quadrados parciais (PLS-DA) e testes univariados. A área sob a curva (AUC), sensibilidade, especificidade e exatidão foram calculadas através da curva característica de operação do recetor (ROC).

Resultados: O perfil volátil da urina (VOCs e VCCs) permitiu discriminar pacientes com BCa de pacientes controlo. Diferenças estatisticamente significativas foram observadas ao nível de 30 metabolitos, demonstrando uma AUC de 0.924, uma sensibilidade de 85%, especificidade de 82% e exatidão de 83% para a deteção de BCa. As classes principais dos compostos identificados foram essencialmente aldeídos, alcanos, cetonas e hidrocarbonetos aromáticos. Considerando a discriminação dos diferentes estadios de BCa, nomeadamente o estadio II, III e IV comparativamente ao estadio 0a, também foram identificadas alterações estatisticamente significativas, demonstrando uma AUC de 0.830, sensibilidade de 50%, sensibilidade de 82% e exatidão de 80%.

Discussão/Conclusões: Este trabalho demonstra o importante papel das análises metabolómicas para a deteção de BCa em amostras de urina. Alguns dos compostos identificados foram também encontrados em trabalhos anteriores presentes na literatura, provando assim que um painel de biomarcadores para a deteção de BCa pode ser estabelecido no futuro como meio de diagnóstico. Diversos desafios ainda existem relativamente à clarificação das vias metabólicas associadas à libertação de compostos voláteis em pacientes com cancro e controlos. Mais estudos usando amostras de urina de pacientes com BCa e controlos são necessários por forma a elucidar o metabolismo de VOCs assim como a sua potencial aplicabilidade em dispositivos médicos portáteis. Desta forma, será possível assegurar um diagnóstico mais precoce, com um tratamento mais adaptado e menos custos.

Palavras-chave: Cancro da bexiga (BCa); compostos orgânicos voláteis (VOCs); compostos carbonílicos voláteis (VCCs); metabolómica; cromatografia gasosa acoplada a espetrometria de massa através de microextração em fase sólida por *headspace* (HS-SPME/GC-MS).

Table of contents

Abstract iv

Index of figures x

Index of tables xii

List of abbreviations xiii

Chapter 1. Introduction 1

1.1. Bladder cancer 2

1.1.1. Incidence and risk factors 2

1.1.2. Classification and diagnosis 2

1.1.3. Staging 4

1.1.4. Treatment 5

1.2. From smell perception to disease recognition 6

1.2.1. Potential of volatile organic compounds as diagnostic markers 8

1.2.2. Metabolomics applied to VOCs profiling 9

1.3. VOCs as potential biomarkers of bladder cancer: state of the art 18

1.4. Aims of this thesis 29

Chapter 2. Experimental methods 31

2.1. Chemicals 31

2.2. Subjects 31

2.3. Sample preparation 32

2.4. GC-MS analysis 33

2.5. Data pre-processing 35

2.6. Statistical analysis 36

Chapter 3. Results 39

3.1. Identification of VOCs and VCCs present in urine 39

3.2. Volatile profile of urine of BCa patients *vs.* cancer-free controls 40

3.3. Volatile profile of urine of BCa patients with different stages 51

Chapter 4. Discussion 56

Chapter 5. Conclusions and future perspectives 61

References 62

Appendix 74

Index of figures

Figure 1 - Classification of BCa according to its staging [Reprinted from Dana-Farber Cancer Institute, 2019].....4
Figure 2 - Two versions of charts used to urine in the 12th and 13th century. [Reprinted from Armstrong, 2007; Cahill, 2015].....7
Figure 3 - Analysis workflow in untargeted metabolomic studies. [Reprinted from Alonso et al., 2015] 11
Figure 4 - Percentage of VOCs measured in different biological matrices based on collected data from cancer-free individuals. [Adapted from Broza et al., 2015].....13
Figure 5 – Scheme of GC-MS components. [Reprinted from Emwas et al., 2015].....16
Figure 6 - Confusion matrix and performance metrics (sensitivity, specificity, and accuracy) calculated. [Adapted from Fawcett, 2006].....38
Figure 7 - Representative HS-SPME-GC/MS full scan chromatograms of **(A)** VOCs and **(B)** VCCs present in urine. **1)** 2-Pentanone; **2)** Hexanal; **3)** 2-Heptanone; **4)** 4-Methyl-2-heptanone; **5)** 1,2,4-Trimethylbenzene; **6)** Nonanal; **7)** α-Terpineol; **8)** 2,5-Dimethylbenzaldehyde; **9)** Formaldehyde; **10)** Acetaldehyde; **11)** Acetone; **12)** 2-Butanone; **13)** 3-Methylbutanal; **14)** 4-Heptanone.....39
Figure 8 - PCA scores scatter plots obtained for **(A)** VOCs and **(B)** VCCs profiles of urine of all subjects included in this study (controls *n*=60 and BC *n*=60, ●) and QCs (*n*=30, ●).40
Figure 9 - (a) PLS-DA scores scatter plots obtained, after variable selection, for **(A)** VOCs and **(B)** VCCs profiles of urine of controls (*n*=60, ●) and BCa patients (*n*=60, ●); **(b)** Permutation plots obtained for the PLS-DA model of VOCs and VCCs; **(c)** ROC curves obtained for VOCs and VCCs profiles.....42
Figure 10 - Boxplots of the compounds the most statistically significant for discrimination (*n*=11) established as a biomarker panel for control (*n*=60, ●*) vs* BCa (*n*=60, ●).....47
Figure 11 - ROC analysis. **A)** ROC curve from both VOCs and VCCs found to be statistically significant for discrimination (identified *n*=16; unknowns *n*=14), with calculated sensitivity (85%), specificity (82%) and accuracy (83%). **B)** Predicted class probabilities of each samples across 100 cross validations. **C)** Confusion matrix. 49
Figure 12 - Heatmap spearman correlation of the statistically significant compounds with a |r| ≥ 0.5 from the comparison of BCa patients and controls 49
Figure 13 - A) PLS-DA scatter plot for VOCs analysis by comparing stage 0a (*n*=29, ●), stage I (*n*=19, ●), and stages II, III and IV (*n*=12, ●) of BCa. **B)** PLS-DA scatter plot for VCCs analysis by comparing stage 0a (*n*=29, ●), stage I (*n*=19, ●), and stages II, III and IV (*n*=12, ●) of BCa.....51

Figure 14 - A) PLS-DA scatter plot for VOCs analysis by comparing stage 0a (*n*=29, ●) *vs* stage II, III and IV (*n*=12, ●). **B)** ROC curve of VOCs by comparing stage 0a *vs* stage II, III and IV with calculated AUC (0.868), sensitivity (75%), specificity (93%) and accuracy (88%).....52

Figure 15 - ROC analysis. **A)** ROC curve of the discriminant VOCs that were found to be statistically significant for discrimination (11 identified metabolites and 2 unknowns) by comparing stage 0a (*n*=29) *vs* stage II, III and IV (*n*=12) with calculated AUC (0.830), sensitivity (50%), specificity (93%) and accuracy (80%). **B)** Predicted class probabilities of each samples across 100 cross validations. **C)** Confusion matrix. 54

Figure 16 - Boxplots of the compounds the most statistically significant for discrimination (*n*=10) established as a biomarker panel for stage 0a (*n*=29, ●) *vs* stages II, III and IV (*n*=12, ●). 55

Index of tables

Table 1 - Percentage of cases and 5-year relative survival by stage at diagnosis of BCa based on SEER database [Adapted from National Cancer Institute, 2018a]. 5
Table 2 - State of the art of BCa biomarker discovery studies focused on VOCs analysis. 20
Table 3 - Data obtained from both BCa and control patients. 32
Table 4 - Analytical conditions for VOCs and VCCs extraction. 34
Table 5 - Pre-processing parameters for VOCs and VCCs analysis. 36
Table 6 - Metabolites that were found to be statistically significant for discrimination, present in urine of BCa patients and controls. 44
Table 7 - Spearman's correlation indexes and corresponding *p*-values obtained for age with the set of metabolites found altered in BCa compared to controls. 48
Table 8 - Metabolites that were found to be statistically significant for discrimination between stage 0a *vs* stage II, III and IV present in urine of BCa patients (*n*=13). 53

List of abbreviations

AJCC	American Joint Committee on Cancer
ANN	Artificial neural network
AUC	Area under the curve
BC	Before Christ
BCa	Bladder cancer
BPH	Benign prostatic hypertrophy
CC	Colon cancer
CDA	Canonical discriminant analysis
CE	Capillary electrophoresis
CI	Chemical ionization
CIS	Carcinoma in situ
CO_2	Carbon dioxide
COD	Cancer odor database
CRC	Colorectal cancer
CT	Computed tomography
CVV	Cross-validation value
EI	Electron impact
FDR	False discovery rate
FID	Flame ionization detector
FPD	Flame photometric detector
GC	Gas chromatography
GC-MS	Gas chromatography coupled with mass spectrometry
HCA	Hierarchical clustering analysis
HNSCC	Head and neck squamous cell carcinoma
HS-SPME	Headspace solid-phase microextraction
IPO	Portuguese Institute of Oncology
LC	Liquid chromatography
LDA	Linear discriminant analysis
LOD	Limit of detection
LOOCV	Leave-one-out cross validation
LOQ	Limit of quantification

m/z	Mass-to-charge ratio
MIBC	Muscle-invasive bladder cancer
MOS	Metal oxide semi-conductor
MOSFET	Metal oxide semi-conductor field-effect transistor
MRI	Magnetic resonance imaging
MS	Mass spectrometry
MSI	Metabolomics standards initiative
MVA	Multivariate analysis
NaCl	Sodium chloride
NIST	National Institute of Standards and Technology
NMIBC	Non-muscle-invasive bladder cancer
NMR	Nuclear magnetic resonance spectroscopy
OPLS-DA	Orthogonal partial least square discriminant analysis
PC	Prostate cancer
PCA	Principal component analysis
PFBHA	O-(2,3,4,5,6-pentafluorobenzyl)hydroxylamine
PLS-DA	Partial least square discriminant analysis
ppb	Part per billion
ppt	Part per trillion
QC	Quality control
QCM	Quartz crystal microbalance
RF	Random forest
ROC	Receiver operating characteristics
ROS	Reactive oxygen species
RSD	Relative standard deviation
SCC	Squamous cell carcinoma
SD	Standard deviation
SEER	Surveillance, Epidemiology, and End results
SIFT-MS	Selected ion flow tube mass spectrometry
SQ	Single quadrupole
SVM	Support vector machine
TCC	Transitional cell carcinoma
TNM	Tumour, nodes, metastasis
TUR-BT	Transurethral resection of bladder tumour

UC	Urothelial carcinoma
UK	United Kingdom
VCCs	Volatile carbonyl compounds
VIP	Variable importance to the projection
VOCs	Volatile organic compounds
VOMC	Volatile organic metabolites of cancer

Chapter 1. Introduction

Cancer represents one of the biggest causes of death every year around the world. The Cancer Research UK reported that there were 17 million new cases of cancer in 2018, while the most common types have been lung, female breast, bowel and prostate cancer (Cancer Research UK, 2018). Regarding the mortality rate, in 2018 it was estimated that 9.6 million deaths were associated with cancer worldwide, for which lung, stomach and bowel were the most common causes of cancer deaths (Cancer Research UK, 2018). In Portugal, cancer mortality accounts for 253.7 deaths per 100.000 of malign tumours, being prostate, lung and bowel the most frequent cancers in men, and breast, bowel and thyroid the most frequent cancers in women (Bento, 2010).

The National Cancer Institute defines cancer, as a collection of diseases in which abnormal cells can divide and spread to nearby tissue (National Cancer Institute, 2018b). Statistical data from 1975 to 2015 has registered some differences toward cancers incidence while some tend to decrease through the years like stomach, larynx, ovary, Hodgkin lymphoma and oesophageal squamous cell carcinoma, others have been increasing, namely liver and intrahepatic bile duct, melanoma, thyroid and myeloma cancer (National Cancer Institute, 2018b). These results can be explained by two main reasons, namely aging and population growth which, consequently, is related with cancer incidence and deaths, once the majority of cancer deaths occur in people over 50 years old (Roser and Ritchie, 2018).

Several studies have been developed over the years to elucidate the causes and mechanisms of cancer itself, the way it develops, grows, and metastasizes. In terms of cancer diagnosis, several techniques have been applied like blood tests, imaging tests (X-ray, magnetic resonance imaging, computed tomography), colonoscopy and mammography, but some of these techniques are invasive, expensive and require skilled technicians for its management (Sun et al., 2016). Besides that, the fact that many cancers do not present any type of symptoms for long periods of time and are only detected in its advanced stage, makes the early diagnosis more difficult to establish and consequently delays the onset of necessary treatments (Roser and Ritchie, 2018).

Despite the increase of global cancer deaths, a decreasing tendency has been observed for several cancers due to earlier detection, improved treatment strategies and more informed population about risk factors (Roser and Ritchie, 2018). Still, more research is needed, particularly to discover more accurate, specific and cost effective methods that can detect cancer in the early stages thus enhancing treatment effectiveness and survival rate (Rodrigues et al., 2018).

1.1. Bladder cancer

1.1.1. Incidence and risk factors

Bladder cancer (BCa) is the tenth most common form of cancer worldwide and the thirteenth with the highest mortality rate worldwide, consisting of about 4% of visceral cancers in developed countries. It affects 3 to 4 times more males than women, with a bigger incidence in those aged with 55 years and older (Saginala et al., 2020). In 2018, according to the Globocan database, BCa accounted for 199.922 deaths worldwide (Manzi et al., 2020). However, even if the cancer-related mortality has decreased for women, it still remains unchanged for men (DeGeorge et al., 2017).

There are several risk factors associated with BCa, such as older age, smoking, pelvic radiation, chronic bladder infection caused by *Schistosoma haematobium*, family history of BCa, among others (Inamura, 2018; Richters et al., 2020). Besides that, the statistics show that white individuals have more probability to develop BCa than African or Hispanic Americans (American Cancer Society, 2019c). Regarding smoking, this risk factor represents one of the most important in BCa, since smoking increases the risk of developing BCa by 3 times compared with non-smokers (American Cancer Society, 2020).

1.1.2. Classification and diagnosis

BCa can be classified into carcinoma and sarcoma. Bladder carcinoma can be classified into transitional cell carcinoma (TCC) or urothelial carcinoma (UC), which are the most frequent, representing 90% of bladder carcinoma. In these types of cancers, the epithelial cells are the main cells in the origin of cancer. Another type of bladder carcinoma is the squamous cell carcinoma (SCC) accounting for 1 to 2% of all BCas (in the US, but can change depending on countries), in which cells can develop in response to inflammation and become cancerous (American Cancer Society, 2016; Health Engine, 2004). Adenocarcinoma represents a rare type of bladder carcinoma accounting for 1% of all BCas. Sarcoma are the rarest types of BCas that begins in the muscle layers of the bladder (American Cancer Society, 2019c; Cancer.net, 2019).

Symptoms of BCa include mainly microscopic haematuria without pain, that can exacerbate with the higher stage of malignancy, urinary voiding associated with irritation and reduced bladder capacity. However, establishing a diagnosis based on such symptoms can be challenging once the amount of blood in the urine might be reduced or not present in the moment of collection and therefore, not visible in urine tests. In those cases, a confirmation by other complementary exams is required and includes tests like urine cytology with microscopic examination, cystoscopy, biochemical/molecular tests, and imaging of upper urinary tract by computed tomography (CT) (Oeyen et al., 2019; Razmaria, 2015). However, BCa diagnosis is challenging once symptoms only appear when the disease is already in its advanced stages. The difficulties associated with some diagnostic methods are due to invasiveness and difficulties on interpretation or even its high costs are also important barriers to BCa diagnosis (Oeyen et al., 2019). For example, cystoscopy, besides having a general good sensitivity, (75 to 100%), is an invasive method that might require anaesthesia in some cases, and misses certain types of tumours as well as it presents risks of infection and urethral damage (Zhu et al., 2019).

On the other hand, urine cytology can be performed non-invasively with an associated low cost and high specificity (73 to 100%), but it presents low sensitivity (20-40%) for low stages of urothelial tumours, and can be affected by different instrumentation artifacts (Goodison et al., 2013; Oeyen et al., 2019; Zhu et al., 2019). In addition, some existing conditions like renal calculi and urinary tract infections can lead to false-positive results in urine cytology tests (DeGeorge et al., 2017).

Biochemical and/or molecular tests include urine marker-based tests, for instance, UroVysion test, ImmunoCyt test or NMP-22 test, that consist on the identification of certain chemical compounds, proteins and modifications in chromosomes in urine can also be used (e.g., detection of aneuploidy for chromosomes 3, 7, and 17). These tests present a wide range of sensitivity (45 to 100%) and can detect low grade tumours. However, the high cost, invasiveness and inter-individual variations are some disadvantages of these techniques (Goodison et al., 2013; Zhu et al., 2019).

Imaging techniques, such as magnetic resonance imaging (MRI), are employed specially for the establishment of staging. MRI presents no prejudice since no radiation is involved, however it is a high cost technique and the process might sometimes cause some discomfort to the patient (Zhu et al., 2019).

CT, just like MRI, is commonly employed for staging when BCa has already been diagnosed. This method presents very high specificity (78 to 100%), as well as a viewing ability of multiple planes. The inconvenient rely on the exposure to radiation and possible allergic reactions associated with the contrast agent (Zhu et al., 2019).

1.1.3. Staging

Based on diagnostic, BCa staging is most often based on the American Joint Committee on Cancer (AJCC) TNM (tumour, nodes and metastasis) system (**Figure *1***) (Dana-Farber Cancer Institute, 2019). TNM comes from tumour, nodes and metastasis and intends to describe how far and whether the tumour has grown through the bladder, how lymph nodes near the bladder have been affected and how distant cancer has metastasized to other organs. Besides that, this information is also complemented by stage grouping that goes from stage 0 to stage IV (American Cancer Society, 2019a). The stage 0 or carcinoma in situ (CIS) consists of cancer cells present in the most superficial layer of the mucosa and is considered non-invasive. On the stage I or non-muscle-invasive bladder cancer (NMIBC), the cancerous cells are already present in the deeper layers of the mucosa which constitutes around 50 to 70% of the cases. In the stage II or muscle-invasive bladder cancer (MIBC), that comprehends 10 to 20% of the cases, the cancer cells have already invaded the muscle layer (Inamura, 2018; Manzi et al., 2020). In stage III, the cancerous cells have spread to the closest organs from the genitourinary tract like the prostate, the seminal vesicles, the uterus, or vagina. Lastly, in the stage IV (metastatic phase), the tumour already spread to the further organs such as lymph nodes, bone, lung, liver and peritoneum (DeGeorge et al., 2017; Instituto CUF de Oncologia, 2017).

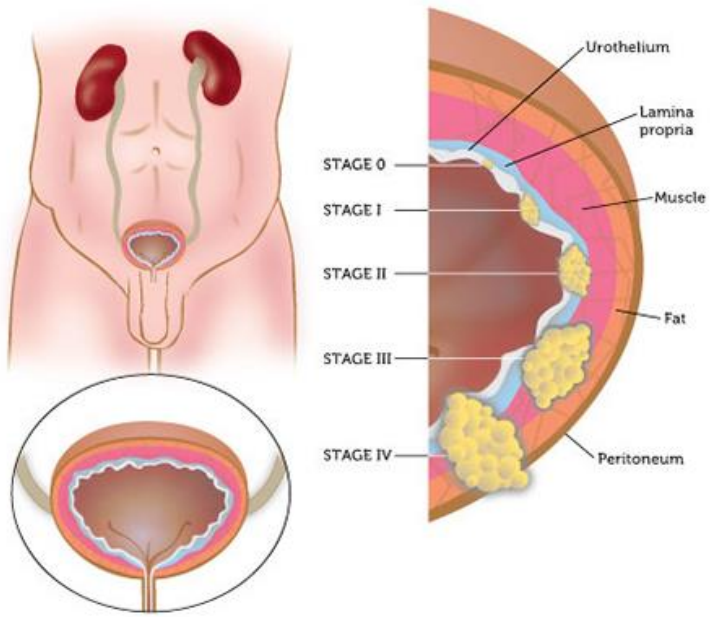


Figure 1 - Classification of BCa according to its staging [Reprinted from Dana-Farber Cancer Institute, 2019].

Another type of classification is adapted by the Surveillance, Epidemiology, and End Results (SEER) database that groups cancer into 3 different categories based on how far the cancer has spread (National Cancer Institute, 2018a). Those categories are: 1) localized, which considers that cancer has not spread outside the bladder; 2) regional, when cancer has spread from the bladder to closed organs; and 3) lymph nodes and distant, in which cancer has spread to more distant organs (American Cancer Society, 2019b). In situ cancer stages represent the biggest percentage of cases with 51%, being the distant stages and the non-classified or unknown stages the ones with the smallest percentage of cases going from 3 to 5%. The 5-year relative survival rate can be calculated by the percentage of patients with BCa alive, five years after diagnosis, divided by the percentage of the general population of corresponding sex and age alive after five years. Considering this 5-year relative survival rate, the in situ stage (95.8%) and the localized stage (69.5%) represent the highest rates of survival, contrary to the regional (36.3%) and the distant stage (4.6%) (**Table *1***) (National Cancer Institute, 2018a). The correct establishment of staging is crucial to plan and initiate the appropriate treatment and consequently a good prognosis.

Table 1 - Percentage of cases and 5-year relative survival by stage at diagnosis of BCa based on SEER database [Adapted from National Cancer Institute, 2018a].

Stage	***Percent of cases (%)***	***5-year relative survival rate***
In situ	51	95.8
Localized	34	69.5
Regional	7	36.3
Distant	5	4.6
Unknown	3	47.3

1.1.4. Treatment

Treatment of BCa is established based on the staging of the disease which means that the sooner it is diagnosed, the better are the chances to treat and consequently the prognosis. When the cancer is still on the first stages (stage 0 and I) and has not invaded

the muscle layers, the removing of the tumour is possible through cystoscopy, followed by intravesical immunotherapy (with bacilli Calmette-Guérrin – BCG) or chemotherapy (mitomycin C, epirubicin or doxorubicin) that allows to eliminate any remaining cancerous tissue and prevent any recurrence or progression. If cancer already invaded the muscle layers, the common treatment employed is complete cystectomy with chemotherapy (cisplatin-based neoadjuvant) and lymphadenectomy. However, for some patients, the practice of chemotherapy and radiation might be a solution rather than bladder removal. For patients with metastatic disease or unresectable BCa, chemotherapy is the treatment employed using a wide range of combination or isolated drugs like gemcitabine, cisplatin, methotrexate, vinblastine and/or doxorubicin (DeGeorge et al., 2017; Razmaria, 2015).

A solution for the substitution of the bladder consists of an ileal conduit, which will introduce an artificial opening (stoma) in the abdominal wall. That conduit will carry urine through the ureters until an external bag attached to the skin (urostomy). Another solution consists on the employment of a portion of intestine to create a "new" bladder capable to contain urine (neobladder) (DeGeorge et al., 2017; Razmaria, 2015). Once BCa has a high rate of recurrence, follow-up after treatment is essential.

Although many diagnostic methods are employed for the detection of BCa, with overall good results, it is clear that it still exists a lack of biomarkers for the establishment of an earlier diagnosis of this type of cancer. For this reason, the study of volatile organic compounds (VOCs) has demonstrated to be a promising approach as diagnostic markers which will be reviewed in the next chapters.

1.2. From smell perception to disease recognition

Since ancient times, men started to establish methodologies that allow to identify the presence of multiple pathologies. In the middle age, urine became a highly studied matrix used as the base of many diagnostics. Hippocrates formulated, in the fourth century BC (before Christ), the humoral theory (blood, phlegm, yellow bile and black bile), which considered that illness had a natural cause and that the achievement of a health state consisted in an equilibrium of those humours. His theory has last for centuries. From 100 BC, 20 different types of urine were described, and some ancient Chinese doctors and Hindu cultures found out that ants were attracted by some urine of patients that presented a sweet flavour (known today as diabetes mellitus). In the 12th and 13th century, physicians started to formulate "urinary charts" (at first only in latin) (***Figure 2***) based on the characteristics of urine, like colour, taste and smell, and related these organoleptic characteristics with diseased patients and healthy individuals. Practice of "uroscopy" was

a guide for physicians during thousands of years, meanwhile, the technological advances have given rise to better approaches over time. This term has changed in the mid-17th-18th century becoming called by urinalysis and practiced in our days as a more efficient way to make diagnosis (Armstrong, 2007; Banday et al., 2011; Lagay, 2002).

