Biospectroscopy towards screening and diagnosis of cancer

A thesis submitted for the degree of
Doctor of Medicine

Faculty of Biological Sciences at Lancaster University,
July 2015

Georgios Theophilou
LRCP & SI MB BCH NUI MRCOG
Acknowledgments

It is true that personal explanations of success do not work. We all owe a big part of what we achieve to parentage and patronage. I have been a beneficiary of an extraordinary opportunity to become a part of a team of remarkable individuals who created close bonds between them. These bonds are not only professional but more importantly, personal.

I would therefore like to thank my laboratory colleagues, namely Rebecca Strong, Kelly Heys, Holly Butler, Junyi Li and Kassio Gomes de Lima for accepting me as a friend and for their continuous help.

I would also like to thank the staff in the pathology department at Royal Preston Hospital in Particular Dr Helen Stringfellow, Dr Da Gama Rose and Mrs Katherine Ashton for their help in retrieving tissues and teaching me techniques needed for my experiments.

I am also grateful to the theatre staff, gynaecologists, nurses and porters for their unselfish help in running the tissue biobank at Royal Preston Hospital.

Most importantly I would like to thank my supervisors Dr Pierre Martin-Hirsch and Professor Francis Martin for their tireless support, encouragement and patience. They have not only offered excellent academic teaching but also provided friendly guidance on a personal basis.

Finally I would like to thank my family, especially my girlfriend Sally for their care, patience and support.

Declaration

I declare that this thesis is my own work and has not been submitted for the award of a higher degree or qualification at this university or elsewhere.
Abstract

Title: Biospectroscopy for screening and early diagnosis of cancer.
Name: Georgios Theophilou
Degree: MBBS
Thesis submitted for the degree of doctor of Medicine
Lancaster University, May 2015

Systems biology is an emerging science that combines high throughput investigation techniques to define the dynamic interplay between different biological regulatory systems in response to internal and external cues. Related technologies, genomics, epigenomics, transcriptomics, proteomics, metabolomics and toponomics have been applied to investigate models of carcinogenesis to identify committing initiating events. Vibrational spectroscopy has the potential to play an integral role within systems biology research approaches, as it is able to identify chemical bond alterations within molecules independent of where these molecules reside. Its integration with current “systems biology” methodologies can contribute in the identification of potential biomarkers of carcinogenesis and assist in their incorporation into clinical practice.

Breast tissue undergoes cyclical and longitudinal molecular and histological alterations that are influenced by environmental factors. These factors may include diet and lifestyle in addition to parity, lactation and menopausal status and are implicated in carcinogenesis. Breast cancer may appear decades after the initial carcinogenic event. Available research in this area is limited to when early histological changes occur due to the difficulties imposed by the molecular and histological diversity of breast tissue. Vibrational spectroscopy in combination with
powerful chemometric techniques has identified spatial and temporal mammary alterations in benign tissue.

Prostate cancer is influenced by environmental factors. Its incidence is higher in populations adopting a Westernised lifestyle and diet and has increased over the past generation. This leads to the assumption that prostatic tissue composition may exhibit chronological alterations. Vibrational spectroscopy techniques were applied to matching prostatic tissues with benign prostatic hyperplasia collected from 1983 to 2013. Significant trans-generational segregation was identified. Spectral areas responsible for this segregation pointed towards epigenetic changes. Immunohistochemical studies for DNA methylation and hypomethylation supported these results.

Vibrational spectroscopy techniques were also implemented to explore molecular changes between normal ovarian tissue, borderline ovarian tumours and malignant ovarian carcinomas. Different chemometric techniques were applied to discriminate cancers from controls. Similar techniques were able to segregate different types of epithelial ovarian carcinomas. The accurate diagnosis obtained using ATR-FTIR spectroscopy demonstrates its potential for development as an assisting tool for histopathological diagnosis.

The endometrial-myometrial junction areas of benign uterine tissues were scrutinised by Synchrotron FTIR and FPA. These techniques in combination with multivariate analysis revealed clear segregation between the functionalis and basalis layers within the uterine crypts. The same techniques illustrated potential areas within these epithelial surfaces where different stem cell types may reside. Targeting the
activation/ inactivation of these stem cells may have applications in the diagnosis and treatment of early uterine cancer.
List of illustrations

<table>
<thead>
<tr>
<th>Title</th>
<th>Page Nº</th>
</tr>
</thead>
<tbody>
<tr>
<td>Figure 1 Wilson Criteria (1968)</td>
<td>9</td>
</tr>
<tr>
<td>Figure 2 Phases of biomarker development</td>
<td>12</td>
</tr>
<tr>
<td>Figure 3 Science integration model for systems biology research</td>
<td>14</td>
</tr>
<tr>
<td>Figure 4 Risk malignancy index (RMI). This algorithm combines</td>
<td>20</td>
</tr>
<tr>
<td>ultrasound findings with menopausal status and CA125 levels</td>
<td></td>
</tr>
<tr>
<td>to identify the risk of malignancy when an ovarian mass is</td>
<td></td>
</tr>
<tr>
<td>identified.</td>
<td></td>
</tr>
<tr>
<td>Figure 5 Ovarian cancer staging</td>
<td>22</td>
</tr>
<tr>
<td>Figure 6 Histology of the prostate</td>
<td>26</td>
</tr>
<tr>
<td>Figure 7 Prostatic secretions</td>
<td>27</td>
</tr>
<tr>
<td>Figure 8 “Prostate risk management program”. Frequently asked</td>
<td>32</td>
</tr>
<tr>
<td>questions and answers to provide information for informed consent</td>
<td></td>
</tr>
<tr>
<td>Figure 9 TNM prostatic cancer staging</td>
<td>33</td>
</tr>
<tr>
<td>Figure 10 Gleason score</td>
<td>34</td>
</tr>
<tr>
<td>Figure 11 Growth factors implicated in prostate cancer</td>
<td>36</td>
</tr>
<tr>
<td>Figure 12 Histopathological features of the normal prostate, benign</td>
<td>37</td>
</tr>
<tr>
<td>prostatic hyperplasia, prostatic intraepithelial neoplasia and</td>
<td></td>
</tr>
<tr>
<td>prostatic adenocarcinoma</td>
<td></td>
</tr>
<tr>
<td>Figure 13 Embryological development of the breast</td>
<td>40</td>
</tr>
<tr>
<td>Figure 14 Breast development stages as proposed by Tanner</td>
<td>41</td>
</tr>
<tr>
<td>Figure 15 Ductal system of the mammary gland</td>
<td>42</td>
</tr>
<tr>
<td>Figure</td>
<td>Description</td>
</tr>
<tr>
<td>----------</td>
<td>-----------------------------------------------------------------------------------------------</td>
</tr>
<tr>
<td>Figure 16</td>
<td>Classification of breast cancer based on anatomical location</td>
</tr>
<tr>
<td>Figure 17</td>
<td>Classification of breast cancer based on the TNM system</td>
</tr>
<tr>
<td>Figure 18</td>
<td>Classification of breast cancer based according to prognosis and correlation with the TNM system</td>
</tr>
<tr>
<td>Figure 19</td>
<td>Diagram of the uterus, cervix, vagina, fallopian tubes and ovaries</td>
</tr>
<tr>
<td>Figure 20</td>
<td>Diagram illustrating the effect of the hormones controlling the menstrual cycle to the endometrium</td>
</tr>
<tr>
<td>Figure 21</td>
<td>Endometrial carcinoma histological subtypes</td>
</tr>
<tr>
<td>Figure 22</td>
<td>Endometrial cancer grade</td>
</tr>
<tr>
<td>Figure 23</td>
<td>Endometrial cancer stage by FIGO</td>
</tr>
<tr>
<td>Figure 24</td>
<td>Molecular alterations associate with endometrial cancer</td>
</tr>
<tr>
<td>Figure 25</td>
<td>The Michelson interferometer</td>
</tr>
<tr>
<td>Figure 26</td>
<td>Conversion of an interferogram to an FTIR spectrum</td>
</tr>
<tr>
<td>Figure 27</td>
<td>Diagram of a Synchrotron facility set-up</td>
</tr>
<tr>
<td>Figure 28</td>
<td>Raman scattering</td>
</tr>
<tr>
<td>Figure 29</td>
<td>Spectrum derived by Raman scattering</td>
</tr>
</tbody>
</table>
# List of Abbreviations

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Full Form</th>
</tr>
</thead>
<tbody>
<tr>
<td>AFP</td>
<td>α-Fetoprotein</td>
</tr>
<tr>
<td>ALND</td>
<td>Axillary Lymph Node Dissection</td>
</tr>
<tr>
<td>AMH</td>
<td>Anti-Mullerian Hormone</td>
</tr>
<tr>
<td>ATP</td>
<td>Adenosine Triphosphate</td>
</tr>
<tr>
<td>ATR</td>
<td>Attenuated Total Reflection</td>
</tr>
<tr>
<td>βHCG</td>
<td>Human Chorionic Gonadotropin</td>
</tr>
<tr>
<td>BMI</td>
<td>Body Mass Index</td>
</tr>
<tr>
<td>BPM</td>
<td>Beats Per Minute</td>
</tr>
<tr>
<td>BRCA</td>
<td>Breast Cancer (gene)</td>
</tr>
<tr>
<td>BUB-1</td>
<td>Budding Uninhibited by Benzodiazepines</td>
</tr>
<tr>
<td>CA 125</td>
<td>Cancer Antigen 125</td>
</tr>
<tr>
<td>CCD</td>
<td>Charge Coupled Device</td>
</tr>
<tr>
<td>CCI</td>
<td>Centre for Clinical Investigations</td>
</tr>
<tr>
<td>CD4</td>
<td>Cluster of Differentiation 4</td>
</tr>
<tr>
<td>CDH 1</td>
<td>Cadherin-1</td>
</tr>
<tr>
<td>CGI</td>
<td>Computer-Generated Imagery</td>
</tr>
<tr>
<td>CMP</td>
<td>Combinational Molecular Phenotype</td>
</tr>
<tr>
<td>CT</td>
<td>Computerised Tomography</td>
</tr>
<tr>
<td>DCIS</td>
<td>Ductal Carcinoma in Situ</td>
</tr>
<tr>
<td>DMT</td>
<td>Dimethyltriptamine</td>
</tr>
<tr>
<td>DNA</td>
<td>Deoxyribonucleic Acid</td>
</tr>
<tr>
<td>EAU</td>
<td>European Association of Urology</td>
</tr>
<tr>
<td>ER</td>
<td>Endoplasmic Reticulum</td>
</tr>
<tr>
<td>ERBB2</td>
<td>Epidermal Growth Factor (HER2)</td>
</tr>
<tr>
<td>FIGO</td>
<td>International Federation of Gynaecology and Obstetrics</td>
</tr>
<tr>
<td>Abbreviation</td>
<td>Full Form</td>
</tr>
<tr>
<td>--------------</td>
<td>-----------</td>
</tr>
<tr>
<td>FSH</td>
<td>Follicle Stimulating Hormone</td>
</tr>
<tr>
<td>FTIR</td>
<td>Fourier Transform Infrared Spectroscopy</td>
</tr>
<tr>
<td>GA</td>
<td>General Anaesthetic</td>
</tr>
<tr>
<td>H &amp; E</td>
<td>Haematoxylin and Eosin</td>
</tr>
<tr>
<td>HAPPI</td>
<td>Human Annotated and Predicted Protein Interactions</td>
</tr>
<tr>
<td>HDL</td>
<td>High Density Lipoprotein</td>
</tr>
<tr>
<td>HE4</td>
<td>Human Epididymis Protein 4</td>
</tr>
<tr>
<td>HER2</td>
<td>Human Epidermal Growth Factor receptor</td>
</tr>
<tr>
<td>HGSC</td>
<td>High Grade Serous Carcinoma</td>
</tr>
<tr>
<td>HiFU</td>
<td>High Frequency Ultrasound</td>
</tr>
<tr>
<td>HNPCC</td>
<td>Hereditary Non-Polyposis Colorectal Cancer</td>
</tr>
<tr>
<td>HPLC</td>
<td>High Performance Liquid Chromatography</td>
</tr>
<tr>
<td>HPV</td>
<td>Human Papilloma Virus</td>
</tr>
<tr>
<td>HRT</td>
<td>Hormone Replacement Therapy</td>
</tr>
<tr>
<td>IGF</td>
<td>Insulin-Like Growth Factor</td>
</tr>
<tr>
<td>IPAD</td>
<td>Integrated Pathway Analysis Pathway</td>
</tr>
<tr>
<td>IR</td>
<td>Infra-red</td>
</tr>
<tr>
<td>K-ras</td>
<td>Kirsten Rat Sarcoma</td>
</tr>
<tr>
<td>KEGG</td>
<td>Kyoto Encyclopaedia of Genes and Genomes</td>
</tr>
<tr>
<td>LCIS</td>
<td>Lobular Carcinoma in Situ</td>
</tr>
<tr>
<td>LD</td>
<td>Linear Discriminants</td>
</tr>
<tr>
<td>LDA</td>
<td>Linear Discriminant analysis</td>
</tr>
<tr>
<td>LDH</td>
<td>Lactate Dehydrogenase</td>
</tr>
<tr>
<td>LGSC</td>
<td>Low Grade Serous Carcinoma</td>
</tr>
<tr>
<td>LH</td>
<td>Luteinising Hormone</td>
</tr>
<tr>
<td>LOH</td>
<td>Loss of Heterozygocity</td>
</tr>
<tr>
<td>Acronym</td>
<td>Definition</td>
</tr>
<tr>
<td>---------</td>
<td>------------</td>
</tr>
<tr>
<td>LUTS</td>
<td>Lower Urinary Tract Symptoms</td>
</tr>
<tr>
<td>MC</td>
<td>Mucinous Carcinoma</td>
</tr>
<tr>
<td>MELC</td>
<td>Multi-epitope Ligand Cartography</td>
</tr>
<tr>
<td>MLH</td>
<td>Mut L Homolog</td>
</tr>
<tr>
<td>MRI</td>
<td>Magnetic resonance imaging</td>
</tr>
<tr>
<td>MRSI</td>
<td>Magnetic resonance spectroscopy imaging</td>
</tr>
<tr>
<td>MS</td>
<td>Mass Spectrometry</td>
</tr>
<tr>
<td>MSH2</td>
<td>Mut S Homolog 2</td>
</tr>
<tr>
<td>MT</td>
<td>Mixed Tumour</td>
</tr>
<tr>
<td>MYC</td>
<td>Myc proto-oncogene</td>
</tr>
<tr>
<td>NGS</td>
<td>Next Generation Sequencing</td>
</tr>
<tr>
<td>NICE</td>
<td>National Institute of Clinical Excellence</td>
</tr>
<tr>
<td>NMR</td>
<td>Nuclear Magnetic Resonance</td>
</tr>
<tr>
<td>NOS</td>
<td>Not Otherwise Specified</td>
</tr>
<tr>
<td>NSC</td>
<td>National Screening Committee</td>
</tr>
<tr>
<td>P53</td>
<td>Phosphoprotein 53</td>
</tr>
<tr>
<td>PALB2</td>
<td>Partner and Localiser of BRCA2</td>
</tr>
<tr>
<td>PC</td>
<td>Principal Component</td>
</tr>
<tr>
<td>PCA</td>
<td>Principal Component Analysis</td>
</tr>
<tr>
<td>PCOS</td>
<td>Polycystic Ovarian Syndrome</td>
</tr>
<tr>
<td>PCR</td>
<td>Polymerase Chain Reaction</td>
</tr>
<tr>
<td>PG2</td>
<td>Prostaglandin 2</td>
</tr>
<tr>
<td>PGR</td>
<td>Progesterone Receptor</td>
</tr>
<tr>
<td>PLCO</td>
<td>Prostate, Lung, Colorectal and Ovarian Cancer Screening Trial</td>
</tr>
<tr>
<td>PSA</td>
<td>Prostatic Specific Antigen</td>
</tr>
<tr>
<td>PTEN</td>
<td>Phosphatase and Tensin Homolog</td>
</tr>
<tr>
<td>QDA</td>
<td>Quadratic Discriminant Analysis</td>
</tr>
<tr>
<td>RMI</td>
<td>Risk Malignancy Index</td>
</tr>
<tr>
<td>Abbreviation</td>
<td>Full Form</td>
</tr>
<tr>
<td>--------------</td>
<td>-----------</td>
</tr>
<tr>
<td>RNA</td>
<td>Ribonucleic Acid</td>
</tr>
<tr>
<td>ROMA</td>
<td>Risk of Ovarian Malignancy Algorithm</td>
</tr>
<tr>
<td>SELDI-TOF</td>
<td>Surface Enhanced Desorption/ Ionisation Time Of Flight</td>
</tr>
<tr>
<td>SPA</td>
<td>Spectral Parameter Analysis</td>
</tr>
<tr>
<td>STI</td>
<td>Sexually Transmitted Disease</td>
</tr>
<tr>
<td>TDLU</td>
<td>Terminal Ductal Lobular Unit</td>
</tr>
<tr>
<td>TNM</td>
<td>Tumour, Nodes, Metastasis</td>
</tr>
<tr>
<td>TP53</td>
<td>Tumour Protein 53</td>
</tr>
<tr>
<td>TURP</td>
<td>Trans-urethral Resection of the Prostate</td>
</tr>
<tr>
<td>USS</td>
<td>Ultrasound Scan</td>
</tr>
<tr>
<td>WHO</td>
<td>World Health Organisation</td>
</tr>
</tbody>
</table>
Contents

Title page I
Acknowledgments II
Declaration II
Abstract III
List of illustrations VI
List of Abbreviations VIII
Contents XII

Chapter 1 General introduction 1
Chapter 2 Extracting biomarkers of commitment to cancer development: potential role of vibrational spectroscopy in systems biology. 101

Georgios Theophilou, Maria Paraskevaidi, Kássio M.G. Lima, Maria Kyrgiou, Pierre L. Martin-Hirsch, Francis L Martin
Expert Reviews in Molecular Diagnostics Early Online, 1-21 (2015)

Chapter 3 Attenuated total reflection infrared spectroscopy coupled with chemometric analysis discriminates normal, borderline and malignant ovarian tissue and classifies subtypes of human ovarian cancer 123

Georgios Theophilou, Kássio M.G. Lima, Pierre L. Martin Hirsch, Helen F. Stringfellow, Francis L. Martin
Analyst DOI: 10.1039/c5an00939a.
<table>
<thead>
<tr>
<th>Chapter 4</th>
<th>A biospectroscopic analysis of human prostate tissue obtained from different time periods points to a trans-generational alteration in spectral phenotype</th>
</tr>
</thead>
</table>
|           | Georgios Theophilou, Kássio M.G. Lima, Matthew Briggs, Pierre L. Martin Hirsch, Helen F. Stringfellow, Francis L. Martin  
|           | Submitted to: “Scientific Reports”                                                                                                                                                               |
| Chapter 5 | Spatial and temporal age-related spectral alterations in benign human breast tissue                                                                                                                                 |
|           | Submitted to: “Journal of Molecular structure”                                                                                                                                                  |
| Chapter 6 | Discussion                                                                                                                                                                                        |
| Appendix 1| Ethical approval, consent forms and patient information leaflets                                                                                                                                 |
| Appendix 2| Other related articles published during the MD                                                                                                                                                   |
| Appendix 3| Abstracts and posters                                                                                                                                                                             |
Chapter 1

General introduction
1.1 Introduction

1.2 Systems biology towards screening biomarkers for cancer
  1.2.1 Carcinogenesis
  1.2.2 Screening
  1.2.3 Biomarkers
  1.2.4 Systems Biology

1.3 The ovary
  1.3.1 Anatomy and physiology
  1.3.2 Ovarian cancer
    1.3.2.1 Epidemiology
    1.3.2.2 Aetiology
    1.3.2.3 Presentation
    1.3.2.4 Screening and diagnosis
    1.3.2.5 Molecular pathology
    1.3.2.6 Management

1.4 The prostate
  1.4.1 Anatomy and physiology
  1.4.2 Benign prostatic hyperplasia
  1.4.3 Prostate cancer
    1.4.3.1 Epidemiology
    1.4.3.2 Aetiology and risk factors
    1.4.3.3 Presentation
    1.4.3.4 Screening and diagnosis
    1.4.3.5 Molecular pathology of prostate adenocarcinoma
    1.4.3.6 Management

1.5 The female mammary gland
  1.5.1 Anatomy and physiology
  1.5.2 Breast cancer
    1.5.2.1 Epidemiology
1.5.2.2 Aetiology
1.5.2.3 Presentation
1.5.2.4 Screening and diagnosis
1.5.2.5 Molecular Pathology
1.5.2.6 Management

1.6 The uterus, cervix and fallopian tube
1.6.1 Anatomy and Physiology
1.6.2 Uterine cancer
1.6.2.1 Epidemiology
1.6.2.2 Aetiology
1.6.2.3 Presentation
1.6.2.4 Screening and Diagnosis
1.6.2.5 Management
1.6.2.6 Molecular pathology

1.7 Infrared spectroscopy
1.7.1 Background
1.7.2 Fourier transform infrared spectroscopy
1.7.3 Synchrotron-based FTIR spectroscopy
1.7.4 Attenuated Total Reflection Fourier Transform Spectroscopy
1.7.5 Raman Spectroscopy
1.7.6 Spectral pre-processing
1.7.7 Multivariate analysis
1.7.8 Evolutionary algorithms
1.7.8.1 Genetic algorithm
1.7.8.2 Sequential progressions algorithm

1.7.9 The role of biospectroscopy in screening and diagnosis of cancer
1.7.9.1 Gastrointestinal cancers
1.7.9.2 Urological cancer
1.7.9.3 Gynaecological cancers

1.8 Aims and objectives
1.1 Introduction

Vibrational spectroscopy techniques can be utilised for the interrogation of biological samples such as tissues and bio-fluids (plasma, serum, urine, saliva) to obtain significant biomolecular information. This spectral information is complex and its analysis requires powerful computational chemometric techniques. The choice of chemometric technique depends on the hypotheses explored and the datasets examined. For example different analysis would be employed for chemical image reconstruction than for classification of sub-populations. Nevertheless spectroscopic techniques coupled with computational analysis have the ability to decipher minute chemical variations within molecules. These variations can be exploited to distinguish benign from neoplastic samples or identify specific cellular lineages within histological areas. These techniques can be developed towards the extraction of specific responsible biomarkers. Other biomarker extraction techniques can then be applied to validate potential results. These biomarkers can be developed for use in clinical practice especially within screening programs.

Vibrational spectroscopy has a direct translatory potential into clinical practice as the techniques associated with it can be developed for use in vitro or in vivo depending on the setting. There is currently extensive research looking at incorporating Raman spectroscopy on endoscopes for the localisation of tissue changes (Bergholt et al., 2010) and FTIR in the classification of cervical cytological samples (Gajjar et al., 2014).

Vibrational spectroscopy is able to identify chemical bond alterations between atoms independent of the molecules they form. It is therefore not specific to the genome, epigenome, transcriptome, proteome and metabolome. It is able to extract information from functional cell lines as well as supporting stroma in tissues. It is also applicable
in cellular smears and fluids. The diversity possessed by biospectroscopic technologies, along with the fact that they do not require extensive or destructive pre-processing of the tested samples makes them prime candidates for incorporation into a “systems biology” setting. Their ability to direct towards chemical alterations that may be associated with functional cellular elements such as proliferation or apoptosis, or signalling between cells may secure them a position amongst “omics” technologies for research into systemic functional alterations in response to exogenous influences. 

This has direct implication is carcinogenesis research as it is now accepted that tumours appear as a consequence of several, cumulative events that cause cellular systems to acquire carcinogenic phenotypical and functional characteristics.

These changes predate histological or cytological neoplastic and pre-neoplastic evidences and may therefore potentially be applied for the identification of very early chemical indicators of disease initiation. Such indicators may provide the stepping-stones for the extraction of biomarkers suitable for utilisation as risk identification tests within population screening programs. In fact even if these techniques do not lead to biomarker extraction, their ability to correctly classify samples according to whether they poses early neoplastic changes or not may be enough to allow them to be incorporated or replace current screening tests.

Within this writing the concepts of carcinogenesis will be explored from a biospectroscopic point of view. Chemical alterations within benign prostatic and mammary tissue that may be influenced by environmental factors will be examined. Also, the potential of vibrational spectroscopy in distinguishing benign from borderline and malignant ovaries will be verified. Uterine tissue will be interrogated to identify chemical variation within its functional layers as well as putative stem cell locations.
1.2 Systems biology towards screening biomarkers for cancer

1.2.1 Carcinogenesis

Carcinogenesis and its synonym oncogenesis are defined as the processes involved in initiating and promoting the development of cancer. These processes stimulate alterations of cellular functions that affect the balance between proliferation and apoptosis causing malignant tumour formation. This deregulation is progressive and implicates all cellular functions including alterations in growth signals and growth-inhibitory signals, evasion of cell death, replicative potential, energy metabolism and immune destruction resistance (Hanahan & Weinberg, 2000). It is believed that the initiation step for carcinogenesis requires the presence of incipient cancer cells, which are genetically unstable (Hahn & Weinberg, 2002; Hanahan & Weinberg, 2000, 2011). Under the influence of endogenous and exogenous carcinogens such as hormonal and dietary influences, smoking, radiation viruses and oxygen free radicals these incipient cancer cells suffer more extensive DNA damage leading to cancer formation. Examples of carcinogens implicated in associated processes include the human papilloma virus (HPV) in cervical cancer, oestrogen in endometrial cancer and polycyclic aromatic hydrocarbons in lung cancer amongst many others.

Many processes involved in carcinogenesis have been identified and extensively described, including some commonalities between all cancers. Several molecular mechanisms, such as point mutations, microsatellite instability, gene silencing, loss of heterozygosity of tumour suppressor genes, altered gene dosage and gene expression via aneuploidy have been postulated as possible important initiating steps. These processes may involve the conversion of proto-oncogenes to oncogenes modifying their expression upsetting the cell cycle homeostasis and therefore increasing
cancerous potential. An example of such an oncogene is the Ras oncogene comprising of H-Ras, K-Ras and N-Ras mutations that are commonly found in human tumours (Bos, 1989). Tumour suppressor genes code for transcription factors that function to control cellular proliferation by suppressing mitosis. Tumour suppressor gene mutation or epigenetic inactivation results in “uncontrolled” proliferation with inadequate repair leading to carcinogenesis. Some of these mutations can be inherited, for example BRCA1 and BRCA2 mutations that are implicated in breast and ovarian cancers. In such cases it is believed that at least two mutations have to be present for cancer development; the inherited tumour suppressor gene mutation and a mutation that causes the inactivation of its allele. This theory has been suggested by Knudson and is known as Knudson two-hit hypothesis (Knudson, 1971). In addition non-mutational changes to DNA can lead to pathogenic alterations in gene expression initiating cancer. Encountered changes are commonly DNA methylation, which is associated with tumour suppressor gene inactivation and histone acetylation, which is associated with oncogene activation (Zardo et al., 2002). The reasons for the genetic and epigenetic alterations that enable carcinogenesis are multivariate and involve all the cellular processes that result in creating an “advantageous” environment for the proliferation of malignant cells in the expense of normal ones (Merlo et al., 2006).

The period from the cancer initiation event to the progression to clinically defined cancer is termed “the latency period”. A number of successive mutations have to accumulate in order for incipient cancer cells to enter and progress through this period (Foulds, 1954; Nowell, 1976). The latency period is different for different tumours and can last for as long as decades (Armitage & Doll, 1954; O’neill et al., 2003; Vogelstein & Kinzler, 1993). The diversity of the mutations occurring during this time explains the considerable genetic and phenotypic heterogeneity in cancer and the
time for their accumulation explains the varied speeds with which different cancers progress.

The pathological deregulation associated with carcinogenesis is induced by dynamic intracellular and extracellular reactions to environmental influences (Hartwell et al., 1999). The function of these alterations is to support the growth and progression of the malignancy.

1.2.2 Screening

The detection of early disease or increased risk of acquiring a disease through screening programs has led to better prevention measures, improved treatments and reduced morbidity and mortality.

In 1951 the CCI conference on Preventative Aspects of Chronic Disease defined screening as “the presumptive identification of unrecognized disease or defect by the application of tests, examinations or other procedures which can be applied rapidly. Screening tests sort out apparently well persons who probably have a disease from those who probably do not. A screening test is not intended to be diagnostic. Persons with positive or suspicious findings must be referred to their physicians for diagnosis and necessary treatment” (Cohart, 1951).

The UK NSC definition today is: “Screening is a process of identifying apparently healthy people who may be at increased risk of a disease or condition. They can then be offered information, further tests and appropriate treatment to reduce their risk and/or any complications arising from the disease or condition.”
Several screening programs are now in place for the identification of those with increased cancer risk or with pre-cancerous or very early cancerous changes. They have varying degrees of success depending on their ability to correctly pick up early disease and their socio-economic acceptability. The implementation of screening programs is a multi-dimensional task whose goal is the early identification of disease with minimal impact to the service provider and the population screened.

The principles of screening were defined by the world health organization in 1968 and are often known as the Wilson Criteria (Wilson & Jungner, 1968). These principles are still valid today (Figure 1).

<table>
<thead>
<tr>
<th>Figure 1: Wilson Criteria (1968)</th>
</tr>
</thead>
<tbody>
<tr>
<td>The condition should be an important health problem.</td>
</tr>
<tr>
<td>There should be a treatment for the condition.</td>
</tr>
<tr>
<td>Facilities for diagnosis and treatment should be available.</td>
</tr>
<tr>
<td>There should be a latent stage of the disease.</td>
</tr>
<tr>
<td>There should be a test or examination for the condition.</td>
</tr>
<tr>
<td>The test should be acceptable to the population.</td>
</tr>
<tr>
<td>The natural history of the disease should be adequately understood.</td>
</tr>
<tr>
<td>There should be an agreed policy on whom to treat.</td>
</tr>
<tr>
<td>The total cost of finding a case should be economically balanced in relation to medical expenditure as a whole.</td>
</tr>
<tr>
<td>Case finding should be a continuing process and not a &quot;once and for all&quot; project.</td>
</tr>
</tbody>
</table>
Developing a suitable screening biomarker is imperative for the establishment of an effective screening program (Gray et al., 2008). Current screening programs make use of potentially harmful or invasive techniques such as endoscopy or transvaginal ultrasonography and evidence regarding their value is continuously changing.