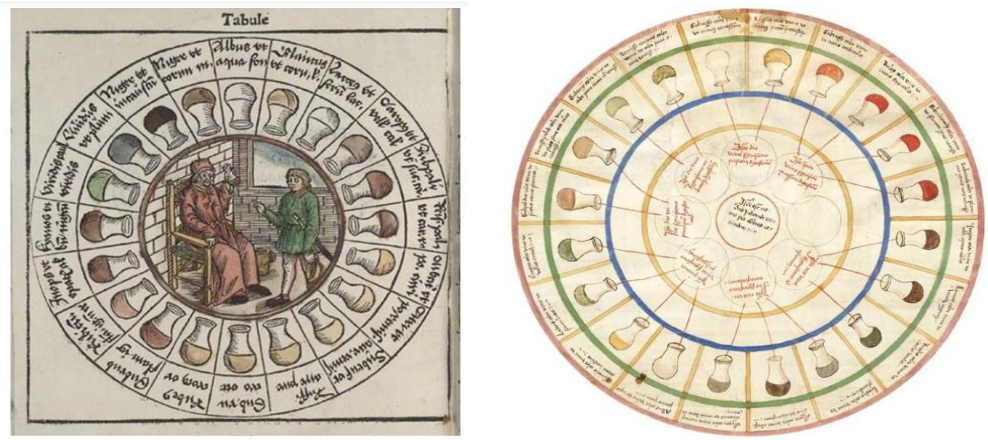


Figure 2 - Two versions of charts used to urine in the 12th and 13th century. [Reprinted from Armstrong, 2007; Cahill, 2015].

Due to their extremely accurate sense of smell, dogs' capacities have been used in hunting and a variety of inspection and security procedures especially by operational teams in the detection of explosives, drug traffic like narcotics, finding missing people in rescue missions and other situations. The ability of dogs to detect VOC concentrations at one part per trillion (ppt) has enhanced the curiosity of many investigators in relation to the possibility to detect diseases (Angle et al., 2016). One of the first reports about dog detection of cancer was published in 1989 when a dog apparently "sniffed at a mole on the leg of the dog handler" which turn out to be a melanoma (Teodoro-Morrison et al., 2014). In other studies where dogs were trained for cancer detection, they were able to identify in different samples, like breath and stool, distinct cancer types (e.g., colon, breast and lung) with impressive sensitivity and specificity going from 88% (minimum) to 99% (maximum) (Angle et al., 2016; Petry et al., 2015). Besides that, other investigators also verified that apart from detecting cancer, dogs could also detect other diseases like *Clostridium difficile* infections (Teodoro-Morrison et al., 2014). These studies have demonstrated that dogs and specially trained dogs present a great ability to detect different kinds of VOCs with high levels of confidence for diagnostic purposes (Angle et al., 2016). Furthermore, other animals like rats and pigs have been pointed in the

literature for their olfactory sensitivity, especially used to detect landmines in the case of rats but also detection of food poisoning, as well as pigs that can localize truffles underground (Teodoro-Morrison et al., 2014). However, dogs present some advantages over other animals regarding medicine and healthcare purposes, based essentially on social acceptance and relative ease of training. Besides that, laboratory equipment capable to detect VOCs is not liable to be transported to the field and consequently, the analysis of those compounds in real time and from its source becomes an obstacle (Angle et al., 2016; Teodoro-Morrison et al., 2014). Although these findings and studies present good results, it is important to keep in mind the difficulties and potential misleading's of animals in such activities.

In this way, VOCs analysis has been presented as a potentially promising diagnostic tool, being the focus of many cancer research studies in the last years (de Boer et al., 2014; el Manouni el Hassani et al., 2018; Schmidt and Podmore, 2015a; Shirasu and Touhara, 2011). It has been shown that patients with cancer tended to exhibit different and specific VOC patterns which were different from cancer free individuals. These findings suggested that VOCs can be used to detect cancer in its early stages (Jin, 2018).

1.2.1. Potential of volatile organic compounds as diagnostic markers

VOCs are molecules with relatively low molecular weight (50-1500 Da) and high vapor pressure (50 to 260 °C) that can be originated from cells (Berenjian et al., 2012; Fleming-Jones and Smith, 2003; Sun et al., 2016). The all set of VOCs produced by an organism is called volatilome (Dragonieri et al., 2016; Shirasu and Touhara, 2011).

The study of VOCs brings great potential for diagnosis of cancer in general, as well as for BCa, since cancerous cells have a different metabolism compared to normal cells, thus producing different compounds that can be released for example by exhaled breath or urine. However, it must also be noted that the study of VOCs might present some confusion effects, for example, in the body, VOCs are emitted naturally and tend to suffer variation depending on diet, gender, environmental influence, lifestyle and physiological modifications. VOCs are released from cells in part as a result of metabolic processes, but not all of them are related to metabolism. The presence of bacteria in different parts of the organism such as mouth, lungs and digestive tract can also produce VOCs (Angle et al.,

2016). The analysis of these compounds offers a perspective of the physiological and pathophysiological mechanisms that occur in healthy and diseased individuals (Jin, 2018).

1.2.2. Metabolomics applied to VOCs profiling

Metabolomics is one of the most representative research fields on the analysis of VOC biomarkers. This “omic” approach comprehends the analysis of all metabolites involved in the metabolic networks of living organisms and their response to pathophysiological conditions or other stimuli (Nicholson et al., 1999). Metabolomic studies can, in general, be developed using several techniques like nuclear magnetic resonance (NMR), liquid chromatography (LC) and gas chromatography (GC) coupled to mass spectrometry (LC-MS and GC-MS), and Capillary Electrophoresis (CE). For VOC profiling, GC-MS is one of the most used analytical techniques due to its characteristics and advantages, namely high-efficient separations, low limits of quantification and high specificity, as well as the identification of sample constituents through the use of mass spectral databases (Fiehn, 2016; Mastrangelo et al., 2015). The identification of sample constituents by its mass spectra is obtained by breaking molecules, using electron impact (EI) and chemical ionization (CI) techniques, leading to the identification of compounds. The equipments used for these analysis are GC-MS with single and/or a triple quadrupole (Chauhan et al., 2014).

1.2.2.1. Metabolomics strategy

The metabolomics workflow (***Figure 3***) includes a variety of steps starting by choosing the biological sample for analysis, that can be blood, urine, exhaled breath or any other biological sample. The next step is focused on the selection of the analytical method to be used, being GC-MS the more appropriated for VOCs detection (Alonso et al., 2015). GC-MS presents high sensitivity and an easy metabolite identification due to the several databases available, however, it’s a destructive technique with low reproducibility (Bujak et al., 2014). This method will traduce information through spectral data that consequently has to be pre-processed in order to manage some corrections inherent to matrix and equipment such as baseline correction, noise filtering, peak detection, deconvolution, alignment and normalization (Alonso et al., 2015).

The aim of data filtering consists to remove noise or non-informative variables in a specific data set. Baseline correction intends to identify and remove those variables that are very small (above the detection level) or close to zero (Chong et al., 2019). Another way of data filtering focus on the removal of variables with high relative standard deviation (RSD=SD/mean), which can be detected by comparing analysis with the quality control (QC) samples (Chong et al., 2019). Following data filtering, normalization is required in order to minimize any systematic bias and promote the consistency of data. Normalization procedures will adjust samples to allow possible comparison with each other. These procedures can be made by different methods whose main ones are: 1) normalization by sum which consists on the division of the area of each peak in the chromatogram by the total sum of all peaks; 2) normalization by reference sample and 3) normalization by reference feature which consist on choosing a specific characteristic like, for example, creatinine in the case of urine samples, to adjust metabolite concentrations due to the inherent variation of dilution among different samples (Chong et al., 2019; Wu and Li, 2016). After that, data scaling is necessary to promote comparability between variables, which can be: 1) auto scaling or unit variance scaling, that gives equal importance to all peaks, however this means that peaks with values close to the level of detection or the baseline will present the same weight as those with higher peaks, which can lead to consider noisy signals and non-informative features, as relevant as the interesting ones; 2) pareto scaling aims to give a relative importance to the different chromatographic peaks taking into account the mean-centered data and 3) range scaling that tends to give most importance to the highest peak areas. However it happens that in some techniques like GC-MS, the peak area is sometimes not proportional to the concentration of the compound, which means that peaks with a smaller area but biologically more significant can be ignored (Sussulini, 2017).

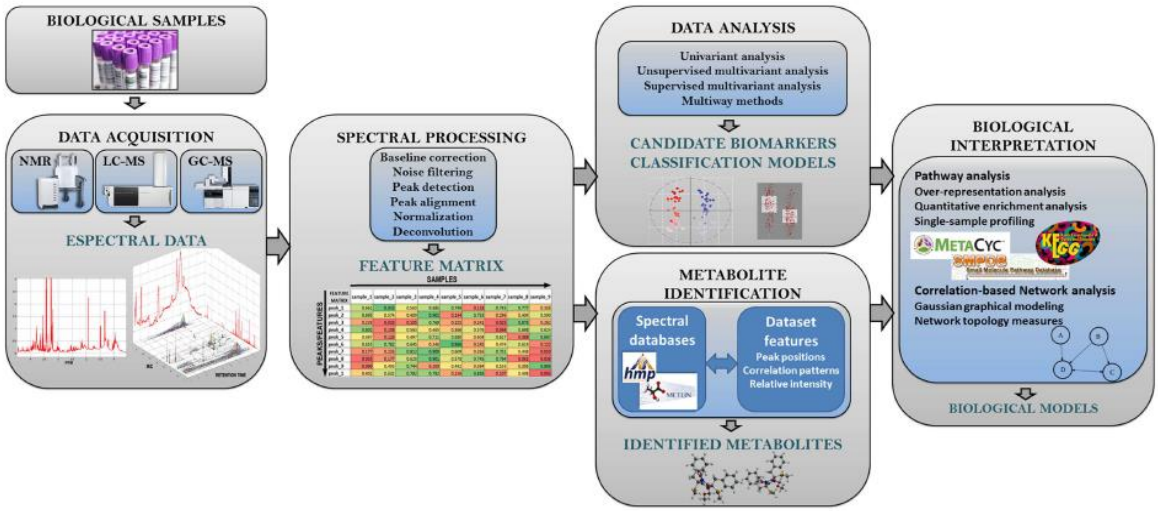


Figure 3 - Analysis workflow in untargeted metabolomic studies. [Reprinted from Alonso et al., 2015]

After sample pre-processing, a huge quantity and complex analytical data is obtained which represents the need to use methods of data analysis. There is a large number of methods for data analysis, like multivariate that can be divided into unsupervised and supervised. With the information obtained by multivariate analysis, it is possible to start to elect biomarkers and elaborate classification models that can lead to further investigations and facilitate metabolite identification (Alonso et al., 2015). Metabolite identification is a very important step in biomarker discovery, which consists on comparison of spectral data with spectral databases created with standard compounds. Thus, spectral databases are essential and vary on the quality of the stored data and the quantity of metabolite spectra available (Alonso et al., 2015). Finally, pathway and network analysis are usually used to understand the dysregulations occurring in metabolic pathways. Pathway analysis employs a prior knowledge about the metabolite identification and analyse metabolite patterns. On the other hand, network analysis takes the correlation existing between the set of measured metabolites and characterizes the complex relationship existing between them (Alonso et al., 2015).

These strategies are highly applied in general metabolomic studies, but also for VOCs profiling. For VOCs identification, the National Institute of Standards and Technology (NIST) database has been widely used. In addition, new databases have emerged due to the increasing number of studies reporting alterations in VOCs related with cancer, namely the cancer odour database (COD) (Janfaza et al., 2019b). The aim of COD is to provide information about volatile organic metabolites of cancer (VOMC) that can turn into

a significant tool for cancer diagnosis, research, and investigation. This information comprehends the collection of results presented in more than 1000 scientific publications which are evaluated and gathered into a structured database. So far, this database includes 446 different compounds and 15 different types of cancer, as well as many other features associated with chemical properties of volatiles, sample matrices, cancer type, among others (Janfaza et al., 2017, 2019b). Currently, other teams are also working in the development of databases with additional information on the VOC profiles, namely for lung cancer, using exhaled breath, colorectal cancer (CRC) and heart failure as well (Jin, 2018). All these strategies aim to achieve a last step that converges in biological interpretation establishing the dysregulated pathways to improve understanding of oncologic processes at metabolic levels (Alonso et al., 2015).

1.2.2.2. Biological matrices used in metabolomics

The detection of VOCs can be performed, as previously discussed, in a variety of samples and body fluids like blood, urine, exhaled breath, faeces, sweat, among others. Each one presents some advantages and disadvantages that must be considered in terms of accessibility, ease of analysis, interfering compounds and complexity, and type of cancer (Nakhleh et al., 2017; Sun et al., 2016).

VOCs are present in different body fluids/matrices like exhaled breath, urine, skin, saliva, faeces, blood and milk. For instance, in a review from Broza et al., 34% of VOCs in a total of 2577 compounds measured in all types of sources (e.g., breath, saliva, faeces, saliva and blood), were identified in exhaled breath, followed by skin with 20% and blood representing the lowest percentage with 6% (***Figure 4***) (Broza et al., 2015). In another review from Lacy Costello et al., a total of 1840 VOCs has been measured, where breath also represents the matrix with the bigger number of identified VOCs (872), followed by skin (532), faeces (381), saliva (359), urine (279) and blood (154) (De Lacy Costello et al., 2014; Filipiak et al., 2016).

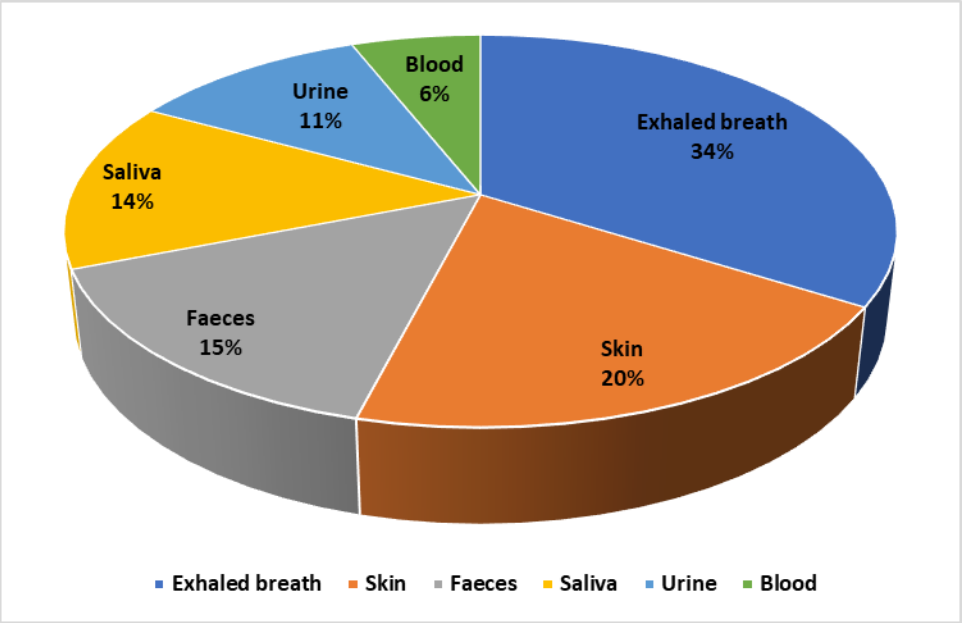


Figure 4 - Percentage of VOCs measured in different biological matrices based on collected data from cancer-free individuals. [Adapted from Broza et al., 2015]

Exhaled breath is composed by many constituents like oxygen, carbon dioxide, nitrogen, water, inert gases and over 3000 different VOCs (Fernandes et al., 2015; Li et al., 2017). Some of those VOCs expelled through breath are basically aromatic, aliphatic and chain hydrocarbons (Fernandes et al., 2015; Oguma et al., 2017). In the past, the different smells of exhaled breath were indicators of many diseases, namely fruity smell was indicative of diabetes or fish-like smell associated with kidney diseases (Fernandes et al., 2015). With the technologic advancement, new methods for collection of breath arisen, making exhaled breath as a non-invasive and accessible sample with low complexity, safer to manage, and suitable for different age groups (Fernandes et al., 2015; Nakhleh et al., 2017; Sun et al., 2016).

Most works developed in the last years are based on exhaled breath as the sample of choice for several cancer detection studies (Jalal et al., 2018). Analysis of exhaled breath can be performed with different types of techniques and equipments, such as electronic nose (e-nose/EN) exclusively, e-nose and GC-MS, and GC-MS exclusively (Krilaviciute et al., 2015). Altomare et al. compared exhaled breath samples of CRC patients and healthy individuals, analysed by GC-MS, and obtained results of sensitivity and specificity of 86% and 83%, respectively, with accuracy levels of 76% (de Boer et al., 2014). From a systematic review elaborated by Krilaviciute et al., the best diagnostic performance was reported in studies on head and neck cancers, malignant mesothelioma and gastric cancer (Krilaviciute et al., 2015). Additionally, breath testing has also been the object of other applications such as urea breath tests for detection of *Helicobacter pylori* and nitric

oxide breath concentrations for evaluation of bronchial asthma severity (Bernabei et al., 2008).

Skin secretions/sweat, for example, account with a wide variety of 532 VOCs including compounds of hydrocarbons, alcohols, esters, amines, ketones, among others. Studies on skin volatiles are applied to different fields like cosmetics, clinical diagnostics, ecology associated to insect vectors, and forensics (Duffy and Morrin, 2019). Besides the great quantity of VOCs, studies on skin volatiles require techniques with high sensitivity and pre-concentration steps to concentrate the sample and consequently allow the detection of the compounds. Additionally, the site of sampling collection, the patient age and daily care products can reveal differences in volatile compounds, thus, methods for sample collection and development of more exhaustive studies should be done to minimize potential interferent factors (Duffy and Morrin, 2019; Gallagher et al., 2008; Jadoon et al., 2015).

Studies on faeces headspace have also been developed, while less than urine or exhaled breath. VOCs detection in faeces has become a very promising approach for early CRC detection. Identification of VOCs in this sample has yet to be carried out once many volatiles associated to CRC are also present in other types of samples like blood, breath, and urine. Thus, as this specimen is the result of processes associated to digestion and are influenced by the gut microbiome, it is necessary to understand in which way volatiles are the result of changes associated with CRC (Chan et al., 2016).

Other samples like saliva have also been the object of some studies for VOCs detection (Jalal et al., 2018; Soini et al., 2010). However, this matrix presents few volatiles compared to exhaled breath or sweat, once it can suffer changes due to several factors associated with diet or consumption of drugs. While saliva is a non-invasive sample to collect, challenges related with this sample focus on distinguishing between endogenous and exogenous compounds, obtain a suitable amount of it, which can be difficult in certain population groups like infants or geriatric patients, or those who have reduced or absent production of saliva. Additionally, the methodology used to collect saliva can also influence the VOC analysis (Jalal et al., 2018; Soini et al., 2010).

Urine has been used since the beginning of times as an attempt to diagnose diseases presenting several utilities and advantages. Through the capacity of kidney filtration, urine is produced and accumulates several residues from blood and other components from distant organs, giving to this sample potential for the development of new diagnostic methodologies. Regarding its advantages, urine is easy to obtain and in large amounts, and it does not require invasive methods for its collection (Bax et al., 2018). Additionally, urine presents a greater diversity of VOCs from different classes (heterocyclic compounds, alcohols, ketones, amines, aldehydes, organic acids and distinct

hydrocarbons) comparing with other matrixes like exhaled air or blood (Jalal et al., 2018; Mazzone et al., 2015). Studies have suggested that it exists high variability among individuals including gender, age, diet, lifestyle and hormonal status (Bax et al., 2018; Mazzone et al., 2015).

Some specific studies based on blood volatiles have also been performed, for example, for CRC, once VOCs are released primarily into de bloodstream before passing to the alveoli and be exhaled (Jalal et al., 2018; Wang et al., 2014). Other studies also intended to evaluate the relation between blood concentrations of VOCs and its correlation with the rate of exhalation once VOCs from exhaled breath should be similar to the ones present in blood, however, exhaled breath can suffer variation in certain cases specially in those patients affected with lung pathologies or having smoking habits (increase of benzene and toluene levels) (Jalal et al., 2018; Jia et al., 2019). Although, blood is an invasive sample to collect, it can give further information about metabolic changes that occur in the organism associated with tumorigenic processes in such specific situations (Jalal et al., 2018; Wang et al., 2014).

1.2.2.3. Analytical techniques and data analysis

In this section, both analytical techniques and data analysis for the determination of volatile compounds will be presented, considering techniques like headspace solid-phase microextraction, GC-MS and data analysis like pre-processing, multivariate and univariate analyses.

Headspace solid-phase microextraction (HS-SPME)

Once some sample matrices (e.g., urine) present a small concentration of VOCs, it is necessary to employ some extraction and pre-concentration techniques to enrich metabolites and facilitate its detection and quantification. HS-SPME is the most used technique for VOCs extraction once samples or volatile compounds can be efficiently partitioned into the headspace (HS) gas volume from either a liquid or solid matrix (Chromacademy, 2019b; Schmidt and Podmore, 2015b). The volatile metabolites present in the sample are released into the HS by heating and agitating processes, as well as the addition of a salt (e.g., NaCl) that will promote the salting out of compounds, until reaching an equilibrium state. Then, a fused silica fiber coated with an absorbent polymer is

introduced into the sample container where it will adsorb the VOCs present into the HS, which will be transferred into the GC injector for analysis (Chromacademy, 2019b; Harvey, 2000).

Gas chromatography coupled with mass spectrometry (GC-MS)

GC-MS is a very widely used technique in metabolomic studies for VOC analysis (Harvey, 2000). GC-MS is constituted by a sample injector, a stationary phase (columns that can be variable), a mobile phase (carrier gas), a MS detector and a computational data system (***Figure 5***). Processing of samples is made by injecting the pre-treated sample into the equipment where it is volatilized and the representative portion (VOC content) is carried onto the column by a carrier gas (usually helium) (Chromacademy, 2012). Once in the column, the sample components are separated by differential partitioning and eluted depending upon the chemical affinity of the metabolite for the stationary phase and the analyte vapor pressure. The metabolites are then, carried into the MS ion source where they are submitted to an electron bombardment and fragmented into molecular ions (usually electron impact ionization) by using electric fields in the vapour phase. Ions formed in the previous step are conducted to the MS detector where they are separated and detected depending on their mass-to-charge (*m/z*) ratio (Chromacademy, 2019a). This signal is then amplified and sent to the computational system where chromatograms are constructed (Chromacademy, 2012).

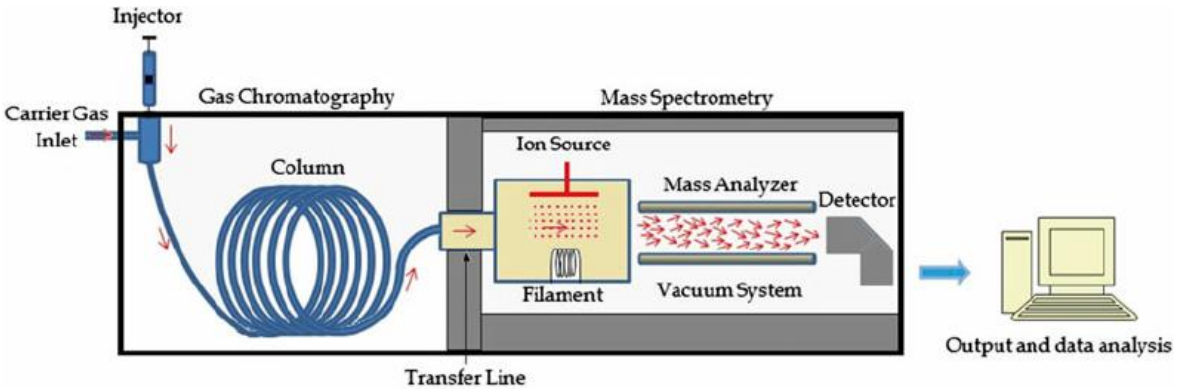


Figure 5 – Scheme of GC-MS components. [Reprinted from Emwas et al., 2015]

As previously mentioned, GC-MS presents several advantages such as, low limit of detection (LOD) and quantification (LOQ) with high resolution separation as MS also provides qualitative and structural information that can help to identify the analytes (Harvey, 2000; Want et al., 2005). On the other hand, the main limitations are long

analysis time, requirement of sample pre-treatment, high costs and need of trained professionals (Harvey, 2000; Want et al., 2005).

Data analysis: pre-processing, multivariate and univariate analyses

The analytical techniques generate a large amount of data, especially when it comes from GC-MS which are very complex. For that reason, methods of pre-processing followed by multivariate and univariate analysis are necessary for better understanding. Pre-processing includes several steps like, 1) peak detection and deconvolution; 2) alignment; and 3) gap filling (Karaman, 2017).

Peak detection consists on the detection of each measured metabolite in a sample and its attribution to the corresponding *m/z* and retention time (RT). In this step, for the deconvolution of the chromatogram peaks, characteristics like the baseline and noise are considered (Karaman, 2017).

The aim of alignment consists on grouping the detected peaks concerning a specific *m/z* and RT window which are next integrated as peak height or peak area and attributed to a feature. Alignment is necessary since the process of chromatographic separation generates some RT shifts that can occur due to mobile or stationary phase, changes in temperature and pressure, or even effects associated with the sample (Karaman, 2017). After alignment, the data table obtained will comprise missing values in some samples. This happens as a result of poorly shaped peaks or peaks with low intensity. In that way, gap filling algorithms will be employed to search those peaks in the raw data based on the defined *m/z* and RT window (Karaman, 2017).

Analytical results obtained from VOCs determination and quantification presents several confounding factors. Those factors include sample composition or complexity as well as characteristics associated with the patient, like age, gender, smoking/drug habits, lifestyle, among others. Besides that, other factors linked to the sample collection such as humidity and temperature can influence the results, thus, the use of multivariate statistical analysis (MVA) can help to determine their influence (Gromski et al., 2014). There are several methods of MVA, which intends to rearrange data, promoting correlation between variables and underlying significant information that appear as a result of metabolomic changes, into a more interpretable and easy set of information (Lubes and Goodarzi, 2017). Such methods can be differentiated into regression or classification. Regression methods can also be separated into linear or non-linear methods. In terms of classification methods, these can be divided into supervised or unsupervised techniques, however, some of those methods can be included in more than one group (Elmasry et al., 2012).

Unsupervised analysis gives information through identification of tendencies, sample grouping and recognition of possible outliers. By the other hand, supervised analysis

promotes classification of groups and enhances potential markers (Mastrangelo et al., 2015). Principal component analysis (PCA) and partial least square discriminant analysis (PLS-DA) are the most commonly used unsupervised and supervised methods, respectively, in metabolomic studies (Engel, 2019).

PCA is considered a fast computing technique that works as a linear unsupervised method, which means that data is assembled and managed that only the most relevant analytical information is obtained, reducing the dimension of data sets (Chen et al., 2013). This method is also used to demonstrate which features originate the greatest variances in the dataset, without considering class labels (Capelli et al., 2016; Weber et al., 2011). Another example of an unsupervised method is the hierarchical clustering analysis (HCA) that can detect non-linear trends in the data that are not conveniently covered by PCA (Alonso et al., 2015).

PLS-DA has also been widely used in medical applications (Zhu et al., 2018b). This technique intends to maximize the covariance between a variable *X* and its corresponding *Y*, allowing to manage highly co-linear and noisy data and providing separation of groups into easily interpretable results (Gromski et al., 2015). Besides that, PLS-DA is mostly applicable to modelling high-dimensional data. PLS-DA is also more flexible than other linear discriminant analysis for example, since it does not assume the data to follow a particular distribution (Lee et al., 2018). Another example of a supervised method is the support vector machines (SVMs), which is able to manage the presence of non-linear relations between the metabolomic data and the variable of interest (Alonso et al., 2015).

Univariate analyses are important when we want to analyse one variable at a time among a group of many others that are measured. The mainly aim of this methodology intends to minimize a possible large number of measured analytes, into one that demonstrates a strongest response under the conditions of the experiment. Examples of univariate methods used include t-test and analysis of variance (ANOVA) (Bartel et al., 2013; Saccenti et al., 2014).

1.3. VOCs as potential biomarkers of bladder cancer: state of the art

In the last years, several works have been developed using urine as a sample for cancer detection studies related to BCa. Studies started by employing trained sniffer dogs to evaluate if volatile biomarkers for BCa existed in urine headspace (Willis et al., 2010). Next, other studies using more advanced techniques like GC-MS conducted both *in vitro* and *ex vivo* experiments to bring knowledge about VOC biomarkers (Cauchi et al., 2016;

Jobu et al., 2012; Khalid et al., 2013; Rodrigues et al., 2018; Spaněl et al., 1999). Then, in the beginning of the 21st century, some groups started to apply the knowledge of metabolomics and its applicability in sensors as a future promising approach for biomedical devices (Bernabei et al., 2008; Heers et al., 2018; Horstmann et al., 2015; van de Goor et al., 2017; Weber et al., 2011; Zhu et al., 2018a) (**Table *2***).

Table 2 - State of the art of BCa biomarker discovery studies focused on VOCs analysis.

Cohort	Sample type	Analytical technique	Data treatment	Results	Ref.
Dogs					
TCC patients (*n*=30); Controls differentiated into 3 groups (C1 - subjects < 33 years, healthy individuals, with no urine abnormality on dipstick analysis; C2 - subjects < 33 years, with any non-cancerous non-urological condition or disease (e.g., psoriasis), and/or one or more positive dipstick finding(s) of a minor nature; C3 - hospital patients of any age with confirmed noncancerous urological disease, with or without urine abnormalities (*n*=180)	Urine	4 Trained dogs	Multilevel logistic regression	Highest sensitivity 73% and 64% when evaluated as a group; 92% specificity when evaluated individually (C1) and 56% specificity for older patients with non-cancerous urological disease.	(Willis et al., 2010)
Metabolomics (GC-MS)					
BCa (*n*=14); PC (*n*=24); Healthy controls (*n*=14)	Urine	SIFT-MS	ANOVA	Higher levels of formaldehyde in the urine of BCa patients (mean 85ppb), followed by PC patients (mean 25ppb) and healthy controls (mean 11ppb)	(Spaněl et al., 1999)

Cohort	Sample type	Analytical technique	Data treatment	Results	Ref.
BCa (*n*=9); Healthy controls (*n*=7)	Urine	GC-MS	PCA	Ethylbenzene, nonanoylchloride, dodecanal, (Z)-2-nonenal and 5-dimethyl-3(2H)-isoxazolone were identified as potential BCa biomarkers	(Jobu et al., 2012)
BCa (*n*=24); Controls (*n*=74)	Urine	GC coupled to a sensor detection system (metal oxide)	LDA; PLS-DA	90% accuracy for BCa and control detection (for PLS-DA); 93% accuracy for BCa and control detection (for LDA)	(Khalid et al., 2013)
TCC of the bladder differentiated into 3 groups (TCC1 – low grade or well differentiated; TCC2 – moderately differentiated; TCC3 – high grade or poorly differentiated) (*n*=72); Control differentiated intro 3 groups (C1 – no urine abnormality on dipstick analysis; C2 – any non-urological non-cancerous condition or disease, and/or one or more positive dipstick findings of a minor nature; C3 – confirmed non-cancerous urological disease, with or without urine dipstick abnormalities (*n*=205)	Urine	GC-MS	PCA; PLS-DA; HCA; RF; SVM	88% accuracy for PLS-DA (C2 vs TCC); 89% accuracy for SVM; 4-heptanone was decreased and 3-hidroxyanthranilic acid increased (C3 vs TCC)	(Cauchi et al., 2016)

Cohort	Sample type	Analytical technique	Data treatment	Results	Ref.
BCa cell lines (Scaber, J82 and 5637); Normal cell lines (SV-HUC-1)	Cell lines	HS-SPME/GC-MS	PCA; PLS-DA; OPLS-DA	2-pentadecanone, γ-dodecalactone and dodecanal seem to be particularly important compounds for BCa detection	(Rodrigues et al., 2018)
Sensors					
BCa (*n*=25); PC (*n*=12); BPH (*n*=29); Other urological pathologies (*n*=33); Control (*n*=18)	Urine	Electronic nose (based on eight QCM gas sensor coated by sensing layers of metalloporphyrins)	PCA; PLS-DA	100% correct classification between ill and healthy group; 100% correct classification between BCa and PC; 100% post-surgery samples recognized as healthy samples	(Bernabei et al., 2008)
TCC patients (*n*=30); Control subjects (*n*=59) differentiated into 3 groups (C1 - individuals aged 18–31, with no urine abnormality on dipstick; C2 – subjects aged 18-32, with any non-cancerous condition or disease, and/or one or more positive dipstick finding; C3 - aged 24-89, with confirmed noncancerous) urological disease, with or without urine dipstick	Urine	e-nose (12 MOS sensors and an array of 10 individual MOSFET sensors together with a capacitance-based humidity sensor and an infrared-based CO_2)	PLS-DA; PCA	65% accuracy, 60% sensitivity and 67% specificity	(Weber et al., 2011)

abnormalities)					
Cohort	**Sample type**	**Analytical technique**	**Data treatment**	**Results**	**Ref.**
Clinical suspicion of primary or recurrent BCa (*n*=15); Patients without BCa but benign urological conditions (*n*=21)	Urine	Metal oxide gas sensor chip with 3 thin oxide layers	PCA; discriminant analysis	Sensitivity 75%, specificity 86%	(Horstmann et al., 2015)
HNSCC patients (*n*=100); BCa patients (*n*=40); CC patients (*n*=28)	Exhaled breath	Aeonose (3 different micro-hotplate metal oxide sensors)	ANN	HNSCC vs CC: sensitivity 79%, specificity 81%; HNSCC vs BCa: sensitivity 80%, specificity 86%; CC vs BCa: sensitivity 88%, specificity 79%	(van de Goor et al., 2017)
Patients with cystoscopically confirmed TUR-BT (*n*=30); Patients with no known disease of the urinary tract (control group, *n*=30)	Urine	Cyranose 320	LDA	93% sensitivity and 87% specificity after sample storage at -20°C; 93% sensitivity and specificity after storage at -80°C	(Heers et al., 2018)
BCa patients (*n*=5); Control group (*n*=5)	Urine	Fluorescence gas-sensor arrays (3 porphyrins, 2 fluorescence dyes, 2 solvatochromic dyes and 1 pH indicator)	CDA; PLS-DA	31/48 samples were correctly detected as BCa (sensitivity 78%, specificity 93%); Potential BCa biomarkers identified ethylbenzene, hexanal, laurie aldehyde and nonanoyl chloride	(Zhu et al., 2018a)

Abbreviations: TCC – Transitional cell carcinoma; **BCa** – bladder cancer; **PC** – prostate cancer; **SIFT-MS** – selected ion flow tube mass spectrometry; **GC-MS** – gas chromatography mass spectrometry; **PCA** – principal component analysis; **ANOVA** – analysis of variance; **LDA** – linear discriminant analysis; **PLS-DA** – partial least square discriminant analysis; **HCA** – hierarchical component analysis; **RF** – random forest; **SVM** – support vector machine; **OPLS-DA** – orthogonal partial least square discriminant analysis; **HS-SPME** – headspace solid phase microextraction; **QCM** – quartz crystal microbalance; **BPH** – benign prostatic hypertrophy; **MOS** – metal oxide semi-conductor; **MOSFET** – metal oxide semi-conductor field-effect transistor; **ANN** – artificial neural network; **HNSCC** – head and neck squamous cell carcinoma; **CC** – colon cancer; **TUR-BT** – transurethral resection of bladder tumour; **CDA** – canonical discriminant analysis.

In 2010, Willis et al. published an article study that aimed to evaluate the sensitivity and specificity which can be achieved by a group of four trained dogs for BCa detection. The experience consisted of 30 double-blind test runs where one urine sample (from 30 patients with recurrent TCC of the bladder) of BCa was placed alongside six control samples (from 180 control subjects differentiated into 3 different groups) (Willis et al., 2010). Results of sensitivity ranged 73% for the best performing dog and 57% for the worst, while specificity ranged from 92% for the best performing dog, correctly identifying control 1, and 56% for the worst performing dog identifying control 3 samples. An important consideration based on the results achieved in other studies, is that accuracy rates were higher for different tumour types (such as ovarian carcinoma and colorectal cancer) other than BCa. One of those reasons focus on the choice of the sample which, on other studies were breath samples and watery stool samples, comparatively to this study where the sample used was urine (Willis et al., 2010).

In 1999, Spaněl et al., performed a study to analyse the presence of formaldehyde in the headspace of urine from patients with BCa and prostate cancer (PC) by using a selected ion flow tube mass spectrometer (SIFT-MS). In this work, the authors included 14 patients with BCa, 24 patients with PC and 14 healthy individuals (controls). The results revealed that the concentrations of formaldehyde were higher in the urine headspace of BCa patients (mean 85 ppb) and lower in the urine headspace of controls (mean 25 ppb), being that formaldehyde was present in 13 out of 14 urine samples from BCa patients and absent from the majority of the urine samples from controls (Spaněl et al., 1999). Concerning the urine headspace of PC patients, formaldehyde was present at an intermediate level between BCa and controls. However, the authors also highlight that the patients from BCa and PC groups had tumours at different stages of the disease and that might influence, with great certainty, the formaldehyde concentrations in the urine of those patients (Spaněl et al., 1999).

Later in 2012, Jobu et al. studied the biochemical profiles of volatile compounds in urine of BCa patients using GC-MS. This study included 9 patients with BCa and 7 non-BCa patients (controls) (no age or gender was referenced). The urine was collected at two different moments, 3 days before surgery and 3-7 days after surgery (enucleation or transurethral resection of bladder tumour – TURBT) (Jobu et al., 2012). The results of the chromatograms showed 12 peaks that was confirmed based on the NIST database of which 5 were detected on BCa patients (ethylbenzene, nonanoyl chloride, dodecanal, (Z)-2-nonenal and 5-dimethyl-3(2H)-isoxazolone). Also, compared with urine cytology, an evident separation was present between controls and pre-operative patients and 7 out of 9 patients presented distinct score plots when PCA was applied (Jobu et al., 2012).

Another study developed by Khalid et al., that combined a GC-sensor device with a statistical model, with the aim of identifying BCa from urine samples, included 24 BCa patients (aged 27-91, mean 71) and 74 controls (aged 29-86, mean 64) (Khalid et al., 2013). The sensor device consisted of a conventional GC oven with a capillary column and a heated metal oxide sensor employed as the detector. The leave-one-out cross validation (LOOCV) was used as statistical validation of robustness and performance within samples, and linear discriminant analysis (LDA) for between groups comparison in order to identify specific differences that can be highly present in those profiles. After that, PLS-DA was employed as a two-group classifier (BCa and controls) (Khalid et al., 2013). The identification proved to be correct in 24/24 BCa samples and 70/74 controls (Khalid et al., 2013).

Later in 2016, a study made by Cauchi et al., aimed to identify BCa from urine headspace by using GC-MS. Their work included 72 patients with TCC of the bladder, both new or recurrent which has donated their urine before surgery, and 205 controls (Cauchi et al., 2016). Regarding the exploratory analysis, the results obtained for both PCA and HCA showed that cancer do not have an influence on class differentiation and thus, are not responsible for the major part of the variance (Cauchi et al., 2016). Overall, 88.5% of the cancer patients and 88.2% of controls were correctly classified when using as classifiers TCC urine samples and samples from the healthy group having other diseases like urinary tract infections (C2) (Cauchi et al., 2016). Some compounds were found to be decreased in controls compared to TCC group like 2-pentanone, 2,3-butanedione, 4-heptanone, dimethyl disulphide, 2-propanol, acetic acid, piperitone and thujone while others were found to be increased like hexanal, benzaldehyde, butyrophenone, 3-hydroxyanthranilic acid, benzoic acid, *trans*-3-hexanoic acid, *cis*-3-hexanoic acid and 2-butanone (Cauchi et al., 2016).

Another work using GC-MS has been developed by Rodrigues at al., which intended to study the volatile metabolomic signature of BCa cell lines. For this purpose the authors have used BCa cell lines being J82 (TCC; grade III/IV, stage pT3), Scaber (SCC; grade III/IV, stage pT4) and 5637 (TCC; grade II), and non-tumorigenic cell lines like SV-HUC-1 (Rodrigues et al., 2018). A clear separation in the metabolome profile has been noticed between the Scaber lines and the SV-HUC-1 lines (Q^2=0.969). The performance was calculated by receiver operating characteristics (ROC) curves, showing very good results (area under the curve, AUC=1) for several compounds such as 2-pentadecanone, α-terpineol, 2-methylbutan-2-ol, and 1-phenylethanol (Rodrigues et al., 2018). When comparing J82 with SV-HUC-1, 40 VOCs has been identified, being 4-methylbenzaldehyde, cyclohexanone, and 2-pentadecanone (AUC=1). At the same conditions, 5637 and SV-HUC-1 lines were also compared, being identified 42 VOCs like

nonanal, 1-phenylethanol, 4-methylnonane, dodecane, and γ-dodecalactone, and unknowns 10 and 24 (AUC=1).

When analysing the significantly altered compounds present in the cancer cell lines, it was possible to find that compounds like 2-nonanone, 4-methylheptan-2-one, dodecane 2,3-dimethylhexane, hexadecane and tetradecane were increased in cancer cells medium compared to SV-HUC-1 medium, and compounds like 2-phenylpropanol and isopentanol were decreased in cancer cells medium (Rodrigues et al., 2018). The main VOCs found in BCa cell lines belong to the classes of ketones and alkanes, while alcohols representing the class with more decreased VOCs in BCa cell lines (Rodrigues et al., 2018). In this study the authors also made a comparison between VOCs released in low grade and high-grade cancer cell lines, concluding that there is effectively an influence on VOCs released according to grade. Compounds that were found to have an impact for discrimination were cyclohexanone, methyl isobutyl ketone, styrene, dodecane, nonanal and benzaldehyde (Rodrigues et al., 2018). Another experience has been made to evaluate VOCs levels between the two subtypes of BCa (TCC and SCC), concluding that there are also differences of VOCs among the two. The compounds found to have influence on discrimination were nonanal, 2-pentadecanone, n-butyl acetate, cyclohexanone, and phenol (Rodrigues et al., 2018).

In 2008, a study developed by Bernabei et al., intended to detect urinary tract cancers from a group including, BCa (*n*=25), PC (*n*=12), benign prostatic hypertrophy (BPH) (*n*=29) and various urological pathologies (*n*=33), with an electronic nose based on eight quartz crystal microbalances (QCMs) gas sensors coated by sensing layers of metalloporphyrins (Bernabei et al., 2008). When comparing data from ill patients and healthy controls, 100% of the samples were correctly classified (with PLS-DA). Prior to that, the authors concluded that discrimination between the two cancers were not complete but sort of gradual (Bernabei et al., 2008). When analysing all the data set and including the group of BCa post-surgery patients (with PCA), the investigators obtained results where all samples from BCa post-surgery patients were recognized as healthy samples, which were initially not expected, once many of the individuals were affected by BPH and the removal of the tumour do not eliminate this condition, and second, some patients were evaluate in an early phase after treatment (1 week) or even after a longer time (3 months) (Bernabei et al., 2008).

Weber et al. developed a study in 2011 to characterize VOCs content of urine from TCC of the bladder, based on e-nose technology and elaborated with different sensors (Weber et al., 2011). This study involved the participation of 30 individuals (aged 50-88) with TCC and 59 controls divided into 3 categories depending on age and other urological conditions. First results obtained by PCA showed that the presence of cancer did not

contribute to the separation and so it does not represent a major part of the variance (Weber et al., 2011). When comparing the most distinct groups (C1 and TCC), the best results were obtained showing 70% of cancer patients and non-cancerous patients correctly identified, however, the results obtained from patients with other non-cancerous urological disease accounted for an accuracy of 65%, sensitivity of 60% and 67% of specificity. As mentioned in other studies, individuals with a more advanced stage of the disease were in general more difficult to classify (Cauchi et al., 2016; Weber et al., 2011).

Later in 2015, Horstmann et al., revealed their first results of a pilot study made with an electronic nose system to detect BCa in urine. This work included 15 individuals with clinical suspicion of primary of recurrent BCa (8 confirmed correctly by histopathology) and 21 individuals without BCa (28 confirmed correctly by histopathology) but with other benign urological conditions (Horstmann et al., 2015). The sensor used consisted on a metal oxide gas sensor chip with three thin oxide layers. The data analyses performed involved PCA and discriminant analysis. Regarding the results obtained by the e-nose, 6/8 BCa patients were correctly identified and 2/8 were non-identified (stage pTa) accounting for a sensitivity of 75%. From the group without BCa, 24/28 were correctly identified showing a specificity of 86% (Horstmann et al., 2015).

Other authors like Van de Goor et al., studied the viability of using e-nose technology for discriminating different cancer types like head and neck, bladder, and colon carcinomas through exhaled breath. This study included 100 patients with primary HNSCC, 40 patients with BCa, and 28 with colon cancer (CC). The device used consisted on a Aeonose composed of three different micro-hotplate metal oxide sensors (van de Goor et al., 2017). Results concerning sensitivity, specificity and accuracy were obtained by double cross-validation and AUC (AUC=0.83 for HNSCC and CC; AUC=0.85 for HNSCC and BCa) (van de Goor et al., 2017). When differentiating HNSCC and CC the results showed a sensitivity of 79%, specificity of 81%, and an accuracy of 81%. When differentiating HNSCC and BCa the results showed a sensitivity of 80%, specificity of 86%, and an accuracy of 84%. And finally, when differentiating BCa and CC the results showed a sensitivity of 88%, specificity of 79%, and an accuracy of 84% (van de Goor et al., 2017).

Another pilot study with an e-nose has been developed more recently by Heers et al., with the aim of detecting bladder tumours through VOCs. A total of 30 patients with bladder tumours and 30 patients without diseases of the urinary tract (controls) were included in this study (Heers et al., 2018). The device used consisted of a Cyranose 320 e-nose which contains 32 polymer sensors, trained throw the measurement of several well categorised samples (Heers et al., 2018). With samples stored at -20°C, the LDA correctly identified 28/30 patients with BCa and 26/30 as non-tumour samples (Heers et al., 2018).

As for sensitivity and specificity, the e-nose results obtained were 93.3% and 86.7% respectively. When samples were stored at -80°C, the results were similar, correctly identifying 28/30 BCa patients and 28/30 control samples. Results based on sensitivity and specificity were both 93.3%. The cross-validation value (CVV) showed close results for both temperatures, being 55% for samples stored at -20°C and 53% for samples stored at -80°C. There were no significant differences between high and low grades of tumour, as well as for the smoking habits or other confounding factors (Heers et al., 2018).

Finally, Zhu et al., also came out with a study in 2018 with the aim of detecting BCa VOCs biomarkers by using an optic fluorescence sensor. Based on a review of the literature, the authors chose five compounds as biomarkers for urinary BCa, namely ethylbenzene, hexanal, laurie aldehyde, and nonanoyl chloride (Zhu et al., 2018a). This study included 5 patients with BCa and 5 control patients as a first attempt to test the device. As statistical data tests, the canonical discriminant analysis (CDA) was performed instead of PCA which is more suitable for sensor arrays. For the prediction model, PLS-DA and then LOOCV was used to evaluate the performance. The analyses performed a sensitivity of 77.75% and a specificity of 93.25% (R^2=0.97, Q^2=0.83) (Zhu et al., 2018a).

1.4. Aims of this thesis

Although many diagnostic methods are employed in our days for the detection of BCa, with overall good performances, not many advances have been made in this field. Actual methods of diagnostic include urine cytology, cystoscopy, imaging techniques like MRI and CT or even biochemical tests (Antunes et al., 2018). However, these tests present some limitations such as invasiveness, low sensitivity especially for detection of low grade tumours, high costs and the need of specialized physicians for interpretation of results. Thus, the development of new techniques and tools for earlier detection of BCa is required (Antunes et al., 2018). Hence, metabolomics has arisen as a promising approach for the diagnosis of cancer, since it is noninvasive, requires small amounts of sample for analysis and presents high analytical sensitivity. The study of metabolites already present in the organism or released in response of a disease like cancer will allow to better understand metabolic processes in the body and even the pathological dynamism and its response to the tumour, promoting an earlier diagnosis and more appropriate treatment (Antunes et al., 2018). So far, only a few studies have been developed concerning BCa

metabolomics, especially concerning VOCs as potential biomarkers. First studies on that field can be initiated by using *in vitro* cell lines, which present a lower complexity in terms of experimental techniques and are not affected by confounding factors such as diet, lifestyle or presence of diseases comparing to other types of samples but still can provide initial information on the metabolic changes that can happen in the organism when BCa is present. However, as the major goal of this work is to develop a strategy that will allow to detect BCa in its early stage and promote the onset of treatment as soon as possible, it is important to start by choosing the most appropriated biological samples since the future diagnostic methodology will be applied to this chosen sample.

The main aims of this thesis are:

1) To evaluate the performance of volatiles (VOCs) in general, and volatile carbonyl compounds (VCCs) in particular, for discrimination of urine samples collected from BCa patients and cancer-free individuals (controls), by headspace solid-phase microextraction/gas chromatography-mass spectrometry (HS-SPME/GC-MS).
2) To identify candidate biomarkers for BCa detection and staging in urine.
3) To define the volatilome alterations occurring in BCa, thus contributing to a better insight into the pathophysiological alterations associated with this type of cancer.