Recent technological advances in high throughput omics technologies have opened new horizons in the search for new biomarkers of disease. The choice of biomarkers is ever expanding and selecting the most suitable involves considering all of the prerequisites that a screening program has to fulfil. Population screening involves a highly heterogeneous sample and cancer is a highly heterogeneous process by nature. Following "reductionist" approaches to search for suitable biomarkers within the genome, epigenome, transcriptome, metabolome and toponome, as well as vibrational spectroscopy, has so far proven imperfect. Systems biology is a promising, novel scientific field that integrates information derived from different high throughput technologies to understand models of biological mechanisms (Kitano, 2002). These models can be used to predict complex behaviours these mechanisms may adopt in response to internal and external stimuli. The structural and functional changes associated with these behaviours may harbour potential cancer biomarkers, which once identified may be applied within population screening programs.
1.2.3 Biomarkers

In 1998, the National Institutes of Health Biomarkers Definitions Working Group defined a biomarker as “a characteristic that is objectively measured and evaluated as an indicator of normal biological processes, pathogenic processes, or pharmacologic responses to a therapeutic intervention.”

International Programme on Chemical Safety, led by the World Health Organization (WHO) and in coordination with the United Nations and the International Labour Organization, has defined a biomarker as “any substance, structure, or process that can be measured in the body or its products and influence or predict the incidence of outcome or disease”.

The usefulness of a particular biomarker lies in its ability to provide early indication of presence or progression of disease (Srinivas et al., 2001). There are two general types of cancer biomarkers: biomarkers that identify risk of developing cancer and biomarkers that detect cancer in the pre-clinical stage. Biomarkers of risk identify individuals with a predisposition to develop cancer before tumourigenesis. They usually detect gene alterations that dictate familial predisposition to cancer development. Examples of such genes include BRCA1 and BRCA2 in breast cancer and MLH and MSH2 in familial non-polyposis coli syndrome.

Biomarkers are designed to identify early disease after the biological onset of carcinogenesis and before symptoms occur. They rely on the occult neoplastic process producing a signature that can be identified and quantified.
Five phases of biomarker development have been proposed each aiming to examine particular desirable attributes of the test without jeopardising safety or wasting resources (Figure 2) (Pepe et al., 2001).

**Figure 2: Phases of biomarker development, adapted form (Pepe et al., 2001)**

The success of a biomarker within a screening program depends on several factors including population compliance, effect of treatment on cancer and economic viability. If any of these elements are poor the chances of failure of the screening program will prevent its establishment.

The most pivotal attribute of an ideal biomarker for use in cancer screening would be its ability to accurately identify the earliest change that defines commitment to disease. Such a biomarker would distinguish those individuals who will eventually get cancer if left untreated from those that will never get cancer despite their being at increased risk. In order to produce such a biomarker the lifecycle of cancer needs to be understood.
1.2.4 Systems Biology

Systems biology involves multidisciplinary approaches for investigating all-inclusive functions within biological systems. It involves the bringing together of diverse sciences including biology, biomedicine, engineering, analytics and computational science to examine specific hypotheses considering biological functions. This concept is in contrast to reductionist approaches where a specific portion of a biological system is examined in isolation usually using only one approach. In systems biology a cycle of operational protocols involving theoretical, analytic and computational models are utilized to propose specific testable hypotheses about a biological system. These hypotheses then undergo experimental validation. This cycle is repeated with specific refinements at each stage depending on the accumulation of knowledge on involved cellular processes. The expansive nature of this type of investigation has the potential to examine dynamic biological systems as whole entities and therefore acquire integral models of there functions (Kolch & Kholodenko, 2013). To achieve this, the methodology uses high throughput investigating technologies to identify the genomic, epigenomic, transcriptomic, proteomic, metabolomic, and toponomic make-up of a functional cellular system amongst others (Figure 3).
<table>
<thead>
<tr>
<th>Discipline</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>Phenomics</td>
<td>Studies of the set of observable characteristics or traits</td>
</tr>
<tr>
<td>Genomics</td>
<td>Studies the complete DNA sequence</td>
</tr>
<tr>
<td>Epigenomics</td>
<td>Studies alterations on gene transcriptional activity that is not related to DNA sequence changes</td>
</tr>
<tr>
<td>Transcriptomics</td>
<td>Studies the complete compliment of RNA generated by a cell</td>
</tr>
<tr>
<td>Proteomics</td>
<td>Studies the full complement of proteins contained in cells</td>
</tr>
<tr>
<td>Metabolomics</td>
<td>Studies end-products of cellular metabolism</td>
</tr>
<tr>
<td>Toponomics</td>
<td>Studies the location of protein networks in relation to phenotypic alterations</td>
</tr>
</tbody>
</table>

**Figure 3: Science integration model for systems biology research**

Bio-informatics is an associated science that brings together data produced by high throughput technologies. Bio-informatics can be distinguished into computer science, computational analysis, biology, chemo-metrics and statistics. These sciences combine to develop powerful computational algorithms that aim to identify cellular components that are implicated in specific cellular functions under investigation. Examples of such algorithms are gene finding, motif recognition, similarity searches, multiple sequence alignment, protein structure prediction and phylogenetic analysis amongst others (Baxevanis, 2001; Mount, 2001). In order for investigators to have access to as diverse datasets as possible data storage facilities are available online. Such services include the Pathway Interaction database (NCI Nature Curated), Reactome, Biocarta, KEGG or Protein Lounge. Several integration software packages
are available to combine these datasets, for example the Human Pathway Database (HPD), Pathlist, PathSys, Integrated Pathway analysis pathway (IPAD), Human annotated and predicted protein interactions (HAPPI) and KEGG converter. Subsequent analysis of the resulting data is dependent on the use of these data. Classification techniques can be used to identify cellular functions that are related with a particular phenotype while feature extraction techniques can detect cellular component alterations that may drive particular cellular functions.

Systems biology is especially useful in carcinogenesis research. Cancer development is a complex and lengthy process that involves several functional alterations in tissue microenvironment. Systems biology has the potential to identify these alterations and create models of carcinogenesis recognizing key elements that may be used as biomarkers for use in screening or diagnosis. Isolating crucial steps in carcinogenic processes may also assist in therapeutics by providing specific targets for the development of pharmacological agents. The ultimate objective of systems biology in cancer is to establish methodologies to identify the committing steps to cancer development. This could lead in the invention of personalized interventions to prevent or halt cancer formation before it surfaces.
1.3 The ovary

1.3.1 Anatomy and physiology

The ovaries (female gonads) are small (4X3X2cm) whitish, paired organs located in the pelvis in the ovarian fossae in close association with the external and internal iliac arteries and the ureter.

They consist of 3 layers. The medulla is composed of supporting stroma that contains neurovascular structures. The cortex contains thousands of primordial follicles supported by connective tissue stroma. The surface is covered by cuboidal epithelium called the germinal epithelium. They are not covered by peritoneum. The suspensory ligament of the ovary attaches it to the pelvic sidewall and transmits its neurovascular bundle. The ovary is also attached to the broad ligament via the mesovarium and the uterine fundus via the ovarian ligament.

Embryologically the ovary is partially formed from the gonadal ridge and partially from the mesonephros. At approximately 5 weeks of gestation the medial part of the mesonephros develops into the urogenital (gonadal) ridge. The coelomic thickening forms the ovarian epithelium while the subcoelomic mesoderm forms the ovarian stroma. The primordial germ cells migrate to the gonadal ridges from the yolk sac. They undergo successive mitotic divisions and organise within the ovarian cortex. By the seven month of foetal life there are about 7 million primordial follicles in the ovarian cortex but by birth only 2 to 4 million remain. Primordial follicles do not undergo meiosis unless they undergo ovulation. The ovaries descend into the pelvis guided by the gubernaculum in the 3rd month of foetal life. The gubernaculum eventually forms the ovarian and round ligaments. Because of origin of the ovaries their blood supply originates from the aorta close to the renal arteries.
The ovarian follicles are the functional units of the ovary. They contain the oocytes and they secrete hormones (oestrogen and progesterone) required to develop the secondary sexual characteristics, control the menstrual (ovarian) cycle and prepare the uterus for ovulation.

1.3.2 Ovarian cancer

1.3.2.1 Epidemiology

Ovarian cancer is the seventh most common cancer in females worldwide but the fifth most common in Europe and the UK. The crude incidence rate in the UK is 22 per 100,000 women. The incidence rises with age from about 35 years and peaks at 80 years. 75% of ovarian cancers are diagnosed in those ages over 55 years old. There is significant variation in the incidence between different ethnicities, with Americans and Europeans having a higher incidence than Japanese, East Indians and Latinos. The incidence in Japan has been steadily increasing since the 1990’s, reflecting the effect of a “Westernised lifestyle” on ovarian cancer risk.

1.3.2.2 Aetiology

Ovarian cancer risk increases with age and as mentioned earlier is commoner in Europe and America than South East Asia and commoner in Caucasians.

It is also genetically driven. The risk of ovarian cancer in those with one affected first degree relative is 5% and those with two affected relatives is 7% (Hoffman, 2012).
This risk also increases with a family history of endometrial cancer, colon cancer and with HNPCC (Aarnio et al., 1999).

Mutations in the BRCA genes are also associated with an increased risk as part of “Hereditary Breast-Ovarian Cancer Syndrome”. The risk of developing ovarian cancer with a BRCA mutation is 5-10% with 52% of those related to BRCA1, 32% to BRCA2 and 16% to yet unidentified genes (Ford et al., 1998). BRCA mutation related ovarian cancers develop on average 15 years earlier than their sporadic counterparts (Aarnio et al., 1999).

Endogenous and exogenous oestrogen is also associated with an increase in ovarian cancer risk. Infertility due to polycystic ovarian syndrome and therefore anovulation has been found to increase this risk (Rossing et al., 1994). Treatment with clomiphene citrate, human menopausal gonadotropin and other fertility medication is associated with an increased risk of ovarian cancer (Shushan et al., 1996). Hormone replacement therapy with unopposed oestrogen has also been implicated in the development of ovarian cancer (Beral et al., 2007).

Westernised lifestyle and diet have been reported to be associated with an increased risk, in several studies (Cramer et al., 1984; Shu et al., 1989). It has been proposed that lactose in dairy products may be linked to ovarian cancer (Cramer et al., 1989). Also high fat dietary intake has a dose response to the increase in malignancy risk.

Certain factors are protective against ovarian cancer. These include oral contraception and pregnancy due to the interruption of oestrogenic influence (Whittemore et al., 1992). Tubal ligation and hysterectomy have also been shown to reduce ovarian cancer risk (Chiaffarino et al., 2005). It is believed that this occurs due to the prevention of ascending migration of potential carcinogens.
1.3.2.3 Presentation

NICE recommends that women presenting repetitively with the following symptoms warrant further investigations: abdominal distension; fullness or loss of appetite; pelvic or abdominal pain; increased urinary frequency or urgency; symptoms suggestive of irritable bowel syndrome in the previous 12 months; unexplained weight loss, fatigue, or changes in bowel habit. They recommend that women with symptoms suggestive of ovarian cancer should have a CA125 test. Those with a concentration of 35IU/mL or more should have an abdominal and pelvic ultrasound scan. The ultrasound scan features, menopausal status and CA125 levels are inputted into an algorithm called the “risk malignancy index” (RMI) to estimate their risk of having ovarian cancer (Jayson et al., 2014). RMI is calculated by multiplying scores given for menopausal status, ultrasound features and the levels of CA125 (Figure 4). Those with a score of 250 or more should be referred to a specialist team.

Other algorithms have been proposed to identify those at high risk of ovarian cancer in primary care but have not been incorporated to clinical practice (Hamilton et al., 2009; Hippisley-Cox & Coupland, 2012).
Figure 4: Risk malignancy index (RMI). This algorithm combines ultrasound findings with menopausal status and CA125 levels to identify the risk of malignancy when an ovarian mass is identified.

<table>
<thead>
<tr>
<th>Feature</th>
<th>Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ultrasound findings:</td>
<td>Ultrasound score:</td>
</tr>
<tr>
<td>• Multilocular cyst</td>
<td>• None = 0</td>
</tr>
<tr>
<td>• Solid areas</td>
<td>• 1 abnormality = 1</td>
</tr>
<tr>
<td>• Ascites</td>
<td>• ≤ abnormalities = 3</td>
</tr>
<tr>
<td>• Metastases</td>
<td></td>
</tr>
<tr>
<td>• Bilateral abnormalities</td>
<td></td>
</tr>
<tr>
<td>Menopausal status</td>
<td>Menopausal score:</td>
</tr>
<tr>
<td></td>
<td>• Premenopausal = 1</td>
</tr>
<tr>
<td>CA125</td>
<td>• Postmenopausal = 2</td>
</tr>
<tr>
<td>Levels (U/mL)</td>
<td></td>
</tr>
</tbody>
</table>

1.3.2.4 Screening and diagnosis

Unfortunately, most ovarian cancers are not symptomatic, until at an advanced stage, usually stage III. RMI has not been proved to be reliable for population screening as it is not sensitive (81%) or specific (75%) enough (Jacobs et al., 1990). Also because of the very low prevalence of ovarian cancer the potential harms of such a non-specific screening tool would outweigh the benefits (Zurawski et al., 1988). Two large randomised control trials have been conducted to evaluate the efficacy of CA125 and ultrasound for population screening. The American PCLO study did not identify any improvement in mortality from yearly screening (Buys et al., 2011). Preliminary results from the British UKTOCS study show that screening has encouraging sensitivity in picking up ovarian cancer (Menon et al., 2009).
It has also been proposed that transvaginal ultrasound performed on a yearly basis may be able to identify ovarian cancer at early stage when it is more successfully treatable (Depriest et al., 1994). This approach is also not viable because of the extensive costs associated with it.

Human epididymis protein 4 (HE4) has also been suggested as an alternative to CA125 but has not shown better sensitivity or specificity (Nolen & Lokshin, 2010). When HE4 and CA125 are used together to form a tumour marker panel for a potential screening test the sensitivity and specificity of the test increases but does not surpass the sensitivity of RMI (Moore et al., 2010).

Several other biomarker panels have been proposed for ovarian cancer screening but have not been employed in a clinical setting (Yurkovetsky et al., 2010).

Ovarian cancer diagnosis is by histological evaluation of biopsy or surgical excision sample. Several types of ovarian cancer exists but these are broadly distinguished to epithelial, stromal and germ cell tumours (these are discussed in the molecular pathology section).

Staging of ovarian cancer is surgical according to the FIGO staging system (Figure 5). To adequately stage these tumours, a total hysterectomy, bilateral salpingooopherectomy, omental biopsy, pelvic lymphnode biopsies and peritoneal washings or biopsies have to be performed. This is because of the potential of micrometastases (30%) even with stage 1 ovarian malignancy (Jayson et al., 2014). Tumours are graded according to their degree of differentiation. Grade 1 cancers are well differentiated, grade 2 are moderately differentiated and grade 3 are poorly differentiated. In addition borderline tumours may be graded as grade 0.
<table>
<thead>
<tr>
<th>Stage</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>IA</strong></td>
<td>Tumour limited to 1 ovary with an intact capsule and no surface tumour and negative peritoneal washings</td>
</tr>
<tr>
<td><strong>IB</strong></td>
<td>Tumour has similar features to IA but involves both ovaries</td>
</tr>
<tr>
<td><strong>IC1</strong></td>
<td>Tumour limited to 1 or both ovaries with surgical spill</td>
</tr>
<tr>
<td><strong>IC2</strong></td>
<td>Tumour limited to 1 or both ovaries with capsule rupture before surgery or tumour on the surface</td>
</tr>
<tr>
<td><strong>IC3</strong></td>
<td>Tumour limited to 1 or both ovaries with malignant cells in peritoneal washings or ascites</td>
</tr>
<tr>
<td><strong>IIA</strong></td>
<td>Extension to uterus and/or fallopian tubes</td>
</tr>
<tr>
<td><strong>IIB</strong></td>
<td>Extension to other pelvic intraperitoneal tissues</td>
</tr>
<tr>
<td><strong>IIIA1</strong></td>
<td>Positive peritoneal lymph nodes</td>
</tr>
<tr>
<td></td>
<td>- IIIA1(i) - metastasis 10mm</td>
</tr>
<tr>
<td></td>
<td>- IIIA1(ii) - metastasis 10mm</td>
</tr>
<tr>
<td><strong>IIIA2</strong></td>
<td>Microscopic, extrapelvic, peritoneal metastasis +/- positive retroperitoneal lymph nodes</td>
</tr>
<tr>
<td><strong>IIIB</strong></td>
<td>Microscopic, extrapelvic, peritoneal metastasis 2cm +/- positive retroperitoneal lymph nodes. Includes extension to the capsule of liver or spleen.</td>
</tr>
<tr>
<td><strong>IIIC</strong></td>
<td>Microscopic, extrapelvic, peritoneal metastasis 2cm +/- positive retroperitoneal lymph nodes. Includes extension to the capsule of liver or spleen.</td>
</tr>
<tr>
<td><strong>IVA</strong></td>
<td>Pleural effusion with positive cytology</td>
</tr>
<tr>
<td><strong>IVB</strong></td>
<td>Hepatic and/or splenic parenchymal metastasis, metastasis to extra-abdominal organs, including inguinal and extra-abdominal positive lymph nodes</td>
</tr>
</tbody>
</table>
1.3.2.5 Molecular pathology

Epithelial ovarian cancer is the commonest type especially in postmenopausal women. It arises from the ovarian cortex. There are several subtypes of epithelial ovarian malignancy: high-grade serous, low-grade serous, mucinous, endometrioid, clear cell, mixed epithelial tumours and carcinosarcomas. In addition borderline tumours are also of epithelial origin but have a lower malignant potential.

The above subtypes have been grouped into two types of epithelial ovarian carcinoma. Type 1 are comprised of low-grade serous, endometrioid, clear cell, mucinous carcinomas and Brenner tumours. They have a generally slow course, as they are genetically stable. They are characterised by specific mutations involving KRAS, BRAF, ERBB2, CTNNB1, PTEN PIK3CA, ARID1A, and PPP1R1A, which target specific cell signaling pathways. Type 2 tumours are comprised of high-grade serous, high-grade endometrioid tumours and carcinosarcomas (Kurman & Shih Ie, 2011). They are mostly associated with TP53 mutations and BRCA mutations(Ahmed et al., 2010; Senturk et al., 2010). Serous ovarian tumours are believed to arise from malignant implants originating in the fallopian tube epithelium (Przybycin et al., 2010). Similarly endometrioid and clear cell ovarian tumours may be associated with retrograde bleeding causing endometriosis (Malins et al, 1997).

Sex cord tumours arise from the stroma of the ovary specifically from the granulosa cells, theca cells and fibrocytes. They usually present at early stages and are usually unilateral and confer an excellent prognosis. They appear in childhood presenting as precocious puberty, in adulthood with ovarian accidents (torsion or hemorrhagic cysts) and postmenopausally with postmenopausal bleeding due their secretion of high amounts of oestrogen.
Germ-cell tumours include dysgerminomas, teratomas, endodermal sinus tumous, and choriocarcinomas. Dysgerminomas have a good prognosis but occur more commonly in younger women. Endodermal sinus tumours have a poor prognosis and choriocarcinomas metastasize early. These tumours have specific tumour markers due to their secretions. Dysgerminomas are detected by LDH and more recently AMH, endodermal sinus tumours by AFP and choriocarcinomas by b-HCG.

1.3.2.6 Management

Surgery is used for treatment and staging. The type of surgery planned depends on pre-staging by imaging usually using MRI. For stage IA cancers unilateral salpingoopherectomy can be performed with peritoneal washings, for women wishing to conserve their fertility. This renders staging incomplete and therefore has a small risk of “missing” micrometastases (Jayson et al., 2014). In postmenopausal women complete staging procedures are recommended independent of preliminary staging. In children who commonly present with dysgerminomas preservation of one ovary is recommended to allow for complete development of sexual characteristics. Advanced stage III and IV ovarian cancer is treated by cytoreductive surgery with or without neo-adjuvant chemotherapy. The radicality of such procedures is currently under review but it is accepted that the more complete the cytoreduction the better the outcomes (Dauplat et al., 2000). Adjuvant chemotherapy is commonly administered postoperatively in the form of carboplatin alone or in combination with taxols (Jayson et al., 2014).

Radiation therapy used to be the preferred treatment for dysgerminomas as they are very radiosensitive, but as this caused infertility chemotherapy is favoured for these tumours also nowadays (Michael, 2012).
1.4 The prostate

1.4.1 Anatomy and physiology

Embryologically the prostate is formed at 13-16 weeks gestation from five paired epithelial buds that project posteriorly from the urethra into the urogenital sinus under the influence of dihydrotestosterone (DHT). The top pairs are derived form mesoderm and form the transformation and peri-urethral zones and the lower pairs are derived from endoderm and form the peripheral zone.

Anatomically the prostate is about 20 to 30ml in volume and weighs approximately 20g (Terris & Stamey, 1991). It surrounds the urethra with is base being in continuity with the bladder and its apex forming the external urethral sphincter. It is related anteriorly with the pubis symphysis, posteriorly with the rectum and inferiorly with the perineal membrane. Its primary blood supply is a branch of the inferior vesical artery, arising from the anterior branch of the internal iliac artery. Additional blood supply is derived from the middle rectal and internal pudendal arteries. Venous drainage is into the dorsal venous complex and the middle rectal veins. Lymphatic channels drain into the internal and external iliac and obturator lymph node groups.

Histologically the prostate is composed by 70% simple columnar and cuboidal epithelium and 30% fibromuscular stroma. The columnar epithelium forms papillary projections into glandular acini and is surrounded by the cuboidal layer. Approximately 70% of the glandular elements reside in the peripheral zone, 25% in the central zone and 5 to 10% in the transitional and perirethral zones (Figure 6). These zones are surrounded by a 3-layered facia and capsule composed of collagen, elastin and smooth muscle.
Figure 6: Histology of the prostate

- Urethra
- Ejaculatory ducts
- Periurethral zone
- Fibromuscular zone
- Transition zone
- Central zone
- Peripheral zone
The epithelial cells of the glandular elements provide secretions that empty through glandular ducts into the urethra to form the major component of the seminal plasma of the ejaculate (Frick & Aulitzky, 1991). The extracellular matrix plays an important role in the development and control of cellular functions including secretion (Figure 7). The muscular prostatic capsule plays a role in ejaculation and micturition as both the urethra and ejaculatory ducts pass through the prostate.

The maintenance of prostatic size and function is testosterone dependent. Testosterone is enzymatically transformed to dihydrotestosterone to activate RNA-polymerase and eventually lead to the production of several proteins amongst others (Frick & Aulitzky, 1991). Endogenous testosterone or other androgens have not been shown to have a relationship with prostate cancer (Morgentaler, 2006)

<table>
<thead>
<tr>
<th>Proteins</th>
<th>Other</th>
</tr>
</thead>
<tbody>
<tr>
<td>Acid Phosphatase</td>
<td>Citrate</td>
</tr>
<tr>
<td>PSA</td>
<td>Spermine</td>
</tr>
<tr>
<td>Leucine aminotransferase</td>
<td>Spermidine</td>
</tr>
<tr>
<td>Diamine oxidase</td>
<td>Putrescine</td>
</tr>
<tr>
<td>B glucuronidase</td>
<td>Zinc</td>
</tr>
<tr>
<td>Plasminogen activator</td>
<td>Myoinositol</td>
</tr>
<tr>
<td>Complement C3 and C4</td>
<td>Cholesterol</td>
</tr>
<tr>
<td>Transferrin</td>
<td></td>
</tr>
<tr>
<td>Growth factors</td>
<td></td>
</tr>
<tr>
<td>Annexin1</td>
<td></td>
</tr>
</tbody>
</table>
1.4.2 Benign prostatic hyperplasia

Over half of men aged over 50 years have BPH. A third of these men present with lower urinary tract symptoms (LUTS) that range from urinary tract infection to urinary retention to renal insufficiency (Thorpe & Neal, 2003). A direct relationship between the BPH and LUTS is not fully evidenced (Lepor, 2005). Although most men will have histological BPH by the age of 80 the term involves prostates with volume more than 25ml (Kaplan et al., 2006). BPH refers to a dysfunctional proliferative process of both the stromal and epithelial elements within the periurethral and transitional zones (Mcneal, 1983). In fact it has been suggested that symptomatic BPH is related to an increased stromal: epithelial ratio as well as percentage smooth muscle (Lepor, 2005). Therefore the mechanisms involved in related urinary symptoms include bladder outlet obstruction and prostate smooth muscle dysfunction.

The treatment of BPH can be medical or surgical. Medical therapies are targeted towards diminishing bladder outlet obstruction by decreasing prostatic volume and relaxing prostatic smooth muscle. A1 adrenergic antagonists, like Doxazosin or 5a-reductase inhibitors like finasteride and their combination have been shown to alleviate BPH related symptoms but not directly through reducing prostatic volume (Kaplan et al., 2006). Surgically both transurethral resection of the prostate and radical prostatectomy have been shown to improve BPH related LUTS (Lepor, 2005). Other minimally invasive therapies such as laser coagulation, vaporisation or enucleation are also recommended by the European Association of Urology (EAU) (Madersbacher et al., 2004).
1.4.3 Prostate cancer

1.4.3.1 Epidemiology

Prostate cancer is the second most commonly diagnosed cancer in males and the sixth leading cause of cancer death worldwide (Torre et al., 2015). In the United Kingdom it is the most commonly diagnosed cancer with 41,700 diagnoses in 2011 and 10,800 deaths from the disease in 2012 (Cancer Reserch Uk, 2014). It develops in men aged more than 50 years. It is estimated that by the age of 80 years about 80% of men will have evidence of prostatic neoplasia (Breslow et al., 1977). In most of these cases, cancer follows an indolent course without any attached symptomatology, therefore it usually doesn’t warrant treatment. Prostate cancer exhibits considerable variation between ethnicities and countries (Breslow et al., 1977). It is most common in black men and least common in Asians. Also it is most common in America and least common in South-East Asia independent of ethnicity (Breslow et al., 1977; Parkin et al., 2010; Parkin et al., 1999). These differences may exist because of genetic variation, lifestyle dissimilarities or differing detection rates (Haas & Sakr, 1997; Potosky et al., 1995).

1.4.3.2 Aetiology and risk factors

The only established risk factors for prostatic cancer are age, family history and as mentioned earlier ethnicity (Bratt, 2002). Men with an affected first-degree relative have double the risk of acquiring prostate cancer later in life (Zeegers et al., 2003). Other associations with increased risk include obesity (Buschemeyer & Freedland, 2007), high blood pressure (Martin et al., 2010) and lack of exercise (Friedenreich et al., 2010).
The exact aetiology of prostatic cancer is not well understood but is believed to result from a complex series of initiation and promotion events under both genetic and environmental influences (Witte, 2009). Some studies have identified an increase in risk amongst men with BRCA1/BRCA2 mutations (Agalliu et al., 2007; Gayther et al., 2000). Genetic variations in the biosynthesis and metabolism of androgens have also been suggested. These involve the activity of androgen receptor within the prostate, which depends on the length on CAG repeats (Gu et al., 2012). Other factors likely influencing prostate cancer development include endogenous hormone imbalances especially considering androgens and oestrogens (Wu et al., 1991). Exogenous influences may include calcium and vitamin D consumption in dairy products (Aune et al., 2015), dietary fat (Park et al., 2007) and phytoestrogens (Barnes, 2001). Prostate cancer has also been linked with sexually transmitted infections (STI) most notably with human papilloma virus (HPV-2, HPV-16 and PHV-18) and is commoner in individuals with many previous sexual partners (Dennis et al., 2009; Dennis & Dawson, 2002).

1.4.3.3 Presentation

Prostate cancer rises from the glandular epithelial cells and is therefore considered an adenocarcinoma. It most usually appears in the peripheral zone. The symptoms it causes are similar to those of BPH; commonly hesitancy, poor urine stream, dysuria, post-micturition dribbling, recurrent urinary tract infections and sexual dysfunction. Additionally it may cause haematuria, lower back and pelvic pain, urinary and faecal incontinence depending on its stage (Miller et al., 2003). On per-rectum examination, positive findings would include asymmetrical lobar enlargement and nodularity.
1.4.3.4 Screening and diagnosis

Prostate screening is controversial. Currently digital rectal exam followed by a blood test for prostate specific androgen (PSA) levels are the most acceptable options. In the UK there is no established population-screening programme but the “Prostate Risk Management” program has been introduced as an informed choice programme (Figure 8) based on the findings of the “European Randomized Study of Screening for Prostate Cancer” (ERSPC). This study assigned individuals to a control group and a group undergoing four yearly screening tests by PSA levels. It concluded that within 9 years of screening there was a reduction in the rate of deaths from prostate cancer by 20% with a number needed to treat of 24 (Roobol et al., 2009). When these numbers were projected to 25 years the number needed to treat falls to 5 men (Gulati et al., 2011). The Prostate, Lung, Colorectal and Ovarian (PCLO) study, conducted in the United States, concluded that regular screening did not differ significantly from opportunistic screening in relation to prostate cancer specific mortality due to the very small number of deaths in the studied population (Andriole, 2014).
Figure 8: “Prostate risk management program”. Frequently asked questions and answers to provide information for informed consent.

<table>
<thead>
<tr>
<th>Benefits</th>
<th>Limitations</th>
</tr>
</thead>
<tbody>
<tr>
<td>It may reassure you if the test result is normal.</td>
<td>It can miss cancer and provide false reassurance.</td>
</tr>
<tr>
<td>It may give you an indication of cancer before symptoms develop.</td>
<td>It may lead to unnecessary worry and medical tests when there is no cancer.</td>
</tr>
<tr>
<td>It may find cancer at an early stage when treatments could be of benefit.</td>
<td>It cannot tell the difference between slow-growing and fast-growing cancer.</td>
</tr>
<tr>
<td>If treatment is successful, the worst possible outcomes of more advanced cancer, including death, are avoided.</td>
<td>It may make you worry by finding slow-growing cancers that may never cause any symptoms or shorten your life.</td>
</tr>
<tr>
<td>Even if the cancer is more advanced and treatment is less successful, it will usually extend life.</td>
<td>48 men will undergo treatment in order to save one life.</td>
</tr>
</tbody>
</table>

Adopted from *PSA (prostate specific antigen) testing for prostate cancer. An information sheet for men considering PSA test*. This patient information sheet was updated by Dr Deborah Burford and Dr Joan Austoker from the Cancer Research UK Primary Care Education Research Group, University of Oxford, and Professor Michael Kirby, Visiting Professor to the Faculty of Health and Human Sciences, University of Hertfordshire. The information on this sheet is based on material from the booklet *Prostate Cancer Risk Management Programme information for primary care; PSA testing in asymptomatic men* (NHS Cancer Screening Programmes, 2009).
The gold standard for the diagnosis of prostate cancer is by histological examination of per rectum core biopsies (Essink-Bot et al., 1998). These biopsies are undertaken under ultrasound (US) or magnetic resonance imaging guidance (MRI) (Natarajan et al., 2011). The tissues are stained with H&E for light microscopy. Staging of prostate cancer is radiological and follows the TNM system (Figure 9). In the presence of neoplasia the Gleason score is most commonly applied to grade the abnormality (Figure 10).

<table>
<thead>
<tr>
<th>Tumour</th>
<th>Lymph nodes</th>
<th>Metastasis</th>
</tr>
</thead>
<tbody>
<tr>
<td>T1</td>
<td>Not palpable</td>
<td>Nx Not checked</td>
</tr>
<tr>
<td>T2a</td>
<td>½ of one lobe</td>
<td>N0 No involvement</td>
</tr>
<tr>
<td>T2b</td>
<td>&gt;½ of one lobe</td>
<td></td>
</tr>
<tr>
<td>T2c</td>
<td>Both lobes</td>
<td></td>
</tr>
<tr>
<td>T3a</td>
<td>Capsule breached</td>
<td>N1 Involved</td>
</tr>
<tr>
<td>T3b</td>
<td>Spread to seminal vesicles</td>
<td></td>
</tr>
<tr>
<td>T4</td>
<td>Metastasis</td>
<td></td>
</tr>
</tbody>
</table>
Figure 10: Gleason score

<table>
<thead>
<tr>
<th>Pattern of differentiation</th>
<th>Small uniform glands</th>
<th>Increased stroma between glands</th>
<th>Infiltration of cells from glands at margins</th>
<th>Irregular masses of neoplastic cells with few glands</th>
<th>Occasional glands between sheets of cells</th>
</tr>
</thead>
<tbody>
<tr>
<td>Primary grade</td>
<td>Dominant pattern (&gt;50%)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Secondary grade</td>
<td>Next most frequent pattern (5&lt;50%)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Tertiary grade</td>
<td>Small component with more aggressive pattern</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Gleason score = primary grade + secondary grade (if only 2 patterns witnessed)

Gleason score = primary grade + tertiary grade (if a tertiary grade is seen)
1.4.3.5 Molecular pathology of prostate adenocarcinoma

Several molecular alterations have been implicated in the development and progression of prostate cancer. The potential of these alterations in determining malignancy has not been confirmed but is under consideration with extensive research being undertaken in these areas. Several molecules that regulate the cell cycle including division and apoptosis have been studied but they have not been directly linked to prostate carcinogenesis. These include inactivation of $P53$ (Osman et al., 1999), variable expression of $P27$ (Macri & Loda, 1998), and positivity of $Bcl-2$ oncoprotein (Keshgegian et al., 1998). $PTEN$ is a tumour suppression gene commonly inactivated in sporadic and high-grade prostate tumours implying close links with their development (Mcmenamin et al., 1999; Whang et al., 1998). Several growth factors have also been linked to the development of prostate malignancy. The control of these growth factors by androgens may be lost causing deregulated cellular proliferation (Chung et al., 1997). Perhaps the closest association is between Insulin-like growth factors (IGF) I and II. Population studies have shown that IGF-I plasma levels increase in prostate cancer and improve the detection rate over PSA alone (Chan et al., 2002). In addition IGF-II levels are increased in malignant prostatic cells. Other growth factors implicated in the development and progression of prostate cancer are listed in figure 11.
**Figure 11: Growth factors implicated in prostate cancer**

<table>
<thead>
<tr>
<th>Growth Factors</th>
</tr>
</thead>
<tbody>
<tr>
<td>Transforming growth factor-b (TGF-β)</td>
</tr>
<tr>
<td>Fibroblast growth factors (FGF)</td>
</tr>
<tr>
<td>Epidermal growth factors (EGF)</td>
</tr>
<tr>
<td>Insulin-like growth factors (IGF)</td>
</tr>
<tr>
<td>Platelet-derived growth factors (PDGF)</td>
</tr>
<tr>
<td>Vascular endothelial growth factors (VEGF)</td>
</tr>
<tr>
<td>Neurotensin</td>
</tr>
<tr>
<td>Endothelins</td>
</tr>
<tr>
<td>Colony-stimulating factors</td>
</tr>
</tbody>
</table>

A. Normal prostate at high magnification showing the columnar epithelial papillae and the “corpora amylacea” which are laminated concretions within the lumen of the acini.

B. Benign prostatic hyperplasia at medium magnification showing preservation of the columnar epithelial arrangement around the lumina.

C. Prostatic intraepithelial neoplasia (high grade) at high magnification with cellular proliferation around a single or a small number of acini. This suggests the presence of overt adenocarcinoma in 50% of cases.

D. Prostatic adenocarcinoma at low power showing a residual glandular element surrounded by densely packed neoplastic cells exhibiting prominent nucleoli and lacking intervening stroma.

Figure 12: Histopathological features of the normal prostate, benign prostatic hyperplasia, prostatic intraepithelial neoplasia and prostatic adenocarcinoma
Adapted from “the internet pathology laboratory for medical education, Mercer University School of Medicine, Savannah (accessed on 11/02/2015)
1.4.3.6 Management

The management options for prostate carcinoma depend on the age of the patient, the symptoms as well as the aggressiveness of the cancer (Gleason score), stage and PSA level. “Active monitoring” is a valid option for older men with comorbidities, slow-growing cancers or those who would not accept potential treatment side effects. This surveillance involves serial PSA level testing and repeated biopsies to identify progression of the disease, which would prompt re-discussion of the management options. Surgical options involve radical prostatectomy, which can be undertaken by retropubic, perineal, and laparoscopic or robotic approaches. The major side effects of such operations are erectile dysfunction and urinary incontinence. Robotic nerve sparing procedures reduce the risk of such side effects but do not eliminate them. Radiotherapy can be used as adjuvant or stand-alone treatment and can be in the form of external beam or brachytherapy. Other options commonly used in early prostate cancer include high intensity focused ultrasound (HIFU) and cryosurgery. For later stages where the tumour has extended beyond the prostate capsule or has exhibited metastases, chemotherapy, immunotherapy or hormonal therapy or a combination thereof is used (Gerritsen, 2012; Yagoda & Petrylak, 1993). Finally palliative care aims to improve quality of life when the disease is deemed terminal. One of the commonest and most severe symptoms is bone pain from spinal metastases. This is usually treated with a combination of opioids, steroids, bisphosphonates and radiation therapy (Thompson et al., 2007).
1.5 The female mammary gland

1.5.1 Anatomy and physiology

The breast consists of glandular lobules drained by 15 to 20 lactiferous glands that open onto the nipple. The nipple is surrounded by the areola, which contains sebaceous glands for lubrication. The glandular lobules are surrounded by adipose tissue and are separated by connective tissue septae. Suspensory ligaments made of connective tissue and called Cooper’s ligaments attach the deep skin layer to the pectoralis facia to provide support. For clinical reasons the breast is divided into four anatomical quadrants: superolateral, superomedial, inferomedial and inferolateral. Blood supply to the breast is primarily by the internal mammary artery. Breast lymphatics drain into the supra-clavicular, infra-clavicular, internal mammary and axillary lymph nodes.

Embryologically, breast tissue is formed from ectoderm and mesenchyme. The epidermal ectoderm develops into the primitive mammary ridges or milk lines by the sixth week of gestation. The mammary ridges of the pectoral area develop into the mammary buds. These incrementally develop into lactiferous ducts that canalise prenatally to give rise to glandular alveoli and ducts. The mesenchyme gives rise to connective tissue and vessels as well as the nipple. The areola develops around the nipple as the epidermal mammary pit that originates from the original mammary ridges (Figure 13).
**Figure 13: Embryological development of the breast**

Adapted from “Embryology and Anatomy of the Breast” (Skandalakis, 2009)

<table>
<thead>
<tr>
<th>Time Period</th>
<th>Development</th>
</tr>
</thead>
<tbody>
<tr>
<td>4-6 weeks of gestation</td>
<td>Development of the ectodermal mammary ridges (milk lines)</td>
</tr>
<tr>
<td>10 weeks of gestation</td>
<td>Development of the pectoral part of the milk lines</td>
</tr>
<tr>
<td>20 weeks of gestation</td>
<td>Development of the areola and 15-20 solid cords</td>
</tr>
<tr>
<td>3rd trimester</td>
<td>Development of the lactiferous ducts from the milk lines</td>
</tr>
<tr>
<td>After birth</td>
<td>Visible nipple</td>
</tr>
<tr>
<td>Puberty</td>
<td>Ducts develop acini in their ends</td>
</tr>
</tbody>
</table>

The mammary glands remain underdeveloped until puberty as Tanner stage 1. At 10 to 11 years of age thelarche starts with the formation of the breast buds and areolar hyperplasia signifying Tanner stage 2. Within approximately 12 months Tanner stage 3 occurs when breast tissue extents beyond the areolae. In stage 4 the nipples and areolae form a secondary mount causing contour separation. In adulthood, stage 5, the breast is fully developed with a single contour (Figure 14) (Marshall & Tanner, 1969).
From a cellular point of view, stromal fatty and fibrous tissue proliferation occurs before ductal elongation and branching under oestrogenic and progestogenic influence (Russo & Russo, 2004). The primary ducts branch into segmental and sub-segmental ducts that lead to terminal duct formation that further branch to several acini. The terminal ductal lobular unit (TDLU) is composed of the acini emerging from a single terminal duct and their surrounding stroma (Figure 15). Also under pubertal oestrogenic influence, the alveolar epithelium becomes bilayered with glandular luminal cells and myoepithelial basal cells (Tiede & Kang, 2011).
The breast reaches maturation by the age of 20 years and consists of lobules with short terminal ducts ending in a small number of alveoli. These lobules are termed type-1 lobules. Type-4 lobules are only seen after pregnancy and lactation and consist of terminal ducts branching into several ductules and containing large numbers of alveoli. Type-2 lobules emerge during pregnancy and type-4 during lactation (Howard & Gusterson, 2000).

The breast functions as a nutritional organ after parturition by secreting milk for breastfeeding. The production of milk (galactopoiesis) is a complex process that starts
in the second trimester of pregnancy under the influence of several hormones. Oestrogen and progesterone stimulate growth and differentiation of the TDLUs (Mohrbacher & Knorr, 2012). Prolactin and to a lesser extent growth hormone are also involved in the TDLU differentiation but also control the production of breast milk. Oxytocin causes the contractions of myoepithelial basal cells for milk ejection. Other hormones involved in breast tissue differentiation and lactation include, human placental lactogen, thyroid stimulating hormone, adrenocorticotrophic hormone, follicle stimulating and luteinizing hormones. The process of menopausal breast involution is divided into the pre-climacteric phase and the postmenopausal phase. These phases are continuous and progressive and result from rises in FSH and oestradiol fluctuations. From approximately 35 years of age the mammary glandular epithelium is gradually replaced by adipose tissue and the stromal connective tissue by condensed collagen. Postmenopausally, only the major branches of the lactiferous ducts can be seen surrounded by few lobules within fibrotic tissue. As the interlobular connective tissue is reduced the breast adopts a shrunken contour (Mansel, 2009).
1.5.2 Breast cancer

1.5.2.1 Epidemiology

Globocan estimates breast cancer incidence to be 1.67 million cases in 2012 worldwide. This makes breast cancer the commonest cancer in women. It is also the most common cause of death from cancer in less developed countries and the second cause of cancer death in developed regions after lung cancer. Incidence rates vary considerably across the world with Western Europe having a rate of 96 per 100,000 in contrast to East Asia that has a rate of 27 per 100,000 (Ferlay et al., 2015). In the UK, breast cancer incidence is the sixth highest in Europe. Data from the “Office of National Statistics” reveal that 49,936 women and 349 men were diagnosed with the disease in 2011. 41% of breast cancers diagnosed in the UK are stage 1, 45% are stage 2, 9% stage 3 and 5% stage 4 (Lyratzopoulos et al., 2012). Breast cancer related mortality in the UK was 9,698 per 100,000 females and 58 per 100,000 males in 2012. Mortality has greatly decreased since the 1970’s due to the introduction of population screening, increased specialisation of care and more effective surgical and medical treatments (Kingsmore et al., 2003).

1.5.2.2 Aetiology

Several risk factors have been implicated in the development of breast cancer. Increased oestrogen exposure features extensively as a predisposing factor (Pike et al., 1993). This is evident by the increased risk associated with early menarche and late menopause, nulliparity, the use of oral contraception and hormone replacement therapy. Also pregnancy at an early age and breast-feeding have a protective effect as they reduce the overall period of oestrogen exposure (Adami et al., 1995). Other risk factors include obesity, lack of exercise, increased alcohol intake and radiation
exposure. Genetic variation plays a major role in the susceptibility for breast cancer. Women with a positive family history are at higher risk especially if they harbour BRCA1 or BRCA2 mutations. Several other mutations have also been identified including TP53, PTEN, STK11, CDH1, PALB2 and other lower penetrance variants (Mavaddat et al., 2010). Of course, even in the case of BRCA carrier families the estimated lifetime risk of developing breast cancer is 65% for BRCA1 and 45% for BRCA2 (Antoniou & Easton, 2006). This indicates that susceptibility to breast cancer is mediated through variants in many genes as well as environmental influences, each conferring an increase in the risk for disease (Yager & Davidson, 2006). There is evidence that women from South-East Asia who have a lower risk of breast cancer than Western Caucasians also have lower urine and blood levels of oestrogen (Wu & Pike, 1995). This has not been related to fat intake or body weight (Martin-Moreno et al., 1994). Also it cannot be attributed to genetic make-up. The rate of breast cancer in Asian migrants to America and Japanese-Americans is 70% of the rate in White Americans (Miller et al., 1989). A similar increase is noticed in urban areas within Japan. These changes were not related to age at menarche, age at first birth, age at menopause, and parity (Nagata et al., 1997).
1.5.2.3 Presentation

The commonest symptom of breast cancer is a new mass. This can be hard or soft, tender or painless, irregular or rounded. A hard irregular painless fast-growing mass has a higher risk of being malignant. Other signs include skin irritation, puckering, swelling, erythema and peau d’orange. The nipple may be thickened, swollen, discoloured or retracted. There may be discharge from the nipple. Also associated with breast cancer is lymphadenopathy with enlarged lymphnodes, usually in the axillary area.

1.5.2.4 Screening and diagnosis

Breast cancer screening can be divided to universal population screening and high-risk screening. One of the methods for population screening is self-examination. This is currently promoted in the UK through the breast awareness campaign, it has not been proven to reduce breast cancer related deaths (Kosters & Gotzsche, 2003). It involves regular inspection and palpation of the breast to notice any changes as described previously and reporting them without delay. Mammography is used for universal screening from the age of 50 to 70 years in three-yearly intervals in the UK and currently being extended in England to include the ages of 47 to 73 years. A mammogram consists of essentially two x-rays of each breast whose images are digitally processed to reveal any masses that are denser than the surrounding tissue. A positive mammogram either leads to a repeat examination, another form of imaging such as MRI, or a biopsy. The effectiveness of a mammogram has been disputed in a 2011 Cochrane review. The authors concluded that although the relative risk reduction is 15% the absolute risk reduction is only 0.5%, and these numbers may be
skewed by self-selection and other biases (Gotzsche & Nielsen, 2011). Also, mammography leads to 30% increase in overdiagnosis and overtreatment, especially for those aged below 50 years. Therefore it was concluded that the risks of mammography outweigh the benefits in younger (<50 years) women (Gotzsche & Nielsen, 2011). For younger women mammography can be used as a diagnostic test. Due to the higher tissue density in younger women, ultrasonography or MRI testing are considered better imaging methods. In fact, MRI is very sensitive and has a high negative predictive value; so a negative MRI should discourage any invasive tests. The disadvantage is that it is less specific than mammography as benign conditions such as fibromas and adenomas may have similar radiological characteristics to malignant lesions (Hrung et al., 1999). It is also more expensive and requires the injection of metal, radiopaque substances such as gadolinium. BRCA testing can be used for screening in high risk families were at least one member of the family has been affected by a BRCA mutation (Gabai-Kapara et al., 2014).

The gold standard for the diagnosis of breast cancer is histological. Biopsies for histological examination can be retrieved by fine needle aspiration, stereotactic biopsy, ultrasound or MRI guided core biopsy, and excisional biopsy.

There are a re several classification methods for breast cancer. One of the easiest ones to understand is based on anatomical location of the tumour as described in (Figure 16). Breast cancer staging is based on TNM classification but also has a prognosis based staging system (Figure 17,18).
**Figure 16: Classification of breast cancer based on anatomical location**

Adapted from “The American Joint Committee on Cancer: the 7th edition of the AJCC cancer staging manual and the future of TNM” (Edge & Compton, 2010)

<table>
<thead>
<tr>
<th>Tumor Location</th>
<th>Histologic Subtype</th>
</tr>
</thead>
<tbody>
<tr>
<td>Carcinoma, NOS</td>
<td></td>
</tr>
<tr>
<td>Ductal</td>
<td>Intraductal (<em>in situ</em>)</td>
</tr>
<tr>
<td></td>
<td>Invasive with predominant component</td>
</tr>
<tr>
<td></td>
<td>Invasive, NOS</td>
</tr>
<tr>
<td></td>
<td>Comedo</td>
</tr>
<tr>
<td></td>
<td>Inflammatory</td>
</tr>
<tr>
<td></td>
<td>Medullary with lymphocytic infiltrate</td>
</tr>
<tr>
<td></td>
<td>Mucinous (colloid)</td>
</tr>
<tr>
<td></td>
<td>Papillary</td>
</tr>
<tr>
<td></td>
<td>Scirrhous</td>
</tr>
<tr>
<td></td>
<td>Tubular</td>
</tr>
<tr>
<td></td>
<td>Other</td>
</tr>
<tr>
<td>Lobular</td>
<td>Invasive with predominant <em>in situ</em> component</td>
</tr>
<tr>
<td></td>
<td>Invasive</td>
</tr>
<tr>
<td>Nipple</td>
<td>Paget disease, NOS</td>
</tr>
<tr>
<td></td>
<td>Paget disease with intraductal carcinoma</td>
</tr>
<tr>
<td></td>
<td>Paget disease with invasive ductal carcinoma</td>
</tr>
<tr>
<td>Other</td>
<td>Undifferentiated carcinoma</td>
</tr>
<tr>
<td></td>
<td>Metaplastic</td>
</tr>
</tbody>
</table>
**Figure 17: Classification of breast cancer based on the TNM system**

Adapted from “The American Joint Committee on Cancer: the 7th edition of the AJCC cancer staging manual and the future of TNM” (Edge & Compton, 2010)

<table>
<thead>
<tr>
<th>TX</th>
<th>Primary tumor cannot be assessed.</th>
</tr>
</thead>
<tbody>
<tr>
<td>T0</td>
<td>No evidence of primary tumor.</td>
</tr>
<tr>
<td>Tis</td>
<td>Carcinoma <em>in situ</em>.</td>
</tr>
<tr>
<td>Tis (DCIS)</td>
<td>DCIS.</td>
</tr>
<tr>
<td>Tis (LCIS)</td>
<td>LCIS.</td>
</tr>
<tr>
<td>Tis (Paget)</td>
<td>Paget disease of the nipple NOT associated with invasive carcinoma and/or carcinoma <em>in situ</em> (DCIS and/or LCIS) in the underlying breast parenchyma. Carcinomas in the breast parenchyma associated with Paget disease are categorized based on the size and characteristics of the parenchymal disease, although the presence of Paget disease should still be noted.</td>
</tr>
<tr>
<td>T1</td>
<td>Tumor ≤20 mm in greatest dimension.</td>
</tr>
<tr>
<td>T1mi</td>
<td>Tumor ≤1 mm in greatest dimension.</td>
</tr>
<tr>
<td>T1a</td>
<td>Tumor &gt;1 mm but ≤5 mm in greatest dimension.</td>
</tr>
<tr>
<td>T1b</td>
<td>Tumor &gt;5 mm but ≤10 mm in greatest dimension.</td>
</tr>
<tr>
<td>T1c</td>
<td>Tumor &gt;10 mm but ≤20 mm in greatest dimension.</td>
</tr>
<tr>
<td>T2</td>
<td>Tumor &gt;20 mm but ≤50 mm in greatest dimension.</td>
</tr>
<tr>
<td>T3</td>
<td>Tumor &gt;50 mm in greatest dimension.</td>
</tr>
<tr>
<td>T4</td>
<td>Tumor of any size with direct extension to the chest wall and/or to the skin (ulceration or skin nodules).</td>
</tr>
<tr>
<td>T4a</td>
<td>Extension to the chest wall, not including only pectoralis muscle adherence/invasion.</td>
</tr>
<tr>
<td>T4b</td>
<td>Ulceration and/or ipsilateral satellite nodules and/or edema (including peau d'orange) of the skin, which do not meet the criteria for inflammatory carcinoma.</td>
</tr>
<tr>
<td>T4c</td>
<td>Both T4a and T4b.</td>
</tr>
<tr>
<td>T4d</td>
<td>Inflammatory carcinoma.</td>
</tr>
<tr>
<td>NX</td>
<td>Regional lymph nodes cannot be assessed (e.g., previously removed).</td>
</tr>
<tr>
<td>-------------</td>
<td>---------------------------------------------------------------</td>
</tr>
<tr>
<td>N0</td>
<td>No regional lymph node metastases.</td>
</tr>
<tr>
<td>N1</td>
<td>Metastases to movable ipsilateral level I, II axillary lymph node(s).</td>
</tr>
<tr>
<td>N2</td>
<td>Metastases in ipsilateral level I, II axillary lymph nodes that are clinically fixed or matted.</td>
</tr>
<tr>
<td></td>
<td>OR</td>
</tr>
<tr>
<td></td>
<td>Metastases in clinically detected ipsilateral internal mammary nodes in the absence of clinically evident axillary lymph node metastases.</td>
</tr>
<tr>
<td>N2a</td>
<td>Metastases in ipsilateral level I, II axillary lymph nodes fixed to one another (matted) or to other structures.</td>
</tr>
<tr>
<td>N2b</td>
<td>Metastases only in clinically detected ipsilateral internal mammary nodes and in the absence of clinically evident level I, II axillary lymph node metastases.</td>
</tr>
<tr>
<td>N3</td>
<td>Metastases in ipsilateral infraclavicular (level III axillary) lymph node(s) with or without level I, II axillary lymph node involvement.</td>
</tr>
<tr>
<td></td>
<td>OR</td>
</tr>
<tr>
<td></td>
<td>Metastases in clinically detected ipsilateral internal mammary lymph node(s) with clinically evident level I, II axillary lymph node metastases.</td>
</tr>
<tr>
<td></td>
<td>OR</td>
</tr>
<tr>
<td></td>
<td>Metastases in ipsilateral supraclavicular lymph node(s) with or without axillary or internal mammary lymph node involvement.</td>
</tr>
<tr>
<td>N3a</td>
<td>Metastases in ipsilateral infraclavicular lymph node(s).</td>
</tr>
<tr>
<td>N3b</td>
<td>Metastases in ipsilateral internal mammary lymph node(s) and axillary lymph node(s).</td>
</tr>
<tr>
<td>N3c</td>
<td>Metastases in ipsilateral supraclavicular lymph node(s).</td>
</tr>
<tr>
<td>M0</td>
<td>No clinical or radiographic evidence of distant metastases.</td>
</tr>
<tr>
<td>cM0 (i+)</td>
<td>No clinical or radiographic evidence of distant metastases, but deposits of molecularly or microscopically detected tumor cells in circulating blood, bone marrow, or other non-regional nodal tissue that are ≤0.2 mm in a patient without symptoms or signs of metastases.</td>
</tr>
<tr>
<td>M1</td>
<td>Distant detectable metastases as determined by classic clinical and radiographic means and/or histologically proven &gt;0.2 mm.</td>
</tr>
</tbody>
</table>
**Figure 18: Classification of breast cancer based according to prognosis and correlation with the TNM system**

Adapted from “The American Joint Committee on Cancer: the 7th edition of the AJCC cancer staging manual and the future of TNM” (Edge & Compton, 2010)

<table>
<thead>
<tr>
<th>Stage</th>
<th>T</th>
<th>N</th>
<th>M</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>Tis</td>
<td>N0</td>
<td>M0</td>
</tr>
<tr>
<td>IA</td>
<td>T1b</td>
<td>N0</td>
<td>M0</td>
</tr>
<tr>
<td>IB</td>
<td>T0</td>
<td>N1mi</td>
<td>M0</td>
</tr>
<tr>
<td></td>
<td>T1b</td>
<td>N1mi</td>
<td>M0</td>
</tr>
<tr>
<td>IIA</td>
<td>T0</td>
<td>N1c</td>
<td>M0</td>
</tr>
<tr>
<td></td>
<td>T1b</td>
<td>N1c</td>
<td>M0</td>
</tr>
<tr>
<td>IIB</td>
<td>T2</td>
<td>N0</td>
<td>M0</td>
</tr>
<tr>
<td></td>
<td>T3</td>
<td>N0</td>
<td>M0</td>
</tr>
<tr>
<td>IIIA</td>
<td>T0</td>
<td>N2</td>
<td>M0</td>
</tr>
<tr>
<td></td>
<td>T1b</td>
<td>N2</td>
<td>M0</td>
</tr>
<tr>
<td></td>
<td>T2</td>
<td>N2</td>
<td>M0</td>
</tr>
<tr>
<td></td>
<td>T3</td>
<td>N1</td>
<td>M0</td>
</tr>
<tr>
<td></td>
<td>T3</td>
<td>N2</td>
<td>M0</td>
</tr>
<tr>
<td>IIIB</td>
<td>T4</td>
<td>N0</td>
<td>M0</td>
</tr>
<tr>
<td></td>
<td>T4</td>
<td>N1</td>
<td>M0</td>
</tr>
<tr>
<td>IIIc</td>
<td>Any T</td>
<td>N3</td>
<td>M0</td>
</tr>
<tr>
<td>IV</td>
<td>Any T</td>
<td>Any N</td>
<td>M1</td>
</tr>
</tbody>
</table>
1.5.2.5 Molecular Pathology

The majority of neoplastic breast diseases arise from the TDLU independent of histological type. Therefore the designations ductal and lobular lesions reflect cellular morphology rather than origin (Simpson et al., 2003). The difference in cellular morphology can be explained by differences in the genetic alterations associated with these lesions. Lobular lesions are associated with the loss of E-Cadherin, a cell adhesion molecule, from a genetic alteration in position 16q22 further associated with promoter methylation. Ductal lesions are not associated with these changes (Roylance et al., 2003; Simpson et al., 2003). In addition pleomorphic lobular carcinomas express genetic changes associated with both aforementioned types and also show overexpression of human epithelial growth factor 2 (Her-2) (Middleton et al., 2000).

Histological classification of breast cancer is not a good predictor of outcome. Grade is a more reliable prognostic factor and also better reflects the associated genotypic and phenotypic patterns (Roylance et al., 1999; Simpson et al., 2003). In fact two pathways have been modelled for breast cancer progression: the “low grade” arm compromises of well-differentiated ductal carcinoma is situ progressing to invasive ductal carcinoma and lobular carcinoma in situ progressing to invasive lobular carcinoma. These tumours are usually oestrogen (ER) and progesterone receptor (PgR) positive and Her-2 negative. Tumours in the “high grade” arm are usually hormone receptor (ER and Pgr) negative and either Her-2 positive or negative. Pleomorphic lobular carcinoma has features of both but is usually high grade (Simpson et al., 2005).
1.5.2.6 Management

The management of breast cancer depends on the stage the disease is at the time of diagnosis. Therefore a positive biopsy prompts a pre-operative staging of the axillary lymph nodes by ultrasound. Any abnormal lymph nodes are biopsied. Also MRI scanning is considered in order to confirm size and extend of the disease. For precancerous lesions such as DCIS conservative surgery is considered involving at least 2mm radial excision margins. For Paget’s disease of the breast, excision of the nipple and surrounding areola is an alternative to total mastectomy. In addition to breast surgery axillary lymph node status has to be evaluated by sentinel lymph node biopsy for all invasive cancers. Patients with involved axillary lymph nodes should be offered axillary lymph node dissection (ALND).

Adjuvant therapy is considered depending on the ER, PgR and HER2 status and other prognostic and predictive factors. There are several therapy types that can be chosen alone or in combination according to the tumour characteristics. Endocrine therapy includes Tamoxifen for ER-positive tumours with or without chemotherapy and ovarian suppression. Aromatase inhibitors like Anastrozole are used as either primary therapy or in addition or as a replacement of Tamoxifen in postmenopausal women. Patients who had conservative surgery for invasive carcinoma should have radiotherapy as should those with a complete mastectomy and are at high risk of recurrence. Axillary radiotherapy can be added if ALND is not possible after sentinel node biopsy. Adjuvant radiotherapy is used in the area if more than three lymph nodes are involved or there are lymph nodes involved and poor prognostic factors. Breast cancer survivors are followed-up by annual mammography or other imaging to detect recurrence at the same time controlling side-effect like lymphoedema of the arms and menopausal symptoms.
1.6 The uterus, cervix and fallopian tube

1.6.1 Anatomy and Physiology

The uterus is a fibromuscular organ situated in the pelvis when not gravid. The uterine corpus is muscular. The fundus forms the proximal part of the uterus and contains two openings termed ostia, leading to the fallopian tubes. The lower part termed the isthmus leads to the fibrous cervix that ends in the vagina (Figure 19).

Figure 19: Diagram of the uterus, cervix, vagina, fallopian tubes and ovaries

The uterine wall consists of three layers: the endometrium, myometrium and serosa. The endometrium consists of the inner stratum functionalis, and stratum basalis. The stratum functionalis develops during the proliferative phase of the menstrual cycle and consists of the stratum compactum and stratum spongialis. The myometrium has
three layers from inside out: the stratum infra-vasculare, stratum vasculare and stratum supra-vasculare that continues to the fallopian tubes and round ligaments. The serosa is the outer peritoneal covering.