Chapter 2. Experimental methods

2.1. Chemicals

Thymol (≥98.5%), benzaldehyde (≥99.5%), 2-butanone (≥99%), cyclohexanone (≥99%), 2-decanone (≥98%), 2,5-dimethylbenzaldhyde (≥99%), heptanal (≥92%), 4-heptanone (≥97%), (E,E)-2,4-hexadienal (≥95%), hexanal (≥98%), 2-hexanone (≥98%), 3-methylbutanal (≥97%), 5-methyl-2-furfural (≥99%), methylglyoxal (40% aqueous solution), nonanal (≥95%), 2-nonanone (≥97%), 2-octanone (≥98%), 3-penten-2-one (≥70%), 3-phenylpropionaldehyde (≥95%), 2-pentanone (≥99.5%), 2-heptanone (≥99%), 1-decanol (≥99.9%), acetaldehyde (≥99%), benzene (≥99.9%), 2,4-dimethylheptane (≥98%), decane (≥99.8%), carvone (≥98.5%), 1-dodecanol (≥98%), propanal (≥98%), isobutanal (≥97%), furfural (≥98.5%), heptanal (≥95%), trans-2-heptenal (≥97%), m-tolualdehyde (≥97%), trans-2-nonenal (≥95%), dimethylglyoxal (≥98%), 2,4-decadienal (≥94%), 2-methyl-1-butanal (≥90%), 2-butenal (≥99%), octanal (≥99%), nonanal (≥95%), decanal (≥95%), (E,E)-2,4-nonadienal (100%), PFBHA (≥98%), and phenylacetaldehyde (≥90%), were purchased from Sigma–Aldrich (Madrid, Spain). Glyoxal (≥95%) was purchased from Fluka (Madrid, Spain). Sodium chloride was obtained from VWR (Leuven, Belgium).

2.2. Subjects

Urine samples of BCa patients and cancer-free individuals (controls) were collected in the morning (without fasting), at the Portuguese Institute of Oncology of Porto (IPO Porto), and frozen at -80°C until analysis. The experimental study was approved by the local Ethics Committee and all subjects signed an informed consent prior to participation.

A total of 120 individuals were included in this study, comprising 60 BCa patients and 60 cancer-free individuals (controls), matched for sex. A match of age was not possible due to the difficulty in finding cancer-free individuals with more than 65 years old. However, this potential confounding factor was considered in the results. Most of the BCa patients were diagnosed in the stage 0a, followed by stage I and only a few at later stages (II, III and IV) (***Table 3***).

Table 3 - Data obtained from both BCa and control patients.

Features	Nº of subjects	Males	Females	Min./Max. age	Mean age ± SD (years)
Cancer free individuals	60	43	17	45/66	51.8 ± 5.1
BCa patients	60	43	17	43/87	68.2 ± 10.6
Stage 0a	30	22	8	43/87	68.9 ± 10.5
Stage I	19	12	7	53/83	70.4 ± 9.7
Stage II	3	3	-	57/69	63.5 ± 5.5
Stage III	4	2	2	43/78	60.3 ± 14.3
Stage IV	5	5	-	51/80	64.6 ± 10.1

The use of quality control samples (QCs) in metabolomics is essential since many processes that result from the equipment analysis, associated with the injector or the column for example, can affect the detection of metabolites. The QC samples will minimize any false results as well as promoting the reliability and reproducibility of the data obtained to meet the predefined criteria.

The QC samples used were prepared of a pool of urine samples and separated in aliquots and injected between the real samples.

2.3. Sample preparation

Urine samples of BCa, controls and QC were thawed at room temperature. For VOCs analysis, 2 mL of urine were added to a 10 mL glass vial containing 0.54 g of NaCl and 30 µL of thymol (2 mg/L) used as internal standard (IS). Thymol was also used to verify and control any possible problems of injection.

For VCCs analysis, 250 µL of urine were added to a 10 mL glass vial with 7.5 µL of the derivatizing agent PFBHA (40 g/L in ultrapure water).

For both VOCs and VCCs analysis, all samples were randomly injected and the QC samples were analysed on every 10 samples.

2.4. GC-MS analysis

The analytical conditions regarding VOCs and VCCs extraction is detailed in ***Table 4***. For VOCs extraction, the autosampler used consisted of a Combi-PAL autosampler type (Varian Pal Autosampler, Switzerland).

For the VCCs extraction, the autosampler consisted of a Bruker CTC PAL-xt (Bruker Daltonics). The chromatograph used consisted of a Scion 436-gas which was coupled to a Bruker single quadrupole (SQ) mass spectrometer and additionally equipped with a Scion SQ ion trap mass detector, and as a workstation software Bruker Daltonics version 8.2.1 and a column Rxi-5Sil MS (30 m× 0.25 mm internal diameter × 0.25 µm) from Restek Corporation (U.S., Bellefonte, Pennsylvania).

As for VCCs analysis, the chromatograph consisted of a 436-GC model coupled to a EVOQ triple quadrupole mass spectrometer (Bruker Daltonics) and as a workstation software Bruker MS version 8.2.1.

Table 4 - Analytical conditions for VOCs and VCCs extraction.

Analytical conditions	VOCs analysis	VCCs analysis
HS-SPME		
Fiber type	50/30 µm divinylbenzene/carboxen/ polydimethylsiloxane (DVB/CAR/PDMS)	65 µm polydimethylsiloxane/divinylbenzene (PDMS/DVB)
Incubation (time / temperature / stirring)	11 min/ 44°C	6 min/ 62°C
Extraction (time / temperature / stirring)	30 min/ 44°C/ 250 rpm	51 min/ 62°C/ 250 rpm
Desorption (time / temperature)	4 min/ 250°C	5 min/ 250°C
GC-MS		
Column	Rxi-5Sil MS (30 m× 0.25 mm internal diameter × 0.25 µm)	Rxi-5Sil MS (30 m× 0.25 mm internal diameter × 0.25 µm)
Carrier gas (flow rate)	helium C-60/ 1 mL/min	helium C-60/ 1 mL/min
Oven temperature	40°C for 1 min/ 250°C for 5 min/ 300°C	40°C for 1 min/ 250°C for 5 min/ 300°C for 1 min
Transfer line temperature	250°C	260°C
Ion source temperature	260°C	270°C
Manifold temperature	41°C	41°
Energy level of electron ionization (EI)	70 eV	70 eV
Mode	Full scan	Full scan
Mass range	40–400 m/z	35–600 m/z
Scan time	500 ms	250 ms

The metabolite identification was achieved by comparing MS spectra with standards, that were injected in the same conditions, whenever they were commercially available. The comparison with the MS data obtained was performed by accessing the National Institute of Standards and Technology (NIST v2.2, 2014) database spectral library, and by comparing the experimental and theoretical Kovats index as well as the reverse match (R-match) and their retention time. Metabolites with variable importance to the projection (VIP) values greater than one were considered discriminants.

As a way to "identify" metabolites with a certain level of confidence, the Chemical Analysis Working Group of the Metabolomics Standards Initiative (MSI) has developed a method of classification which comprehends 4 levels of confidence, being included on level 1 (L1) all "identified metabolites" meaning that those are identified based on RT, accurate mass and fragmentation data or even by using a pattern; levels 2 and 3 (L2 and L3) which are "putatively annotated compounds" and "putatively characterised compound classes" respectively, relates to those which can be identified through public or commercial libraries like NIST database for example; finally level 4 (L4) or "unknown", concerns those metabolites that can be detected and quantified but cannot be qualified by none of these methods (Viant et al., 2017).

2.5. Data pre-processing

Before statistical analysis, the data were first pre-processed using MZmine-2.52 (Pluskal et al., 2010), including filtering, peak detection, chromatogram builder, deconvolution and alignment. The parameters used for VOCs and VCCs pre-processing are summarized in ***Table 5***. The artefact peaks from the chromatographic column (e.g., siloxanes, cyclosiloxanes and phthalates) were manually removed from the final data matrix.

Table 5 - Pre-processing parameters for VOCs and VCCs analysis.

Pre-processing parameters	VOCs	VCCs
Filtering (RT range / *m/z* range)	2-34 min / 50-250	9.8-46 min / 50-300
Peak detection (noise level)	1x104	1x105
Chromatogram builder (intensity threshold / *m/z* tolerance)	5x104 / 0.07	5x106 / 0.1
Deconvolution (peak range / baseline level)	0.03-0.5 min / 1x104	0.03-0.5 min / 5x105
Alignment (*m/z* tolerance / RT tolerance)	0.07 / 0.2	0.1 / 0.2

Finally, the data matrix was normalized by the total area of the chromatograms and then scaled to pareto. In addition, a variable selection method was applied to remove any irrelevant variables using a Mann-Witney test. After obtaining the data matrix by those methods, all variables with a *p*-value > 0.05 were removed.

2.6. Statistical analysis

Both multivariate and univariate statistical tests were used for VOCs and VCCs data analyses. MVA was applied using PCA and PLS-DA using SIMCA-P software (version 15, Umetrics, Sweden). The robustness of the PLS-DA models was confirmed through 7-fold cross validation and permutation test (1000 random permutation of Y-observations, 2 components) specifically R^2 and Q^2 (SIMCA-P software version 15, Umetrics, Sweden). The R^2 represents the percentage of variation, measuring how well the model fits the data. In general, the closest the R^2 is to 1, the better is the model, however that might not always happen. A good value of R^2 will be influenced by a poor reproducibility and high levels of noise. The Q^2 also represents the percentage of variation, however this parameter is predicted according to cross-validation and shows how good the model estimates new data. The closest the Q^2 value gets to 0.5, the better is the predictivity. Those results might be influenced by the presence of noise or existence of outliers.

For validation, ROC, AUC, sensitivity, specificity, and accuracy were performed for PLS-DA models of VOCs and VCCs, using MetaboAnalyst. After employing MVAs, the metabolites with a VIP>1 were submitted to univariate analysis and a Shapiro-Wilk normality test to visualize if the data followed a Gaussian distribution. When data presented a normal distribution (both BCa and control samples present normal distribution), an unpaired Student's t-test with Welch correction was used. On the other hand, when data presented a non-normal distribution (at least one group of BCa or control samples present a non-normal distribution), an unpaired Mann-Whitney test was used. Parameters like the effect size (ES) and the standard error were also calculated by applying mathematical formulas in which the ES can be calculated as the standardized mean difference (ES_{smd}) between the two group means divided by their pooled standard deviation (S_{pooled}) (BCa and control) (**Equation 1**). Being that, S_{pooled} and the degrees of freedom (df) are calculated through **Equation 2** and **Equation 3**. Next, to calculate the confidence interval for the standardized mean difference it is necessary to calculate its variance (V_{smd}), as found in **Equation 4.** Through this result it is possible to calculate the standard error of the standardized mean difference (SE_{smd}) by applying the square root of its variance (**Equation 5**). Finally the 95% confidence interval surrounding the standardized mean differences is calculated through **Equation 6** (Berben et al., 2012). The identified compounds were considered statistically significant when *p*-value < 0.05 (confidence level 95%), which was corrected by the method of false discovery rate (FDR). Those *p*-values and FDR adjusted *p*-value intends to determine if a result is statistically significant (p<0.05), however they can be influenced by the sample size which is more relevant with larger group samples than smaller group samples (Berben et al., 2012).

The normalized areas of the metabolites were computed using GraphPad Prism version 6 (GraphPad Software, San Diego, CA, USA).

$$ES_{smd} = \frac{\bar{X}_1 - \bar{X}_2}{S_{pooled}} \qquad \text{Equation 1}$$

$$S_{pooled} = \sqrt{(n_1 - 1)S_1^2 + (n_2 - 1)s_2^2/df} \qquad \text{Equation 2}$$

$$df = n_1 + n_1 - 2 \qquad \text{Equation 3}$$

$$V_{smd} = \frac{n_1 + n_2}{n_1 n_2} + \frac{ES_{smd}^2}{2(n_1 + n_2)} \qquad \text{Equation 4}$$

$$SE_{smd} = \sqrt{V_{smd}}$$ Equation 5

$$ES_{smd} \pm 1.96SE_{smd}$$ Equation 6

To calculate sensitivity, specificity, and accuracy a confusion matrix can also be obtained which allow to identify a true class and a hypothesized class and so, true positives (TP) and negatives (TN) as well as false positives (FP) and negatives (FN) (***Figure 6***). With these results, it is possible to calculate throw mathematical formulas, the sensitivity, specificity and accuracy (Fawcett, 2006). The percentage of sensitivity, specificity and accuracy are calculated just like described in **Equation 7**, **Equation 8,** and **Equation 9** respectively.

		True classes	
		0	**1**
Predicted classes	**0**	TN	FN
	1	FP	TP

1 = Positive; **0 =** Negative; **TN –** true negatives; **FN –** false negatives; **FP –** false positives; **TP -** true positives.

Figure 6 - Confusion matrix and performance metrics (sensitivity, specificity, and accuracy) calculated. [Adapted from Fawcett, 2006]

$$Sensitivity\,(\%) = \frac{TP}{TP+FN}\ x\ 100$$ Equation 7

$$Specificity\,(\%) = \frac{TN}{TN+FP}\ x\ 100$$ Equation 8

$$Accuracy\,(\%) = \frac{TP+TN}{TP+TN+FP+FN}\ x\ 100$$ Equation 9

Chapter 3. Results

3.1. Identification of VOCs and VCCs present in urine

The aim of this work was to investigate the potential of VOCs, in general, and VCCs for BCa detection in urine by HS-SPME/GC-MS. A representative chromatogram of VOCs and VCCs is shown in ***Figure 7***. The list of all identified compounds is described in ***Table A1*** and ***Table A2*** (see in appendix).

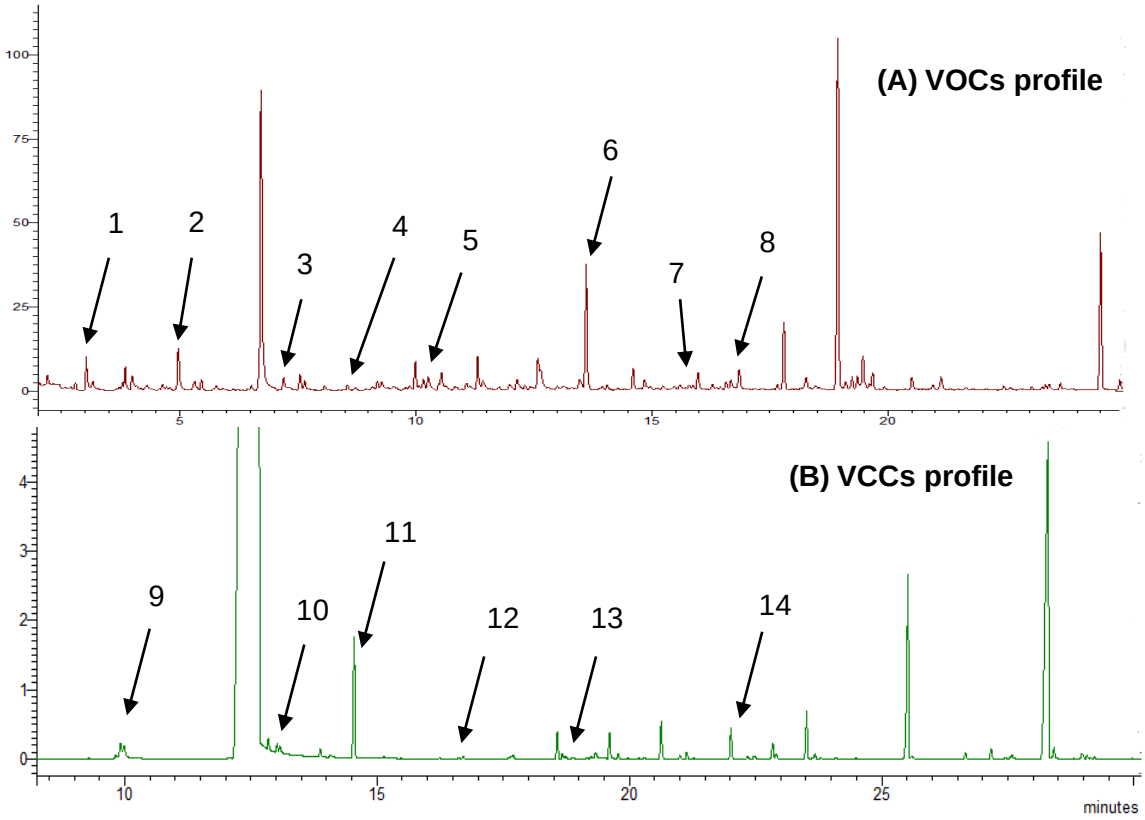


Figure 7 - Representative HS-SPME-GC/MS full scan chromatograms of **(A)** VOCs and **(B)** VCCs present in urine. **1)** 2-pentanone; **2)** Hexanal; **3)** 2-heptanone; **4)** 4-methyl-2-heptanone; **5)** 1,2,4-trimethylbenzene; **6)** Nonanal; **7)** α-terpineol; **8)** 2,5-dimethylbenzaldehyde; **9)** Formaldehyde; **10)** Acetaldehyde; **11)** Acetone; **12)** 2-Butanone; **13)** 3-methylbutanal; **14)** 4-heptanone.

Concerning the identification of VOCs, a total of 103 compounds has been identified being 17 formally identified (L1) and 53 putatively identified (L2). The major class compounds were aldehydes, ketones, aromatic hydrocarbons, alkanes and alcohols, with minor class compounds being monoterpenes, fatty acids, phenols, aromatic

hydrocarbons, sulfones and imidazoles. However, no formal or putative identification was possible for 33 compounds, which were classified as “unknown” (L4).

Regarding VCCs identification, a total of 107 compounds (being 16 in common with the previous VOCs protocol) has been identified being 53 formally identified (L1) and 15 putatively identified (L2). The major class compounds were aldehydes and ketones, with minor class compounds being alcohols, fatty acids, phenols and esters. Thus, 39 compounds were not able to identify and have been classified as “unknown” (L4).

3.2. Volatile profile of urine of BCa patients *vs.* cancer-free controls

Next step was to evaluate the volatile profile of urine of BCa patients and cancer-free controls to determine the presence and differences concerning the metabolites in both groups. To verify the reproducibility of the method, a pool of samples was analysed (QCs). All the BCa (*n*=60) and control samples (AC) (*n*=60) were included on the PCA as well as QC samples (*n*=30) which were closely clustered and centred in the PCA score scatter plot (***Figure 8***), demonstrating the reproducibility of the method.

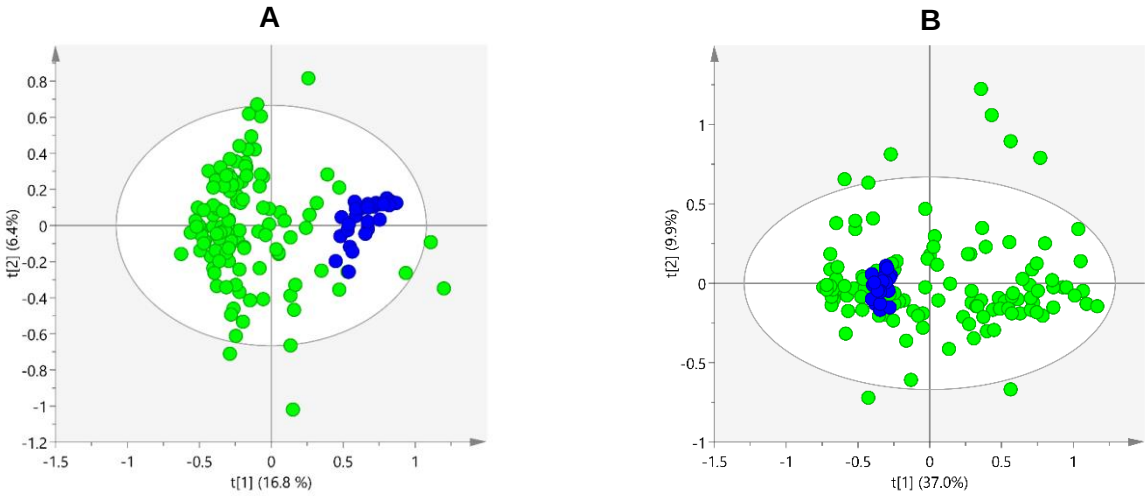


Figure 8 - PCA scores scatter plots obtained for **(A)** VOCs and **(B)** VCCs profiles of urine of all subjects included in this study (controls *n*=60 and BCa *n*=60, ●) and QCs (*n*=30, ●).

After PCA, a variable selection method was established for both VOCs and VCCs analysis to enhance the prediction power. In ***Figure 9*** it is possible to visualize the discriminant capability of the PLS-DA models, especially concerning VOCs (VOCs model

– LV=2; R^2X= 0.268; R^2Y= 0.492; Q^2= 0.218. VCCs model - LV=2; R^2X= 0.347; R^2Y= 0.437; Q^2= 0.250). ***Figure 9b*** shows the permutation testing that was used to confirm the model robustness. The ROC curve obtained for both VOCs and VCCs is demonstrated in ***Figure 9c*** as well as the results concerning the AUC, sensibility, specificity and accuracy (VOCs model: AUC=0.853; sensitivity=72%; specificity=90%; accuracy=81%. VCCs model: AUC=0773; sensitivity=72%; specificity=77%; accuracy=74%).

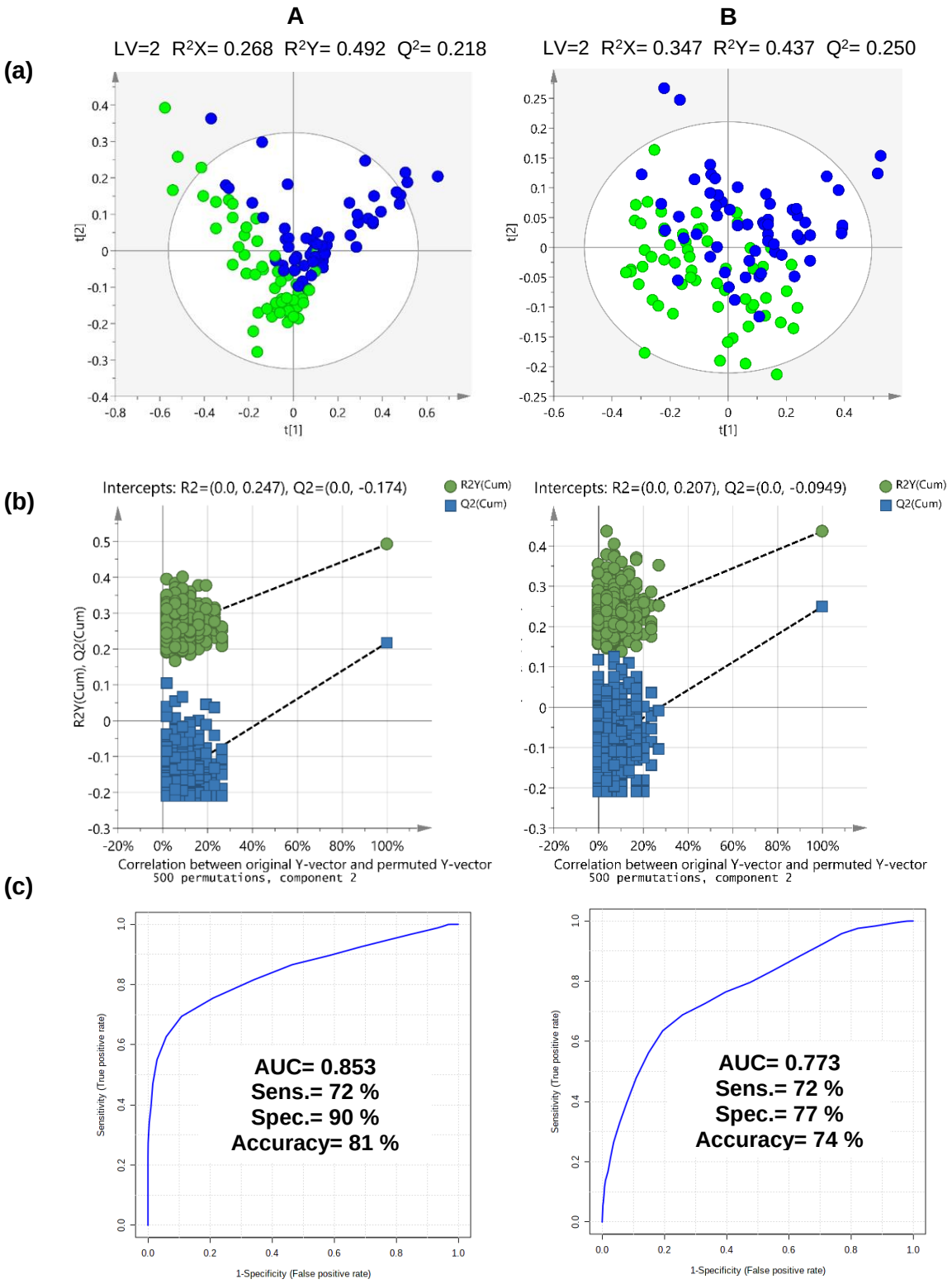


Figure 9 - (a) PLS-DA scores scatter plots obtained, after variable selection, for **(A)** VOCs and **(B)** VCCs profiles of urine of controls (*n*=60, ●) and BCa patients (*n*=60, ●); **(b)** Permutation plots obtained for the PLS-DA model of VOCs and VCCs; **(c)** ROC curves obtained for VOCs and VCCs profiles.