The fallopian tube emerges from the uterus at the isthmus continuing to the ambula and ending at the fimbriated infundibulum. The fallopian tube has three layers: the serosa, the muscularis mucosa and the mucosa. The mucosa consists of ciliated cells and secretory peg cells.

The cervix measures 2.5 to 3cm in length in the adult nulligravida. The vaginal portion is termed the exocervix and contains the external cervical os that connects to the uterine cavity via the cervical canal. The cervical stroma is a mixture of fibrous, muscular and elastic tissue. The cervical canal is lined by columnar epithelium while the exocervix is lined by squamous epithelium. The junction of the two epithelia where squamous metaplasia occurs is called the transformation zone.

Embryologically, the uterus develops from the paramesonephric (Müllerian) ducts. Each duct develops on the side of their respective Wolffian duct. Their common origin develops into the fallopian tube ostium. In the third month of gestation the caudal part of the Müllerian ducts fuse to form the uterus and vagina. The proximal parts remain separate to form the fallopian tubes. The cervix develops in the fifth month as a constriction at the distal uterus (Moore Kl, 1998).

The endometrium and to some extend the myometrium are under cyclical hormonal control by LH, FSH, oestrogen and progesterone. In the proliferative stage FSH causes endometrial thickening due to connective tissue proliferation and development of glandular structures. Oestrogen drives vascular proliferation. Ovulation marks the start of the secretory phase that is governed by LH and progesterone secretion. This
causes the glandular crypts to become corkscrew shaped and secrete glycogen. Menstruation is signalled by a reduction in LH and progesterone. Here the stratum functionalis degenerates due to spiral arteriole retraction and dislodges due to their rupture. The stratum basalis remains largely unaffected (Figure 20).

---

**Figure 20**: Diagram illustrating the effect of the hormones controlling the menstrual cycle to the endometrium

Adopted from: http://www.netwellness.org/healthtopics/pregnancy/pregmenstrualcycle.cfm
1.6.2 Uterine cancer

1.6.2.1 Epidemiology

Uterine cancer was the 6th most common cancer in women worldwide and the 4th most common in the UK in 2011 with a crude incidence rate of 26 per 100,000 women (Ferlay et al., 2013). Its incidence is highest in North America and lowest in South-East Asia (Ferlay et al., 2013). In the United Kingdom there is no significant variation in incidence between different ethnicities.

73% of uterine cancers are diagnosed between the ages of 40 and 74 years with the incidence rising sharply from 40 years, peaking at 70 to 74 years and declining thereafter.

1.6.2.2 Aetiology

Unopposed oestrogen, either endogenous or exogenous, has been implicated in the increase of endometrial cancer risk (Evans et al., 2011). Early menarche, late menopause, nulliparity and anovulation increase oestrogen exposure leading to increased cancer risk (Allen et al., 2008). Hormone replacement therapy given for post-menopausal symptoms causes an estimated 1% of uterine cancer (Parkin, 2011). Only unopposed oestrogen HRT has been shown to be associated endometrial hyperplasia and neoplasia and it is dose and duration dependent (Furness et al., 2012; Grady et al., 1995). Combined or cyclic sequential HRT with adequately opposed oestrogen is not associated with increased risk (Brinton & Felix, 2014). Obesity is of special interest as a risk factor due to its increasing prevalence. The risk of endometrial cancer is 3.3 times higher in women with a BMI of more than 25 and an extra 2.5 times higher in those with BMI of more than 30 (Zhang et al., 2014).
Endometrial cancer risk is associated with diabetes as in this population obesity is commoner. Polycystic ovarian syndrome also increases the risk because it causes anovulation and therefore prolonged oestrogen exposure. Tamoxifen administration in breast cancer survivors increases uterine cancer risk by three times (Early Breast Cancer Trialists' Collaborative et al., 2011). Diethylstilbestrol exposure in utero has been linked to endometrial cancer also. Endometrial cancer risk has been shown to be lower in women who have used the combined oral contraceptive pill for more than 5 years (Mueck et al., 2010).

Hereditary non-polyposis coli (HNPCC) or Lynch syndrome is an autosomal dominant condition associated with a 40-60% risk of uterine malignancy in addition to other colonic and extra-colonic cancers (Hoffman BL, 2012). Most uterine carcinomas associated with HNPCC are endometrioid (>80%) the rest being clear cell carcinomas, carcinosarcomas and uterine papillary carcinomas (Broaddus et al., 2006).

1.6.2.3 Presentation

Most women with endometrial cancer have early-stage disease at presentation. This is because of the early appearance of related symptoms. The most frequent symptom is postmenopausal or abnormal uterine bleeding. About 10% of women presenting with postmenopausal bleeding will be diagnosed with endometrial cancer (Gredmark et al., 1995). For pre-menopausal women the risk of endometrial cancer presenting as abnormal bleeding increases with the presence of associated risk factors (tamoxifen, HRT, PCOS etc.). The commonest investigation protocols in the UK for women with suspected endometrial cancer involve transvaginal ultrasound scanning for endometrial thickness and presence of masses impinging the endometrial cavity. An endometrial thickness of less than 3-5mm is associated with very low risk of
malignancy. If the endometrium is thicker than 5mm a biopsy is warranted (Dijkhuizen et al., 1996; Smith-Bindman et al., 1998). This can be performed using a Pipelle biopsy as a blind procedure or a biopsy under hysteroscopic view. A Pipelle biopsy has a high sensitivity and specificity for endometrial hyperplasia or neoplasia and a hysteroscopy can aid visualisation of other bleeding causes like submucous fibroids and polyps (Dijkhuizen et al., 2000).

1.6.2.4 Screening and Diagnosis

Diagnosis of endometrial cancer is performed by histological examination of biopsy samples. There is currently no well population-screening program for endometrial cancer as investigating high-risk individuals with symptoms usually identifies endometrial cancer at an early stage. Occasionally endometrial pathology may be identified on a cervical screening cytology sample but this is not part of the official cervical screening program.

Histological type and grade of endometrial cancer are important prognostic factors. There are two general types with different pathogenic origins: type 1 develops from complex atypical hyperplasia of the endometrium due to excess/ unopposed oestrogen. Type 2 develops from atrophic endometrium, is not oestrogen dependent and has a worse prognosis due to higher rate of recurrence, and a more aggressive pattern of spread than type 1 (Figure 21).

Uterine cancers can also arise from the stromal or muscular layers, termed stromal sarcomas or leimyosarcomas respectively.
Although endometrial cancer is surgically staged radiological staging is used to assist with planning for the appropriate surgical procedure. MRI is used to estimate the extent of the disease into soft tissues while CT has a better sensitivity for involved lymphnodes and distant metastases. Surgical staging is used for treatment purposes and to plan further treatment in terms of radiotherapy or chemotherapy (Figure 22, 23).

<table>
<thead>
<tr>
<th>Figure 21: Endometrial carcinoma histological subtypes</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Type 1</strong></td>
</tr>
<tr>
<td>Endometrioid (75%–80%)</td>
</tr>
<tr>
<td>Ciliated adenocarcinoma</td>
</tr>
<tr>
<td>Secretary adenocarcinoma</td>
</tr>
<tr>
<td>Papillary or villoglandular</td>
</tr>
<tr>
<td>Adenocarcinoma with squamous differentiation. (Adenoacanthoma, Adenosquamous)</td>
</tr>
<tr>
<td>Mucinous (1%)</td>
</tr>
<tr>
<td>Squamous cell (&lt;1%)</td>
</tr>
<tr>
<td>Mixed (10%)</td>
</tr>
<tr>
<td><strong>Type 2</strong></td>
</tr>
<tr>
<td>Uterine papillary serous (&lt;10%)</td>
</tr>
<tr>
<td>Clear cell (4%)</td>
</tr>
<tr>
<td>Poorly differentiated</td>
</tr>
<tr>
<td>Endometriod Grade 3</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Figure 22: Endometrial cancer grade</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Grade 1</strong></td>
</tr>
<tr>
<td>Gland forming tumour with &lt;5% neoplastic cells forming solid sheets</td>
</tr>
<tr>
<td><strong>Grade 2</strong></td>
</tr>
<tr>
<td>5-50% of neoplastic cells form solid sheets</td>
</tr>
<tr>
<td><strong>Grade 3</strong></td>
</tr>
<tr>
<td>&gt; 50% of neoplastic cells form solid sheets</td>
</tr>
</tbody>
</table>

* In tumours showing squamous differentiation, the squamous elements should be excluded from the architectural assessment
* Presence of notable nuclear atypia, inappropriate for the architectural grade should prompt upgrading from grade 1 to grade 2.
1.6.2.5 Management

Treatment options include surgery, radiotherapy and systemic therapy. Surgery can be performed via the open or laparoscopic routes. Low stage and grade, type 1 endometrial cancer, up to stage IA and grade 1 can be treated by hysterectomy, bilateral salpingooopherectomy and trachelectomy. Higher stage or grade cancers are usually treated with additional pelvic and/or paraaortic lymphadenectomy, omentectomy, peritoneal biopsy and peritoneal fluid cytology. Type 2 endometrial cancer regularly spreads to the pelvic and paraaortic lymphatics and therefore lymphadenectomy in these cases is imperative (Thomas et al., 2007). High stage uterine cancers of any type can be treated with surgical cytoreduction followed by adjuvant chemotherapy (Bristow et al., 2001). Radiotherapy, either external or brachytherapy, is also used as adjuvant therapy to treat lymphatic micrometastases.
This treatment modality is added to surgery when stage is more than II and grade is 3, when lymphadenectomy has been performed. Vaginal brachytherapy may also be considered for localised recurrence (Pearcey & Petereit, 2000). Adjuvant chemotherapy has not been shown to provide better outcomes than surgery alone for early stage cancers (Morrow et al., 1990). It is used in a palliative setting where it has been shown to have a small survival benefit (Randall et al., 2006). Progesterone, or progesterone alternating with Tamoxifen have been shown by phase II trials to result in prolonged complete response and are used when performance status does not allow surgical input (Fiorica et al., 2004).

1.6.2.6 Molecular pathology

Type 1 endometrial cancer usually arises on a background of endometrial hyperplasia, and is preceded by atypical hyperplasia, also called endometrial intraepithelial neoplasia. The risk of developing cancer in women diagnosed with endometrial hyperplasia is 1%, 3%, 8% and 28%, respectively, in cases of simple hyperplasia without atypia, complex hyperplasia without atypia, simple hyperplasia with atypia and complex hyperplasia with atypia, respectively (Kurman et al., 1985).

Type 2 endometrial carcinoma is unrelated to oestrogen and may be associated with “serous endometrial intraepithelial carcinoma” arising from atrophic endometrium or polyps.

Different molecular alterations are associated with the 2 types of endometrial malignancy. Type 1 is associated with microsatellite instability and mutations in the PTEN, K-RAS, PIK3CA and CTNNB1 (beta-catenin) genes. Type 2 cancers exhibit
alterations of p53, loss of heterozygosity (LOH) on several chromosomes and less commonly molecular alterations (STK15, p16, E-cadherin and c-erb-B2) (Prat et al., 2007).

*PTEN* loses its function in 50% of endometrioid carcinomas leading to an up-regulation of PI3k/Akt/mTOR pathway of cellular proliferation (Colombo et al., 2013). *PTEN* mutation is associated with 20% of endometrial hyperplasia suggesting it’s involvement in the early stages of carcinogenesis (Latta & Chapman, 2002).

*K-RAS* gene mutations also lead to uncontrolled proliferation through abnormal signaling pathways. They are also associated with endometrial hyperplasia leading to type 1 endometrial cancer [97].

Beta-catenin gene mutations are found in 14 to 44% of type 1 endometrial cancers. These mutations cause alterations in cellular differentiation and signaling pathways (Koul et al., 2002).

*PIK3CA* is a commonly encountered mutation in the *PI3K* (phosphatidylinositol 3-kinase) enzyme that is involved in the regulation of proliferation and cellular metabolism (Suh et al., 2014).

Microsatellite instability is common in women with Lynch syndrome (HNPCC) and therefore associated with type-1 endometrial cancers. It is secondary to MLH-promoter hypermethylation, which causes the accumulation of mutations within these short segments of DNA where some important coding sequences for endometrial carcinogenesis reside (Colombo et al., 2013).

*P53* mutations are commonly associated with type 2 tumours (90%), especially serous carcinomas and confer a poor prognosis. They are present in about 10% of type 1
grade 3 tumours (Tashiro et al., 1997). The P53 pathway can either be activated or deactivated leading to tumour progression (Catasus et al., 2009) (Figure 24).

Loss of heterozygosity is commonly found in type 2 tumours. Chromosomal gains or losses occur due to up regulation of genes such as STK15, BUB1 and CCNB2 that are involved in the mitotic spindle checkpoints and therefore essential for chromosome segregation (Tritz et al., 1997). Amplification of c-erb-B2 (HER-2) and Reduced expression of E-cadherin also cause deregulation of spindle formation causing chromosomal instability.

**Figure 24: Molecular alterations associate with endometrial cancer**

The commonest molecular alterations associated with Type 1 (EEC) and Type 2 (NEEC) endometrial cancers (Matias-Guiu & Prat, 2013).
1.7 Infrared spectroscopy

1.7.1 Background

Spectroscopic techniques are used to identify the chemical consistency of materials. They exploit the fact that materials absorb, emit or scatter electromagnetic radiation at specific frequencies when atomic bonds within them vibrate and resonate. These resonance frequencies are characteristic and representative of changes in dipole moment that depend on atomic mass, bond strength and bond symmetry. They produce a characteristic spectrum of intensity of absorbance, transmittance or scatter, depending on the technique used, against the wavelength or wavenumber (frequency) where resonance occurs. The features of an IR spectrum (number of infrared absorption bands, their intensities and their shapes) are directly related to the molecular structure of a compound. Each bond in these molecules can vibrate in different ways: stretch, scissor, rock, wag or twist. The more complex the molecule the more complex the vibrational model it possesses and therefore the more complex the characteristics of the electromagnetic spectrum produced. Biospectroscopy is the utilisation of vibrational spectroscopy techniques for the study of biological materials. To examine these compounds the mid-IR spectral region (1800cm\(^{-1}\) to 90cm\(^{-1}\)) is used as it provides relevant information about the chemical structures they consist of. This spectral region is termed the “biochemical fingerprint” region. The complexity of biological materials requires specific instrumentation and computation in order to extract sensible and applicable information (Kelly et al., 2011).

There are two commonly used spectroscopic approaches: Fourier-Transform infrared spectroscopy and Raman spectroscopy.
1.7.2 Fourier transform infrared spectroscopy

Infrared spectroscopy was revolutionised by the advent of the “Michelson interferometer” (Figure 25). This instrument has made the procedure faster, more precise and more sensitive. It allows the simultaneous measuring of all relevant infrared frequencies. It uses a beam splitter to divide the infrared beam into two. One beam is reflected on a stationary mirror and the other on a moving one. The beams are made to intersect and their interaction causes interference. The resulting signal is termed the interferogram, which contains information about every infrared frequency contained within the beam emitted from the globar infrared source. The resulting interferogram is internally calibrated against very stable laser light. To interpret the interferogram into a graph where the y-axis represents absorption or transmittance intensity and the x-axis the corresponding wavenumber a Fourier transformation algorithm is used (Figure 26). A sample under examination will reflect or transmit specific electromagnetic frequencies depending on the chemical bonds it contains. Background frequencies, measured without the sample are subtracted. The remaining frequencies undergo Fourier transformation to form the infrared spectrum.

Focal plain array infrared spectroscopy is essentially the same technique but instead of using one sensor to detect the spectral signals a panel of sensors is used. These sensors are usually arranged in a $128 \times 128$ or $256 \times 256$ pattern (CCD). The advantage of this technique is the rapid acquisition of the complete spectrum at each spatial point of the area of sample under investigation. This results in the formation of a “hypercube” which contains information in two spatial dimensions (pseudo-image) and one spectral dimension (the spectrum for each point of that image).
Figure 25: The Michelson interferometer

Figure 26: Conversion of an interferogram to an FTIR spectrum
1.7.3 Synchrotron-based FTIR spectroscopy

Synchrotron-based FTIR is commonly used in transmission mode as this gives more information than transflection. It makes use of a highly collimated, intense beam arising by accelerating electrons using powerful magnets in a circular arrangement and extracting the resulting infrared light by using optical gates (Whelan et al., 2013). This beam is then directed similarly to a bench-top spectrometer onto the sample (Figure 27). The beam characteristics allow for much narrower apertures and therefore a much higher definition of the spectrum with resolutions as small as 2µm. The benefit of this is that chemical bond alterations can be localised within different cellular compartments when processing tissue or bio-fluids (Tobin et al., 2004).

Figure 27: Diagram of a Synchrotron facility set-up
1.7.4 Attenuated Total Reflection Fourier Transform Spectroscopy

Attenuated total reflection Fourier transform spectroscopy (ATR-FTIR) is based on the production of evanescent waves arising from the infrared beam when it is reflected at the interface between a high refractive index diamond crystal and a lower refractive index sample. The evanescent wave penetrates the sample by 0.2µm to 5µm causing a shift in the energy state of the chemical bonds within the sample molecules. This energy change is detected and quantified similarly to transmission or transflection FTIR (Kazarian & Chan, 2013). The advantages of ATR-FTIR include simple instrumentation, relatively short measurement times, sampling of large areas (250µm × 250µm), minimal non-destructive preparation to samples, and minimal use of consumables.
1.7.5 Raman Spectroscopy

Raman spectroscopy is based on inelastic light scattering. This effect was discovered by Raman in 1928 (Raman, 1928). When a monochromatic laser light is directed onto a material it can cause a shift of the system energy into a virtual state. This energy can return to a normal vibrational state causing elastic (Rayleigh) light scattering. If the energy of the resulting photons is altered after this interaction inelastic light scattering (Raman) is produced. Stokes Raman scattering happens when the energy emitted is lower than the incoming photon while anti-Stokes scattering happens when the energy emitted is higher (Figure 28).

Raman spectroscopy is complimentary to FTIR. It can analyse symmetrical molecules as it can polarise the electrons of the symmetrical chemical bonds within them.

For biological samples a laser of 785nm is commonly used as it has the appropriate frequency and wavelength to penetrate the sample enough while achieving high spectral resolution. The resulting spectrum provides a similar biological fingerprint to other methods of biospectroscopy (Wachsmann-Hogiu et al., 2009) (Figure 29). It is unaffected by water, and noes not damage tissue in the right settings and therefore can be used in vivo.

The Raman effect can be very weak especially when analysing biological fluids and large molecules. To enhance the Raman signal, gold or silver nanoparticles are added to the sample (Nie & Emory, 1997). Their close proximity to the molecules has an enhancing effect in the electronic excitations therefore improving the resulting spectrum.
Figure 28: Raman scattering

Virtual energy states

Vibrational energy states

- Infrared absorption
- Rayleigh scattering
- Stokes Raman scattering
- Anti-Stokes Raman scattering

Figure 29: Spectrum derived by Raman scattering

Wavenumber (cm\(^{-1}\))

Intensity/au

Amide I (1650 cm\(^{-1}\))
Protein and lipids (1445 cm\(^{-1}\))
Phenylalanine (1003 cm\(^{-1}\))
Collagen (1314 cm\(^{-1}\))
\(v_{as}PO_2^-\) (1080 cm\(^{-1}\))
Nucleic acids (781 cm\(^{-1}\))
1.7.6 Spectral pre-processing

There are several factors interfering with the spectrum acquired with any of the acquisition methods. These include, the instrumentation itself, the sample preparation and minute changes in background environment. In order to minimise the influence of these parameters on the spectrum they are pre-processed. The type of pre-processing depends not only on the interference that needs to be removed but also the information that is required to be extracted. Spectra from FTIR techniques are usually cut to retain the biochemical fingerprint region only between 1800 and 900 cm\(^{-1}\). The baseline of the spectra is corrected to remove artefacts due to sloping or oscillation caused by infrared dispersion or Mie scattering. These corrections are made by methods such as ‘rubberband’ and “Mie scattering” correction so that the baselines of spectra are of similar intensities (Bassan et al., 2010).

Normalisation aims to correct spectra in the horizontal dimension. This is usually done by scaling spectra to a specific peak (usually Amide I) or by vector normalization according to their individual Euclidian norms. This processing is performed to counteract variability that arises due to differences in sample thickness or concentration.

Raman data are usually cut between 1,750 and 800 cm\(^{-1}\). Background fluorescence is subtracted from the spectra using an automatic baseline correction method (Whittaker filter) (Trevisan et al., 2012).
1.7.7 Multivariate analysis

Spectra extracted from biological specimens are very complex due to the amount of different molecules they contain. This complexity does not allow exploration by univariate analysis but requires multivariate analytical approaches that examine the potential subtle but multiple biochemical discrepancies between the samples compared (Wang & Mizaikoff, 2008).

Principal component analysis (PCA) is one of the most commonly used multivariate analysis approaches. It is important as a biological exploratory tool as its operation can be thought of as revealing the internal structure of the data in a way that best explains the variance within the data. PCA is mathematically defined as an orthogonal linear transformation that transforms data to a new coordinate system such that the greatest variance by some projection of the data comes to lie on the first coordinate (called the first principal component (PC)), the second greatest variance on the second coordinate, and so on. This causes multidimensional data to become reduced to a single point relating to their variability called the “score”. The second score is derived by the same method after all correlation to the first PC has been subtracted from the system. PCA scores, where each axis represents a PC, can be plotted for the most relevant PCs. Scores that co-cluster are considered similar and scores that segregate are dissimilar. Scores plots can be one, two or three-dimensional and use different PC combinations depending on their relevance. Each PC has a corresponding loadings plot that corresponds to the coefficients of the linear combination that formed each PC. These coefficients are derived from a line that passes through the multidimensional mean and minimizes the sum of squares of the distances of the points from that line. PCA is an unsupervised technique and therefore will only detect significant variability if it exists. One of the disadvantages that it has when used on biological samples is
that an excessive number of PCs can reveal arbitrary variability that is not relevant to
the hypothesis.

Linear discriminant analysis (LDA) is a generalization of Fisher’s linear discriminant.
It expresses observations within classes as linear combinations of their features called
linear discriminants (LDs). It is a supervised technique requiring pre-assignment of
the classes to be compared. The resulting combination of characteristics can be used
as a linear classifier to identify intra-class cohesion and inter-class variation. A
loadings vector exists for each LD, which identifies the wavenumbers responsible for
discrimination between classes (Martin et al., 2007). LDA can be internally cross
validated when it can be utilised as a stand-alone technique.

PCA and LDA can be combined to reduce the number of variables to points in
hyperspace and arrange them so that similar ones co-cluster while dissimilar ones
segregate. Loadings plots can then identify the features (wavenumbers) most
responsible for segregation.

1.7.8 Evolutionary algorithms

1.7.8.1 Genetic algorithm

Genetic algorithm is a heuristic that mimics Mendelian genetics. Here the properties
of candidate solutions to a problem are represented similarly to chromosomes making
up the phenotype of an individual. A number of these candidate solutions are selected
and their fitness to solve the problem at hand is tested. The fittest of those candidates
are selected to be tested in the next generation. The properties of the fittest candidates
can be modified by crossover or mutation; to optimise them before entering the next
generation were the process is repeated. The algorithm continues until a satisfactory fitness level has been reached for the population or the maximum number of generations has been produced. In biospectroscopy a number of wavenumbers is chosen that best classify the population. Successive generations form different and more relevant wavenumber combinations, which better solve the problem of classification. Once sufficient classification is reached the algorithm stops and its output represents the most relevant wavenumbers.

1.7.8.2 Sequential progressions algorithm

This algorithm utilises a bottom-up search procedure were a number of features is added to an initial one until a final feature set is reached. Each number of features has to be “fit for purpose” i.e. be able to help with the problem solution. The initial feature set is termed the null set. Each additional feature set added is aimed to maximise the selection criteria. When adding new features makes the feature set worse fitting than the previous one the algorithm terminates. If that doesn't happen a maximum set of features is set by a cost versus function algorithm. SPA can be cross-validated at every stage. Its main disadvantage is that it does not include a mechanism for excluding features from the feature sets after they have been added (Andrew R. Webb, 2011).
1.7.9 The role of biospectroscopy in screening and diagnosis of cancer

The molecular alterations that are involved in cancer development contain associated chemical bond alterations. These alterations can be exploited using biospectroscopic techniques to attempt classification of subject populations for screening or diagnostic purposes. The same methods have the ability to extrapolate the types of molecules most responsible for classification therefore identifying potential biomarkers. These biomarkers can be validated by other technologies and can be applied in clinical settings using simple spectroscopic or other instrumentation.

1.7.9.1 Gastrointestinal cancers

Vibrational spectroscopy technologies have been used to study areas of the gastrointestinal tract including the colon, stomach and oesophagus.

*In vitro* studies have shown that Raman spectroscopy can differentiate between normal mucosa, metaplastic and adenomatous polyps and adenocarcinomas arising in the colonic epithelium with sensitivities and specificities of more than 90% (Mahadevan-Jansen & Richards-Kortum, 1996). Similar studies have been performed *in vivo* with the help of fibre-optics through a colonoscopic port to achieve accuracies of 95% (Molckovsky et al., 2003). Also, FTIR micro-spectroscopy was used to differentiate normal from adenomatous polyps and malignant cells for *in vitro* colonic tissue with a percentage success of 89%, 81% and 83% respectively. In this study the major discriminating factors were associated with spectral differences in the phospholipid, phosphate stretching and DNA/RNA bands (Argov et al., 2002), which could prove to be valuable biomarkers within a potential screening test. Mackanos et al performed a similar study intending to classify colonic mucosa as normal, hyperplastic or dysplastic (Mackanos et al., 2009). This group demonstrated correct
classification with a sensitivity of 96%, specificity of 92%, accuracy of 93% and predictive value of 82%.

Efforts have also been made to identify rapidly dividing cells within the colonic epithelial crypts where carcinogenesis may occur using synchrotron based FTIR. Walsh et al (Walsh et al., 2009) identified positions where stem cells may reside within bowel crypts and their relationship with nearby epithelial cells. The proposed stem cells are located at the base of the crypts and exhibit genetic changes associated with colon cancer (Van De Wetering et al., 2002). These cells where differentiated from surrounding cells by wavelengths corresponding to protein phosphorylation (970cm\(^{-1}\)) and symmetric (1080cm\(^{-1}\)) and asymmetric (1250cm\(^{-1}\)) phosphate stretching. These DNA conformational changes may be important markers for identifying cells with a central role in carcinogenic processes.

Barrett’s oesophagus is a premalignant condition defined by dysplasia of the oesophageal epithelium. Raman techniques have successfully been used to identify biomarkers that discriminate normal and inflamed epithelium from dysplastic and cancerous epithelium in rats, both in vitro and in vivo (Boere et al., 2003; Hattori et al., 2007). Mazia et al has demonstrated the ability of FTIR techniques to identify cellular molecular changes that differentiate normal from cancerous cells in the oesophagus (Maziak et al., 2007).

The potential of both Raman and FTIR technologies for the identification of molecular changes associated with evolution of gastric cancer has also been studied (Li et al., 2005)
1.7.9.2 Urological cancer

Vibrational spectroscopy techniques have been used in the past to identify biomarkers that discriminate bladder inflammation from bladder cancer. Crow et al were able to classify benign and malignant bladder tissue with 84% accuracy. They used Raman fibre-optics to examine 29 snap-frozen bladder tissue samples and identified differences in amide I and II, C-C stretch and CH2 bend and twist (Crow et al., 2005). A similar study using FTIR on samples from 10 patients detected obvious spectroscopic discriminatory changes in the proteins (1650, 1550 cm\(^{-1}\)), lipids (2925, 2850 cm\(^{-1}\)) and nucleic acids (1080, 1236 cm\(^{-1}\)) (Nafie A. Al- Muslet, 2011). Biomarkers of commitment to cancer may reside within these spectral differences as biochemical changes happen before morphological changes.

Prostatic cancer has received substantial attention due to its prevalence and the inadequacies of current proposed screening biomarkers (Mistry & Cable, 2003). Crow et al recorded Raman spectra from 38 snap-frozen prostate samples from trans-urethral resections. They found that Raman fibre-optic techniques are able to distinguish benign prostate samples from malignant ones with 86% accuracy (173). A recent study by Patel at al compared benign prostate samples derived from two separate demographic cohorts (India and United Kingdom) with different incidence of invasive prostatic cancer by both Raman and infrared techniques. This study has shown that biochemical variations within prostatic DNA and protein structure, which may account for differences in cancer risk between the two cohorts, are identifiable by both techniques in association with multivariate analysis (Patel et al., 2011). Future similar studies may identify similar biomarkers within a more homogenous population that may be used a part of screening programs.
1.7.9.3 Gynaecological cancers:

Cervical cancer screening technologies have been studied extensively because of the availability of an effective treatment in its pre-invasive stage (Quinn et al., 1999). Current screening programs are effective in reducing mortality from cervical cancer (Peto et al., 2004) but suffer from a low specificity leading to over-treatment (Nanda et al., 2000). FTIR and Raman approaches have proved to be valuable for the exploration of potential screening biomarkers of cervical dysplasia (Walsh et al., 2007). A recent study by Gajjar et al has demonstrated that ATR-FTIR spectroscopy identified cytological atypia more consistently than conventional cytological examination when both techniques were correlated to histological diagnosis (Gajjar et al., 2014). This may be due to the fact that biochemical alterations happen before morphological changes. The spectroscopic variations responsible for identifying different grades of atypia may potentially provide biomarkers that can identify cervical pre-malignant disease even earlier and with higher specificity than current tests.

Ovarian cancer is of special interest as the five-year survival associated with it remains at around 30% and no effective screening tool has so far been identified. Current ovarian screening has several limitations including very low positive predictive value of combined Ca125 and ovarian ultrasonic picture for early stage disease (Menon et al., 2009). Both Raman and FTIR approaches have been used to extract biomarkers to discriminate normal and cancerous ovarian tissue (Krishna et al., 2007; Mehrotra et al., 2010). Krishna et al identified differences between malignant and benign ovarian tissue within the DNA, lipid, amide I and amide II spectral regions using FTIR. When using a Raman approach they identified changes within similar bands including amide I and III. They also noticed that ovarian tissue with
benign disease had separate spectroscopic differences within different wavelengths involving polysaccharide and protein bands (Krishna et al., 2007).

Endometrial cancer has an excellent prognosis when identified and treated in the early stages. Current screening involves high risk and symptomatic individuals and incorporates trans-vaginal ultrasound testing, hysteroscopy and endometrial biopsy. This is time and resource consuming, intrusive and disturbing to the patients. The discovery of biomarkers that could be used as part of a population-screening program would alleviate these issues and probably reduce the mortality associated with the disease even further. A study of 76 patients undergoing hysterectomy identified spectral biomarkers able to distinguish between benign endometrium and endometrioid and non-endometrioid cancer. The major discriminating wavenumbers included amide I and II, symmetric and asymmetric PO$_2$- stretching, carbohydrate, glycogen and protein phosphorylation bands (Taylor et al., 2011).

A more recent study by Gajjar et al interrogated serum and plasma using ATR-FTIR coupled with a classification machine to discover potential biomarkers for the diagnosis of ovarian and endometrial cancer with very promising accuracies. Extraction of biomarkers using blood products can be used effectively and is easily translated into practice. Also classification machines may be of value in cases where one classification method does not provide robust answers.

Vibrational spectroscopy has the ability to be used for cancer diagnosis and screening over conventional approaches. The resulting data can be used to form hypothetical initiating carcinogenic processes. Specific events within these processes can be extracted as biomarkers of disease and be incorporated successfully into clinical practice.
1.8 Aims and objectives

The main focus of this thesis is to evaluate the potential of biospectroscopy as a group of novel techniques for screening and diagnosis of cancer. In addition, the associated methods, ATR-FTIR, synchrotron-based FTIR and Raman spectroscopy, in conjunction with suitable analytical techniques have been utilised to investigate potential differences between benign tissues which may in turn harbour putative biomarkers of increased risk of cancer. The same techniques have been used in an effort to identify stem cell locations to be targeted for diagnosis and treatment.

The objectives of this thesis are therefore the following:

• To explore the role that biospectroscopy can play in a “systems biology” setting for the exploration of carcinogenic processes and the identification of related putative biomarkers.

• To investigate the use of biospectroscopy methods in discriminating between benign, borderline and malignant ovarian tumours as well as differentiating the subtypes of ovarian carcinomas.

• To identify similarities and differences that may exist within the histological compartments of mammary glands of women of similar ages and with advancing age. Also to identify the potential locations of stem cells within the terminal ductal lobular units.

• To investigate potential trans-generational differences in benign human prostatic tissue harvested from similarly aged men. To further evaluate those differences utilising immunohistochemical techniques.
References


Martin F L, German M J, Wit E, Fearn T, Ragavan N & Pollock H M. (2007). Identifying variables responsible for clustering in discriminant analysis of data from


Chapter 2

Extracting biomarkers of commitment to cancer development: potential role of vibrational spectroscopy in systems biology

Georgios Theophilou, Maria Paraskevaidi, Kássio M.G. Lima, Maria Kyrgiou, Pierre L. Martin-Hirsch, Francis L Martin

Expert Reviews in Molecular Diagnostics Early Online, 1-21 (2015)

Contribution:
I performed the literature review.
I wrote the first draft of the article

Professor F. L Martin  G. Theophilou
Extracting biomarkers of commitment to cancer development: potential role of vibrational spectroscopy in systems biology


Georgios Theophilou¹,², Maria Paraskevadi³, Kássio MG Lima¹,⁴, Maria Kyrgiou⁵, Pierre L Martin-Hirsch¹,² and Francis L Martin*¹
¹Centre for Biophotonics, LEC, Lancaster University, Lancaster LA1 4YQ, UK
²Department of Obstetrics and Gynaecology, Central Lancashire Teaching Hospitals NHS Foundation Trust, Preston, UK
³Department of Chemistry, University of Ioannina, Ioannina, Greece
⁴Institute of Chemistry, Biological Chemistry and Chemometrics, Federal University of Rio Grande do Norte, Natal, 59072-970, RN, Brazil
⁵Department of Surgery and Cancer, Institute of Reproductive and Developmental Biology, Faculty of Medicine, Imperial College, London W12 0NN, UK
*Author for correspondence: Tel.: +44 (0) 152 451 0206 f.martin@lancaster.ac.uk

The complex processes driving cancer have so far impeded the discovery of dichotomous biomarkers associated with its initiation and progression. Reductionist approaches utilizing ‘omics’ technologies have met some success in identifying molecular alterations associated with carcinogenesis. Systems biology is an emerging science that combines high-throughput investigation techniques to define the dynamic interplay between regulatory biological systems in response to internal and external cues. Vibrational spectroscopy has the potential to play an integral role within systems biology research approaches. It is capable of examining global models of carcinogenesis by scrutinizing chemical bond alterations within molecules. The application of infrared or Raman spectroscopic approaches coupled with computational analysis under the systems biology umbrella can assist the transition of biomarker research from the molecular level to the system level. The comprehensive representation of carcinogenesis as a multilevel biological process will inevitably revolutionize cancer-related healthcare by personalizing risk prediction and prevention.

Keywords: carcinogenesis • computational techniques for data fusion • omics technologies • screening biomarkers • systems biology • vibrational spectroscopy

Biomarkers for cancer screening

Recent technological advances in high throughput, ‘omics’ technologies have opened new horizons in the search for novel screening biomarkers for cancer. The choice of biomarkers is ever expanding, and selecting the most suitable ones involves considering all of the prerequisites that a screening program has to fulfill, known as the Wilson Criteria (Box 1) [1]. The usefulness of a particular biomarker lies in its ability to provide early indication of presence or progression of disease [2]. Five phases of biomarker development have been proposed, each aiming to examine particular desirable attributes of the test without jeopardizing safety or wasting resources (Box 1) [3]. Several biomarkers have been proposed for potential use in clinical practice, but very few have been relevant. In fact out of approximately 1500 biomarkers identified annually, only some 30 will be incorporated into clinical practice [2]. Examples of such biomarkers are cancer antigen 125 (CA125), approved for use in the prediction of ovarian cancer risk in postmenopausal women with ovarian cysts and prostate specific antigen (PSA), used in the prediction of prostate cancer. These biomarkers have been used as a primary screening tool; unfortunately, they have not resulted in a reduction in mortality when used in this setting [2].

The most pivotal attribute of an ideal biomarker for use in cancer screening is its ability to accurately identify the earliest change that defines commitment to disease. Such a biomarker would distinguish those individuals who will eventually get cancer if left untreated from those that will never get cancer despite their being at increased risk. In order to produce such a biomarker, the lifecycle of cancer...
Box 1. Wilson criteria (1968).

- The condition should be an important health problem
- There should be a treatment for the condition
- Facilities for diagnosis and treatment should be available
- There should be a latent stage of the disease
- There should be a test or examination for the condition
- The test should be acceptable to the population
- The natural history of the disease should be adequately understood
- There should be an agreed policy on whom to treat
- The total cost of finding a case should be economically balanced in relation to medical expenditure as a whole
- Case finding should be a continuing process and not a ‘once and for all’ project

Needs to be understood. So far, following ‘reductionist’ approaches to search for suitable biomarkers within the genome, epigenome, transcriptome, metabolome and topome, as well as vibrational spectroscopy, have proven imperfect. Systems biology is a promising, novel scientific field that integrates information derived from different high-throughput technologies to understand models of biological mechanisms [4]. These mechanisms may harbor potential cancer biomarkers, which once identified may be applied within population screening programs.

In this article, we discuss current technologies for extracting biomarkers of early disease for use in potential screening programs for cancer. We comment on their role within a ‘systems biology’ research context. We argue that vibrational spectroscopy has the potential to play a central part in this approach. Finally, we suggest that systems biology approaches involving biospectroscopy can be applied in the search for screening biomarkers to be employed in population screening tests.

Carcinogenesis

Carcinogenesis is a multistep process during which phenotypically normal cells are transformed into cancer cells. This transformation involves multiple genetic and epigenetic changes whose diversity accounts for the vast heterogeneity that exists between different types of cancer [5]. The definition of these processes is currently the focus of extensive research, as recognizing potential common processes involved in the initiation of most or all cancers, would assist with the identification of cancer biomarkers. Therefore, knowledge of the suggested models that govern carcinogenesis is a vital component of a structured approach in the search for potential screening biomarkers.

One of the most widely accepted models of carcinogenesis was suggested by Hanahan and Weinberg [6]. Their model proposes that at least six essential alterations in cell physiology dictate malignant growth: self-sufficiency in growth signals, insensitivity to growth-inhibitory signals, evasion of cell death, limitless replicative potential, sustained angiogenesis, tissue invasion and metastasis [6]. Later added to this model were energy metabolism re-programming and immune destruction resistance [7]. They stipulate that these concepts are applicable to the development of most cancers and that studying them may assist in simplification of the diagnosis and treatment of cancer (Figure 2) [67].

It has also been proposed that the initiation step for carcinogenesis requires the presence of incipient cancer cells or cancer stem cells, which become tumorigenic and ultimately malignant, as they acquire specific traits [68]. It has been hypothesized that these cells possess a ‘mutator phenotype’, which arises from random DNA damaging events causing these cells to be genetically unstable [9]. This genetic instability, given the right circumstances, may progress through an ever-increasing mutation sequence toward the evolution of cancer [71]. Favorable conditions for cancer initiation may be caused by exogenous carcinogens including ionizing radiation, oncogenic microorganisms and chemicals [12]. These can activate extracellular or intracellular signals that promote unrestricted cellular proliferation involving a cascade of events leading to malignancy [13–15]. The driving forces for these events are genetic, epigenetic and cellular alterations that disrupt the normal balance between proliferation, differentiation, senescence, DNA repair.

Figure 1. The five phases of biomarker development.

ROC: Receiver operating characteristic.
Data taken from [3] by permission of Oxford University Press [3].

doi: 10.1586/14737159.2015.1028372
and apoptosis [26,27]. Epigenetic changes may involve reversible addition of methyl groups to DNA bases or acetyl groups to histone proteins. Also micro-RNAs may bind to the untranslated regions of mRNA to alter their gene expression modulation function. Further downstream, protein stability may be affected by folding/unfolding, conformational changes, hydrophobic interactions, hydrogen bonding, redox potential, pH, temperature and metal binding [9]. Another significant factor is the recruitment of nearby cells that actively participate in the development and expression of malignant characteristics forming a tumor microenvironment. The interplay between cancer cells and normal stromal cells supports the proliferation of cancer cells by disturbing the equilibrium between cellular proliferation and destruction [18]. All these processes lead to the expression of proteins, growth factors, cytokines, immune suppressors and others, which may be detectable by a variety of tests [14,16,17].

Some of the molecular features of these alterations have already been identified as potential cancer biomarkers and are currently in use. The same processes provide a huge field into which biomarker research can expand with the advent and use of novel high-throughput technologies in a systems biology setting. So far, we have been unable to determine the committing steps required for the initiation of cancer development. The reason may be the significant emphasis given to reductionist, pathway-centric research approaches. Looking for these steps in one carcinogenesis pathway in isolation is incomplete due to the extensive crosstalk among cell regulatory pathways [19]. These pathways are interconnected, and any change in one will directly or indirectly affect the others [20,21]. It is therefore logical to consider an approach that attempts integration of as many pathways as possible to determine the pathophysiological processes that are associated with the early steps toward cancer to identify potential screening biomarkers.

**Systems biology**

As discussed, cancer is a highly heterogeneous process involving several events that disrupt normal cellular functions. The progression to cancer is therefore a dynamic process involving a network of promoting signals that may arise from genetic perturbations and environmental cues and affect any part of the biological system including the genome, epigenome, transcriptome, proteome and metabolome. In order to understand the fundamental processes driving carcinogenesis, one may consider a ‘modular approach’, where, in addition to studying some of its components individually, the behavior and function of the corresponding network are studied as a whole [19,22]. Systems biology is a scientific discipline that endeavors to quantify all the functional elements of a biological system. It integrates information concerning their interactions to explain emerging behaviors [23]. The complex interplay of cellular activities associated with carcinogenesis requires such a global approach. An understanding of just one element of this system is not sufficient. Identifying all the different elements of the system each in isolation is also deficient as it is not the individual components within the system that may drive carcinogenesis but their complex interactions [24]. Integrating this data into network models may generate detailed models of cellular regulation for different environmental situations [25,26]. This in turn allows the formation of coherent hypotheses for carcinogenesis. It is possible that the relevant abundances within a biological system of DNA, RNA, proteins, metabolites and other chemicals and the interaction between their different combinations may provide an opportunity to identify suitable biomarkers for cancer screening.

Systems biology approaches typically involve three steps [27]: the generation of global data after a particular perturbation, the integration of resulting data into network models and the generation of experimentally testable hypotheses based on the integrated data. These approaches aim to identify mechanisms, by which interactions modulate tissue properties and how these interactions change relative to different environmental stimuli [19,24]. Bioinformatics science provides a conceptual framework for the integration of the steps above to enable a global perspective of biological functions. It requires the unification of biology, statistics and information technology into one discipline. Algorithms have been developed to identify...
cellular components and define their functions and interactions according to endogenous and exogenous influences. These include gene finding, motif recognition, similarity searches, multiple sequence alignment, protein structure prediction and use differential equations to develop dynamic models of biological processes, enabling precise predictions of phenotypes based on quantitative information.

These complex mathematical methodologies are collectively termed computational techniques for data fusion. Applying computational techniques for data fusion to a combination of data sources aims to provide more knowledge about the investigated system than what would have been obtained by analyzing the individual data sources separately. The choice of computational technique or combination thereof depends on the data source, the original analysis method and data storage system. The scope of the investigation also dictates the methodology choices. These may include unsupervised learning, learning algorithms, quality assessment, biomarker discovery, classification models and feature selection models.

Unsupervised learning methods, such as clustering, organize data into structures that may be useful for developing genomics studies. Cluster analysis is an exploratory data analysis tool, which aims to classify objects so that their association is maximal if they belong to the same category and minimal if not. Different phylogenetic analysis among others.

Several databases and simulation tools now exist to integrate data from diverse sources to attempt modeling of biological processes. These global data can be organized and presented in syntactic, semantic and schematic modes and stored. Several such facilities exist, for example Pathway Interaction database (NCI Nature Curated), Reactome, Biocarta, KEGG or Protein Lounge. These data collections are available for organization and reduction to deliver useful information based on the questions asked. Several integration models have been suggested and are available to use for example: Human Pathway Database, Pathlist, PathSys, Integrated Pathway analysis pathway, Human annotated and predicted protein interactions and KEGG converter. Integration models utilize mathematical approaches. For example, Boolean networks combine network analysis with logic models to provide links between molecular interactions and cellular phenotype. Another approach is to

Figure 3. Multidisciplinary model of cancer initiation. Carcinogenesis is a complex process involving diverse linked structural and functional alterations. Systems biology has the potential to integrate data from high-throughput methodologies to extract cancer screening biomarkers.

Figure 4. Biomarker extraction for cancer screening can be likened to the ‘needle in the haystack problem’. In this case, each haystack represents a different aspect of cellular biophysics. The field where the haystacks rest links them together so they become a system and are affected equally by environmental alterations. The human body can be considered as such a system affected by environmental changes that promote changes in any or all of its compartments. These changes can be examined under the ‘systems biology’ umbrella.
clustering techniques achieve classification by the application of various similarity measures and grouping algorithms. Resultant data configuration depends on the technique used and can be very variable. Therefore clustering is used to generate hypotheses about data structure rather than to test such hypotheses. Cluster analysis is the principal analytical tool for data fusion based on individual unsupervised studies. Available software implementations include K-means clustering [33], hierarchical clustering [34], two-way clustering [35] and self-organizing map [36].

Data fusion in systems biology can be applied for biomarker selection [37]. To achieve this, holistic measurement methods simultaneously generate results for many variables. Biomarker discovery involves the identification of an optimal subset of variables that significantly differentiates separate classes and can be used for accurate prediction of the class membership. Selection and identification of these biomarkers is essential for obtaining an understanding of the complex biological processes under study. In the context of biospectroscopic analysis, a biomarker involves the 'most important' wavenumber variables used in the modeling process [38]. The search for a multivariate biomarker can be seen as an optimization problem in high-dimensional space. Useful techniques for this approach are regularized discriminant analysis, principal component analysis-linear discriminant analysis (PCA-LDA) [39], partial least squares-discriminant analysis [40] and the related variable importance of projection values [41].

One of the inherent issues with such multivariate approaches is selecting a suitable threshold to achieve class discrimination without the introduction of arbitrary separation. To solve this problem, feature selection techniques are applied. These are heuristic processes or algorithms that identify optimal subsets of variables that can significantly separate the differentiated populations [42].

Some well-established classification systems include discriminant analysis [43], support vector machines [44], random forests [45], k-nearest neighbors [46] and artificial neural networks [47]. These represent the newest and most promising trends in 'omics' sciences. LDA, for example, LDA-linear and QDA-quadratic, represent classical parametric methods that should be in the portfolio of any data miner and bioinformatician. Integrating data from different high-throughput technologies is not without challenges. The methods of integration may require extensive adjustment of dataset parameters. The estimation of these parameters may be difficult, as in these situations there are no gold standards to measure against. Omics technologies require different experimental conditions, which may affect the responses identified within the biological system, making parameter adjustment even more complex. Also, each technology presents inherent confounding factors that produce systematic biases (e.g., mass spectrometry [MS] favors the identification of highly abundant proteins, see below). These biases may be introduced and amplified by dataset fusion causing prediction errors. Moreover high-throughput techniques have inherently high false-positive and false-negative rates [48]. Integration of datasets with different statistical types, sizes and powers may be difficult to validate internally or externally. The limited availability of quantitative information for molecular and cellular processes that have been proved to be involved in carcinogenesis inevitably reduces the accuracy of such integrated biological models.

Despite these challenges, global understanding of the biological processes involved in carcinogenesis may lead to better appreciation of the interplay between genotype and environmental influences toward overall cancer risk definition. This bioinformatics approach to systems biology may ultimately provide algorithms that act as screening biomarkers that assign individuals to simple categories according to risk. The advances of high-throughput technologies including genomics, epigenomics, transcriptomics, proteomics, metabolomics and toponomics are central to the contact of systems biology studies. These approaches have identified putative biomarkers of cancer, some of which are currently in use with varying
success (Table 1). We discuss these technologies and propose a novel technique that may be valuable as part of a systems biology setup.

Genomics

Genomics can be defined as the generation and analysis of individual genes or genomes to determine their structure or function by the application of recombinant DNA and DNA sequencing methods [46,50]. The application of genomics to cancer biology has enabled the identification and characterization of several individual genes and patterns of gene expression that are able to distinguish malignant and premalignant cells from their normal counterparts [31]. Mutations associated with cancer initiation or propagation may include aneuploidies, chromosomal translocations, amplifications or deletions as well as single nucleotide polymorphisms [52,53]. The completion of projects such as the Human Genome Project, Celera and ENCODE have facilitated the isolation of specific genes involved in genetic predisposition to cancer as well as spontaneous polymorphisms associated with an increased risk of cancer. Current genomics technologies enable simultaneous evaluation of gene expression alterations that may represent commitment within the processes that lead to malignancy. These alterations have been utilized by different technologies to identify screening biomarkers for cancer.

The advent of robotics has allowed the manufacture of chip technology arrays, able to process thousands of cDNA sequences rapidly. These chips are able to detect differences in gene expression profiles between different cells and cells placed in different conditions [54]. The same chips assisted in the identification of insertions or deletions in tumor suppressor genes [55]. This function has been a valuable tool in the study of carcinogenesis and the detection of potential biomarkers of cancer. Several arrays are currently in use. An example is Genechip® produced by Affymetrix [56]. This uses synthetic oligonucleotides as probes, which are produced by a combination of photolithography and light-directed solid-phase DNA synthesis directly onto glass wafer. These oligonucleotides are complimentary to genes on the array and allow the measurement of particular gene expression profiles with high sensitivity. This array is unable to predict and analyze unknown genes, as their sequences need to be known for the manufacture of the

<table>
<thead>
<tr>
<th>Table 1. Putative biomarkers for cancer identified by ‘omics’ technologies.</th>
<th>Prognostics</th>
<th>Epigenomics</th>
<th>Transcriptomics</th>
<th>Proteomics</th>
<th>Metabolomics</th>
<th>Toponomics</th>
</tr>
</thead>
<tbody>
<tr>
<td>Prostate</td>
<td>PSA, PSMA</td>
<td>EZH2, FOXA1, AS, ZHX2 mutations</td>
<td>BCP, BCRA1, BCRA2 mutations</td>
<td>MPT1, PSMA, AIP1, and EPCAM</td>
<td>HIF-2α, CPT1A, HIF-1α, and CDH1</td>
<td>FOXM1, CDH2, and RHOA</td>
</tr>
<tr>
<td>Ovarian</td>
<td>BRCA1, BRCA2 mutations</td>
<td>G-ClnG island hypermethylation (BCRCA)</td>
<td>E-cadherin, E-cadherin receptor</td>
<td>BCRCA</td>
<td>RASSF1A, SIS2, and E-cadherin methylation</td>
<td>RASSF1A, SIS2, and E-cadherin methylation</td>
</tr>
<tr>
<td>Breast</td>
<td>BRCA1, BRCA2 mutations</td>
<td>G-ClnG island hypermethylation (BCRCA)</td>
<td>E-cadherin, E-cadherin receptor</td>
<td>BCRCA</td>
<td>RASSF1A, SIS2, and E-cadherin methylation</td>
<td>RASSF1A, SIS2, and E-cadherin methylation</td>
</tr>
<tr>
<td>Pancreatic</td>
<td>Paclitaxel</td>
<td>G-ClnG island hypermethylation (BCRCA)</td>
<td>E-cadherin, E-cadherin receptor</td>
<td>BCRCA</td>
<td>RASSF1A, SIS2, and E-cadherin methylation</td>
<td>RASSF1A, SIS2, and E-cadherin methylation</td>
</tr>
<tr>
<td>Gastric</td>
<td>PTEN, PI3K</td>
<td>G-ClnG island hypermethylation (BCRCA)</td>
<td>E-cadherin, E-cadherin receptor</td>
<td>BCRCA</td>
<td>RASSF1A, SIS2, and E-cadherin methylation</td>
<td>RASSF1A, SIS2, and E-cadherin methylation</td>
</tr>
<tr>
<td>Hepatocellular</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Colorectal</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
oligonucleotide probes. Also, it lacks sensitivity for transcripts of low abundance [57].

Serial analysis gene expression is an alternative technique that is equally sensitive but can also analyze gene expression within uncharacterized genomes [58]. It achieves this by cloning small tags of the cDNA, which are then cleaved and sequenced allowing the analysis of multiple transcripts. The amount of time the cloned cDNA is observed corresponds to the genetic abundance for that particular gene. This enables serial analysis gene expression to measure the expression of thousands of genes quantitatively without prior knowledge of the gene sequence. This feature can be used in the search for cancer biomarkers by detecting differentially expressed genetic patterns [59].

Both microarray technologies have facilitated the identification of several alterations in gene expression that are in use as cancer biomarkers. HER2 alterations, 17q23 genomic amplifications and cycloxygenase-2 protein expression are associated with breast cancer screening and prognosis [60]. Myc and AIB1 protein expression have been used for the prediction of hepatocellular carcinoma [61] and insulin-like growth factor-binding protein expression alterations for the prediction of prostatic carcinoma [62].

Next-generation sequencing (NGS) technology has revolutionized the extraction of genetic information from biological systems. It enables rapid sequencing of entire genomes by sequentially identifying signals emitted by parallel re-synthesizing of a massive number of DNA fragments. These DNA fragments are re-assembled to reveal the entire genomic sequence of a DNA sample. NGS has enabled the rapid and relatively inexpensive sequencing of entire genomes allowing fast hypothesis testing as well as the establishment of massive genomic databases [63]. Cataloguing mutations in multiple cancer types using NGS is likely to play a vital role in the compilation of datasets for integration into a 'systems biology' setup.

Although several projects have assemled genes associated with increase in cancer risk, their function within perturbed regulatory networks has not been established. Such an endeavor would require the availability of software packages that integrate heterogeneous data involving thousands of genes for simultaneous analysis. This system should not only be able to cope with high dimensional data but also provide different (meta) views on the data that therefore have to be cross-linked [64]. For this purpose, several integrative bioinformatics software have been proposed, for example BRIDGEP [65], which is based on three web-based applications. Although these software advances have not yet been extensively applied in cancer research, they have the capability to facilitate the analysis of the genome in relation to biological process variations.

Epigenomics

Epigenetic changes are functionally relevant alterations in gene transcriptional activity, which occur in the absence of DNA sequence changes. These alterations may be long term and heritable through mitotic generations. It is now accepted that it is the accumulation and functional cooperation between genetic and epigenetic changes that drives carcinogenesis rather than accumulating mutations alone [66,67]. It has even been postulated that epigenetic changes might occur before genetic mutations, causing genetic instability responsible for the formation of the 'Mutator Phenotype' mentioned earlier [68,69]. Epigenetic processes relevant to cancer risk and tumor formation include chromatin re-modeling, DNA methylation, gene imprinting, RNA interference or RNA-directed epigenetic activation and silence, phosphorylation, ubiquitination, sumoylation, ribosylation and ATP-modifying proteins [70-72]. They cause genomic instability, triggering the activation of the silent allele of an oncogene or the silencing of a tumor suppressor gene [73].

Epigenomics is the global analysis of epigenetic modifications. It aims to identify the complete set of epigenetic changes across the entire genome. Similarly to genomics, epigenomics relies on high-throughput technologies and bioinformatics to identify putative disease-associated biomarkers [73,74]. The most studied epigenetic alterations are DNA methylation and histone modification, which are the alterations most associated with carcinogenesis due to their involvement in tumor suppressor gene inactivation and oncogene activation, respectively [75]. Clearly a much broader range of genes are affected by epigenetic changes [76].

One of the mainstream technologies used in epigenomics is methylated DNA immunoprecipitation. It involves a modified DNA microarray technology that uses monoclonal antibody-labeled probes to identify areas of methylation (5-methylcytosine) within the DNA sequence [77,78]. Methylated DNA immunoprecipitation can be coupled with sequencing technologies to analyze methylation within the genome [81]. This technology has been used in the past to search for screening biomarkers for prostate, colon [82] and breast cancer [83]. The probes used in these arrays are predefined and as such carry the risk of 'missing' exact methylation sites that can be important in cancer initiation.

Bisulfite genomic sequencing has a far higher resolution down to a single 5-methylcytosine. Bisulfite causes only unmethylated cytosine to be converted into uracil. DNA treated with bisulfite can then be amplified by a PCR technique that amplifies 5-methylcytosine-containing sequences allowing the identification of DNA regions harboring hypermethylation [84].

Histone modifications are important epigenetic changes that can also be associated with cancer initiation [85]. The main modifications involve acetylation, methylation and phosphorylation of their N-terminal regions. Monoclonal antibodies have been used to identify specific histone modification sites within the DNA sequence using PCR and immunofluorescence coupled to gene array technologies such as ChIP [86,87]. Although this method was sensitive, it was plagued with cross-reactivity between different modification sites and the requirement for a specific reagent for every different modification. MS is the preferred technique as it allows the quantification of standalone alterations or combinations of histone modifications [88]. It utilizes charge-to-mass ratios to identify specific molecular fingerprints. A complication of this technique is the fact proteins
have to be enzymatically digested (usually by trypsin) before the employment of MS. This approach may cause unwanted modifications resulting in erroneous findings. Top-down approaches may have the ability to process intact histones [89].

Epigenomics technologies have been able to identify putative biomarkers for the early detection of prostate and breast cancer [90], colorectal and endometrial cancer [91] among others. Similar approaches have been used to explore the activation/inactivation of oncogenes associated with the anti-neoplastic activity of tamoxifen in breast and its carcinogenic effect in uterine tissue [92]. As a systems biology tool for biomarker research, epigenomics has the potential to identify most or all of the epigenetic changes that may be associated with carcinogenesis. Mapping the epigenome may deliver a more rounded view of pre-translational modifications based on environmental stimuli. The cellular network interaction patterns that result from these relationships may be used to provide a better understanding of carcinogenesis lifecycle and the identification of biomarkers of cancer initiation to be used in screening programs.

Transcriptomics

Transcriptomics is the study of the ‘complete complement of RNA generated by a cell or population of cells’ [93]. It aims to interpret the functional elements of the genome by determining transcriptional and post-transcriptional modifications under different conditions. The transcriptome encompasses different types of RNA including mRNA, non-coding RNA and small RNA. The genome is relatively stable when compared to the transcriptome. Gene expression changes by transcription in response to different cellular functions and environmental conditions. This is particularly important when investigating the committing step to carcinogenesis as it may be related to a particular pattern of gene expression [94]. Studying global transcriptome variations such as alternative splicing, novel transcript and gene fusion provides information not only on the actively transcribing parts of the genome but also on the control of the transcription process. Micro-RNA, for example, is not translated but controls the expression of target mRNA; thus it influences protein translation and defines cellular processes that may be related with carcinogenesis [95]. Since a single micro-RNA can bind to 100 different target transcripts, it has been estimated that micro-RNAs may be able to regulate up to 30% of the protein-coding genes in the human genome. Micro-RNA, in association with changes in transcription, may have a role in biomarker discovery and has already been associated with several cancers [96-97]. In addition to micro-RNA, long non-coding RNAs (lncRNA) regulate gene transcription, either locally (cis-regulation) or distally (trans-regulation). These lncRNAs are either involved in maintaining a consistent gene expression (imprinting lncRNA) or facilitate epigenetic regulation of gene expression through mRNA processing and translation. Dysregulation of lncRNA expression is associated with carcinogenesis and is cancer and tissue specific [98]. Unfortunately to date, it is unclear to what extent lncRNAs are responsible for gene expression alterations. They are involved in disruption of cellular processes and therefore become prime candidates for cancer biomarker development.

Various technologies are in use to characterize the transcriptome including specialized microarray techniques that utilize fluorescence hybridization and sequencing approaches [94]. Hybridization techniques are relatively inexpensive and provide high resolution but rely on existing genome knowledge and are limited by complicated normalization methods and cross-hybridization [99]. RNA sequencing is able to both determine and quantify transcriptomes without prior knowledge of the genome sequence in most cases. Similar to genome sequencing, RNA is converted to cDNA fragments with adaptors attached to their ends. These are sequenced to produce 30–400 bp (i.e., base pairs) and then aligned to either to reference genomes or transcripts [94]. RNA sequencing can also determine transcripts that do not correspond to an existing genome. It can identify the precise position of introns or exons therefore revealing not only the transcription sequence between them but also the association between adjacent regions. This technique has very high resolution and is reproducible [100]. It also has limitations that include the multistep production of cDNA libraries and the creation of biases by fragmenting and reorienting the investigated transcripts [101].