Then, VIPs higher than 1 in PLS-DA models were considered as relevant for group discrimination. Of the 1071 variables of VOCs, 193 had a VIP>1. Based on the NIST mass spectra library and comparison with standards it was possible to identify formally 5 compounds (L1), to identify 5 compounds putatively (L2), and 6 compounds were unidentified being classified as “unknown 1,2,3…” (L4). Of the 3113 variables of VCCs, 668 had a VIP>1. Based on the NIST mass spectra library and comparison with standards it was possible to identify formally 8 compounds (L1) and 8 compounds were unidentified being classified as “unknown 1,2,3…” (L4).

Next, the AUC of all compounds was also calculated as well as the ES and the standard error. As a result, 12 compounds were found increased in BCa compared to controls, such as benzaldehyde, 2,4-dimethylhexane, 2-methylnonane, 1-methylnaphthalene, 2-methylnaphthalene, and 4-methylphenol (along with other 6 unknown compounds), while 18 were found decreased in BCa compared to controls such as, 2-furaldehyde, formaldehyde, glyoxal, hexanal, methylglyoxal, phenylacetaldehyde, 2-butanone, 2-pentanone, 4-heptanone, and carvone (along with other 8 unknown compounds) (***Table 6***).

Figure 10 represents the boxplots of the 11 compounds the most statistically significant for discrimination established as a biomarker panel for control *vs* BCa.

Table 6 - Metabolites that were found to be statistically significant for discrimination, present in urine of BCa patients and controls.

Chemical name (IUPAC)	*p*-value original	*p*-value (FDR)	AUC	Effect size	Standard error	Variation	Identification level	Metabolic pathways
Aldehydes								
2-Furaldehyde	0.0001	0.0001	0.792	-0.67	0.37	↓	L1	-
Benzaldehyde	0.0440	0.0460	0.607	0.41	0.36	↑	L1	-
Formaldehyde	0.0009	0.0041	0.675	-0.60	0.37	↓	L1	-
Glyoxal	0.0063	0.0150	0.601	-0.17	0.36	↓	L1	Peroxidation of polyunsaturated fatty acids
Hexanal	0.0055	0.0150	0.647	-0.55	0.36	↓	L1	Steroid hormone biosynthesis
Methylglyoxal	0.0430	0.0430	0.608	-0.40	0.36	↓	L1	Pyruvate metabolism; glycine, serine and threonine metabolism
Phenylacetaldehyde	0.0450	0.0460	0.606	-0.46	0.36	↓	L1	Phenylalanine metabolism
Alkanes								
2,4-Dimethylhexane	0.0001	0.0005	0.702	0.74	0.37	↑	L2	-
2-Methylnonane	0.0300	0.0400	0.615	0.54	0.36	↑	L2	-
Aromatic hydrocarbon								
1-Methylnaphthalene	0.0092	0.0160	0.638	0.60	0.37	↑	L2	-
2-Methylnaphthalene	0.0140	0.0220	0.631	0.59	0.37	↑	L2	-
4-Methylphenol	0.0001	0.0005	0.706	0.51	0.36	↑	L1	-

Chemical name (IUPAC)	*p*-value original	*p*-value (FDR)	AUC	Effect size	Standard error	Variation	Identification level	Metabolic pathways
Ketones								
2-Butanone	0.0270	0.0290	0.617	-0.31	0.36	↓	L1	Fatty acid and carbohydrate metabolism
2-Pentanone	0.0240	0.0280	0.620	-0.32	0.36	↓	L1	Fatty acid metabolism
4-Heptanone	0.0001	0.0012	0.708	-0.30	0.36	↓	L1	Fatty acid metabolism
Monoterpene								
Carvone	<0.0001	0.0001	0.735	-0.64	0.37	↓	L1	-
Unidentified								
Unknown 5	0.0025	0.0050	0.660	-0.45	0.36	↓	L4	-
Unknown 8	0.0290	0.0400	0.615	0.41	0.36	↑	L4	-
Unknown 9	<0.0001	0.0001	0.730	0.70	0.37	↑	L4	-
Unknown 16	0.0004	0.0012	0.689	0.40	0.36	↑	L4	-
Unknown 39	0.0190	0.0260	0.624	-0.45	0.36	↓	L4	-
Unknown 40	0.0230	<0.0001	0.620	-0.47	0.36	↓	L4	-
Unknown 43	0.0180	0.0260	0.626	-0.42	0.36	↓	L4	-
Unknown 19	0.0010	0.0027	0.674	0.67	0.37	↑	L4	-
Unknown 45	0.0200	0.0260	0.623	0.42	0.36	↑	L4	-
Unknown 21	0.0022	0.0050	0.662	0.64	0.37	↑	L4	-
Unknown 51	0.0120	0.0230	0.633	-0.56	0.36	↓	L4	-
Unknown 64	0.0068	0.0150	0.643	-0.53	0.36	↓	L4	-

Chemical name (IUPAC)	*p*-value original	*p*-value (FDR)	AUC	Effect size	Standard error	Variation	Identification level	Metabolic pathways
Unknown 65	0.0007	0.0041	0.681	-0.59	0.37	↓	L4	-
Unknown 72	0.0001	<0.0001	0.745	-0.96	0.38	↓	L4	-

L1 – identified compounds, through comparison with a chemical reference standard. **L2** – putatively annotated compounds identified through the comparison with public or commercial libraries. **L4** – unknown compounds that can be detected and quantified but cannot be identified by either the previous methods.

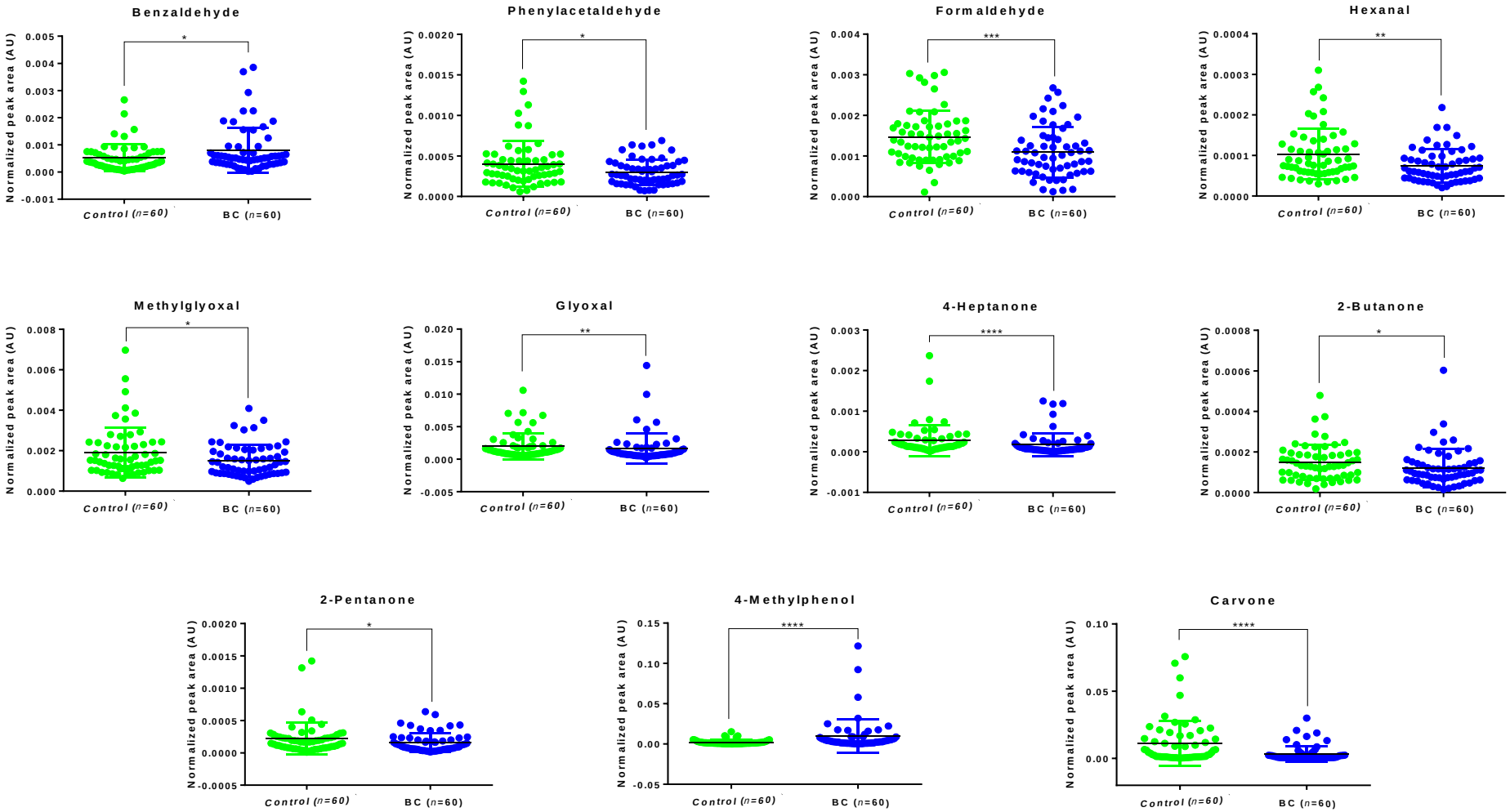


Figure 10 - Boxplots of the compounds the most statistically significant for discrimination (*n*=11) established as a biomarker panel for control (*n*=60, ●*) vs* BCa (*n*=60, ●).

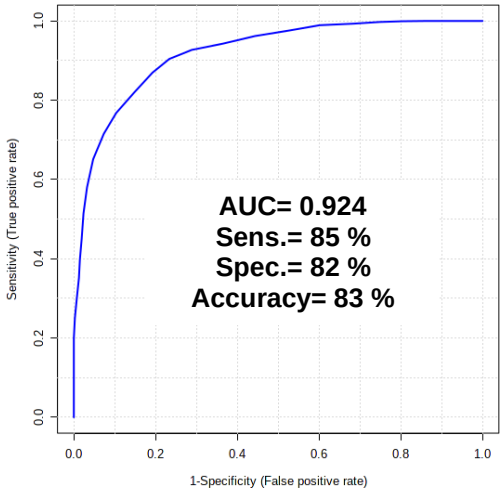
To investigate if the identified compounds had a relation with age of the patients, a Spearman correlation was computed and detailed in ***Table 7***. The results showed a poor correlation with a $|r| \leq 0.43$ which means that age has no significant correlation with the metabolites found altered in BCa compared to controls.

Table 7 - Spearman's correlation indexes and corresponding *p*-values obtained for age with the set of metabolites found altered in BCa compared to controls.

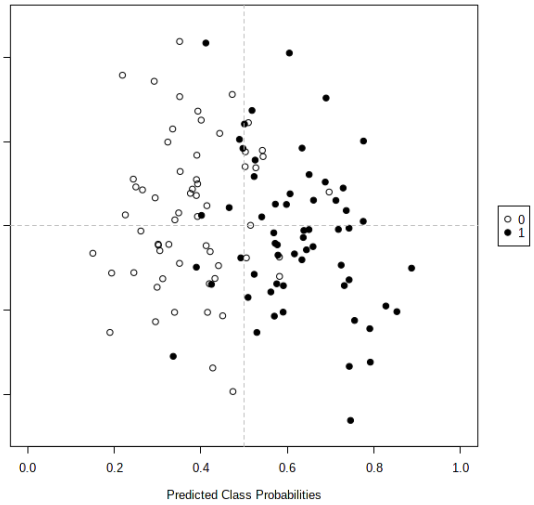
Compound	*r*	*p*
2-Furaldehyde	-0.43	<0.0001
Benzaldehyde	0.18	0.054
Formaldehyde	-0.27	0.003
Glyoxal	-0.24	0.0078
Hexanal	-0.19	0.0355
Methylglyoxal	-0.21	0.0207
Phenylacetaldehyde	-0.18	0.0505
2,4-Dimethylhexane	0.37	<0.0001
2-Methylnonane	0.2	0.0325
1-Methylnaphthalene	0.25	0.0063
2-Methylnaphthalene	0.21	0.0232
4-Methylphenol	0.27	0.003
2-Butanone	-0.22	0.0164
2-Pentanone	-0.19	0.0333
4-Heptanone	-0.31	0.0006
Carvone	-0.39	<0.0001
Unknown 5	-0.39	<0.0001
Unknown 8	0.15	0.0969
Unknown 9	0.3	0.0008
Unknown 16	0.26	0.0046
Unknown 19	0.34	0.0001
Unknown 21	0.31	0.0005
Unknown 39	-0.31	0.0006
Unknown 40	-0.19	0.0385
Unknown 43	-0.43	<0.0001
Unknown 45	-0.28	0.0021
Unknown 51	-0.15	0.1001
Unknown 64	-0.27	0.0027
Unknown 65	-0.25	0.0058
Unknown 72	-0.31	0.0006

The ROC curve, from both VOCs and VCCs, found to be statistically significant for discrimination was also calculated, obtaining an AUC of 0.924, a sensitivity of 85%, a specificity of 82%, and an accuracy of 83% (***Figure 11A***). The predictive class probabilities (***Figure 11B***) and the confusion matrix was also calculated (***Figure 11C***). These results are calculated based on formulas from the literature as described previously on chapter 2.6.

A



B



C

		True classes	
		0	**1**
Predicted classes	**0**	49	9
	1	11	51

Figure 11 - ROC analysis. **A)** ROC curve from both VOCs and VCCs found to be statistically significant for discrimination (identified *n*=16; unknowns *n*=14), with calculated sensitivity (85%), specificity (82%) and accuracy (83%). **B)** Predicted class probabilities of each samples across 100 cross validations. **C)** Confusion matrix.

To establish a correlation comparison between the 30 statistically significant compounds found altered in BCa compared with controls, a Spearman's correlation coefficient was employed for all the statistically significant compounds with a |r| ≥ 0.5 and represented in a heatmap (***Figure 12***). Considering the results, it was possible to identify a strong correlation between the compound Unknown 39 (RT 19.61 *m/z* 71) and methylglyoxal, as well as the compound Unknown 9 (RT 12.16 *m/z* 57) with 2-methylnonane, 1-methylnaphtalene, 2,4-dimethylhexane and 2-methylnaphtalene, which suggests a possible relationship between these compounds in terms of biochemical pathways.

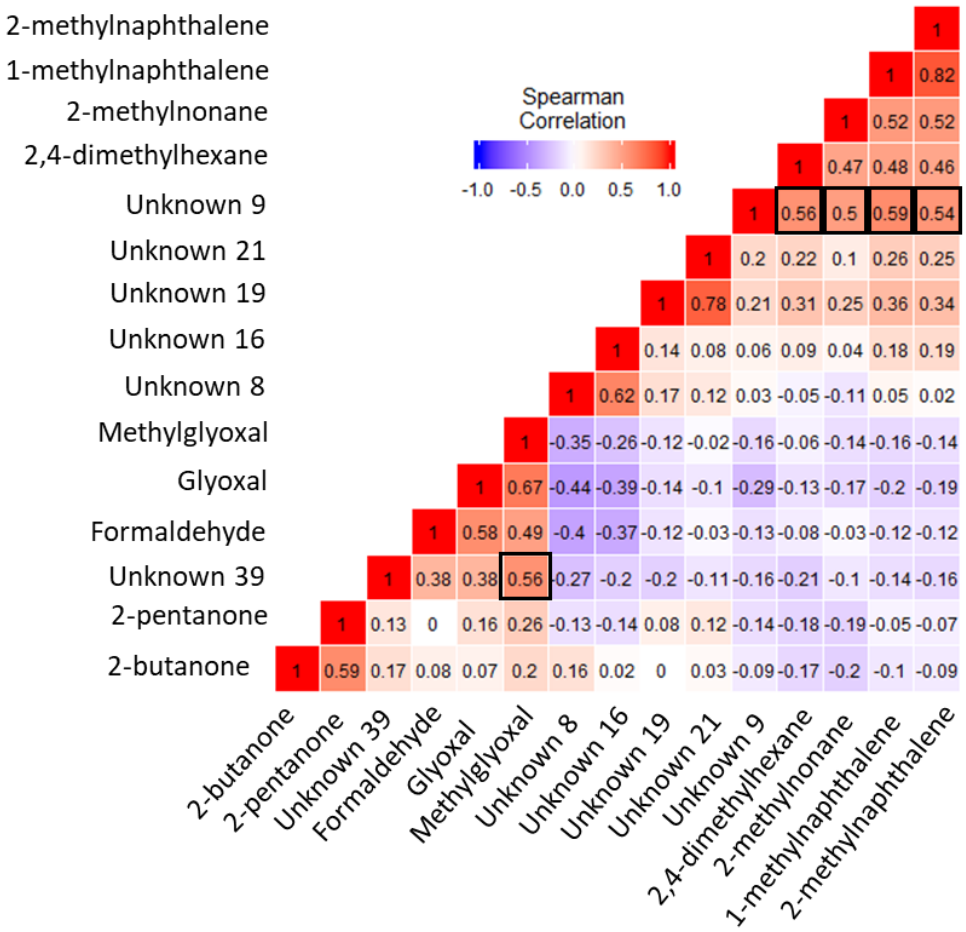


Figure 12 - Heatmap spearman correlation of the statistically significant compounds with a |r| ≥ 0.5 from the comparison of BCa patients and controls.

3.3. Volatile profile of urine of BCa patients with different stages

Additionally, MVAs was used to know the profile of VOCs and VCCs for the different stages of BCa. A PLS-DA was employed from a matrix including stage 0a (*n*=29), stage I (*n*=19) and stage II, III and IV all gathered in the same group (*n*=12) (***Figure 13***). The results showed that when comparing the stage 0a, stage I and stages II, III and IV the PLS-DA analyses showed negative values of Q^2 for both VOCs and VCCs which means that the established model does not have a predictive relevance.

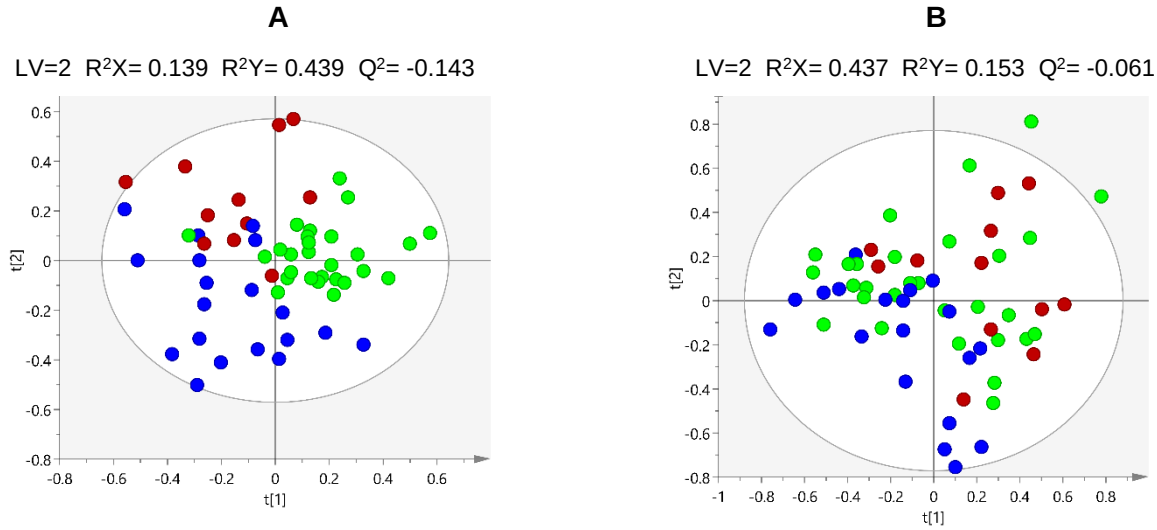


Figure 13 - A) PLS-DA scatter plot for VOCs analysis by comparing stage 0a (*n*=29, ●), stage I (*n*=19, ●), and stages II, III and IV (*n*=12, ●) of BCa. **B)** PLS-DA scatter plot for VCCs analysis by comparing stage 0a (*n*=29, ●), stage I (*n*=19, ●), and stages II, III and IV (*n*=12, ●) of BCa.

Next, a pairwise comparison was performed by comparing stage 0a *vs* stage I, stage 0a *vs* stage II, III and IV, and stage I *vs* II, III and IV. However, a group separation was only found for VOCs concerning stage 0a *vs* stage II, III and IV (***Figure 14***).

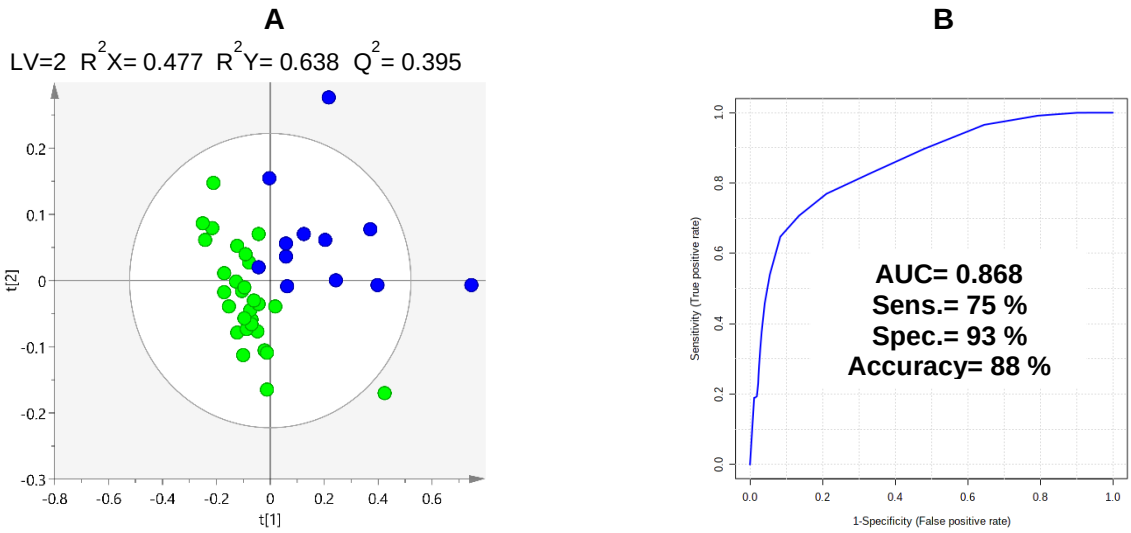


Figure 14 - A) PLS-DA scatter plot for VOCs analysis by comparing stage 0a (*n*=29, ●) *vs* stage II, III and IV (*n*=12, ●). **B)** ROC curve of VOCs by comparing stage 0a *vs* stage II, III and IV with calculated AUC (0.868), sensitivity (75%), specificity (93%) and accuracy (88%).

After PLS-DA, a VIP superior to 1 was considered as relevant for group discrimination. It was possible to identify a total of 13 compounds, being 3 formally identified metabolites (L1), 7 putatively identified (L2) and 2 unidentified compounds defined as “unknown 1,2,3…” (L4). A univariate t-test was also employed to obtain *p*-values and FDR adjusted *p*-value, as well as the ES (***Table 8***).

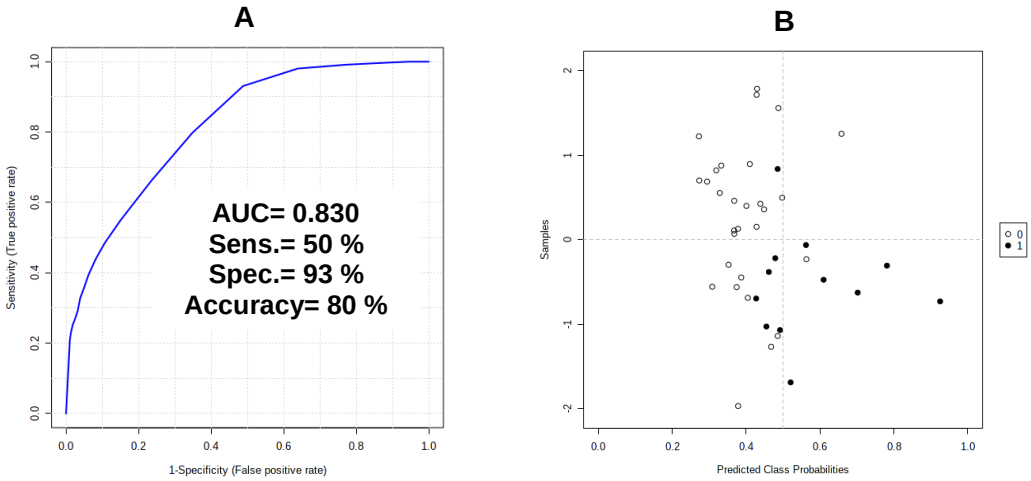
Table 8 - Metabolites that were found to be statistically significant for discrimination between stage 0a *vs* stage II, III and IV present in urine of BCa patients (*n*=13).