Current transcriptomics techniques provide a large amount of complex data that unfortunately lack documentation of the associated experimental and biological variations [102]. Systems biology hypothesizes are based on genetic and environmental influences that may alter genomic signals whose downstream effects regulate the transcriptome for protein production. Controlling experimental conditions coupled with the ability to visualize the transcriptome in its complexity can provide powerful information that can be integrated with other omics approaches to study biological systems.

Proteomics

Both genomics and transcriptomics technologies are capable of rapidly identifying gene and mRNA alterations that may be associated with cancer initiation or progression. These alterations do not always correlate with downstream changes in protein expression, which directly affect cellular function [103]. Macromolecular interactions such as protein–protein, protein–DNA and protein–RNA interactions are crucial for most biological processes. Cellular changes that occur with carcinogenesis may cause post-translational protein modifications in their quantity, activity, interaction and localization, so-called ‘proteomic fingerprints’. These can be considered as fundamental members of systems biology, making the proteome a prime candidate in the search potential cancer screening biomarkers [104,105].

Recent advances in MS have made possible the development of high-throughput methods used in the search for such biomarkers [106]. 2D gel electrophoresis represents the foundation of proteomic approaches [107]. It separates proteins by the sequential use of two different electrophoretic techniques based on two
different protein properties. One technique separates proteins according to their isoelectric point and the other according to their molecular weights. Representation of these differences on perpendicular axes provides a template for visual comparison between proteins. Proteins involved in changes associated with malignant transformation can then be identified by MS [108]. This technique has been used extensively to identify putative biomarkers for breast, prostate, bladder and liver cancer among others [108-112]. It is an effective technique in terms of resolution but has low sensitivity and is labor intensive. Also, it cannot resolve proteins with extremes in molecular weight with the potential of overlooking smaller protein biomarkers.

Matrix-assisted laser desorption/ionization in conjunction with MS is another technique used to detect proteins and peptide fragments. This is also a multiplex procedure. It involves trypsin digestion of proteins followed by separation of the resultant peptides by reverse-phase high performance liquid chromatography (HPLC). The separated peptides are analyzed by MS and once identified they are matched to potential parent proteins by searching pre-populated databases. Matrix-assisted laser-desorption/ionization-MS has been used to identify a variety of potential screening biomarkers for breast, prostate and other cancers using mainly biofluids including, blood, urine, saliva and nipple aspirate [113-118]. The high-throughput and relative simplicity of this technique has made it so attractive that it has been extensively used over the past two decades, although it is not without shortcomings. These include difficulties with sample collection and pre-processing standardization, which affect reproducibility of the results [119].

Surface-enhanced laser desorption/ionization, time-of-flight (SELDI-TOF) with tandem MS is an approach that aims to overcome the requirement for purification and separation of proteins prior to MS analysis [120]. It combines chromatography with MS to provide rapid protein expression profiles from a variety of biological and clinical samples [121]. It utilizes protein capture and separation on a solid-phase protein chip surface by exploiting adsorption, partition, electrostatic interaction and affinity chromatography [121]. These proteins can then be ionized and identified by measuring their weight using MS. The main advantage of SELDI-TOF MS is that it uses an array format for high-throughput analysis requiring small amounts of biological sample with little preparation [122]. Its main limitation is that its resolution and therefore sensitivity reduces dramatically for larger peptides. It is, though, more sensitive than the previously discussed techniques [123]. Moreover, this technology can be used to identify protein-protein and protein-DNA interactions enabling the detection of post-translationally modified proteins that may be involved in cellular signaling pathways or other hallmarks of carcinogenesis [123].

As many as 20,000 proteins have been estimated to populate human serum alone [124]. Single proteins or protein panels that are either uniquely or differentially expressed when studying different perturbations secure a place for proteomics within systems biology [104,125]. The integration of proteomic data for systems biology analysis is supported by analytical software. For example Integrative Proteomics Data Analysis Pipeline (IPDAP) [109] is based on two computational proteomics platforms, namely Computational Proteomics Laboratory Database (CPAS), Systems Biology Analysis Management System (SBEAMS) [126] among other tools. The integration of these analytic approaches allows the evaluation of proteomic variability as a consequence of cellular network operations under the influence of genetic and environmental signals.

Metabolomics

Metabolomics science studies changes in cellular metabolism presented as alterations in cellular phenotype. End products of cellular metabolism are assessed in the context of genetic regulation, enzymatic activity as well as the immediate surrounding microenvironment [127,128]. This allows the assessment of specific metabolic pathways that may be true indicators of the processes related to the initiation or progression of cancer [127,129]. Specific metabolic requirements of cells involved in deranged proliferation and apoptosis can also be used to evaluate upstream events to allow the discovery of genetic alterations associated with carcinogenesis [130]. As cancer development is a uniquely heterogeneous process, identification of specific metabolites, metabolite panels or metabolite quantities may be applied in the research for screening biomarkers [131,132]. The two most accepted methods for metabolomic detection and quantification are nuclear magnetic resonance (NMR) spectroscopy and MS [133]. NMR exploits the behavior of molecules when placed in a magnetic field, allowing the identification of different nuclei based on their resonant frequency. MS, as discussed earlier, determines the composition of molecules based on the mass-to-charge ratio in charged particles. The resultant metabolite detection and quantification is acquired as a dataset called a spectrum [134,135]. Both techniques can be adapted to examine tissue or biofluid samples for the simultaneous detection of hundreds of endogenous metabolites that may potentially be used as cancer biomarkers [136,137]. Due to the high quantity and complexity of the data acquired, these techniques are used in combination with informatics based multivariate statistics. Similarities and differences between compared groups can therefore be evaluated [138-140]. The responsible metabolites can then be identified, by inputting the data acquired into large databases, like the Human Metabolome Database [141]. Using these methods, the cancer metabolome is beginning to be characterized. For example, NMR-based studies have identified over 30 metabolites, including elevated choline containing compounds, in breast tissue associated with breast cancer prediction or diagnosis with considerable reliability when compared with histology [142-144]. Differentiation of benign tissue from cancer tissue in situ is also possible using MRI-related techniques like magnetic resonance spectroscopy imaging (MRSI) for the detection of choline compounds, with excellent sensitivity and specificity [145]. Similarly, the detection of elevated spermine and reduced citrate levels in prostatic fluid using NMR correlates with the presence of cancer [146,147]. NMR and MS studies on serum samples have also been shown
to distinguish benign ovarian disease from borderline and cancerous tumors in both pre- and postmenopausal women [148,149]. Metabolomics techniques have proved to be valuable in other types of cancer as well including brain colon and hepatocellular carcinoma [154,155].

Organizing metabolomics data to input in a systems biology format requires reproducible purification of the metabolites and identification of their chemical constituents and quantities for different environmental settings. This is essential as diet, smoking or even exercise can alter metabolic pathways. Also, handling of samples after harvesting requires specific techniques to arrest metabolism until analysis [156]. Classification of these alterations depending on experimental variations and tested hypotheses need the employment of chemometric methods. Such tools have already started to appear, for example, ‘metabolomics spectral formatting, alignment and conversion tools’ by Sunner et al. [155] and ‘high speed peak alignment algorithm’ by Johnson et al. [156].

Toponomics

Toponomics is a recently developed discipline that studies cellular functions by observing the location of specific protein networks within morphologically intact cells [157]. It has been suggested that the position of a particular protein within its microenvironment influences its function [157]. For example, a protein close to the plasma membrane may be associated with receptor activation while the same protein close to the nucleus may be associated with the transport of transcription factors into the nucleus [158,159]. This method can be used to explore the hierarchical position occupied by a protein that is implicated in the process of disease [160]. Topological protein alterations may be involved in carcinogenesis. Abnormally differentiating cells may rearrange their protein network architecture in response to upstream signals, simultaneously signaling for the activation of further carcinogenic cellular processes. The toponome has been marked as the ‘next generation’ in systems biology research as it can be cell and disease type specific [161]. Multi-epitope-ligand cartography (MELC) is the most usual method of ‘mapping’ protein associations within cells or tissues. MELC is able to associate spatial and temporal protein colocalization and/or exclusion with pathological processes therefore revealing their functionality [162]. MELC is an automated multidimensional fluorescence imaging technology that can explore the toponome in a quantitative manner. Using fluorescent dye-labeled tags (monoclonal antibodies), directed against specific proteins, it can locate up to 100 proteins in both cells and tissues [163,164]. The fluorescence images associated with each tag can be superimposed on each other to represent the presence or absence of particular proteins within each image pixel and therefore compose a protein matrix termed the ‘combinational molecular phenotype (CMP)’ [165,166]. A CMP may contain particular biomarkers, for example, CD4. CMPs are summarized into CMP-motifs, which are binarized for computerized analysis. The CMP-motifs are applied to the whole image to identify particular protein-cluster frequency in the different parts of the visual field. The CMP-motifs of greatest interest can be color coded and superimposed on the corresponding biological structures to create toponome maps [162]. Toponome maps of sample groups can be compared to identify potential differences. The toponome has been used to describe protein topology and function in an attempt to identify biomarkers associated with bowel cancer [167] and prostate cancer [168] among others. MELC use is limited by the availability of fluorescent dye-labeled tags, which bind to certain protein complexes. These tags have to be highly calibrated to prevent interference, for example, sterical hindrance where a tag is incorrectly directed to a neighboring protein complex [169]. Also, this is a lengthy technique, with each cycle taking more than 1 hr with the risk of either antibody tag denaturation or protein complex disintegration [162]. Its capability to resolve protein networks in a spatiotemporal fashion can be applied synergistically with other omics technologies to identify models of carcinogenesis and deduce relevant screening biomarkers.

Biospectroscopy

Biospectroscopy encompasses all sciences that utilize optical techniques to study tissues, cells and biological molecules. These techniques range from macroscopic approaches, for example, diffuse optical imaging and tomography, to microscopic approaches, for example, confocal and fluorescence microscopy. Vibrational spectroscopy is one of the most useful branches of biophotonics when it comes to biomarker research as it has the ability to acquire information beyond cellular or subcellular architectural structure; it has the ability to extrapolate the molecular or even elemental makeup of a sample under investigation [170]. Biospectroscopy can be divided into four major branches based on the technologies used to obtain information from biological tissues: Raman, infrared (IR), fluorescence and X-ray spectroscopy. The most commonly currently used technologies for biomarker research is Raman and IR spectroscopy.

There are several reasons for which biospectroscopy can become an integral part of a systems biology approach in conjunction with the technologies described earlier in this document. A search for a biomarker or panel of biomarkers that may be associated with a particular event or process that may occur as part of one of the hypothesized carcinogenesis routes using any single omic approach would be incomplete [171]. Complex biomolecules such as proteins, lipids, carbohydrates and nucleic acids leave distinct vibrational signatures that reflect molecular structure and conformation and can be picked up by spectroscopic approaches [172-174]. Therefore, vibrational spectroscopy can be utilized to identify and quantify changes within these cellular constituents. Spectroscopic data may be integrated with data from other technologies to form an omics data compendium. The study of the relationships and interactions between the different parts of this compendium may lead to a better understanding of carcinogenesis as a global process.

This section of the review will explore the advantages and disadvantages of vibrational spectroscopy applications. We will discuss currently used biospectroscopic instrumentation. Finally,
we will give examples revealing the potential of such technologies to be integrated into a systems biology approach for cancer biomarker research.

**Biospectroscopy techniques**

**IR spectroscopy**

IR spectroscopy utilizes the fact that when chemical bonds between atoms within molecules are excited due to exposure to IR radiation, they alter their vibration status. This vibration alteration happens at specific energy levels and is associated with the absorption of IR radiation at specific wavelengths. The resulting absorption spectrum is specific to the chemical bonds present within a particular sample. The region within the IR spectrum that provides most spectral information about cellular molecules is the mid-IR region ($\approx 4000$–$400$ cm$^{-1}$). This region contains most information on the cellular chemical constituents termed the "cellular biochemical fingerprint" ($\approx 1800$–$900$ cm$^{-1}$). This region of the IR spectrum is known to be absorbed by a number of biochemicals, which correspond to identifiable peak absorption frequencies, including Amide I (1650 cm$^{-1}$), Amide II (1550 cm$^{-1}$), Amide III (1260 cm$^{-1}$), carbohydrates (1155 cm$^{-1}$), glycerol (1030 cm$^{-1}$), lipids (1750 cm$^{-1}$), asymmetric phosphate stretching vibrations ($v_{\text{as}}$PO$_2^{–}$; 1225 cm$^{-1}$), symmetric phosphate stretching vibrations ($v_{\text{sym}}$PO$_2^{–}$; 1080 cm$^{-1}$), and protein phosphorylation (970 cm$^{-1}$). Alterations within the different regions of the IR spectrum may correspond to biochemical composition changes and provide biomarker information. For example, carbohydrate changes may be associated with hormonal influences while phosphate stretching may be associated with DNA changes. A combination of these changes may be detectable by IR spectroscopy and may represent microenvironmental influences causing mutations as part of a potential carcinogenic process.

**Fourier-transform infrared spectroscopy in transmission or transfection mode**

Fourier-transform infrared (FTIR) spectroscopy uses a silicon carbide thermal mid-IR source (Global) (Section 6A). The IR from this source is channeled through a Michelson interferometer. The modulated IR beam reaches the sample under examination causing the chemical bonds between the present molecules to vibrate. These vibrations cause changes in the dipole moments of the molecules at specific IR wavelengths. The resulting interferogram undergoes an optical inverse Fourier transformation. The degree of IR absorption at different wavelengths by the different molecules within the sample can be quantified [175,176].

**Attenuated total reflection FTIR spectroscopy**

Attenuated total reflection FTIR (ATR-FTIR) spectroscopy is based on the attenuation effect of light when it is reflected at the interface between a high refractive index material (e.g., diamond) and an IR absorbing low refractive index material (biological sample) (Section 6B). Therefore, in this case, the sample has to be in contact with the crystal (Section 6C). While the IR wave is reflected at the contact interface, an evanescent IR wave penetrates the sample by 0.2–5 μm and is absorbed by the sample detected as a specific spectrum depending on the chemical composition of the sample [177,178].

**Synchrotron-based FTIR spectroscopy**

Synchrotron-based FTIR spectroscopy is a superior tool to conventional FTIR spectroscopy for the detection of molecular biomarkers within cells. The reason for this is that a synchrotron emits a collimated light beam that is much more brilliant than that of a bench top spectrometer allowing the fast acquisition of spectra at diffraction-limited apertures [179]. This provides an excellent signal to noise ratio that is up to 1000-times greater than that of conventional IR spectrometry and allows spatial resolutions as small as 2 μm [180,181]. To produce a mid-IR radiation beam, electrons accelerate and enter a circular path by the use of strong electromagnets. This produces a broad electromagnetic spectrum of which different wavelength bands can be isolated and directed to a specific beam-line for use in spectroscopy [182].

**FTIR spectroscopy in cancer research**

FTIR spectroscopy has the potential to detect biochemical changes that probably contain the committing steps that occur during cancer initiation and precede any histological or cytological morphological changes. Box 2 gives some examples of some cancer biomarkers identified by FTIR spectroscopy. Several studies have exploited FTIR to identify putative screening or diagnostic cancer biomarkers using biofluids or tissues. For example, a UK pilot study has demonstrated that blood plasma- or serum-based ATR-FTIR spectroscopy with a classification ensemble has the potential to discriminate ovarian cancers from controls [183]. Another study revealed the potential of ATR-FTIR in segregating benign from malignant brain tumors identifying relevant biomarkers at the same time [184]. A third study has demonstrated that ATR-FTIR spectroscopy of cervical liquid-based cytology is better in identifying underlying atypia or disease than conventional screening when compared with histology [185].

Synchrotron-based FTIR spectroscopy has made possible the identification of subcellular biomolecular information with a high signal to noise ratio [182]. This enables the localization, identification and tracking of specific chemical events within an individual living cell [186,187]. Imaging at subcellular resolution may enable the localization of stem cells from which potential carcinogenic processes may be initiated [186].

The advent of portable and handheld FTIR instruments using globar sources has facilitated the examination of tissues in vivo. Fiber optic technology is relatively inexpensive and can be adapted easily to facilitate examination of different areas of the body. For example, it has been used to detect colonic dysplasia by attaching an FTIR probe to a colonoscope [189]. The portable, non-destructive, painless nature of such probes and their ease of use qualify their application in studies to identify temporal spectroscopic changes that can be used as part of a systems biology approach.
Raman spectroscopy

Raman spectroscopy (Figure 6B) is not limited to detecting molecules encompassing polar vibrational modes. It can analyze molecules containing symmetrically vibrating chemical bonds, for example, double and triple carbon–carbon bonds in aromatic hydrocarbons. Its function is based on inelastic light scattering [190]. A laser beam, of for instance 785 nm, can be used to excite molecules from their ground state to a virtual energy state using energy ranging from 200 to 4000 cm⁻¹. The molecules return to their original state by releasing photons. In about 1 of 10⁷ of these energy transfers, a molecule does not return to its original resting state causing an emitted photon to have to alter its energy level to keep the energy content within the system constant. This energy shift is termed ‘Stokes shift’ when the photon enters a lower frequency level and ‘anti-Stokes shift’ when the photon enters a higher frequency level. The change of frequency levels results in ‘inelastic light scattering’, This is detected and analysed by specially designed software to produce a spectrum containing the intensities of the wavelengths where inelastic scattering occurred.

The Raman signal can also be enhanced using add-on techniques. Surface-enhanced Raman spectroscopy commonly uses

Figure 6. Vibrational spectroscopy techniques. These can be used as part of a systems biology setup to understand carcinogenic processes and extract biomarkers of cancer. (A) Fourier-transform infrared (FTIR) spectroscopy. The IR source can be globar or synchrotron-based. (B) Attenuated total reflection FTIR spectroscopy. (C) Representation of the diamond, which comes into close proximity with the sample allowing the emergence of evanescent waves that cause chemical bond vibrations. (D) Raman spectroscopy setup.

Box 2. Fourier-transform infrared biomarkers of carcinogenesis.

- DNA structural alterations (hydrogen bonding of nucleic acids)
- Reduced glycan level
- Increased hydrogen bonding of phosphodiester groups of nucleic acids
- Decreased hydrogen bonding in C–OH groups of proteins
- Increased hypomethylation
- Disorder of methylene chains in membrane lipids
gold or silver nanoparticles to increase the electrical potentials that are in close proximity to the investigated molecules therefore intensifying the measured signals by up to $10^{11}$ [191]. Resonance Raman spectroscopy uses a light source that is ‘tuned’ into the electronic absorption band. When such energy is applied at the excitation wavelengths, it causes resonance therefore greatly enhancing the excited electronic state. This enhances the Raman signal and is useful for studying large molecules [192]. Raman can be used in conjunction with IR spectroscopy to cover each other’s weaknesses, which may appear as weak bands in either the IR or Raman spectrum [193].

Raman spectroscopy in cancer research

Most biological molecules are Raman active and have specific biological fingerprints [193-195]. Raman can therefore provide an excellent tool for the detection of subtle spectral changes that may be involved with the early stages of carcinogenesis independent on whether these changes involve nucleotides, proteins or other by-products. Also the concentration of specific molecular constituents can be calculated from the intensity of the relative wavelength peak [190]. This can be important as it allows correlation of the amount of biochemical change to the stage of carcinogenesis. Raman spectroscopy, similarly to other vibrational spectroscopy techniques, can be performed both in vitro and in vivo [195]. In vitro, it can be used to analyze fresh unfixed and unstained tissue as well as tissue embedded in paraffin after minimal pre-processing by de-waxing and drying. This allows the use of both historical and recently collected samples from a diverse group of patients with cancer diagnosis to be compared with control samples from healthy patients. The sample required for Raman spectroscopy is only as big as the laser source footprint so small biopsies including fine needle aspirates can be processed [196]. Raman probes using fiber optics are also currently being developed allowing the use of Raman spectroscopy in vivo [197-199].

This along with the non-destructive nature of the Raman beam [200] enables the conduction of longitudinal studies to identify environmental influences that may affect the network of interactions between the different components of a biological system.

Raman is also excellent for analyzing fluids due to the weak Raman scatter caused by water [201]. It has been shown that biofluids may contain information on carcinogenic processes [202-205]. They also require minimal processing before undergoing Raman or surface-enhanced Raman spectroscopy, and their analysis can be applied on a large and automated scale. Preliminary studies have suggested that Raman spectroscopy can be used to identify screening biomarkers. For example, when screening high-risk patients with liver cirrhosis, it distinguishes patients with hepatocellular carcinoma from those without the disease with 91% accuracy [206]. Another similar study has suggested that Raman spectroscopy coupled with chemometric techniques can be used for real-time non-invasive diagnosis of cervical cancer from serum samples [207].

One of the main difficulties with Raman is that ambient light spectra can be superimposed over Raman spectra. It is therefore prudent for Raman measurements to be taken in a dark environment to avoid unwanted noise spectra. Also tissues exhibit autofluorescence, which can also mask Raman spectra. This issue is more obvious when using Raman to examine tissues in vivo, where the higher depth of radiation and consequently larger sample volume causes increased fluorescence. Significant fluorescence can be avoided by using near IR laser light between 70 and 900 nm for molecular excitation [208]. Laser at 785 nm is therefore commonly used. Another issue encountered when using fiber-optic probes for in vivo investigations is the collection of Raman signals generated by the walls of the fibers themselves. Several methods have been proposed to alleviate this including different materials for the fiber optics, lens and filter systems and modulating the wavelength of the laser beam [208], fluorescence and ambient light suppression by modulated wavelength spectroscopy. The high absorption of the collimated laser energy by tissues may cause local destruction if applied for longer lengths of time but current portable probe systems have the ability to acquire Raman spectra within seconds avoiding pain and related tissue damage.

Both FTIR and Raman spectroscopy approaches have been employed for the identification of potential biomarkers for cancer (Figure 7). We propose a potential role of objective optical techniques as part of a systems biology approach for biomarker research.

Expert commentary

As discussed earlier, carcinogenesis involves several interconnected and interdependent processes. Understanding these processes requires the integration of all relevant molecular, cellular and bioinformatics knowledge [209]. The availability of bioinformatics technologies has allowed the compilation of vast amounts of data that can be intercalated and examined in unison to appreciate the cooperative events leading to cancers and thus identify early biomarkers of disease [20]. Vibrational spectroscopy data from different sources can easily be integrated with currently available datasets to form biological systems models. Several resources are currently available containing experimental data obtained by high-throughput omics technologies [20]. Several ways of integrating these databases have been proposed that range from the assembly of various biological data sources into a single database, termed ‘data warehousing’, to the creation of web links to relevant databases, termed ‘link integration’ [20]. Vibrational spectroscopy data lend themselves to the creation of such integrated databases as they possess a well-defined nomenclature with each wavenumber being associated with specific biochemical fingerprints [202,203]. Datasets concerning vibrational spectroscopy that have been obtained by different labs may be compared, validated and integrated to establish extensive databases for use within a systems biology structure. One issue with inter-laboratory data integration is the standardization of analytical techniques used to classify potential biochemical changes associated with different perturbations. Data derived by such consistent methods can be used to form hypothetical initiating carcinogenic processes. Specific events within these processes can be extracted as biomarkers of
disease. Identifying biomarkers related to specific initiating events related to carcinogenesis can be very difficult due to the innate heterogeneity of the disease and the population it affects. The establishment of the promising field of systems biology aims to solve any interoperability problems related to the integration of the related omics technologies and would therefore provide the infrastructure for the standardization of biospectroscopic data within this environment.

**Five-year view**

Systems biology has the potential to revolutionize cancer screening and diagnosis. We now have better understanding of the complexity underlying cancer initiation and progression. Cancer arises as a result of multiple genetic and environmental signals, which perturb normal cellular networking structures resulting in multiple biochemical alterations [23]. Defining ways of studying these alterations on a global level is realized through the integration of high-throughput ‘omics’ technologies [212,215]. Vibrational spectroscopy is such a technology and therefore has the potential to be utilized as an integral part of systems biology research. Although this approach faces many challenges with regard to complexities of data handling [214,215], it also has significant potential benefits as it can be adapted to identify biomolecules involved in, or are the consequence of carcinogenesis. The chemometric analysis methods used in vibrational spectroscopy are similar to the analysis used for other omics technologies. Therefore, data generated is compatible and potentially available for immediate integration with existing ‘systems biology’ data pools. This information can be collated with existing or developing data pools from other technologies to create ‘global’ knowledge foundations that may reinforce hypotheses of carcinogenesis, identify relevant functional biomarkers and

doi: 10.1586/f14737159.2015.1028372
ultimately recognize pathways of risk classification for the development of population screening tests.

Vibrational spectroscopy techniques are undergoing intensive research at the moment as their value to biomedical applications has been realized over the past two decades. The adaptation of its capabilities within systems biology can assist the exploration of early carcinogenic processes and the identification of potential biomarkers to be used in predictive medicine. Such biomarkers may revolutionize health care by moving from reactionary diagnostics and therapeutics to personalized preventative measures to fight cancer in its very early preclinical stages.

**Acknowledgements**

The authors were supported by the Lancashire Teaching Hospitals NHS Trust and the Rosemere Cancer Foundation.

**Financial & competing interests disclosure**

The authors have no relevant affiliations or financial involvement with any organization or entity in a financial conflict of interest in the subject matter or materials discussed in the manuscript. This includes employment, consultancies, honouraria, stock ownership or options, expert testimony, grants or patents received or pending, or royalties.

**Key issues**

- Population screening program development for cancer is difficult due to the heterogeneity of the population screened and the variability of the carcinogenesis process.
- Carcinogenesis is a complex process. Biomarkers for carcinogenesis are more useful when they identify very early functional or phenotypical changes that may be associated with committing steps in carcinogenesis.
- Systems biology is a science that aims to integrate omics technologies derived data to understand complex biological mechanisms.
- Computational data fusion techniques are used to integrate data derived from different omics technologies (genomics, epigenomics, transcriptomics, proteomics, metabolomics, toponomics and potentially biophotonics) to test specific hypotheses.
- Integration of omics datasets is important in cancer biomarker mining due to the complexity of disease initiation involving genetics, environmental influences and intercellular signaling.
- Vibrational spectroscopy techniques detect molecular alterations by identifying changes within their chemical bonds. The main technologies used are Fourier-transform infrared spectroscopy and Raman spectroscopy.
- Large amounts of data produced are analyzed using similar chemometric methods to omics technologies. This allows the use of existing data fusion methods to integrate vibrational technology datasets in a systems biology environment.
- Validation of these techniques in isolation and in conjunction with each other is paramount for the development of robust cancer biomarkers for their use in potential population screening programs.

**References**

Papers of special note have been highlighted as:

- of interest
- of considerable interest

4. Proposes a systematic biomarker development process for cancer screening.
11. Provides an outline of carcinogenic processes.
22. Hanahan D, Campos LM. Accessory to the crime: functions of cells recruited to the...
39. One of the first studies to use biospectroscopy to examine cellular metabolism.
52. Outlines technologies that have been applied in the analysis of the genome, transcriptome and proteome in the context of carcinogenesis.

doi:10.1586/14737159.2015.1028372


• Identifies the importance of epigenetic processes for the initiation of neoplasia and outlines how this can be assessed.


87. Hebbel TR, Thorne AW, Crane-Robinson C. A direct link between core histone acetylation and transcriptionally active chromatin. EMBO J 1988(75): 1295-402


• The first paper hypothesizing a transgenerational increase in hormone-driven cancers arising from environmental influences.


131. Ghude K, Sarkova NJ. Therapeutic targets and biomarkers identified in cancer choline phospholipid metabolism. Pharmacoepigenomics 2006;7(7):1109-23


Vibrational spectroscopy & biomarkers of cancer development

Review

120


- Demonstrates the potential of infrared spectroscopy to classify different types of cells when coupled to computational analysis.
177. Kazarin SG, Chan KL. Applications of ATR-FITR spectroscopic imaging to biomedical samples. Biochim Biophys Acta 2000;1578(7):858-864
178. Kazarin SG, Chan KL. ATR-FITR spectroscopic imaging: recent advances and applications to biological systems. Analyst 2013;138(7):1940-51
- Demonstrates that infrared spectroscopy has the potential to identify putative stem cell regions. These stem cells may play a vital role in carcinogenesis.

DOI: 10.1586/14737159.2015.1028372


Chapter 3

Attenuated total reflection infrared spectroscopy coupled with chemometric analysis discriminates normal, borderline and malignant ovarian tissue and classifies subtypes of human ovarian cancer

Georgios Theophilou, Kássio M.G. Lima, Pierre L. Martin Hirsch, Helen F. Stringfellow, Francis L. Martin

Contribution:
I designed the study.
I retrieved, pre-processed and processed the ovarian tissue samples by ATR-FTIR.
I performed a portion of the data analysis.
I have written the first draft of the manuscript.

Professor F. L Martin                      G. Theophilou

……………………………………………………………………………………………………
Attenuated total reflection infrared spectroscopy coupled with chemometric analysis discriminates normal, borderline and malignant ovarian tissue and classifies subtypes of human ovarian cancer

Georgios Theofilou¹, ², Kássio M.G. Lima¹, ³, Pierre L. Martin Hirsch¹, ², Helen F. Stringfellow², Francis L. Martin¹*

¹Centre for Biophotonics, LEC, Lancaster University, Lancaster LA1 4YQ, UK,
²Department of Obstetrics and Gynaecology, Central Lancashire Teaching Hospitals NHS Foundation Trust, Preston, UK,
³Institute of Chemistry, Biological Chemistry and Chemometrics, Federal University of Rio Grande do Norte, Natal 59072-970, RN-Brazil

*Corresponding author: email: f.martin@lancaster.ac.uk; phone: +44 (0)1524 510206; fax: +44 (0)1524 510217
Abstract

Surgical management of ovarian tumours largely depends on their histo-pathological diagnosis. Currently, screening for ovarian malignancy with tumour markers in conjunction with radiological investigations has a low specificity for discriminating benign from malignant tumours. Also, pre-operative biopsy of ovarian masses increases the risk of intra-peritoneal dissemination of malignancy. Intra-operative frozen section, although sufficiently accurate in differentiating tumours according to their histological type, increases operation times. This results in increased surgery-related risks to the patient and additional burden to resource allocation.