Chemical name (IUPAC)	*p*-value original	*p*-value (FDR)	AUC	Effect size	Standard error	Variation	Identification level	Metabolic pathways
2,4-Dimethylheptane	<0.0001	0.0002	0.888	1.82	0.78	↑	L2	-
1,3-Dimethylheptane	<0.0001	0.0002	0.888	1.40	0.74	↑	L2	-
Decane	0.0006	0.0019	0.830	1.32	0.73	↑	L1	-
Undecane	0.0260	0.0290	0.733	1.28	0.73	↑	L1	-
1,2,4-Trimethylbenzene	0.0230	0.0270	0.727	0.71	0.69	↑	L2	-
1,2,3-Trimethylbenzene	0.0340	0.0340	0.713	0.69	0.69	↑	L2	-
1,3-Dimethyl-2-ethylbenzene	0.0062	0.0160	0.770	1.04	0.71	↑	L2	-
1,2,4,5-Tetramethylbenzene	0.0082	0.0180	0.761	1.05	0.71	↑	L2	-
2,5-Dimethylbenzaldehyde	0.0130	0.0240	0.658	-0.68	0.69	↓	L2	Alcohols and fatty acids metabolism
Benzene-like compound	0.0210	0.0270	0.730	0.88	0.70	↑	L3	-
Carvone	0.0180	0.0260	0.736	-0.51	0.68	↓	L1	-
Unknown 7	0.0150	0.0250	0.741	1.01	0.71	↑	L4	-
Unknown 9	0.0006	0.0019	0.830	1.34	0.73	↑	L4	-

L1 – identified compounds, through comparison with a chemical reference standard. **L2** – putatively annotated compounds identified through the comparison with public or commercial libraries. **L3 –** putatively characterised compound classes identified based on spectral and/or physicochemical properties consistent with a particular class of organic compounds. **L4** – unknown compounds that can be detected and quantified but cannot be identified by either the previous methods.

The ROC curve from the 13 discriminant VOCs (11 identified metabolites and 2 unknowns) that were found to be statistically significant for discrimination by comparing stage 0a *vs* stage II, III and IV is presented in **Figure *15A***, as well as the predicted class probabilities (**Figure *15B***) and the confusion matrix (**Figure *15C***).

Figure 16 represents the boxplots of the 10 compounds the most statistically significant for discrimination established as a biomarker panel for stage 0a *vs* stages II, III and IV.



C

		True classes	
		0	**1**
Predicted classes	**0**	27	6
	1	2	6

Figure 15 - ROC analysis. **A)** ROC curve of the discriminant VOCs that were found to be statistically significant for discrimination (11 identified metabolites and 2 unknowns) by comparing stage 0a (*n*=29) *vs* stage II, III and IV (*n*=12) with calculated AUC (0.830), sensitivity (50%), specificity (93%) and accuracy (80%). **B)** Predicted class probabilities of each samples across 100 cross validations. **C)** Confusion matrix.

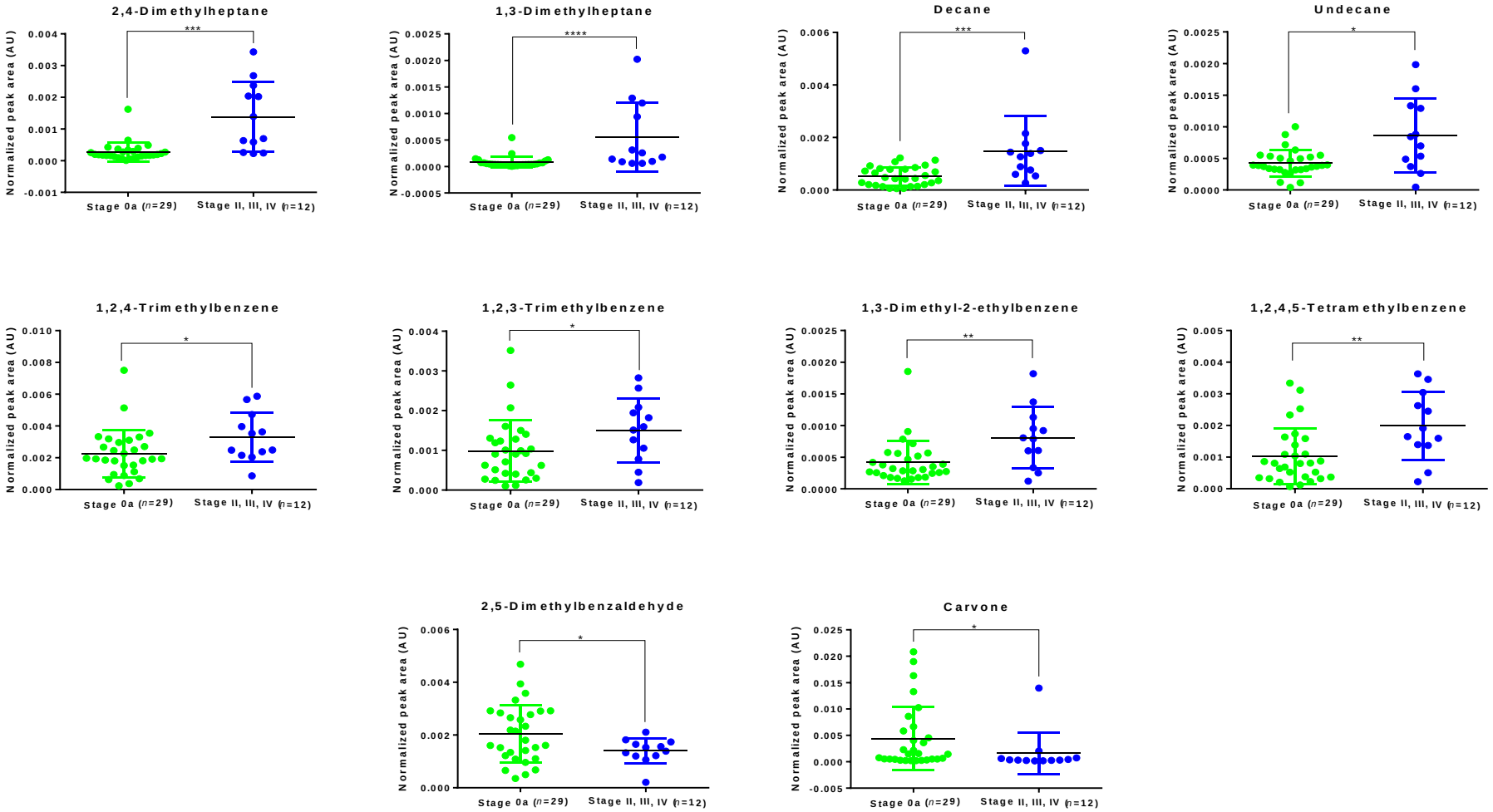


Figure 16 - Boxplots of the compounds the most statistically significant for discrimination (*n*=10) established as a biomarker panel for stage 0a (*n*=29, ●) *vs* stages II, III and IV (*n*=12, ●).

Chapter 4. Discussion

This work unveiled the potential of HS-SPME/GC-MS to evaluate the presence of discriminant VOCs and VCCs on the urinary headspace of BCa patients compared to cancer-free individuals (Cauchi et al., 2016; Jobu et al., 2012). A total of 103 VOCs and 107 VCCs (16 in common) were identified in urine being mainly aldehydes, ketones, aromatic hydrocarbons, and alkanes.

Considering the study of the correlation between the age of patients and the volatiles present in BCa and controls, the data obtained in this work demonstrated a poor correlation comparing the two groups. Some studies employing other matrices like breath, found that the levels of aldehydes had no association with both age and gender (Janfaza et al., 2019a). A study developed by Lima et al., also showed that the age of the patients had no significant influence to discriminate between PC patients and controls (Lima et al., 2019). In a study from Monteiro et al., the results obtained through the identification of a potential VOC-biomarker panel in the urine showed that there were only a small effect on the differences on age to the classification of RCC samples, which might be even negligible or a possible bias (Monteiro et al., 2017).

The comparison of metabolites present in urine of BCa patients between the stages 0a *vs* stages II, III and IV showed a set of 13 statistically significant compounds obtaining a sensitivity of 50%, specificity of 82% and accuracy of 80%. To our knowledge, it is the first time that a comparison of volatile urinary profile is performed for the different stages of the disease. However, some studies found that, for instance, benzaldehyde was one of the compounds that contributed for the discrimination between high grade and low grade cancer cell lines (Rodrigues et al., 2018).

Compared to other studies using the same or similar techniques (Jobu et al., 2012; Khalid et al., 2013; Spaněl et al., 1999), our study was made with an appreciable sample size (BCa samples, *n*=60; control samples, *n*=60) which represents a strong point of this work. Multivariate and univariate analyses enabled to establish a candidate biomarker panel of 30 volatile compounds present in urine of BCa patients compared to controls, unveiling a sensitivity of 85%, specificity of 82% and accuracy of 83%. These results were similar to the results obtained by Cauchi et al. (sensitivity 88.5%, specificity 88.2% and accuracy 88.4% with PLS-DA) and better than the ones obtained by Jobu et al. (55.5% were correctly identified with BCa), even though the sample size in this study was smaller (BCa samples, *n*=9; control samples, *n*=7) (Cauchi et al., 2016; Jobu et al., 2012).

Different processes occur in the body as a result of cancer, one of them is lipid peroxidation, polyunsaturated fatty acids present in the phospholipid membrane underlies oxidation reactions, both enzymatic (like cytochrome P450 – CYP450) and non-enzymatic (Janfaza et al., 2019a; Lima et al., 2019; Medeiros, 2019). Such events tend to produce phospholipid compounds that can decompose into electrophilic derivatives like aldehydes. Additionally, the decomposition of lipid hydroperoxides can also generate radical and non-radical products like alkanes, ketones and other compounds (Medeiros, 2019).

An aldehyde that has been found altered in urine of patients with several cancer types is hexanal, being correlated with 8 different cancer types, which proves that it can be a potential biomarker for cancer detection (e.g., lung cancer and renal cancer) (Dator et al., 2019; Janfaza et al., 2019a; Rocha, 2019). In a study developed by Cauchi et al. the authors identified that hexanal was increased in BCa urine samples compared with controls (Cauchi et al., 2016). These results are in contrast with our study, where hexanal was found to be decreased in BCa samples. Moreover, Cauchi et al. found benzaldehyde as being increased in BCa samples, which has also been identified as being increased in our BCa samples (Cauchi et al., 2016). A study developed by Rodrigues et al., also found benzaldehyde significantly increased within the cancer cells medium in comparison to the non-tumorigenic medium (Rodrigues et al., 2018).

Another compound found in our work and already associated to BCa is formaldehyde. This compound has been the target of a study developed in 1999 by Spaněl et al. where its levels were found increased in the headspace of urine of BCa patients compared with controls, and also at higher levels when compared with PC patients (Spaněl et al., 1999). Though, in our work, formaldehyde was found to be decreased in BCa samples compared with controls. A possible explanation mentioned by Spaněl et al. is that both patients with BCa and PC were at diverse stages of the disease and some of them were already in remission, which might have influenced the concentrations of formaldehyde in the urine of the patients (Spaněl et al., 1999).

In the present study, the aldehydes that were found to change in BCa for the first time included 2-furaldehyde, glyoxal, methylglyoxal and phenylacetaldehyde unveiling down-regulated levels in BCa compared to controls.

Despite being identified for the first time in this work as candidate biomarkers of BCa, some compounds like glyoxal and methylglyoxal have already been found to be decreased, in other types of cancers like renal and prostate (Lima et al., 2019; Rocha, 2019). As for phenylacetaldehyde, this compound has also been associated to PC, being significantly increased in urine of cancer patients compared with controls, which is in contrast with this work, where it was found decreased (Lima et al., 2019). However, it has also been reported that the concentration of aldehydes can be both reduced or increased

in association with cancerous diseases. In some studies, done with cells, it has been noticed that the aldehydes present around the cancer cells were decreased compared with the control medium, suggesting that those aldehydes can be consumed by the cancerous cells. Another explanation for decreased aldehyde levels might be due to higher activity of aldehyde dehydrogenases (ALDH) in cancer cells since it has been associated with different types of cancer cells like lung cancer and oesophageal cancer (Lima et al., 2019).

The association of BCa and the smoking habits of patients was not conducted however, it is known that cigarette smoke contain carcinogenic products for example acetaldehyde, formaldehyde, benzene and naphthalene derivatives which represent many of the compounds found in this work (American Lung Association, 2020) However, concerning aldehydes, more studies are necessary to understand its formation in several types of cancer. Part of this is due to the fact that some aldehydes like formaldehyde and acetaldehyde can originate from exogenous sources like environment of even being a product of metabolism processes like the generation of acetaldehyde from ethanol by the gut flora (Janfaza et al., 2019a). In this work, compared to the literature, most of the discriminant aldehydes were found decreased in BCa compared to controls, except for benzaldehyde which was also found increased in other types of cancers like lung cancer and breast cancer, and tends to increase as reactive oxygen species (ROS) rise as a result of inflammation and oxidative stress (Janfaza et al., 2019a; Rodrigues et al., 2018; Woollam et al., 2019; Zimmermann et al., 2007). Besides that, the decreased concentrations of aldehydes found in this work might also be the result of their low concentrations, in general, in biological samples as long as they have propensity to rapidly react with other compounds (Janfaza et al., 2019a).

Alkanes or saturated hydrocarbons can be originated from lipid peroxidation. However, their presence might be associated with their solubility in the different biological matrices. A possible explanation for the increased levels of alkanes and methylated alkanes in cancer, can be related with the variable activity of CYP450. Yet, the origin of these compounds is not clearly understood since some authors consider it as secondary products of oxidative stress while others assume that it can originate from the mevalonic acid pathway of cholesterol synthesis (Janfaza et al., 2019a; Silva et al., 2011). The alkanes found in this work, namely 2,4-dimethylhexane and 2-methylnonane, were both up-regulated in BCa samples compared with controls and associated with BCa for the first time in this work.

Aromatic hydrocarbons, such as benzene derivatives, specially found significantly altered at more advanced stages of the disease in this work, were also found increased in the headspace of cancerous cells of lung cancer as mentioned in other studies (Schmidt

and Podmore, 2015a). The origin of some other compounds, like the ones derived from naphthalene, is not fully understood, possibly being associated to processes that originate from the degradation products of steroids (Silva et al., 2011). Nevertheless, it must also be considered that such compounds are ubiquitously present in the environment, having strong carcinogenic effects, and being associated to several types of cancers such as lung, breast, stomach, colon, bladder cancer, and others (Rengarajan et al., 2015). In this study these compounds included 1-methylnaphtalene, 2-methylnapthalene and 4-methylphenol, which were all up-regulated in BCa samples compared with controls and were found to change in BCa for the first time.

The existence of cancer also takes the organism to produce ketones and alcohols, this happen due to the local hypoxia, that generates over-proliferation of cells, which leads to anaerobic respiration that in its turn, will promote the release of these compounds through the glycolytic pathway (Amor et al., 2019). Some very common ketones already associated to BCa are 2-butanone, 2-pentanone and 4-heptanone. The work developed by Cauchi et al. also found that compounds like 2-pentanone (also associated to CRC) and 4-heptanone was found to be decreased in BCa samples as in our work. By the contrary, 2-butanone was found to be increased in the author's work, which do not happen in our work, where this compound was found to be decreased (Cauchi et al., 2016). Still, another work developed by Rodrigues et al., found that 2-pentanone was increased in the Scaber cell line (representing SCC and high-grade BCa) compared with the non-tumorigenic SV-HUC-1 cell line (Rodrigues, 2016). Apart from being already associated with BCa, compounds like 2-butanone and 2-pentanone have already been found in other types of cancers like PC; 2-pentanone was also found to be present in BCa cell lines (Lima et al., 2019; Woollam et al., 2019); and 4-heptanone was already associated to RCC (Wang et al., 2016).

Some compounds found, namely the unknown compounds, also were considered responsible for group discrimination. Among these, 6 compounds were found to be increased in BCa samples compared with controls and 8 were found to be decreased.

Urine can represent a very complex matrix to analyse, since it corresponds to the reservoir of all end up products of metabolism and its composition is influenced by several exogenous sources like food intake, environment, smoking habits, and others (Gao and Lee, 2019). For this reason, it is expectable to find several compounds that cannot be identified in a first approach, however, an attempt of identification of these compounds is possible by using NMR and/or MS techniques (Boiteau et al., 2018). This identification can be achieved by comparison of peaks and its *m/z* and RT values by considering the unknown compound and a commercial standard considered similar. Nevertheless, it is important to consider that these methods used for the correct identification of an unknown

compound are costly and require time and effort, for that reason, other approaches have been developed (Bowen and Northen, 2010). One of those approaches focus on the combination of MS and NMR called SUMMIT MS/NMR, which combines data obtained by both techniques and chemical informatics, and attempt to reduce the time of identification and be suitable for a variety of complex mixtures (Leggett et al., 2020). Another approach relies on the application of MS/MS to find a more rightful standard for identification. Also, implementation of isotopic marking or chemical derivatization can help to confirm metabolite structures based on the changes in mass in particular fragments or even create associations of metabolic pathways (Boiteau et al., 2018; Bowen and Northen, 2010). Additionally, the heatmaps of correlation allow us to identify compounds that are biochemically similar or involved in the same pathways.

Chapter 5. Conclusions and future perspectives

This work showed the importance and the applicability of HS-SPME/GC-MS on the evaluation of the presence of discriminant VOCs and VCCs in the urine of BCa patients compared with cancer-free controls. The results obtained demonstrated a satisfactory overall performance of the PLS-DA models to discriminate BCa patients from controls, enabling the definition of a set of 30 candidate biomarkers (85% sensitivity, 82% specificity and 83% accuracy) that can be used in the future as a non-invasive tool for BCa detection. A stage comparison was performed for the first time in this study and a discrimination between advanced stages and early stages was accomplished based on the alterations observed in the levels of 13 volatile metabolites (50% sensitivity, 82% specificity and 80% accuracy). The identification of the unknown compounds still represents a challenge in metabolomic studies. Its correct identification is necessary since it will allow a better understanding of biochemical pathways as well as their role in mechanisms associated with cancer.

Future studies should rely on the selection of a smaller panel of candidate biomarkers for both BCa detection and staging, as well as the validation of these results in a larger cohort. Besides that, a possible applicability of these results into clinics using a volatile sensor-based approach may help to improve cancer related mortality and more effective therapies.

References

Alonso, A., Marsal, S., and Julià, A. (2015). Analytical methods in untargeted metabolomics: State of the art in 2015. *Front. Bioeng. Biotechnol.* 3, 1–20. doi:10.3389/fbioe.2015.00023.

American Cancer Society (2016). About bladder cancer. Available at: http://www.cancer.org/cancer/bladdercancer/detailedguide/bladder-cancer-what-is-bladder-cancer.

American Cancer Society (2019a). Bladder cancer stages. *Early Detect. diagnosis, staging*. Available at: https://www.cancer.org/cancer/bladder-cancer/detection-diagnosis-staging/staging.html [Accessed October 9, 2019].

American Cancer Society (2019b). Survival rates for bladder cancer. *Early Detect. diagnosis, staging*. Available at: https://www.cancer.org/cancer/bladder-cancer/detection-diagnosis-staging/survival-rates.html [Accessed October 11, 2019].

American Cancer Society (2019c). What Is Bladder Cancer? *About Bl. cancer*. Available at: https://www.cancer.org/cancer/bladder-cancer/about/what-is-bladder-cancer.html.

American Cancer Society (2020). Bladder cancer causes, risk factors, and prevention.

American Lung Association (2020). What's in a cigarette. *Smok. Facts*. Available at: https://www.lung.org/quit-smoking/smoking-facts/whats-in-a-cigarette [Accessed August 13, 2020].

Amor, R. E., Nakhleh, M. K., Barash, O., and Haick, H. (2019). Breath analysis of cancer in the present and the future. *Eur. Respir. Rev.* 28, 1–10. doi:10.1183/16000617.0002-2019.

Angle, C., Waggoner, L. P., Ferrando, A., Haney, P., and Passler, T. (2016). Canine detection of the volatilome: A review of implications for pathogen and disease detection. *Front. Vet. Sci.* 3, 1–7. doi:10.3389/fvets.2016.00047.

Antunes, H., Tavares-da-silva, E., Marques, I., Carvalho, J., Abrantes, A. M., and Alves, M. (2018). Metabolomics and Bladder Cancer : Current State and Future Perspectives. *Acta urológica Port.* 35, 46–51.

Armstrong, J. A. (2007). Urinalysis in western culture: A brief history. *Kidney Int.* 71, 384–387. doi:10.1038/sj.ki.5002057.

Banday, K. M., Nanda, R. K., and Chauhan, V. S. (2011). Tuberculosis diagnosis by volatile organic compounds in urine. *Labome* 1. doi:http://dx.doi.org/10.13070/mm.en.1.84.

Bartel, J., Krumsiek, J., and Theis, F. J. (2013). Statistical methods for the analysis of high-throughput metabolomics data. *Comput. Struct. Biotechnol. J.* 4, e201301009. doi:10.5936/csbj.201301009.

Bax, C., Taverna, G., Eusebio, L., Sironi, S., Grizzi, F., Guazzoni, G., et al. (2018). Innovative diagnostic methods for early prostate cancer detection through urine analysis: A review. *Cancers (Basel).* 10, 1–29. doi:10.3390/cancers10040123.

Bento, M. J. (2010). Registo Oncológico Nacional 2010.

Berben, L., Sereika, S. M., and Engberg, S. (2012). Effect size estimation: Methods and examples. *Int. J. Nurs. Stud.* 49, 1039–1047. doi:10.1016/j.ijnurstu.2012.01.015.

Berenjian, A., Chan, N., and Malmiri, H. J. (2012). Volatile organic compounds removal methods: A review. *Am. J. Biochem. Biotechnol.* 8, 220–229. doi:10.3844/ajbbsp.2012.220.229.

Bernabei, M., Pennazza, G., Santonico, M., Corsi, C., Roscioni, C., Paolesse, R., et al. (2008). A preliminary study on the possibility to diagnose urinary tract cancers by an electronic nose. *Sensors Actuators, B Chem.* 131, 1–4. doi:10.1016/j.snb.2007.12.030.

Boiteau, R. M., Hoyt, D. W., Nicora, C. D., Kinmonth-Schultz, H. A., Ward, J. K., and Bingol, K. (2018). Structure elucidation of unknown metabolites in metabolomics by combined NMR and MS/MS prediction. *Metabolites* 8. doi:10.3390/metabo8010008.

Bowen, B. P., and Northen, T. R. (2010). Dealing with the unknown: Metabolomics and metabolite atlases. *J. Am. Soc. Mass Spectrom.* 21, 1471–1476. doi:10.1016/j.jasms.2010.04.003.

Broza, Y. Y., Mochalski, P., Ruzsanyi, V., Amann, A., and Haick, H. (2015). Hybrid volatolomics and disease detection. *Angew. Chemie - Int. Ed.* 54, 11036–11048. doi:10.1002/anie.201500153.

Bujak, R., Struck-Lewicka, W., Markuszewski, M. J., and Kaliszan, R. (2014). Metabolomics for laboratory diagnostics. *J. Pharm. Biomed. Anal.* 113, 108–120. doi:10.1016/j.jpba.2014.12.017.

Cahill, M. (2015). What medieval doctors used to do with urine was once disgusting,

cutting edge medicine. *Viralnova*. Available at: http://www.viralnova.com/urine-doctors/.

Cancer.net (2019). Bladder Cancer: Introduction. *Types Bl. cancer*. Available at: https://www.cancer.net/cancer-types/bladder-cancer/introduction.

Cancer Research UK (2018). Worldwide cancer statistics. *Cancer Res. UK*. Available at: https://www.cancerresearchuk.org/health-professional/cancer-statistics/worldwide-cancer#heading-One.

Capelli, L., Taverna, G., Bellini, A., Eusebio, L., Buffi, N., Lazzeri, M., et al. (2016). Application and uses of electronic noses for clinical diagnosis on urine samples: A review. *Sensors (Switzerland)* 16, 1–23. doi:10.3390/s16101708.

Cauchi, M., Weber, C. M., Bolt, B. J., Spratt, P. B., Bessant, C., Turner, D. C., et al. (2016). Evaluation of gas chromatography mass spectrometry and pattern recognition for the identification of bladder cancer from urine headspace. *Anal. Methods* 8, 4037–4046. doi:10.1039/c6ay00400h.

Chan, D. K., Leggett, C. L., and Wang, K. K. (2016). Diagnosing gastrointestinal illnesses using fecal headspace volatile organic compounds. *World J. Gastroenterol.* 22, 1639–1649. doi:10.3748/wjg.v22.i4.1639.

Chauhan, A., Goyal, M. K., and Chauhan, P. (2014). GC-MS technique and its analytical applications in science and technology. *J. Anal. Bioanal. Tech.* 5, 1000222. doi:10.4172/2155-9872.1000222.

Chen, S., Wang, Y., and Choi, S. (2013). Applications and technology of electronic nose for clinical diagnosis. *Open J. Appl. Biosens.* 2, 39–50. doi:10.4236/ojab.2013.22005.

Chong, J., Wishart, D. S., and Xia, J. (2019). Using MetaboAnalyst 4.0 for Comprehensive and Integrative Metabolomics Data Analysis. *Curr. Protoc. Bioinforma.* 68, 1–128. doi:10.1002/cpbi.86.

Chromacademy (2012). Theory and instrumentation of GC sample introduction. Available at: www.chromacademy.com.

Chromacademy (2019a). GC-MS process. *Chromacademy*. Available at: https://www.chromacademy.com/frameset-chromacademy.html?fChannel=3&fCourse=5&fSco=101&fPath=sco101/gcms_C1_008.asp.

Chromacademy (2019b). The why, what, and how of headspace GC. Available at:

https://www.chromacademy.com/chromatography-headspace-GC.html.

Dana-Farber Cancer Institute (2019). Bladder cancer. *How we diagnose Bl. cancer*. Available at: https://www.dana-farber.org/bladder-cancer/diagnosis/ [Accessed October 7, 2019].

Dator, R. P., Solivio, M. J., Villalta, P. W., and Balbo, S. (2019). Bioanalytical and mass spectrometric methods for aldehyde profiling in biological fluids. *Toxics* 7, 1–35. doi:10.3390/TOXICS7020032.

de Boer, N. K. H., de Meij, T. G. J., Oort, F. A., Ben Larbi, I., Mulder, C. J. J., van Bodegraven, A. A., et al. (2014). The scent of colorectal cancer: Detection by volatile organic compound analysis. *Clin. Gastroenterol. Hepatol.* 12, 1085–1089. doi:10.1016/j.cgh.2014.05.005.

De Lacy Costello, B., Amann, A., Al-Kateb, H., Flynn, C., Filipiak, W., Khalid, T., et al. (2014). A review of the volatiles from the healthy human body. *J. Breath Res.* 8, 1–29. doi:10.1088/1752-7155/8/1/014001.

DeGeorge, K. C., Holt, H. R., and Hodges, S. C. (2017). Bladder Cancer: Diagnosis and Treatment. *Am. Fam. Physician* 96, 507–514.

Dragonieri, S., Quaranta, V. N., Carratu, P., Ranieri, T., and Resta, O. (2016). Influence of age and gender on the profile of exhaled volatile organic compounds analyzed by an electronic nose. *J. Bras. Pneumol.* 42, 143–145. doi:10.1590/s1806-37562015000000195.

Duffy, E., and Morrin, A. (2019). Endogenous and microbial volatile organic compounds in cutaneous health and disease. *TrAC - Trends Anal. Chem.* 111, 163–172. doi:10.1016/j.trac.2018.12.012.

el Manouni el Hassani, S., Berkhout, D., Bosch, S., Benninga, M., de Boer, N., and de Meij, T. (2018). Application of fecal volatile organic compound analysis in clinical practice: Current state and future perspectives. *Chemosensors* 6, 29. doi:10.3390/chemosensors6030029.

Elmasry, G., Kamruzzaman, M., Sun, D. W., and Allen, P. (2012). Principles and applications of hyperspectral imaging in quality evaluation of agro-food products: A review. *Crit. Rev. Food Sci. Nutr.* 52, 999–1023. doi:10.1080/10408398.2010.543495.

Engel, J. (2019). Multivariate statistical analysis. *Futur. Learn*. Available at: https://www.futurelearn.com/courses/metabolomics/0/steps/10717.

Fawcett, T. (2006). An introduction to ROC analysis. *Pattern Recognit. Lett.* 27, 861–874. doi:10.1016/j.patrec.2005.10.010.

Fernandes, M. P., Venkatesh, S., and Sudarshan, B. G. (2015). Early detection of lung cancer using nano-nose - A review. *Open Biomed. Eng. J.* 9, 228–233. Available at: http://www.ncbi.nlm.nih.gov/pubmed/16060297.

Fiehn, O. (2016). Metabolomics by gas chromatography-mass spectrometry: the combination of targeted and untargeted profiling. *Curr. Protoc. Mol. Biol.* 114, 1–32. doi:10.1186/s40945-017-0033-9.Using.

Filipiak, W., Mochalski, P., Filipiak, A., Ager, C., Cumeras, R., Davis, C. E., et al. (2016). A compendium of volatile organic compounds (VOCs) released by human cell lines. *Curr. Med. Chem.* 23, 2112–2131. doi:10.2174/09.

Fleming-Jones, M. E., and Smith, R. E. (2003). Volatile organic compounds in foods: A five year study. *J. Agric. Food Chem.* 51, 8120–8127. doi:10.1021/jf0303159.

Gallagher, M., Wysocki, C. J., Leyden, J. J., Spielman, A. I., Sun, X., Preti, G., et al. (2008). Analyses of volatile organic compounds from human skin. *Br. J. Dermatol.* 159, 780–791. doi:10.1111/j.1365-2133.2008.08748.x.

Gao, Q., and Lee, W.-Y. (2019). Urinary metabolites for urological cancer detection: a review on the application of volatile organic compounds for cancers. *Am. J. Clin. Exp. Urol.* 7, 232–248. Available at: http://www.ncbi.nlm.nih.gov/pubmed/31511830%0Ahttp://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC6734043.

Goodison, S., Rosser, C. J., and Urquidi, V. (2013). Bladder cancer detection and monitoring: Assessment of urine- and blood-based marker tests. *Mol. Diagnosis Ther.* 17, 71–84. doi:10.1007/s40291-013-0023-x.

Gromski, P., Muhamadali, H., Ellis, D., Xu, Y., Correa, E., Turner, M., et al. (2015). A tutorial review: Metabolomics and partial least squares-discriminant analysis - a marriage of convenience or a shotgun wedding. *Anal. Chim. Acta* 879, 10–23. doi:10.1016/j.aca.2015.02.012.

Gromski, P., Xu, Y., Kotze, H., Correa, E., Ellis, D., Armitage, E., et al. (2014). Influence of missing values substitutes on multivariate analysis of metabolomics data. *Metabolites* 4, 433–452. doi:10.3390/metabo4020433.

Harvey, D. (2000). *Modern analytical chemistry*. Available at: www.mhhe.com.

Health Engine (2004). Bladder cancer: Squamous cell carcinoma of the bladder. *Stat. Bl. Cancer (Squamous Cell Carcinoma Bl.* Available at: https://healthengine.com.au/info/bladder-cancer-squamous-cell-carcinoma-of-the-bladder.

Heers, H., Gut, J. M., Hegele, A., Hofmann, R., Boeselt, T., Hattesohl, A., et al. (2018). Non-invasive detection of bladder tumors through volatile organic compounds: A pilot study with an electronic nose. *Anticancer Res.* 38, 833–837. doi:10.21873/anticanres.12291.

Horstmann, M., Steinbach, D., Fischer, C., Enkelmann, A., Grimm, M.-O., and Voss, A. (2015). Pd25-03 an electronic nose system detects bladder cancer in urine specimen: First results of a pilot study. *J. Urol.* 193, e560–e561. doi:10.1016/j.juro.2015.02.1652.

Inamura, K. (2018). Bladder cancer: New insights into its molecular pathology. *Cancers (Basel).* 10. doi:10.3390/cancers10040100.

Instituto CUF de oncologia (2017). Cancro da bexiga. Available at: https://www.saudecuf.pt/oncologia/o-cancro/cancro-da-bexiga [Accessed October 1, 2019].

Jadoon, S., Karim, S., Akram, M. R., Kalsoom Khan, A., Zia, M. A., Siddiqi, A. R., et al. (2015). Recent developments in sweat analysis and its applications. *Int. J. Anal. Chem.* 2015, 1–7. doi:10.1155/2015/164974.

Jalal, A. H., Alam, F., Roychoudhury, S., Umasankar, Y., Pala, N., and Bhansali, S. (2018). Prospects and challenges of volatile organic compound sensors in human healthcare. *ACS Sensors* 3, 1246–1263. doi:10.1021/acssensors.8b00400.

Janfaza, S., Khorsand, B., Nikkhah, M., and Zahiri, J. (2019a). Digging deeper into volatile organic compounds associated with cancer. *Biol. Methods Protoc.* 4, 1–11. doi:10.1093/biomethods/bpz014.

Janfaza, S., Khorsand, B., Nikkhah, M., Zahiri, J., and Nojavani, M. B. (2019b). Cancer Odor Database (COD). Available at: http://bioinf.modares.ac.ir/software/COD/index.html.

Janfaza, S., Nojavani, M., Khorsand, B., Nikkhah, M., and Zahiri, J. (2017). Cancer Odor Database (COD): a critical databank for cancer diagnosis research. *Database (Oxford).* 2017, 1–6. doi:10.1093/database/bax055.

Jia, Z., Patra, A., Kutty, V. K., and Venkatesan, T. (2019). Critical review of volatile

organic compound analysis in breath and in vitro cell culture for detection of lung cancer. *Metabolites* 9, 1–17. doi:10.3390/metabo9030052.

Jin, H. (2018). Volatile organic compound sensors – the future of cancer detection? *Res. Featur.*, 1–4. doi:10.26904/rf-125-2225.

Jobu, K., Sun, C., Yoshioka, S., Yokota, J., Onogawa, M., Kawada, C., et al. (2012). Metabolomics study on the biochemical profiles of odor elements in urine of human with bladder cancer. *Biol. Pharm. Bull.* 35, 639–642. doi:10.1248/bpb.35.639.

Karaman, I. (2017). "Prepocessing and pretreatment of metabolomics data for statistical analysis," in *Metabolomics: From Fundamentals to Clinical Applications*, ed. A. Sussulini, 145–161. doi:10.1007/978-3-319-47656-8.

Khalid, T., White, P., De Lacy Costello, B., Persad, R., Ewen, R., Johnson, E., et al. (2013). A Pilot Study Combining a GC-Sensor Device with a Statistical Model for the Identification of Bladder Cancer from Urine Headspace. *PLoS One* 8, 1–8. doi:10.1371/journal.pone.0069602.

Krilaviciute, A., Heiss, J. A., Leja, M., Kupcinskas, J., Haick, H., and Brenner, H. (2015). Detection of cancer through exhaled breath: a systematic review. *Oncotarget* 6, 38643–38657. doi:10.18632/oncotarget.5938.

Lagay, F. (2002). The legacy of humoral medicine. *AMA J. Ethics* 4. doi:10.1001/virtualmentor.2002.4.7.mhst1-0207.

Lee, L. C., Liong, C. Y., and Jemain, A. A. (2018). Partial least squares-discriminant analysis (PLS-DA) for classification of high-dimensional (HD) data: A review of contemporary practice strategies and knowledge gaps. *Analyst* 143, 3526–3539. doi:10.1039/c8an00599k.

Leggett, A., Wang, C., Li, D.-W., Somogyi, A., Bruschweiler-Li, L., and Brüschweiler, R. (2020). Identification of unknown metabolomics mixture compounds by combining NMR, MS and cheminformatics. *Methods Enzymol.* 615, 407–422.

Li, W., Liu, H., Xie, D., He, Z., and Pi, X. (2017). Lung cancer screening based on type-different sensor arrays. *Sci. Rep.* 7, 1–12. doi:10.1038/s41598-017-02154-9.

Lima, A. R., Pinto, J., Azevedo, A. I., Barros-Silva, D., Jerónimo, C., Henrique, R., et al. (2019). Identification of a biomarker panel for improvement of prostate cancer diagnosis by volatile metabolic profiling of urine. *Br. J. Cancer* 121, 857–868. doi:10.1038/s41416-019-0585-4.

Lubes, G., and Goodarzi, M. (2017). Analysis of volatile compounds by advanced analytical techniques and multivariate chemometrics. *Chem. Rev.* 117, 6399–6422. doi:10.1021/acs.chemrev.6b00698.

Manzi, M., Riquelme, G., Zabalegui, N., and Monge, M. E. (2020). Improving diagnosis of genitourinary cancers: Improving diagnosis of genitourinary cancers: Biomarker discovery strategies through mass spectrometry-based metabolomics. *J. Pharm. Biomed. Anal.* 178. doi:10.1016/j.jpba.2019.112905.

Mastrangelo, A., Ferrarini, A., Rey-Stolle, F., García, A., and Barbas, C. (2015). From sample treatment to biomarker discovery: A tutorial for untargeted metabolomics based on GC-(EI)-Q-MS. *Anal. Chim. Acta* 900, 21–35. doi:10.1016/j.aca.2015.10.001.

Mazzone, P. J., Wang, X. F., Lim, S., Choi, H., Jett, J., Vachani, A., et al. (2015). Accuracy of volatile urine biomarkers for the detection and characterization of lung cancer. *BMC Cancer* 15, 15–20. doi:10.1186/s12885-015-1996-0.

Medeiros, M. H. G. (2019). DNA damage by endogenous and exogenous aldehydes. *J. Braz. Chem. Soc.* 30, 2000–2009. doi:10.21577/0103-5053.20190056.

Monteiro, M., Moreira, N., Pinto, J., Pires-Luís, A. S., Henrique, R., Jerónimo, C., et al. (2017). GC-MS metabolomics-based approach for the identification of a potential VOC-biomarker panel in the urine of renal cell carcinoma patients. *J. Cell. Mol. Med.* 21, 2092–2105. doi:10.1111/jcmm.13132.

Nakhleh, M. K., Amal, H., Jeries, R., Broza, Y. Y., Aboud, M., Gharra, A., et al. (2017). Diagnosis and classification of 17 diseases from 1404 subjects via pattern analysis of exhaled molecules. *ACS Nano* 11, 112–125. doi:10.1021/acsnano.6b04930.

National Cancer Institute (2018a). Cancer Stat Facts: Bladder Cancer. *Cancer Stat.* Available at: https://seer.cancer.gov/statfacts/html/urinb.html [Accessed October 31, 2019].

National Cancer Institute (2018b). Online Summary of Trends in US Cancer Control Measures. *Natl. Cancer Inst. Cancer Trends Prog. Rep.* Available at: https://progressreport.cancer.gov/diagnosis/incidence.

Nicholson, J., Lindon, J., and Holmes, E. (1999). “Metabonomics”: understanding the metabolic responses of living systems to pathophysiological stimuli via multivariate statistical analysis of biological NMR spectroscopic data. *Xenobiotica* 29, 1187–1189. doi:10.1080/004982599238047.

Oeyen, E., Hoekx, L., De Wachter, S., Baldewijns, M., Ameye, F., and Mertens, I. (2019). Bladder cancer diagnosis and follow-up: The current status and possible role of extracellular vesicles. *Int. J. Mol. Sci.* 20. doi:10.3390/ijms20040821.

Oguma, T., Nagaoka, T., Kurahashi, M., Kobayashi, N., Yamamori, S., Tsuji, C., et al. (2017). Clinical contributions of exhaled volatile organic compounds in the diagnosis of lung cancer. *PLoS One* 12, 1–10. doi:10.1371/journal.pone.0174802.

Petry, N. M., Wagner, J. A., Rash, C. J., and Hood, K. K. (2015). Perceptions about professionally and non-professionally trained hypoglycemia detection dogs. *Diabetes Res. Clin. Pract.* 109, 389–396. doi:10.1097/CCM.0b013e31823da96d.Hydrogen.

Pluskal, T., Castillo, S., Villar-Briones, A., and Orešič, M. (2010). MZmine 2: Modular framework for processing, visualizing, and analyzing mass spectrometry-based molecular profile data. *BMC Bioinformatics* 11. doi:10.1186/1471-2105-11-395.

Razmaria, A. (2015). Bladder cancer. *J. Am. Med. Assoc.* 314, 1886.

Rengarajan, T., Rajendran, P., Nandakumar, N., Lokeshkumar, B., Rajendran, P., and Nishigaki, I. (2015). Exposure to polycyclic aromatic hydrocarbons with special focus on cancer. *Asian Pac. J. Trop. Biomed.* 5, 182–189. doi:10.1016/S2221-1691(15)30003-4.

Richters, A., Aben, K. K. H., and Kiemeney, L. A. L. M. (2020). The global burden of urinary bladder cancer: an update. *World J. Urol.* 38, 1895–1904. doi:10.1007/s00345-019-02984-4.

Rocha, S. F. (2019). Metabolomic studies in human renal cell carcinoma cell lines using gas chromatography/mass spectrometry.

Rodrigues, D. F. G. (2016). Volatile Organic Compounds for the Detection of Bladder Cancer : an In Vitro Metabolomic Approach.

Rodrigues, D., Pinto, J., Araújo, A. M., Monteiro-Reis, S., Jerónimo, C., Henrique, R., et al. (2018). Volatile metabolomic signature of bladder cancer cell lines based on gas chromatography–mass spectrometry. *Metabolomics* 14, 1–15. doi:10.1007/s11306-018-1361-9.

Roser, M., and Ritchie, H. (2018). Cancer. *Our World Data*. Available at: https://ourworldindata.org/cancer.

Saccenti, E., Hoefsloot, H. C. J., Smilde, A. K., Westerhuis, J. A., and Hendriks, M. M. W. B. (2014). Reflections on univariate and multivariate analysis of metabolomics data.

Metabolomics 10, 361–374. doi:10.1007/s11306-013-0598-6.

Saginala, K., Barsouk, A., Aluru, J. S., Rawla, P., Padala, S. A., and Barsouk, A. (2020). Epidemiology of Bladder Cancer. *Med. Sci.* 8, 1–12.

Schmidt, K., and Podmore, I. (2015a). Current challenges in volatile organic compounds analysis as potential biomarkers of cancer. *J. Biomarkers* 2015, 1–16. doi:10.1155/2015/981458.

Schmidt, K., and Podmore, I. (2015b). Solid Phase Microextraction (SPME) Method Development in Analysis of Volatile Organic Compounds (VOCS) as Potential Biomarkers of Cancer. *J. Mol. Biomark. Diagn.* 06. doi:10.4172/2155-9929.1000253.

Shirasu, M., and Touhara, K. (2011). The scent of disease: Volatile organic compounds of the human body related to disease and disorder. *J. Biochem.* 150, 257–266. doi:10.1093/jb/mvr090.

Silva, C. L., Passos, M., and Cmara, J. S. (2011). Investigation of urinary volatile organic metabolites as potential cancer biomarkers by solid-phase microextraction in combination with gas chromatography-mass spectrometry. *Br. J. Cancer* 105, 1894–1904. doi:10.1038/bjc.2011.437.

Soini, H. A., Klouckova, I., Wiesler, D., Oberzaucher, E., Grammer, K., Dixon, S. J., et al. (2010). Analysis of volatile organic compounds in human saliva by a static sorptive extraction method and gas chromatography-mass spectrometry. *J. Chem. Ecol.* 36, 1035–1042. doi:10.1007/s10886-010-9846-7.

Spaněl, P., Smith, D., Holland, T. A., Singary, W. Al, and Elder, J. B. (1999). Analysis of formaldehyde in the headspace of urine from bladder and prostate cancer patients using selected ion flow tube mass spectrometry. *Rapid Commun. Mass Spectrom.* 13, 1354–1359. doi:10.1002/(sici)1097-0231(19990730)13:14<1354::aid-rcm641>3.3.co;2-a.

Sun, X., Shao, K., and Wang, T. (2016). Detection of volatile organic compounds (VOCs) from exhaled breath as noninvasive methods for cancer diagnosis Young Investigators in Analytical and Bioanalytical Science. *Anal. Bioanal. Chem.* 408, 2759–2780. doi:10.1007/s00216-015-9200-6.

Sussulini, A. (2017). *Metabolomics: From Fundamentals to Clinical Applications*. Springer.

Teodoro-Morrison, T., Diamandis, E. P., Rifai, N., Weetjens, B., Pennazza, G., de Boer, N. K., et al. (2014). Animal olfactory detection of disease : Promises and pitfalls. *Am. Assoc. Clin. Chem.* 60, 1473–1479. doi:10.1373/clinchem.2014.231282.

van de Goor, R. M. G. E., Leunis, N., van Hooren, M. R. A., Francisca, E., Masclee, A., Kremer, B., et al. (2017). Feasibility of electronic nose technology for discriminating between head and neck, bladder, and colon carcinomas. *Eur. Arch. Oto-Rhino-Laryngology* 274, 1053–1060. doi:10.1007/s00405-016-4320-y.

Viant, M. R., Kurland, I. J., Jones, M. R., and Dunn, W. B. (2017). How close are we to complete annotation of metabolomes? *Curr. Opin. Chem. Biol.* 36, 64–69. doi:10.1016/j.cbpa.2017.01.001.

Wang, C., Li, P., Lian, A., Sun, B., Wang, X., Guo, L., et al. (2014). Blood volatile compounds as biomarkers for colorectal cancer. *Cancer Biol. Ther.* 15, 200–206. doi:10.4161/cbt.26723.

Wang, D., Wang, C., Pi, X., Guo, L., Wang, Y., Li, M., et al. (2016). Urinary volatile organic compounds as potential biomarkers for renal cell carcinoma. *Biomed. Reports* 5, 68–72. doi:10.3892/br.2016.686.

Want, E. J., Cravatt, B. F., and Siuzdak, G. (2005). The expanding role of mass spectrometry in metabolite profiling and characterization. *ChemBioChem* 6, 1941–1951. doi:10.1002/cbic.200500151.

Weber, C. M., Cauchi, M., Patel, M., Bessant, C., Turner, C., Britton, L. E., et al. (2011). Evaluation of a gas sensor array and pattern recognition for the identification of bladder cancer from urine headspace. *Analyst* 136, 359–364. doi:10.1039/c0an00382d.

Willis, C. M., Britton, L. E., Harris, R., Wallace, J., and Guest, C. M. (2010). Volatile organic compounds as biomarkers of bladder cancer: Sensitivity and specificity using trained sniffer dogs. *Cancer Biomarkers* 8, 145–153. doi:10.3233/CBM-2011-0208.

Woollam, M., Teli, M., Angarita-Rivera, P., Liu, S., Siegel, A. P., Yokota, H., et al. (2019). Detection of volatile organic compounds (VOCs) in urine via gas chromatography-mass spectrometry QTOF to differentiate between localized and metastatic models of breast cancer. *Sci. Rep.* 9, 1–12. doi:10.1038/s41598-019-38920-0.

Wu, Y., and Li, L. (2016). Sample normalization methods in quantitative metabolomics. *J. Chromatogr. A* 1430, 80–95. doi:10.1016/j.chroma.2015.12.007.

Zhu, C. Z., Ting, H. N., Ng, K. H., and Ong, T. A. (2019). A review on the accuracy of bladder cancer detection methods. *J. Cancer* 10, 4038–4044. doi:10.7150/jca.28989.

Zhu, S., Corsetti, S., Wang, Q., Li, C., Huang, Z., and Nabi, G. (2018a). Optical sensory arrays for the detection of urinary bladder cancer-related volatile organic compounds.

J. Biophotonics 12, 1–10. doi:10.1002/jbio.201800165.

Zhu, S., Corsetti, S., Wang, Q., Li, C., Huang, Z., and Nabi, G. (2018b). Optical sensory arrays for the detection of urinary bladder cancer related volatile organic compounds (VOCs). *J. Biophotonics* e201800165, 1–16. doi:10.1002/jbio.201800165.

Zimmermann, D., Hartmann, M., Moyer, M. P., Nolte, J., and Baumbach, J. I. (2007). Determination of volatile products of human colon cell line metabolism by GC/MS analysis. *Metabolomics* 3, 13–17. doi:10.1007/s11306-006-0038-y.