We set out to determine whether attenuated total reflection Fourier-transform infrared (ATR-FTIR) spectroscopy, combined with multivariate analysis can be applied to discriminate between normal, borderline and malignant ovarian tumours and classify ovarian carcinoma subtypes according to the unique spectral signatures of their molecular composition.

Formalin fixed, paraffin-embedded ovarian tissue blocks were de-waxed, mounted on Low-E slides and desiccated before being analysed using ATR-FTIR spectroscopy. Chemometric analysis in the form of principal component analysis (PCA), successive projection algorithm (SPA) and genetic algorithm (GA), followed by linear discriminant analysis (LDA) of the obtained spectra revealed clear segregation between benign versus borderline versus malignant tumours as well as segregation between different histological tumour subtypes, when these approaches are used in combination.

ATR-FTIR spectroscopy coupled with chemometric analysis has the potential to provide a novel diagnostic approach in the accurate diagnosis of ovarian tumours assisting surgical decision making to avoid under-treatment or over-treatment, with minimal impact to the patient.
**Keywords:** ATR-FTIR spectroscopy; Biospectroscopy; Classification analysis; Ovarian cancer; Wavenumber selection
Introduction

Ovarian cancer is the fifth most common gynaecological cancer (incidence of 18 per 100,000) and the fourth most common cause of cancer death (mortality of 8.8 per 100,00) in women in the UK (Cancer Research Uk, 2011). In 2009, 5,900 women were diagnosed with ovarian cancer; 3,500 died from the disease the year after. The high related mortality is a consequence of late presentation and diagnosis at stage III or IV resulting in five-year survival rates of 20% and 6% respectively (Cancer Research Uk, 2011).

Ovarian cancer refers to a heterogeneous group of tumours, as indicated by differences in epidemiological and genetic risk factors, precursor lesions, patterns of spread, molecular events during oncogenesis, response to chemotherapy and prognosis (Prat, 2012). 90% of ovarian cancers are malignant epithelial tumours termed carcinomas, the remainder being germ cell and sex cord-stromal tumours (Lee Kr, 2003). The commonest types of ovarian carcinomas are high-grade serous carcinoma (HGSC), low-grade serous carcinoma (LGSC), mucinous carcinoma (MC), endometrioid carcinoma (EC), clear cell carcinoma (CCC), carcinosarcoma (CS) and mixed tumours (MT). Ovarian carcinomas are graded according to their cellular differentiation from normal tissue. This does not apply to HGSC and LGSC as they are considered different entities (Prat, 2012). Borderline epithelial tumours comprise approximately 15% of epithelial ovarian tumours and have a good prognosis. With the implementation of the “international classification of diseases for oncology (ICD-O-3)”, these tumours are no longer considered malignant (Cadron et al., 2006).

The complexity and heterogeneity of ovarian cancer with regards to risk factors, precursor lesions, morphological and clinical manifestations has hindered the development of robust population-based screening programs. Currently, in the UK, the assessment for ovarian cancer is based on “Risk Malignancy Index” (RMI 1 or 2), which encompasses menopausal
status, ultrasonographic ovarian presentation and blood levels of the tumour marker Ca125 (see E.S.I. Table S1) (Jacobs et al., 1990; Tingulstad et al., 1996). Other blood derived biomarkers with similar accuracy to RMI have been suggested, for example HE4 and ROMA, but have not been established in practice (Karlsen et al., 2012). Also, magnetic resonance imaging (MRI), Doppler ultrasound and computed tomography (CT) have shown accuracies of 80% for the diagnosis of malignancy and 80-90% for the detection of abdominal spread (Brown et al., 1994; Buist et al., 1994; Outwater et al., 1996).

Prognostic factors for ovarian cancer include stage and grade of cancer at diagnosis and residual disease after primary staging surgery (Brun et al., 2000). Histopathological tumour type is important when considering personalised treatment options. This plays a major role in chemotherapy responsiveness and therefore in overall survival rate (Brun et al., 2000; Sugiyama et al., 2000). For example HGSC demonstrates much better response to platinum based chemotherapy than CCC. This results in CCC having a lower 5-year survival than HGSC (Du Bois et al., 2003).

It is obvious that current methods of ovarian cancer diagnosis and management have significant limitations. This is therefore a field that can benefit from research to identify novel methods of detecting and categorizing ovarian cancer to aid personalized intra- and post-operative management while also minimizing patient risk and resource expenditure.

Vibrational spectroscopy is a bio-analytical tool that has the potential to classify normal and pathological tissue according to their chemical and molecular differences (Martin F L, 2011) Related techniques including Fourier-transform infrared (FTIR) and Raman spectroscopy have been utilised in the past few years to detect structural alterations that occur in molecules within cells according to their chemical bonds. (Baker et al., 2014; H. Y. Holman et al., 2000; Jackson et al., 1997; Mourant et al., 2003; Taillandier & Liquier, 1992) The
resulting spectral differences may be used to distinguish benign from cancerous processes and classify cancer subtypes. Examples of areas studied by these methods include breast (Kelly et al., 2011), endometrial (Taylor et al., 2011), cervical (Gajjar et al., 2014), prostatic (Baker et al., 2009) and brain cancers (Gajjar et al., 2012).

We utilised ATR-FTIR to interrogate ovarian tissue harvested from women undergoing oophorectomies for several reasons including pelvic pain, postmenopausal bleeding, menorrhagia or dysfunctional uterine bleeding, premenstrual tension, risk reduction due to breast cancer or positive family history and imaging revealing ovarian cysts/ masses. We hypothesized that interrogation of ovarian tissue with ATR-FTIR spectroscopy will allow diagnostic segregation of benign, borderline and cancerous tumours. Additionally, this method will allow classification of epithelial ovarian cancer subtypes. The complex resulting spectral datasets were analysed using multivariate analysis in the form of principal component analysis followed by linear discriminant analysis (PCA-LDA) and variable selection techniques in the form of successive projection algorithm (SPA) and genetic analysis (GA) again followed by linear discriminant analysis (SPA-LDA, GA-LDA). These chemometric techniques intended in reducing the complexity of the spectral datasets and allowing visual representation. They were combined to form a classification machine capable of significant classification. Additionally, we analysed our spectral datasets using multivariate control charts based on principal component analysis (PCA) to examine whether biospectroscopy could correctly classify normal, borderline and cancerous ovaries.
Methods

Tissue collection and preparation
Ovarian specimens were acquired from the Royal Preston Hospital bio-bank with appropriate ethics clearance (REC reference 10/H0308/75). They included 35 histologically benign ovarian samples, 30 samples containing borderline ovarian tumours and 106 samples with a diagnosis of epithelial carcinoma. The ovarian carcinomas were further subdivided to HGSC (n=46), LGSC (n=9), EC (n=15), MC (n=12), CCC (n=13), CS (n=7) and MT (n=4). Table 1 lists the specific histological diagnoses for these samples. The tissue samples were embedded in paraffin. 10-µm-thick tissue sections were floated onto Low-E IR reflective slides (Kevley Technologies, Chesterland, OH, USA) slides. These were de-waxed by serial immersion in three sequential fresh xylene baths for five minutes and washed in an acetone bath for a further five minutes (57). The resulting samples were allowed to air dry and placed in a desiccator until analysis. 4 µm thick parallel tissue sections were floated to glass slides and stained with hematoxylin and eosin for histological comparison when needed.

Classification of ovarian tissues according to their histopathological characteristics

Figure 1 shows a benign ovarian tumour (mucinous cystadenoma) (Figure 1a), a borderline tumour (Figure 1b) and different ovarian carcinoma subtypes stained (Figure 1 c to i) with H&E. The World Health Organization (WHO) criteria for classification of epithelial ovarian tumours are based on optical microscopy after H&E staining. They describe the tissues these carcinomas resemble and how they differ from each other in general terms. The World Health Organization (WHO) lists general criteria to assist the differentiation between the different subtypes (see E.S.I. Table S2) (Chen et al., 2003; Lalwani et al., 2011).
Figure 1: Benign and malignant ovarian tumour examples stained with H&E. (a) Mucinous cystadenoma (benign) (b) Mucinous borderline tumour (c) High-grade serous carcinoma (d) endometrioid carcinoma (e) Low-grade serous carcinoma (f) Carcinosarcoma (g) serous borderline carcinoma (h) Mucinous carcinoma (i). Clear cell carcinoma
**ATR-FTIR spectroscopy**

IR spectra were obtained using a Bruker Vector 27 FTIR spectrometer with a Helios ATR attachment containing a diamond crystal (≈250µm × 250µm sampling area) (Bruker Optics Ltd., Coventry, UK). Spectra were acquired from 10 different locations across each specimen. A new background measurement was taken for every sample processed. The ATR crystal was cleaned with distilled water and dried with dry tissue paper before the acquisition of spectral background. The spectral resolution was 8cm⁻¹ with 2X zero filling of the interferogram giving data spacing of 4cm⁻¹. Spectra were co-added for 32 scans; these were converted into absorbance by Bruker OPUS software. Absorbance spectral images were converted to suitable digital files (.txt) for input to Matlab software.

**Computational analysis**

The ATR-FTIR datasets were processed using an in-house produced toolbox (iRootlab) (Trevisan et al., 2013) and PLS toolbox 7.8 (Eigenvector Research, Inc.3905 West Eaglerock Drive, Wenatchee, WA 98801) within a MATLAB R2014a environment (Mathworks Inc, Natick, MA, USA). The wavenumber regions inputted were between 4,000cm⁻¹ and 600cm⁻¹. Spectra were then cut to include the regions between 1,800-900cm⁻¹. PCA-LDA reduces the complex spectral dataset into single points in hyperspace, while maximizing inter-class variation and minimizing intra-class variation. The disadvantage of this method is the potential over fitting of spectra causing arbitrary separation and therefore positive results. This can be counteracted by using large spectral datasets of more than five times the number of variables.
For PCA-LDA, SPA-LDA and GA-LDA model, the samples were divided into training (70%), validation (15%) and prediction sets (15%) by applying the classic Kennard-Stone (KS) uniform sampling algorithm to the IR spectra (Kennard, 1969). Training samples were used in the modelling procedure (including variable selection for LDA), whereas the prediction set was only used in the final evaluation of the classification. The optimum number of variables for SPA-LDA and GA-LDA was determined from the minimum cost function $G$ calculated for a given validation dataset:

$$G = \frac{1}{N_V} \sum_{n=1}^{N_V} g_n, \quad (1)$$

where $g_n$ is defined as

$$g_n = \frac{r^2(x_n, m_{I(n)})}{\min_{I(m) \neq I(n)} r^2(x_n, m_{I(m)}) \ r^2(x_n, m_{I(n)})} \quad (2)$$

and $I(n)$ is the index of the true class for the $n^{th}$ validation object $x_n$. $g_n$ is defined as risk of misclassification of the $n$th validation object $x_n$, $n = 1, ..., N_V$. In this definition, the numerator is the squared Mahalanobis distance between object $x_n$ (of class index $I_n$) and the sample mean $m_{I(n)}$ of its true class. The denominator in Eq. (2) corresponds to the squared Mahalanobis distance between object $x_n$ and the center of the closest wrong class.

The GA routine was carried out during 100 generations with 200 chromosomes each. Crossover and mutation probabilities were set to 60% and 10%, respectively. Moreover, the algorithm was repeated three times, starting from different random initial populations. The best solution (in terms of the fitness value) resulting from the three realizations of the GA was employed. For this study, LDA scores, loadings, and discriminant function (DF) values were obtained for the specimens. The first LDA factor (LD1) was used to visualize the alterations in the sample in 1-dimensional (1D) score plots that indicate the main biochemical alterations.
Multivariate control charts were based on PCA. When the PCA model is applied on data collected when only common use variation is present, the future data behavior can be referenced against this “in-control” model. In this sense, new multivariate observations can be projected onto the plane defined by the PCA loading vectors to obtain their scores ($t_{inew} = p_i^T y_{new}$) and the residuals $e_{new} = y_{new} - \hat{y}_{new}$, where $\hat{y}_{new} = PA_{A,new}$, $t_{A,new}$ is the $(A \times 1)$ vector of scores from the model and $P_A$ is the $(q \times A)$ matrix of loadings. The presence of samples within the ±2s control limits in the Shewhart control chart built using the relevant PC scores. Trends and systematic behaviors in the score plot are clear indications of “out-of-control” processes (in this case, normal ovarian tissue, borderline ovarian tissue and different ovarian carcinoma subtypes).
Results

Classification of normal ovaries, borderline ovarian tumours and ovarian carcinomas using PCA-LDA, SPA-LDA and GA-LDA following ATR-FTIR

To discriminate between normal ovaries, ovaries with borderline tumours and ovaries with carcinomas, the spectral dataset was pre-processed using 1st order Savitzky-Golay smoothing, (Order 2; Window 15) (Figure 2). Overall the IR spectra appear to overlap in the biochemical fingerprint area (1800cm\(^{-1}\) to 900cm\(^{-1}\)). On closer inspection there are subtle but significant differences identified in the regions 1150-1000cm\(^{-1}\) (glycogen and nucleic acids), 1300-1200cm\(^{-1}\) (asymmetric phosphate and Amide III), 1550-1450cm\(^{-1}\) (protein moieties), 1600 to 1540cm\(^{-1}\) (Amide I region and DNA base region) and 1730-1630cm\(^{-1}\) (phospholipids and other lipids).
Figure 2: Analysing ovarian tissues by ATR-FTIR and pre-processing resulting datasets.
(a) 10µm thick ovarian tissue sections (b) Sample in close proximity with the ATR diamond (c) Unprocessed spectra (d) Resulting spectra after pre-processing.

To examine these visible differences and attempt classification of the 3 categories, 3 types of chemometric analysis were used. Classification was achieved using PCA-LDA, SPA-LDA and GA-LDA. 70% of the spectra were used to train the algorithm, 15% to test it internally and 15% to validate it externally (see E.S.I. Table S3). On comparing the spectra using PCA-LDA, 7 principal components were used as this number provided significant classification (P<0.001) without the introduction of arbitrary separation. Figure 3
c shows the 2-D scores plot derived by PCA-LDA. It reveals segregation of cancerous tissue from normal and borderline tumours, with the latter classes completely overlapping. The majority of the difference between the normal and cancerous ovaries was attributed to Amide I (1674 cm\(^{-1}\)), nucleic acids (1620 cm\(^{-1}\)), different conformations of phenyl rings (1585 cm\(^{-1}\), 1504 cm\(^{-1}\)), polysaccharides (1431 cm\(^{-1}\)) and symmetric phosphate stretching (1096 cm\(^{-1}\)) (Figure 3a). The chemometric technique that classified the three classes most successfully [66.4\%] was GA-LDA using 29 variables determined from the minimum cost function G (Figure 3 g & h) (see E.S.I. Table S4). The related 2-D scores plot illustrates that spectral points from different classes dissociate while spectral points from the same class co-cluster (Figure 3i). SPA-LDA also achieved considerable classification [55.9\%] with separation of classes on a 2-D scores plot (Figure 3f) when applied using 23 variables (Figure 3d) again using the minimum cost function G (Figure 3e) (see E.S.I. Table S4). All three techniques identified differences that aided classification within similar spectral regions. These differences were tentatively identified in the spectral regions of \(\nu = 1400\text{ cm}^{-1}\) (protein), \(\nu = 1740\text{ cm}^{-1}\) (lipid), \(\nu = 1045\text{ cm}^{-1}\) (phosphate), \(\nu = 1545\text{ cm}^{-1}\) (carbohydrate).
Figure 3: Classification of benign, borderline and malignant ovarian tissue by spectral analysis using PCA-LDA, SPA-LDA and GA-LDA. (red=cancer, green= borderline, blue= benign) (a) Loadings plot identifying the major discriminant wavenumbers for the three classes. The X-axis is cm\(^{-1}\) and the Y-axis represents absorbance coefficient. The five wavenumbers contributing to the most segregation were derived from the points furthest away from the X-axis. (b) Cost/ function plot identifying the optimal number of PCs to be used for PCA (c) Scores plot graphically representing classification by PCA-LDA. The X-axis represents LD1 and the Y-axis LD2 (d) Wavenumber selection for SPA-LDA. (e) Cost/ function plot identifying the optimal number of wavenumbers to be used for the SPA algorithm (f) Scores plot graphically representing classification by SPA-LDA. The X-axis represents LD1 and the Y-axis LD2 (g) Wavenumber selection for GA-LDA. (h) Cost/ function plot identifying the optimal number of wavenumbers to be used for the GA algorithm. (i) Scores plot graphically representing classification by GA-LDA. The X-axis represents LD1 and the Y-axis LD2.
**Lipid to protein ratio, phosphate to carbohydrate ratio and RNA to DNA ratio**

To further evaluate the importance of the above spectral regions in classifying the ovarian tumours to benign, borderline and malignant, intensity ratios of important for classification spectral areas when using the pre-mentioned approaches were measured.

Figure 4a shows the lipid to protein ratio, which is obtained by calculating the ratio of band intensities at 1750cm\(^{-1}\) to 1730cm\(^{-1}\) (lipids) and 1410cm\(^{-1}\) to 1390cm\(^{-1}\) (protein). The lipid to protein ratio was higher in neoplastic tissue and lower in borderline and benign tissue. Normal and benign tissue exhibited similar ratios.

The parameters used for the tentatively assigned phosphate to carbohydrate ratio in each IR spectrum were derived from the intensity of phosphate at 1055cm\(^{-1}\) to 1045cm\(^{-1}\) and of carbohydrate at 1555cm\(^{-1}\) to 1535cm\(^{-1}\). The phosphate to carbohydrate ratio is mildly increased in ovarian carcinomas relative to borderline and benign tissue and the difference is also significant (P<0.0001) (Figure 4b).

Similarly, when comparing the intensity ratios of RNA (1111cm\(^{-1}\) to 1131cm\(^{-1}\)) to DNA (1010cm\(^{-1}\) to 1030cm\(^{-1}\)), the ovarian carcinomas exhibited slightly lower ratio (Figure 4c).
Figure 4: Classification of ovarian tumours to benign, borderline and malignant using spectral intensity ratios. (a) Intensity ratio of lipid to protein. (b) Intensity ratio of phosphate to carbohydrate. (c) Intensity ratio of RNA to DNA.

Classification of normal ovaries, ovaries with borderline tumours and ovaries with ovarian carcinoma using multivariate control charts were based on PCA following ATR-FTIR

Multivariate control charts are commonly used in industry for quality control of chemical substances. A similar approach may be used in biospectroscopy. Tissue from benign ovaries can act as “control tissue” against which borderline and neoplastic tissues are compared. The control is represented by a line at zero, and another line is drawn usually at 2 standard deviations. How far from normal this line is from 0 depends on the variability that exists
within the examined tissue. When comparing borderline and malignant tissue with benign control tissue everything outside the standard deviation lines is considered abnormal.

Interestingly, control charts, derived from the PCA analysis already performed are able to distinguish between normal and neoplastic ovaries (Figure 5a) and normal and ovaries with borderline tumours (Figure 5b).

**Figure 5:** Classification of ovarian tumours to benign, borderline and malignant using Shewhart control charts after PCA. (a) Benign ovarian tissue VS malignant tissue (b) Benign ovarian tissue VS borderline tissue. (The blue dotted line is drawn at 2 standard deviations).
Classification of ovarian carcinoma subtypes using PCA-LDA, SPA-LDA and GA-LDA following ATR-FTIR

Similar chemometric techniques have been used to classify epithelial ovarian carcinomas according to their subtypes. The aforementioned pre-processing of the spectral datasets was applied in this case also. 7 PCs were used for PCA (6b), 23 wavenumbers for SPA (Figure 6d) and 44 (Figure 6g) wavenumbers for GA (see E.S.I table S5). The number of wavenumbers to be used was again determined by the minimum cost function G (Figure 6 e & h). PCA, SPA and GA followed by LDA were not adequately successful when comparing the spectral datasets of all cancer subtypes together as revealed by the associated 3-D scores plots (Figure 6 c, f & i). There was however visible separation between clear cell carcinoma (cyan), carcinosarcoma (pink) and high-grade serous carcinoma (blue) subtypes when analysed SPA-LDA and separation between clear cell carcinoma (cyan), carcinosarcoma (pink) spectral classes when analysed by GA-LDA. Unfortunately there was not adequate visual separation between classes with PCA-LDA.
Figure 6: Classification of ovarian carcinoma subtypes by spectral analysis using PCA-LDA, SPA-LDA and GA-LDA. (blue= high grade serous, red= low grade serous, black= endometrioid carcinoma, yellow= mixed, green= mucinous, cyan= clear cell, pink=carcinosarcoma) (a) Pre-processed spectral dataset. Each colour represents a particular neoplastic subtype. (b) Cost/ function plot identifying the optimal number of PCs to be used for PCA (c) Scores plot graphically representing classification by PCA-LDA. The X-axis represents LD1 and the Y-axis LD2 (d) Wavenumber selection for SPA-LDA. (e) Cost/ function plot identifying the optimal number of wavenumbers to be used for the SPA algorithm (f) Scores plot graphically representing classification by SPA-LDA. The X-axis represents LD1 and the Y-axis LD2 (g) Wavenumber selection for GA-LDA. (h) Cost/ function plot identifying the optimal number of wavenumbers to be used for the GA algorithm. (i) Scores plot graphically representing classification by GA-LDA. The X-axis represents LD1 and the Y-axis LD2.
Two category discriminant analysis of ovarian carcinoma subtypes using PCA-LDA, SPA-LDA and GA-LDA following ATR-FTIR

To increase the classification success rate, spectral datasets representing different epithelial tumour subtypes were compared in pairs. The three chemometric techniques previously mentioned where utilised again. Similar validation methods were used with 70% of the data being used to train the system, 15% for internal validation and 15% for external validation. The optimum number of principal components for PCA and variables for SPA-LDA and GA-LDA was determined by power versus cost calculation using the minimum cost function $G$ (See E.S.I. Figure S1). Electronic supplementary information (E.S.I.) Figures S2, S3 and S4 represent graphically the 2-D scores plots derived by PCA-LDA, SPA-LDA and GA-LDA respectively following comparison of all the carcinoma subcategories after processing by ATR-FTIR. The three analytical techniques were not equally successful at distinguishing between the categories compared. Figure 7 presents the percentage success for classification with each method. In general distinguishing between the different carcinoma subclasses was more successful when using GA-LDA.
Figure 7: Percentage successful classification of ovarian carcinoma subtypes when compared in pairs using 3 chemometric analyses: PCA-LDA, SPA-LDA, GA-LDA. Red boxes represent the most successful technique for each particular pair analysed. Amber colour represents the second most successful technique and green the least.

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>84.44</td>
<td>89.55</td>
<td>89.20</td>
<td>88.62</td>
<td>73.68</td>
<td>89.47</td>
</tr>
<tr>
<td>1</td>
<td>85.88</td>
<td>85.88</td>
<td>100.00</td>
<td>65.00</td>
<td>100.00</td>
<td>100.00</td>
</tr>
<tr>
<td>2</td>
<td>68.00</td>
<td>68.00</td>
<td>86.00</td>
<td>77.50</td>
<td>60.00</td>
<td>62.50</td>
</tr>
<tr>
<td>3</td>
<td>86.00</td>
<td>84.00</td>
<td>80.00</td>
<td>73.00</td>
<td>90.00</td>
<td>90.00</td>
</tr>
<tr>
<td>4</td>
<td>68.00</td>
<td>68.00</td>
<td>82.22</td>
<td>62.22</td>
<td>62.22</td>
<td>62.22</td>
</tr>
<tr>
<td>5</td>
<td>68.00</td>
<td>68.00</td>
<td>71.43</td>
<td>94.28</td>
<td>94.28</td>
<td>94.28</td>
</tr>
</tbody>
</table>
Discussion

Our study demonstrates that ATR-FTIR in conjunction with powerful chemometric approaches has the potential to distinguish between normal, borderline and neoplastic ovarian tissue. It also has the potential to distinguish between different ovarian epithelial carcinoma subtypes.

The most conspicuous differences are between normal ovaries and overt carcinoma as identified by the mentioned chemometric techniques. This finding is significant due to its potential for translation into clinical practice. Currently, histological identification is the gold standard in the diagnosis of ovarian cancer and therefore essential for surgical decision-making. Benign ovarian masses do not require extensive surgery while ovarian carcinomas will usually be managed by “staging” surgery involving a bilateral salpingooopherectomy, hysterectomy, omentectomy and pelvic lymphadenectomy. Pre-operative biopsy methods have been suggested to obtain histological diagnosis before embarking in major surgery. For example, image guided fine needle aspiration cytology (FNAC) and core biopsy using ultrasound, CT or MRI imaging have been shown to be effective with a diagnostic accuracy of 80.9% and 93% respectively (Mehdi et al., 2010; Spencer et al., 2006). These diagnostic modalities are usually preserved for women with co-morbidities that prohibit primary staging surgery or where imaging has revealed potentially inoperable disease. The reason for this is the risk of upstaging the disease by causing intra-peritoneal spillage of cancerous cells. Where there is a high clinical suspicion of ovarian cancer, a “staging procedure” is performed, which includes bilateral salpingo-oopherectomy, hysterectomy, omentectomy and pelvic lymphadenectomy.

In cases where clinical suspicion alone is not enough to embark in staging surgery, intra operative consultation by a pathologist is pursued. This utilises “frozen section” of the
specimen, which is then stained, usually with haematoxylin and eosin (H&E) and is examined by optical microscopy. Frozen section distinguishes benign from malignant tumours very accurately, but is less accurate for borderline tumours (Brun et al., 2008; Medeiros et al., 2005). It prevents morbidity associated with surgical staging procedures in benign cases and under-treatment of malignant tumours, which would otherwise require restaging surgery or chemotherapy. Frozen section has several limitations that include sampling difficulties, interpretation errors and communication breakdown (Jaafar, 2006). It also causes increases in surgical times, with resultant morbidity to the patient. ATR-FTIR in conjunction with chemometric analysis allows the identification of molecular biomarkers that can be adapted for easy discrimination between benign and neoplastic tissue during surgery. Indeterminate ovarian masses that would otherwise require a frozen section may be processed using ATR-FTIR. Multivariate control charts may be used to distinguish the ovarian tumours would require extensive surgery from those that will not.

The clinical importance for the diagnosis of ovarian carcinoma subtypes lays with their implications in immediate and subsequent management, medical or surgical, their follow-up and genetic counselling. Patients with early stage (1a) mucinous or endometrioid carcinoma can be treated with surgery alone. Patients with high-grade serous carcinoma will routinely have adjuvant chemotherapy. Patients with mucinous, endometrioid, and clear cell carcinomas may have adjuvant or neo adjuvant combination radiotherapy and chemotherapy. High grade serous adenocarcinoma is also associated with BRCA mutation therefore patients may be referred for genetic testing and if proven positive their families would be screened. ATR-FTIR coupled with a chemometric machine has the potential of being adopted as an assisting tool for pathological interpretation of ovarian carcinomas.
Conclusion

The purpose of this study was to identify spectral differences within ovarian tissues with the capability of classifying the in accordance to their histopathological status. Utilising ATR-FTIR n=171 ovarian tissues were examined. Morphological and molecular alterations within these tissues have already been associated with neoplasia. Spectroscopic analysis of these tissues reveals specific molecular alterations linked to malignancy. The responsible molecular changes for this segregation were primarily alterations in the tentatively assigned lipid (1740 cm\(^{-1}\)) to protein (1400 cm\(^{-1}\)) ratio with a marked increase associated with carcinomas. IR spectroscopy coupled with chemometric analysis has the potential to differentiate not only neoplastic from borderline and benign tissues but also distinguish between different carcinoma subtypes. Further validation of these approaches exploiting other biospectroscopy techniques and using larger architecturally robust datasets is required.
References


Table 1: Histopathology of the ovarian tissues interrogated: Borderline tumours and malignant epithelial carcinomas are similarly staged according to FIGO Ovarian Cancer Staging 2014

<table>
<thead>
<tr>
<th>Benign Ovaries (n=35)</th>
<th>Follicular cysts</th>
<th>Endometriosis</th>
<th>Involved in adhesions</th>
<th>Normal</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>29</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Borderline ovaries (n=30)</th>
<th>Stage/Subtype</th>
<th>Serous</th>
<th>Sero-mucinous</th>
<th>Mixed epithelial</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1a</td>
<td>10</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(2× micro-invasion)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>1b</td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>1c</td>
<td>10</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>2a</td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>3a</td>
<td>4</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>(1× micro-papillary)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>3b</td>
<td>2</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>(1× micro-papillary)</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Ovarian cancer (n=109)</th>
<th>Stage/Subtype</th>
<th>High grade serous</th>
<th>Low grade serous</th>
<th>Endometrial</th>
<th>Mixed ovarian tumour</th>
<th>Mucinous</th>
<th>Clear cell carcinoma</th>
<th>Carcino-sarcoma</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1a</td>
<td>2</td>
<td>2</td>
<td>3</td>
<td>6</td>
<td>4</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>1b</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>1c</td>
<td>15</td>
<td>2</td>
<td>8</td>
<td>6</td>
<td>3</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td></td>
<td>2a</td>
<td>2</td>
<td>2</td>
<td>1</td>
<td></td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>2b</td>
<td></td>
<td></td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>2c</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>3a</td>
<td>2</td>
<td>2</td>
<td>1</td>
<td></td>
<td>3</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>3b</td>
<td>5</td>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>3c</td>
<td>18</td>
<td>3</td>
<td>1</td>
<td></td>
<td>1</td>
<td></td>
<td>4</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Electronic Supplementary Information:

Attenuated total reflection infrared spectroscopy coupled with chemometric analysis discriminates normal, borderline and malignant ovarian tissue and classifies subtypes of human ovarian cancer

Georgios Theophilou1, 2, Kássio M.G. Lima1, 3, Pierre L. Martin Hirsch1, 2, Helen F. Stringfellow2, Francis L. Martin1*

1Centre for Biophotonics, LEC, Lancaster University, Lancaster LA1 4YQ, UK, 2Department of Obstetrics and Gynaecology, Central Lancashire Teaching Hospitals NHS Foundation Trust, Preston, UK,

3Institute of Chemistry, Biological Chemistry and Chemometrics, Federal University of Rio Grande do Norte, Natal 59072-970, RN-Brazil

*Corresponding author: email: f.martin@lancaster.ac.uk; phone: +44 (0)1524 510206; fax: +44 (0)1524 510217

Abbreviations:

LGSC: low grade serous carcinoma  CS: Carcinosarcoma
HGSC: high grade serous carcinoma  DF: Discriminant function
EC: Endometrioid carcinoma  RMI: Risk malignancy index
MC: Mucinous carcinoma
MT: Mixed tumour
CCC: Clear cell carcinoma
Figure S1: Optimization of Principal Component and Wavenumber selection for each of the analytical methods for paired comparison. This example uses the HGSC and LGSC classes.
Figure S2: 1-Dimensional PCA-LDA derived scores plots obtained from paired comparison of ovarian carcinoma subtypes.
Figure S2: 1-Dimensional PCA-LDA derived scores plots obtained from paired comparison of ovarian carcinoma subtypes.
Figure S2: 1-Dimensional PCA-LDA derived scores plots obtained from paired comparison of ovarian carcinoma subtypes.
**Figure S2:** 1-Dimensional PCA-LDA derived scores plots obtained from paired comparison of ovarian carcinoma subtypes.
Figure S3: 1-Dimensional SPA-LDA derived scores plots obtained from paired comparison of ovarian carcinoma subtypes.
Figure S3: 1-Dimensional SPA-LDA derived scores plots obtained from paired comparison of ovarian carcinoma subtypes.
Figure S3: 1-Dimensional SPA-LDA derived scores plots obtained from paired comparison of ovarian carcinoma subtypes.
Figure S3: 1-Dimensional SPA-LDA derived scores plots obtained from paired comparison of ovarian carcinoma subtypes.
Figure S4: 1-Dimensional GA-LDA derived scores plots obtained from paired comparison of ovarian carcinoma subtypes.
Figure S4: 1-Dimensional GA-LDA derived scores plots obtained from paired comparison of ovarian carcinoma subtypes.
Figure S4: 1-Dimensional GA-LDA derived scores plots obtained from paired comparison of ovarian carcinoma subtypes.
Figure S4: 1-Dimensional GA-LDA derived scores plots obtained from paired comparison of ovarian carcinoma subtypes.
Table S1. Risk Malignancy index (RMI): Women with ovarian cysts or vague abdominal symptoms undergo screening using the “Risk malignancy index”. This predicts the risk of an ovarian mass being malignant and dictates further surgical or medical management.