Appendix

Table A1 - List of all identified VOCs obtained from the chromatograms by using HS-SPME/GC-MS analysis**.**

Compounds	Retention time	m/z	CAS#	Experimental KI	KI from literature	R-match	Level identification
2-Butanone	2.20	57/72	78-93-3	598	-	942	L1
Benzene	2.57	78/77/51/50/52/79/74/76	71-43-2	654	-	938	L1
2-Pentanone	3.03	58/71/86	107-87-9	685	-	939	L1
Pentanal	3.17	58/57	110-62-3	699	-	858	L2
2-Methyl-2-pentanol	3.72	59/87/69	590-36-3	694	-	791	L2
4-Methylpentan-2-one	3.80	58/85	108-10-1	735	-	862	L2
Pyrrole	3.99	52/67	109-97-7	755	-	917	L2
2,3-Dimethylhexane	4.16	55/70	584-94-1	760	-	927	L2
4-Methylheptane	4.28	70/71/55/57/84/85/114	589-53-7	767	-	901	L2
3-Methylheptane	4.41	57/85	589-81-1	773	-	909	L2
3-Hexanone	4.63	71/100	589-38-8	784	-	922	L2
2-Hexanone	4.73	58/57/100	591-78-6	790	-	885	L1
Unknown 1	4.80	59/71/57/55/70/84	-	-	-	-	L4
Hexanal	4.97	56/57/72/55/99	66-25-1	800	803	918	L1
2,4-Dimethylheptane	5.47	57/71/85/91	2213-23-2	821	823	918	L1
Unknown 2	5.77	57/85/114	-	-	834	-	L4
2,2-dimethyl-3-pentanone	5.90	57/114/56/85/58	564-04-5	843	839	915	L2
2,4-Dimethyl-1-heptene	5.93	55/70/83/69	19549-87-2	836	840	902	L2
2-Methyl-cyclopentanone	6.13	69/83/98	1120-72-5	846	848	829	L2
2,3-Dimethylheptane	6.33	57/71/84	3074-71-3	855	856	896	L2
4-Methyloctane	6.53	57/71/85/98/128	2216-34-4	863	863	909	L2
4-Heptanone	6.72	50/58/71/86/99/114	123-19-3	871	871	941	L1

2-Heptanone	7.20	58/71	110-43-0	891	889	893	L1
Xylene	7.29	91/106/105/77 /51/92/7/65/103	95-47-6	887	891	856	L2
Unknown 3	7.43	133/151	-	-	898	-	L4
Heptanal	7.55	55/70/81	111-71-7	901	902	907	L1
Dimethyl sulfone	7.94	79/94	67-71-0	922	915	854	L2
Unknown 4	8.43	55/69/59/71/58/56/75	-	-	932	-	L4
4-Methyl-2-heptanone	8.56	58/85	6137-06-0	943	936	926	L2
Unknown 5	8.73	72/82/100	-	-	942	-	L4
Unknown 6	9.09	77/91/119	-	-	954	-	L4
3-Ethylcyclopentanone	9.18	55/70/83/112	10264-55-8	962	957	916	L2
Benzaldehyde	9.28	51/77/105	100-52-7	962	960	889	L1
2-Methylnonane	9.42	57/98/71/56/70/55	7146-60-3	964	964	871	L2
1-Octen-3-ol	9.86	57/72	3391-86-4	980	979	757	L2
1,2,4-Trimethylbenzene	10.26	51/105/120	95-63-6	990	993	877	L2
Decane	10.49	57/71	124-18-5	1000	1000	925	L1
Octanal	10.55	57/69/84	124-13-0	1003	1002	956	L1
Unknown 7	10.85	57/91	-	-	1012	-	L4
1,2,3-Trimethylbenzene	11.07	51/105/120	526-73-8	1013	1019	932	L2
2,6-Dimethylnonane	11.15	85/113	17302-28-2	1018	1022	885	L2
1-Isopropyl-3-methylbenzene	11.21	91/119/134	535-77-3	1023	1024	899	L2
2-Ethylhexan-1-ol	11.31	57/70/98	104-76-7	1030	1027	918	L2
Eucalyptol	11.44	55/81/108/139	470-82-6	1032	1032	923	L2
3,5-Octadien-2-ol	11.58	55/97/111/126	69668-82-2	1038	1036	811	L2
Phenylacetaldehyde	11.77	91/65/92/120/51/63/50/89	122-78-1	1045	1043	891	L1
Unknown 8	11.98	68/105/116/69/61/134	-	-	1049	-	L4
Unknown 9	12.15	57/71/85/70/56/55/84/69	-	-	1055	-	L4
Unknown 10	12.31	71/85/113	-	-	1060	-	L4

Unknown 11	12.41	77/105/120	-	-	1063	-	L4
1-Octanol	12.58	56/69/70/84	111-87-5	1071	1069	885	L2
4-Methylphenol	12.67	51/63/68/77/90/107	106-44-5	1077	1071	950	L2
1,3-Dimethyl-2-ethylbenzene	12.99	119/134/91/77	934-74-7	1080	1083	904	L2
2-Methoxyphenol	13.01	53/81/109	90-05-1	1090	1083	877	L2
Unknown 12	13.11	95/110/55/67/57/85/111/83	-	-	1086	-	L4
2-Nonanone	13.19	58/71	821-55-6	1092	1089	886	L2
1-Methyl-4-isopropenylbenzene	13.19	132/91/117/115/92/63/131	1195-32-0	1090	1089	796	L2
Linalool	13.47	71/80/93/121	78-70-6	1099	1098	887	L2
Undecane	13.51	57/71/85	1120-21-4	1100	1100	800	L2
Nonanal	13.60	57/70/82/98/114	124-19-6	1104	1103	892	L1
1,2,4,5-Tetramethylbenzene	14.05	119/134/91	95-93-2	1116	1118	901	L2
Unknown 13	14.84	91/109/119	-	-	1145	-	L4
trans-(-)-p-Menthan-3-one	15.15	112/69/55/139/70/97/154/111	14073-97-3	1148	1155	916	L2
Unknown 14	15.24	55/93/121/136	-	-	1158	-	L4
p-Mentha-1,5-dien-8-ol	15.59	59/94/79/91/93/77/119/92/65	1686-20-0	1167	1170	819	L2
Levomenthol	15.78	55/81/95/123	2216-51-5	1175	1177	946	L2
Terpinen-4-ol	15.87	59/ 93/ 121/ 136/ 67/81/68/79/91	562-74-3	1177	1180	835	L2
Naphthalene	15.97	51/102/128	91-20-3	1182	1183	890	L2
α-Terpineol	16.28	59/67/93/121	98-55-5	1189	1193	894	L2
Unknown 15	16.44	84/91/121/93/119/83/79/57	-	-	1199	-	L4
Unknown 16	16.57	135/164/91/55/136/79/77	-	-	1203	-	L4
2,5-Dimethylbenzaldehyde	16.84	51/105/133	5779-94-2	1208	1213	915	L2
4-Isopropylbenzaldehyde	17.38	77/119/148	122-03-2	1239	1232	790	L2
Carvone	17.66	54/82/93/ 108/ 107/ 53/ 106	99-49-0	1242	1242	923	L1
Nonanoic acid	18.26	60/115/158	112-05-0	1273	1264	896	L2
1-Decanol	18.45	70/69/83	112-30-1	1273	1270	867	L1

2-Methylnaphthalene	19.11	142/141/115/143/139/71/63/70/57	9-1-57-6	1298	1295	846	L2
Unknown 17	19.24	82/96/138	-	-	1299	-	L4
2-Methoxy-4-vinylphenol	19.46	77/89/107/135	7786-61-0	1317	1307	862	L2
1-Methylnaphthalene	19.53	142/141/115	90-12-0	1307	1311	848	L2
4-Hydroxy-3-methylacetophenone	19.92	77/105/135/150	876-02-8	1323	1324	850	L2
Unknown 18	20.37	193/95/109/138/83/57/137/124/67/110	-	-	1341	-	L4
Eugenol	20.60	55/77/103/149/164	97-53-0	1357	1349	905	L2
n-Decanoic acid	20.85	60/129	334-48-5	1373	1359	789	L2
Unknown 19	20.95	115/145/160	-	--	1363	-	L4
Unknown 20	21.02	58/117/181	-	-	1366	-	L4
Unknown 21	22.58	91/119/147/162	-	-	1426	-	L4
Unknown 22	22.64	91/161/176	-	-	1428	-	L4
1-Dodecanol	23.73	55/83/111	112-53-8	1473	1472	883	L1
2-Benzyl-4,5-dihydro-1H-imidazole	23.87	91/131/159	59-98-3	1477	1478	702	L2
Unknown 23	24.03	55/93/91/77/105/79/145/80/121/92	-	-	1484	-	L4
Unknown 24	24.37	79/107	-	-	1497	-	L4
Unknown 25	24.62	67/96	-	-	1508	-	L4
4βH,5α-Eremophila-1,9,11-triene	24.78	119/161/202	5090-61-9	1511	1515	751	L2
Unknown 26	25.09	55/56/191/57/69/112/83/84/68	-	-	1528	-	L4
Unknown 27	25.66	115/139	-	-	1552	-	L4
Benzophenone	27.35	105/77/182/51/50/181/106/76/183/78	119-61-9	1635	1626	894	L2
Unknown 28	27.66	132/91/133/77/145/119/115/105/55	-	-	1637	-	L4
Unknown 29	27.83	83/153/156/96/132/69/67/82	-	-	1643	-	L4
Unknown 30	28.13	119/121/191/149/107/159/109/57/147	-	-	1658	-	L4
Unknown 31	29.22	149/135/55/83/69/91/53/107/94	-	-	1706	-	L4

Unknown 32	29.74	125/55/69/81/115/171/57/133/97	-	-	1730	-	L4
Unknown 33	30.10	83/54/101/67/82/111/56	-	-	1746	-	L4

L1 – identified compounds, through comparison with a chemical reference standard. **L2** – putatively annotated compounds identified through the comparison with public or commercial libraries. **L4** – unknown compounds that can be detected and quantified but cannot be identified by either the previous methods.

Table A2 - List of all identified VCCs obtained from the chromatograms by using HS-SPME/GC-MS analysis.

Compounds	Retention time	m/z	Standard	Original CAS#	CAS#	Experimental KI	KI from literature	R-match	Identification level
Pentafluorobenzaldehyde	6,88	167/196	-	benzaldehyde 100-52-7	653-37-2	943	876	827	L2
2,3,4,5,6-Pentafluorobenzyl alcohol	9,82	99/177	-	benzyl alcohol 100-51-6	440-60-8	911	978	932	L2
Formaldehyde, O-[(pentafluorophenyl)methyl]oxime	9,92	99/117/161/181/195	-	formaldehyde 50-00-0	-	1041	981	952	L2
Acetaldehyde oxime, o-[(pentafluorophenyl)methyl]-	12,85	181/182/209/161/117/195/99/167/93/119	yes	acetaldehyde/ethanal 75-07-0	-	1112	1078	925	L1
Acetaldehyde oxime, o-[(pentafluorophenyl)methyl]-	13,00	181/182/209/161/117/195/99/167/93/119	yes	acetaldehyde/ethanal 75-07-0	-	1112	1082	932	L1
Unknown 34	13,12	75/143/163/181	-	-	-	-	1086	-	L4
Acetone, (O-pentafluorobenzyl)oxime	14,56	161/181/206	-	acetone 67-64-1	-	1167	1135	929	L2
Levomenthol	15,02	123/81/ 71/ 95/ 55/ 67/ 82/ 69/ 57	-	-	2216-51-5	1175	1151	899	L2
Propionaldehyde, (O-pentafluorobenzyl)oxime, (Z) or (E)-	15,14	181/195/236	yes	propanal 123-38-6	-	1193	1154	881	L1
Propionaldehyde, (O-pentafluorobenzyl)oxime, (Z) or (E)-	15,33	181/195/236	yes	propanal 123-38-6	-	1193	1161	864	L1
Isobutanal O-pentafluorophenylmethyl-oxime	16,25	181/195/250	yes	isobutanal 78-84-2	-	1293	1192	861	L1
2-Butanone oxime, o-[(pentafluorophenyl)methyl]-	16,61	56/181/250	yes	2-butanone 78-93-3	-	1335	1205	764	L1

2-Butanone oxime, o-[(pentafluorophenyl)methyl]-	16,70	56/181/250	yes	2-butanone 78-93-3	-	-	1208	833	L1
Unknown 35	17,55	181/239/117/182/168/119/74	-	-	-	-	1238	-	L4
Nonanoic acid	17,67	60/115/57/73	-	-	112-05-0	1273	1242	891	L2
3-Methyl-2-butanone oxime, o-[(pentafluorophenyl)methyl]-	17,86	58/181/253	yes	3-methyl-2-butanone 563-80-4	-	1370	1249	758	L1
Unknown 36	18,47	181/260/72/161/70/55/182/56	-	-	-	-	1271	-	L4
Butanal, 2-methyl-O-[(pentafluorophenyl)methyl]oxime	18,64	57/181/239	yes	2-methyl-1-butanal 96-17-3	-	1393	1277	833	L1
2-Methoxy-4-vinylphenol	18,67	51/77/107	-	-	7786-61-0	1317	1278	928	L2
2-Pentanone oxime, o-[(pentafluorophenyl)methyl]-	18,76	72/100/181/253	yes	2-pentanone 107-87-9	-	1434	1281	875	L1
5-Methoxy-2-pentanone oxime, o-[(pentafluorophenyl)methyl]-	18,87	45/72/181/253	-	5-methoxy-2-pentanone 17429-04-8	-	1609	1285	769	L2
3-Methylbutanal oxime, o-[(pentafluorophenyl)methyl]-	18,90	181/195/239	yes	3-methylbutanal 590-86-3	-	1328	1287	856	L1
3-Methylbutanal oxime, o-[(pentafluorophenyl)methyl]-	19,11	181/239	yes	3-methylbutanal 590-86-3	-	1328	1294	830	L1
Unknown 37	19,16	181/135/150/77/72/107/51/253	-	-	-	-	1295	-	L4
Crotonaldehyde O-pentafluorophenylmethyl-oxime	19,25	181/250	yes	2-butenal 123-73-9	-	1339	1299	779	L1
Crotonaldehyde O-pentafluorophenylmethyl-oxime	19,32	181/250	yes	2-butenal 123-73-9	-	1339	1302	726	L1
Unknown 38	19,50	181/57/161/182/99/195/362/117/69	-	-	-	-	1308	-	L4

Unknown 39	19,61	181/71/226/182/161/195/99/117	-	-	-	-	1313	-	L4
Unknown 40	19,79	181/253/ 72/ 71/ 55/ 182/ 254	-	-	-	1393	-	-	L4
unknown 41	19,97	181/57/268/69/239/87	-	-	-	-	1326	-	L4
Pentanal oxime, o-[(pentafluorophenyl)methyl]-	20,14	181/239/182/100/57/281/195/266	-	Pentanal	-	1369	1332	822	L2
Unknown 42	20,20	181/60/73/55/57/69/246	-	-	-	-	1335	-	L4
Unknown 43	20,63	181,56,264,67,53,182	-	-	-	-	1351	-	L4
3-Penten-2-one oxime, O-[(pentafluorophenyl)methyl]-	20,79	181,56,264,67,53,182	yes	3-penten-2-one 625-33-2	-	1442	1357	616	L1
Unknown 44	20,92	181/59/195/99/57/75/79/161	-	-	-	-	1362	-	L4
2-Hexanone oxime, o-[(pentafluorophenyl)methyl]-	20,99	72/181/253	yes	2-hexanone 591-78-6	-	1533	1364	612	L1
Unknown 45	21.02	58/117/161/181	-	-	-	-	1366	-	L4
Unknown 46	21,13	181/58/71/161/182/117/99/195	-	-	-	-	1370	-	L4
Unknown 47	21,26	181/58/57/55/240/56	-	-	-	-	1375	-	L4
4-Heptanone oxime, o-[(pentafluorophenyl)methyl]-	22,01	128/181/195/309	yes	4-heptanone 123-19-3	-	1633	1403	851	L1
unknown 48	22,20	181/55/72/85/362/161/56/57	-	-	-	-	1411	-	L4
n-Hexanal, o-[(pentafluorophenyl)methyl]oxime	22,34	114/181/239	yes	hexanal 66-25-1	-	1460	1416	838	L1
n-Hexanal, o-[(pentafluorophenyl)methyl]oxime	22,47	114/181/239	yes	hexanal 66-25-1	-	1460	1422	702	L1
3-Methylpentanal oxime, o-[(pentafluorophenyl)methyl]-	22,67	181/61/57/85/239/55/56	-	3-methylpentanal 15877-57-3	-	1492	1430	768	L2

2-Heptanone oxime, o-[(pentafluorophenyl)methyl]-	22,91	72/181/253	yes	2-heptanone 110-43-0	-	1633	1439	706	L1
2-Furaldehyde O-pentafluorophenylmethyl-oxime	23,16	83/117/181/248/291	yes	furfural 98-01-1	-	1510	1449	955	L1
Unknown 49	23,51	181/161/117/182/167/93/195/180/99/162	-	-	-	-	1463	-	L4
2-Furaldehyde O-pentafluorophenylmethyl-oxime	23,63	83/117/248/291	yes	furfural 98-01-1	-	1510	1468	947	L1
Unknown 50 (não derivatizado)	23,80	67/55/96/57/53/83/69/81	-	-	-	-	1474	-	L4
Cyclohexanone oxime, o-[(pentafluorophenyl)methyl]-	24,11	82/112/181	yes	cyclohexanone 108-94-1	-	1635	1487	803	L1
3,3-Dimethylbutanal oxime, o-[(pentafluorophenyl)methyl]-	24,36	57/181/239/55/182/280/82	-	3,3-dimethylbutanal 2987-16-8	-	1472	1497	700	L2
4-Hydroxy-2-butanone oxime, o-[(pentafluorophenyl)methyl]-	24,50	181/57/86/182/72/56/58	-	4-hydroxy-2-butanone 590-90-9	-	1577	1502	699	L2
n-Heptanal, o-[(pentafluorophenyl)methyl]oxime	24,61	181/239	yes	heptanal 111-71-7	-	1558	1507	767	L1
2-Methyl-3-hexanone oxime, o-[(pentafluorophenyl)methyl]-	25,01	86/181/128/309/281/70/69/100	-	2-methyl-3-hexanone 7379-12-6	-	1569	1524	604	L2
2-Octanone oxime, o-[(pentafluorophenyl)methyl]-	25,24	181/253/72/55/142/69/57	yes	2-octanone 111-13-7	-	1588	1534	682	L1
2,4-Hexadienal oxime, o-[(pentafluorophenyl)methyl]-	25,53	69/110/181/276	yes	142-83-6	-	1599	1546	-	L1
2,4-Hexadienal oxime, o-[(pentafluorophenyl)methyl]-	25,74	69/110/181/276	yes	(E,E)-2,4-hexadienal 142-83-6	-	1599	1551	-	L1
5-(Hydroxymethyl)Furfural O-pentafluorophenylmethyl-oxime	25,84	79/94/97/181/305	yes	5-methyl-2-furfural 620-02-0	-	-	1559	-	L1

5-(Hydroxymethyl)Furfural O-pentafluorophenylmethyl-oxime	26,02	79/94/97/181/305	yes	5-methyl-2-furfural 620-02-0	-	-	1566	-	L1
trans-2-heptenal, o-[(pentafluorophenyl)methyl]oxime	26,34	181/250	yes	trans-2-heptenal 18829-55-5	-	1522	1580	670	L1
n-Octanal, o-[(pentafluorophenyl)methyl]oxime	26,82	181/239	yes	octanal 124-13-0	-	1654	1601	760	L1
n-Octanal, o-[(pentafluorophenyl)methyl]oxime	26,89	69/181/239	yes	octanal 124-13-0	-	1654	1604	651	L1
Unknown 51	27,16	54/72/181/253	-	-	-	-	1615	-	L4
Benzaldehyde, o-[(pentafluorophenyl)methyl]oxime	27,46	181/225/271/301	yes	benzaldehyde 100-52-7	-	2285	1628	840	L1
Unknown 52	27,58	85/98/181/264	-	-	-	-	1633	-	L4
Unknown 53	28,24	181/264/87/85/61/71/67	-	-	-	-	1663	-	L4
Unknown 54	28,41	181/67/53/68/182/139/161/70	-	-	-	-	-	-	L4
Phenylacetaldehyde O-pentafluorophenylmethyl-oxime	28,46	65/91/117/181/297	yes	phenylacetaldehyde 122-78-1	-	1832	1672	895	L1
Phenylacetaldehyde O-pentafluorophenylmethyl-oxime	28,64	65/91/117/181/297	yes	phenylacetaldehyde 122-78-1	-	1832	1680	709	L1
n-Nonanal, o-[(pentafluorophenyl)methyl]oxime	28,97	110/181/239	yes	nonanal 124-19-6	-	2047	1694	835	L1
2-decanone oxime, o-[(pentafluorophenyl)methyl]-	29,07	72/181/253	yes	2-decanone 2548-87-0	-	1383	1698	-	L1
Unknown 55	29,14	181/61/75/146/246/71	-	-	-	-	1702	-	L4
Unknown 56	29,23	181/55/67/54/83/82/57	-	-	-	-	1706	-	L4

m-Tolualdehyde O-pentafluorophenylmethyl-oxime	30,05	181/ 315/77/91/182/65/79/285/78/89	yes	m-Tolualdehyde 620-23-5	-	1721	1743	-	L1
3-Phenylpropionaldehyde O-pentafluorophenylmethyl-oxime	30,57	181/91/104/105/ 117/77/103/65/271/130	yes	3-Phenylpropionaldehyde 104-53-0	-	1931	1767	600	L1
trans-2-nonenal oxime, o-[(pentafluorophenyl)methyl]-	30,68	181/250	yes	trans-2-nonenal 18829-56-6	-	-	1773	-	L1
Isopropyl myristate (não derivatizado)	30,75	60/73/102/228	-	-	110-27-0	1827	1776	798	L2
Unknown 56	30,93	181/278/54/52/53/75/69	-	-	-	-	1784	-	L4
n-Decanal, o-[(pentafluorophenyl)methyl]oxime	31,01	181/239/170/182/55/57/69/240	yes	n-Decanal 112-31-2	-	2143	1788	670	L1
(E,E)-2,4-nonadienal	31,70	181/276/333	yes	6750-03-4	-	-	1822	-	L1
1-Hexadecanol	31,89	69/97/224	-	-	36653-82-4	1880	1832	930	L2
Unknown 58	32,01	181/182/195/99/ 117/161/167/177/198	-	-	-	-	1837	-	L4
(E,E)-2,4-nonadienal	32,19	181/276/333	yes	6750-03-4	-	-	1846	-	L1
2,5-Dimethylbenzaldehyde O-pentafluorophenylmethyl-oxime	32,31	181/ 132/ 77/ 117/ 329/ 91/ 103/ 130/ 148/ 131/ 329	yes	2,5-Dimethylbenzaldehyde 5779-94-2	-	1959	1853	837	L1
Glyoxal (Glycoldial), bis-O-pentafluorobenzyloxime	32,75	181/235	yes	glyoxal 107-22-2	-	1935	1875	971	L1
Glyoxal (Glycoldial), bis-O-pentafluorobenzyloxime	32,84	181/182/161/195/167/117/99/448/93/119	yes	glyoxal 107-22-2	-	1935	1879	921	L1
Glyoxal (Glycoldial), bis-O-pentafluorobenzyloxime	32,91	181/182/161/195/167/117/99/448/	yes	glyoxal 107-22-2	-	1935	1883	938	L1

		93/119							
Methylglyoxal bis-(O-pentafluorophenylmethyl-oxime)	33,02	181/265/432	yes	methylglyoxal 78-98-8	-	2174	1888	912	L1
Methylglyoxal bis-(O-pentafluorophenylmethyl-oxime)	33,21	181/265/432	yes	methylglyoxal 78-98-8	-	2174	1899	912	L1
Methylglyoxal bis-(O-pentafluorophenylmethyl-oxime)	33,52	181/265/432	yes	methylglyoxal 78-98-8	-	2174	1915	938	L1
Unknown 59	33,83	181/70/250/182/161/195/167/99/117	-	-	-	-	1930	-	L4
Unknown 60	33,97	181/182/161/279/195/167/117/99/81	-	-	-	-	1937	-	L4
Dimethylglyoxal dioxime, O,O'-bis[(pentafluorophenyl)methyl]-	34,06	99/181/279	yes	dimethylglyoxal 431-03-8	-	1344	1942	829	L1
2,4-decadienal	34,17	55/181/276	yes	2,4-decadienal 25152-84-5	-	-	1948	-	L1
2,4-decadienal	34,31	55/181/276	yes	2,4-decadienal 25152-84-5	-	-	1955	-	L1
Unknown 61	34,41	181/250/70/182/161/195/99/167/117	-	-	-	-	1960	-	L4
Unknown 62	34,49	181/250/70/182/161/195/99/167/117	-	-	-	-	1964	-	L4
Unknown 63 / Isopropyl palmitate?	34,65	60/57/181/55/61/73/69/102	-	-	-	2013	1972	726	L4/L2
Unknown 64	34,95	181/87/182/311/161/268/195/99/117/295	-	-	-	-	1988	-	L4
Unknown 65	35,78	181/182/57/161/195/99	-	-	-	-	2032	-	L4

Unknown 66	36,18	181/55/96/69/195/117/93/68	-	-	-	-	2053	-	L4
Unknown 67	37,45	181/182/195/117/99/161/167/197/54	-	-	-	-	2122	-	L4
Unknown 68	37,54	181/278/348/55/79/67/351/80	-	-	-	-	2127	-	L4
Unknown 69	37,84	181/250/69/182/195/161/167/119	-	-	-	-	2145	-	L4
Unknown 70	38,18	181/182/161/195/99/117/52/167/51	-	-	-	-	2163	-	L4
Unknown 71	38,40	181/182/161/195/99/69/117/82/167	-	-	-	-	2175	-	L4
Unknown 72	38,62	181/182/195/161/197/471/167/99/117	-	-	-	-	2188	-	L4

L1 – identified compounds, through comparison with a chemical reference standard. **L2** – putatively annotated compounds identified through the comparison with public or commercial libraries. **L4** – unknown compounds that can be detected and quantified but cannot be identified by either the previous methods.