<table>
<thead>
<tr>
<th>Feature</th>
<th>RMI 1</th>
<th>RMI 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ultrasonic:</td>
<td>• No positive ultrasound features= 0</td>
<td>• No positive ultrasound features= 0</td>
</tr>
<tr>
<td></td>
<td>• 1 abnormality= 1</td>
<td>• 1 abnormality= 1</td>
</tr>
<tr>
<td></td>
<td>• 2 abnormalities= 2</td>
<td>• 2 abnormalities= 2</td>
</tr>
<tr>
<td>Bilateral lesions</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ascities</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Multilocular cysts</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Solid areas</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Metastases</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Premenopausal</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Postmenopausal</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>Ca125</td>
<td>U/ml</td>
<td>U/ml</td>
</tr>
</tbody>
</table>

**RMI= Ultrasound score × Menopausal score × Ca125 in U/ML**

<table>
<thead>
<tr>
<th>RMI</th>
<th>Risk</th>
<th>Women (%)</th>
<th>Risk of cancer (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>&lt;25</td>
<td>Low</td>
<td>40</td>
<td>&lt;3</td>
</tr>
<tr>
<td>25-250</td>
<td>Moderate</td>
<td>30</td>
<td>20</td>
</tr>
<tr>
<td>&gt;250</td>
<td>High</td>
<td>30</td>
<td>75</td>
</tr>
</tbody>
</table>
Table S2. Histopathological classification of ovarian epithelial tumours: Descriptive criteria for classification of ovarian carcinomas according to the World Health organization (WHO), 2003. Carcinosarcoma is not included in the main five categories due to its rarity.

<table>
<thead>
<tr>
<th>Carcinoma subtype</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Serous</td>
<td>Composed of cells ranging in appearance from those resembling fallopian tube epithelium in well-differentiated tumours to anaplastic epithelial cells with severe nuclear atypia in poorly differentiated tumours</td>
</tr>
<tr>
<td></td>
<td><strong>Low grade</strong></td>
</tr>
<tr>
<td></td>
<td>Uniform nuclei</td>
</tr>
<tr>
<td></td>
<td>&lt;13/10 high field powers mitotic figures</td>
</tr>
<tr>
<td></td>
<td>Prominent nucleoli</td>
</tr>
<tr>
<td></td>
<td>Differentiated architecture with papillary growth</td>
</tr>
<tr>
<td></td>
<td>Numerous psammoma bodies</td>
</tr>
<tr>
<td>Mucinous</td>
<td>Resembles intestinal or endocervical epithelium</td>
</tr>
<tr>
<td>Endometrioid</td>
<td>Closely resembles the common variant of endometrioid carcinoma of the uterine corpus</td>
</tr>
<tr>
<td>Clear cell</td>
<td>Composed of glycogen-containing clear cells and hobnail cells and occasionally other histological types</td>
</tr>
<tr>
<td>Mixed surface</td>
<td>Composed of an admixture of two or more of the five major histological types, and the minor component(s) must comprise alone or together at least 10% of the tumour</td>
</tr>
<tr>
<td>Carcinosarcoma</td>
<td>Composed of both malignant epithelial and homologous (similar to Mullerian duct system) or heterologous (e.g. cartilage, bone, muscle) stromal elements</td>
</tr>
</tbody>
</table>
**Table S3: Internal and external algorithm validation:** 70% of the spectra were used to train the algorithm, 15% to test it internally and 15% to validate it externally.

<table>
<thead>
<tr>
<th></th>
<th>Normal</th>
<th>Borderline</th>
<th>Cancer</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Train</td>
<td>239 × 235</td>
<td>207 × 235</td>
<td>778 × 235</td>
<td>1224 × 235</td>
</tr>
<tr>
<td>Validation</td>
<td>55 × 235</td>
<td>45 × 235</td>
<td>165 × 235</td>
<td>265 × 235</td>
</tr>
<tr>
<td>Test</td>
<td>55 × 235</td>
<td>45 × 235</td>
<td>165 × 235</td>
<td>265 × 235</td>
</tr>
</tbody>
</table>

**Table S4: Selected wavenumbers for SPA-LDA and GA-LDA.** These wavenumbers were used to achieve classification of normal, borderline and malignant ovarian tissue.

<table>
<thead>
<tr>
<th>Classification into normal, borderline and malignant ovaries</th>
<th>Wavenumbers selected</th>
</tr>
</thead>
<tbody>
<tr>
<td>Chemometric analysis</td>
<td></td>
</tr>
<tr>
<td>SPA-LDA</td>
<td>900, 995, 1026, 1068, 1111, 1165, 1230, 1377, 1404, 1446, 1462, 1512, 1543, 1554, 1562, 1604, 1620, 1643, 1658, 1681, 1747, 1800</td>
</tr>
<tr>
<td>GA-LDA</td>
<td>952, 983, 987, 1041, 1049, 1084, 1099, 1122, 1141, 1168, 1203, 1219, 1346, 1365, 1419, 1446, 1450, 1512, 1527, 1539, 1546, 1558, 1593, 1604, 1631, 1643, 1647, 1720, 759</td>
</tr>
</tbody>
</table>

**Table S5: Selected wavenumbers for SPA-LDA and GA-LDA.** These wavenumbers were used to achieve classification of ovarian carcinoma subtypes.

<table>
<thead>
<tr>
<th>Ovarian carcinoma subtype classification</th>
<th>Wavenumbers selected</th>
</tr>
</thead>
<tbody>
<tr>
<td>Chemometric analysis</td>
<td></td>
</tr>
<tr>
<td>SPA-LDA</td>
<td>964, 991, 1018, 1037, 1068, 1111, 1153, 1219, 1334, 1415, 1458, 1485, 1504, 1539, 1562, 1597, 1624, 1635, 1654, 1662, 1697, 1724, 1800</td>
</tr>
<tr>
<td>GA-LDA</td>
<td>902, 941, 964, 999, 1003, 1018, 1022, 1084, 1099, 1103, 1122, 1192, 1222, 1311, 1381, 1392, 1400, 1408, 1423, 1438, 1469, 1481, 1485, 1489, 1492, 1504, 1516, 1531, 1535, 1554, 1562, 1570, 1589, 1593, 1624, 1627, 1654, 1658, 1674, 1685, 1712, 1732, 1747, 1782</td>
</tr>
</tbody>
</table>
Chapter 4

A biospectroscopic analysis of human prostate tissue obtained from different time periods points to a trans-generational alteration in spectral phenotype

Georgios Theophilou, Kássio M.G. Lima, Matthew Briggs, Pierre L. Martin Hirsch, Helen F. Stringfellow, Francis L. Martin

Contribution:
I designed the study
I retrieved, pre-processed and processed the ovarian tissue samples by ATR-FTIR and Raman and performed immunochemistry.
I performed a portion of the data analysis.
I have written the first draft of the manuscript.

Professor F. L Martin
G. Theophilou
A biospectroscopic analysis of human prostate tissue obtained from different time periods points to a trans-generational alteration in spectral phenotype

Georgios Theophilou¹, ², Kássio M.G. Lima¹, ³, Matthew Briggs², Pierre L. Martin Hirsch¹, ², Helen F. Stringfellow², Francis L. Martin¹*

¹Centre for Biophotonics, LEC, Lancaster University, Lancaster LA1 4YQ, UK

²Department of Obstetrics and Gynaecology, Central Lancashire Teaching Hospitals NHS Foundation Trust, Preston, UK

³Institute of Chemistry, Biological Chemistry and Chemometrics, Federal University of Rio Grande do Norte, Natal 59072-970, RN-Brazil

*Corresponding Author: Prof Francis L Martin PhD, Centre for Biophotonics, LEC, Lancaster University, Lancaster LA1, 4YQ, UK; Tel.: +44(0) 1524 510206; Email: f.martin@lancaster.ac.uk
Abstract

Prostate cancer is the most commonly diagnosed malignancy in males worldwide; however, there is marked geographic variation in its incidence. This may be associated with adoption of a Westernised lifestyle. We set out to determine whether attenuated total reflection Fourier-transform infrared (ATR-FTIR) spectroscopy or Raman spectroscopy combined with principal component analysis-linear discriminant analysis (PCA-LDA) or variable selection techniques employing genetic algorithm (GA) or successive projection algorithm (SPA) could be utilised to explore differences between prostate tissues obtained from differing years. In total, 156 prostate tissue samples from transurethral resection of the prostate (TURP) procedures for benign prostatic hyperplasia (BPH) were collected from 1983 to 2013. These were distributed according to the year of collection to form seven categories: 1983-1984 \((n=20), 1988-1989 \((n=25), 1993-1994 \((n=21), 1998-1999 \((n=21), 2003-2004 \((n=21), 2008-2009 \((n=20)\) and 2012-2013 \((n=21)\). Ten-\(\mu\)m-thick tissue sections were floated onto Low-E (IR-reflective) slides for ATR-FTIR or Raman spectroscopy. Resulting scores plots for PCA-LDA, SPA-LDA or GA-LDA from ATR-FTIR data revealed marked segregation between the seven categories. In fact, there was a chronological development of prostate tissue spectroscopic alterations with successive categories. This classification was less evident following Raman spectroscopy but here also, a significant separation between categories was identified. Moreover, examination of the two categories that are at least one generation (30 years) apart indicated highly significant segregation, especially at spectral regions containing DNA and RNA bands \((\approx 1,000-1,490\text{cm}^{-1})\), involving nucleic acids, phosphate and deoxyribose modifications. This may point towards alterations that have occurred through chemical genotoxicity or through epigenetic modification of
chromatin structure. Immunohistochemical studies for global DNA methylation supported the results obtained by vibrational spectroscopy. This study points to a trans-generational phenotypic change in human prostate as a function of spectral alterations.

**Keywords:** ATR-FTIR spectroscopy; Biospectroscopy; Classification analysis; Prostate cancer; Raman spectroscopy; Wavenumber selection
Introduction

Prostate cancer is the most commonly diagnosed male malignancy in the world with an incidence rate of 214 cases per 100,000 and a mortality rate from associated metastatic disease of 30 in 100,000 (Jemal et al., 2008; Nichol et al., 2005). The percentage of prostate cancer amongst all male cancers is much higher in developed countries (15%) than in developing ones (4%), but there are also large regional differences in incidence rates (Ferlay et al., 2010; Parkin et al., 2010; Parkin et al., 1999).

The only established risk factors for PC are increasing age, ethnic origin and heredity (Bratt et al., 2002; Dunsmuir et al., 1998; Ho et al., 2006). However, the effects of environment and lifestyle appear to be important towards its development (Alberti, 2010; Cancel-Tassin & Cussenot, 2005). The age-adjusted incidence trends for prostate cancer in the 20-year period from 1973 to 1992 were found to increase consistently in 15 countries (Hsing et al., 2000). Associated temporal lifestyle variations may include diet and exercise, with related factors in the prevalence of obesity, diabetes and metabolic syndromes, tobacco smoking and alcohol intake (Buschemeyer & Freedland, 2007; Giovannucci et al., 1993; Martin et al., 2009; Nilsen et al., 2006; Putnam et al., 2000; R. A. La Vallee, 2014; Rohrmann et al., 2007; Whittemore et al., 1995).

Working on the assumption that lifestyle changes are major players in the initiation and development of prostate cancer and that lifestyle and especially diet have changed dramatically in the past 20 y (within one generation), we set out to explore differences that may exist between prostates from different individuals obtained over a 30-y period.
Tissue from transurethral resection of the prostate (TURP) procedures for benign prostatic hyperplasia (BPH) provided the opportunity to study these temporal differences (Figure 1). The cancer risk in this population is comparable or marginally increased relative to the general population (Holman et al., 1999; Karlsson et al., 2011). Although prostate tissue from TURP procedures may be histologically benign, it could harbour early molecular alterations that contribute to prostate cancer development.

In the search for such molecular alterations, biospectroscopy may play an important role as it can identify structural alterations of cellular molecules based on chemical bonds (H. Y. Holman et al., 2000; Jackson et al., 1997; Mourant et al., 2003; Taillandier & Liquier, 1992). Recent studies have also examined its potential in identifying biomarkers for cancer screening (Fung Kee Fung et al., 1997; Harris et al., 2009; Walsh et al., 2007). Attenuated total reflection Fourier-transform infrared spectroscopy (ATR-FTIR) and Raman spectroscopy were used to interrogate prostatic tissue. The resulting spectral data were analyzed using multivariate analysis in the form of principal component analysis followed by linear discriminant analysis (PCA-LDA) and variable selection techniques in the form of sequential progression algorithm (SPA) or genetic analysis (GA), again followed by LDA (SPA-LDA, GA-LDA).

Currently there is a lack of research evaluating potential prostatic molecular changes that have occurred in the past 30 y (>1 generation). This study set out to determine if spectral differences in prostate tissue of men of similar ages have occurred from the 1980’s to the present day. This could lend insights into distinct associations between modern adopted lifestyle and risk of prostate cancer.
Methods

Tissue collection

Archival benign prostate tissue specimens from TURP procedures were collected from one centre. They comprised of prostatic tissue chippings that were formalin-fixed, dehydrated and paraffin-embedded. These tissue blocks \(n=156\) were obtained from the Royal Preston Hospital bio-bank. Appropriate ethics clearance (REC reference 10/H0308/75) was obtained. These specimens were matched for age between sixty and sixty-nine years old. They were also matched for ethnicity with all being “British Caucasian”. Histology of these samples identified only BPH and no other abnormality. The specimens were examined using routine histopathology procedures and found to be free from prostate cancer and other abnormalities other than BPH.

In total 156 specimens were collected from 1983 to 2013. These samples were distributed according to the year of collection to form seven categories: 1983-1984 \(n=20\), 1988-1989 \(n=25\), 1993-1994 \(n=21\), 1998-1999 \(n=21\), 2003-2004 \(n=21\), 2008-2009 \(n=20\) and 2012-2013 \(n=21\). Ten-µm-thick tissue sections were floated onto Low-E IR reflective slides (Kevley Technologies, Chesterland, OH, USA) slides for ATR-FTIR spectroscopy. These were de-waxed by serial immersion in three sequential fresh xylenes baths for five minutes and washed in an acetone bath for another five minutes (Matthew J. Baker et al., 2014). The resulting samples were allowed to air dry and then placed in a desiccator until analysis (Fig. 1B, 1C). Parallel H&E sections were obtained for histological comparison to ensure relevant areas were examined (Fig. 1D).
**ATR-FTIR spectroscopy**

IR spectra were obtained using a Bruker Vector 27 FTIR spectrometer with a Helios ATR attachment containing a diamond crystal (≈250 µm × 250 µm sampling area) (Bruker Optics Ltd., Coventry, UK). Spectra were acquired from 10 different locations across each specimen with a new background taken for every new sample. The ATR crystal was cleaned with distilled water and dried with dry tissue paper before the acquisition of spectral background. The spectral resolution was 8 cm\(^{-1}\) giving data spacing of 4 cm\(^{-1}\). Spectra were co-added for 32 scans; these were converted into absorbance by Bruker OPUS software (Fig. 1E).

**Raman spectroscopy**

Raman spectra were acquired using an InVia Renishaw Raman spectrometer (Renishaw plc, Gloucestershire, UK). Its laser diode, operating at 35 mW, emits a mid-IR beam, whose exact wavelength is 785 nm. This was passed through a Rayleigh holographic edge filter. The spectrometer’s entrance slit of 50 mm combined with a diffraction grating of 1,200 lines per mm achieved a spatial resolution of 1 cm\(^{-1}\). Raman scatter signals were directed onto a Master Renishaw Pelletier cooled charged couple detector (CCD). Spectra were acquired using a Leica microscope via a ×50 objective lens with a numerical aperture of 0.75, giving a spatial resolution of approximately 1 mm. A white light camera mounted on the microscope allowed the use of dark-field visualization of the locations of interest. The Renishaw system was calibrated with a Renishaw silicon calibration source for wavenumber shifts every time the spectrometer was turned on. Ten spectra were acquired from independent locations from each sample. A total of 1,437 spectra were acquired using 100% laser power with an exposure period of 25 seconds and four repeat acquisitions.
(Fig. 1F). Raman spectroscopy was always performed on the same tissue sections following ATR-FTIR spectroscopy. Independent regions were targeted to minimize any confounding influences due to tissue compression by the ATR crystal.

**Computational analysis**

The importing and pre-treatment of the spectral data and the construction of chemometric classification models were executed using PLS toolbox 7.8 (Eigenvector Research, Inc. 3905 West Eaglerock Drive, Wenatchee, WA 98801) and in-house written scripts (irootlab) (Trevisan et al., 2013) within a MATLAB R2013a environment (Mathworks Inc, Natick, MA, USA).

ATR-FTIR spectra were cut to include wavelengths between 1,800 and 900 cm\(^{-1}\) (235 wavenumbers at 3.84 cm\(^{-1}\) spectral resolution); the area associated with the biological spectral fingerprints. The resulting dataset was rubber band baseline-corrected and normalized to the Amide I peak (\(i.e., \approx 1,650\) cm\(^{-1}\)) (Matthew J. Baker et al., 2014; Trevisan et al., 2012).

Raman spectra contained cosmic rays, which were removed using an in-house tool for Matlab. This algorithm excluded cosmic rays by statistically evaluating the whole spectral dataset (all samples) to identify abnormally high ‘spikes’ that did not present repeatedly. The spectral areas containing these spikes were replaced by appropriate values calculated as a function of intensities for the concerned areas for the rest of the data. The abrangence factor (\(k = 5\)) was adjusted to increase the sensitivity of the tool for spike removal. The resulting spectra were cut to include 1,750-800 cm\(^{-1}\) (692 data points). Subtraction of biological tissue auto-fluorescence was carried out using an automatic baseline correction method (Whittaker filter) (Trevisan et al., 2012).
Computational analysis consisted of three models: principal component analysis (PCA), sequential progression algorithm (SPA) and genetic algorithm (GA). All models were followed by linear discriminant analysis (LDA) (Matthew J. Baker et al., 2014). Before applying each analytical model, spectral data were divided into training (70%), validation (15%) and prediction (15%) sets by applying the classic Kennard-Stone (KS) uniform sampling algorithm (Kennard R. W., 1969). The number of samples colonising each set is presented in Electronic Supplementary Information (ESI) Tables S1 and S2 for ATR-FTIR and Raman, respectively. The training datasets were used in the modelling procedures (including variable selection for LDA), whereas the prediction dataset was only used for the final classification evaluation. The optimum number of variables for SPA-LDA and GA-LDA was determined from the minimum cost function \( G \) calculated for a given validation dataset:

\[
G = \frac{1}{N_v} \sum_{n=1}^{N_v} g_n \quad (1)
\]

where \( g_n \) is defined as

\[
g_n = \frac{r^2(x_n, m_{I(n)})}{\min_{l(m)=I(n)} r^2(x_n, m_{l(m)})} \quad (2)
\]

and \( I(n) \) is the index of the true class for the \( n \)th validation object \( x_n \).

PCA is a multivariate analysis technique that aims to reduce the number of variables present in the spectral dataset. Principal components (PCs) can capture most of the variance (>95%) present in the original dataset. A power versus cost calculation identifies the number of PCs that correctly identifies variance within the dataset.
without presenting artificial separation between the different classes. This optimum number was applied to classify the prostates depending on the year they were excised.

SPA is a forward selection method (Martens & Næs, 1989). Its purpose is to select wavelengths whose information content is minimally redundant to solve co-linearity problems. The model starts with one wavelength, then incorporates a new one at each iteration until it reaches a specified number N of wavelengths (Soares et al., 2011). SPA does not modify the original data vectors as PCA does. In this case projections are used only for selection purposes. Thus, the relation between spectral variables and data vectors is preserved.

Genetic algorithms (GA) are combinational algorithms inspired by Mendelian genetics. They use a combination of selection, recombination and mutation to evolve a solution to a problem. They treat data as chromosomes allocating reproductive opportunities in such a way that those chromosomes, which represent a better solution to the target problem are given more chances to “reproduce” than those, which represent poorer solutions (Whitley, 1994). The GA routine was carried out utilising 100 generations containing 200 chromosomes each. Crossover and mutation probabilities were set to 60% and 10%, respectively. Moreover, the algorithm was repeated three times, starting from different random initial populations. The best solution (in terms of the fitness value) resulting from the three realizations of the GA was employed.

LDA was performed following the application of each of the analytical models. LDA scores, loadings, and discriminant function (DF) values were obtained. Usually, the first LDA factor (LD1) is used to visualize the main biochemical alterations within the sample on a 1-dimensional (D) scores plot.
**Immunohistochemistry**

4mm thick, parallel sections of prostatic samples from the 1983-1984 (n=10) and the 2012-2013 (n=10) classes were de-waxed in xylene and taken to absolute alcohol. They were then placed in a warm Tris/EDTA (Trizma Base, Sigma, T1503; Citric acid crystals, BDH277804L; Sodium hydroxide, BDH301675N) buffer and heated under pressure at 900W in a microwave for 4 minutes. They were then cooled under running water and rinsed with Tris buffer. They were treated with hydrogen peroxide blocking agent (Dako) for five minutes, drained and rinsed with Tris buffer. Normal blocking serum was then placed on the sections for twenty minutes followed by 5-methylene cytosine as the primary antiserum (5-mc antibody, dilution 1:400; Genetext: GT4111) for sixty minutes. They were then rinsed with Tris buffer before adding the secondary antibody (Vectastain Universal Elite ABC Kit) for thirty minutes. After another wash with Tris buffer they were incubated in Strept-ABCComplex/HRP (Vectastain Elite ABC Reagent) solution for thirty minutes and then washed again. One drop of chromogen was added to 1ml of Dako and placed on the sections for 10 minutes. They were then washed under running water before counterstaining with haematoxylin for 5 minutes, dehydrated in alcohol, cleared in xylene and mounted in styrolite.
Results

ATR-FTIR spectral dataset


In order to attempt classification of the prostate samples according to year of collection and to determine the biochemical markers responsible for any such classification, it is necessary to apply chemometric analysis techniques. PCA-LDA, SPA-LDA and GA-LDA were therefore adopted to systematically identify spectral differences between the pre-assigned categories.

Fig. 2B shows a scores plot derived following PCA-LDA of the ATR-FTIR spectra. This model was carried out using the first six PCs, which account for >90% of the variance within the sample population. Scores plots identify the similarities and dissimilarities between different categories and present them as clusters of points. Loadings plots identify the distinguishing wavenumbers (as weightings). It is obvious that most spectral classes form a single cluster. It is also obvious that there is separation between the 1983-1984 (blue) and the 2012-2013 (pink) categories, which are >one generation (30 y) apart. This separation is significant \((P <0.0001)\). The loadings plot (Fig. 2C) derived from PCA-LDA identifies the six primary wavenumbers, which are important for separation of the different age groups. These
include 1,227, 1,400, 1,574, 1,624, 1,674 and 1,720 cm⁻¹. ESI: Table S3 lists the molecular entities associated with these wavenumbers.

SPA-LDA was applied to the dataset using the optimum number of variables derived by identifying the minimum cost from function G (Fig. 3B). The twenty-three wavenumbers selected were: 968, 1,018, 1,053, 1,153, 1,234, 1,315, 1,392, 1,415, 1,446, 1,462, 1,489, 1,512, 1,539, 1,562, 1,593, 1,620, 1,631, 1,651, 1,666, 1,693, 1,716, 1,735 and 1,797 cm⁻¹ [Fig. 3A; see (ESI) Table S4]. The resulting 3-D scores plot (Fig. 3C) identified significant segregation between categories (P < 0.05). Spectral points from the same category tend to co-cluster and differing classes segregate. There is a clear progression with time, with categories separated by one generation being furthest apart.

Fig. 4C displays the scores plot for classification achieved utilising GA-LDA. The GA model was built based on the selection of 32 wavenumbers (Fig. 4A; see ESI Table S5) out of the available 234, determined by function G (Fig. 4B). These included wavenumbers: 987, 999, 1,002, 1,026, 1,029, 1,072, 1,191, 1,199, 1,299, 1,303, 1,350, 1,353, 1,365, 1,373, 1,381, 1,388, 1,392, 1,404, 1,415, 1,458, 1,496, 1,504, 1,512, 1,543, 1,554, 1,562, 1,589, 1,600, 1,647, 1,708, 1,720 and 1,751 cm⁻¹. Again, there is separation between the different categories that is significant (P < 0.05).

When comparing the two categories separated by 28 y (1983-1984 and 2012-2013), the distinction between them is much clearer. Fig. 5A shows the pre-processed ATR-FTIR spectra used for analysis applying the three previously mentioned techniques. The 2-D scores plot derived from PCA-LDA of these two categories identifies significant segregation between them (P < 0.0001) (Fig. 5B). The associated
loadings plot (Fig. 5C) identifies the 6 principal segregating wavenumbers. The molecular entities assigned to these are listed in E.S.I. Table S6.

Similarly SPA-LDA identified significant separation \((P <0.05)\) between the two categories as shown by the related scores plot (Fig. 6C). This approach used four wavenumbers: 1,504, 1,620, 1,647 and 1,728 cm\(^{-1}\) (Fig. 6A; see ESI Table S7), as determined by the min cost of function \(G\) (Fig. 6B). GA-LDA produced the best separation (Fig. 7C) using 17 variables, selected at the cost function minimum point (Fig. 7B). These were: 1,049, 1,053, 1,253, 1,415, 1,419, 1,423, 1,500, 1,504, 1,512, 1,516, 1,519, 1,527, 1,531, 1,535, 1,539, 1,543 and 1,546 cm\(^{-1}\) (Fig. 7A; see ESI Table S8). This separation was also significant \((P <0.05)\).

**Raman spectral dataset**

Fig. 2D shows the pre-processed Raman spectra. Each colour represents a different category based on the year of collection. Similar to ATR-FTIR spectra, discrimination of categories requires reduction of the complexity of the spectral dataset. Therefore, PCA-LDA, SPA-LDA and GA-LDA were applied to segregate prostatic tissues based on their Raman spectra.

The PCA-LDA models (Fig. 2E), using six PC scores accounting for >90% of variance, did not reveal any substantial separation (although \(P <0.05\)) and there was a large degree of overlap between all categories. The first six wavenumbers responsible for separation were identified by the associated loadings curve. They included 1,418, 1,457, 1,576, 1,657, 1,704 and 1,739 cm\(^{-1}\) (Fig. 2F; see E.S.I. Table S9).

Fig. 3F shows the SPA-LDA derived scores plot. This approach also exhibited limited segregation of the categories. The cost function minimum point was obtained at 17 wavenumbers (Fig. 3E). These included: 1,000, 1,001, 1,004, 1,062, 1,109,
GA-LDA generated only a slight segregation between categories (Fig. 4F), when 49 selected wavenumbers were used, as directed by the cost function minimum point (Fig. 4E): 842, 845, 874, 892, 920, 946, 965, 967, 971, 997, 998, 1,010, 1,022, 1,067, 1,087, 1,168, 1,182, 1,185, 1,201, 1,251, 1,265, 1,271, 1,272, 1,310, 1,342, 1,373, 1,405, 1,421, 1,423, 1,457, 1,483, 1,496, 1,499, 1,507, 1,518, 1,560, 1,575, 1,629, 1,652, 1,657, 1,660, 1,666, 1,673, 1,700, 1,710, 1,729, 1,733, 1,741 and 1,745 cm\(^{-1}\) (Fig. 4D; see ESI Table S11). There was a slight improvement in separation in comparison with PCA-LDA and SPA-LDA approaches (\(P < 0.05\)).

Analysis of the Raman dataset for categories: 1983-1984 and 2012-2013 by the application of PCA-LDA, SPA-LDA and GA-LDA identified between-category segregation. PCA-LDA using the first six PCs revealed significant separation (\(P < 0.0001\)) (Fig. 5E). The derived loadings plot shows the main segregating wavenumbers: 1,419, 1,459, 1,567, 1,654 and 1,742 cm\(^{-1}\) (Fig. 5F). E.S.I. Table S12 lists their tentative assignments. SPA-LDA analysis using three wavenumbers, as directed by the minimum cost of function G (Fig. 6E): 891, 1,001 and 1,295 cm\(^{-1}\) (Fig. 6D; see ESI Table S13) also revealed between-category segregation (Fig. 6F) (\(P < 0.05\)). GA-LDA of the same dataset generated similar results (Fig. 7F), which were also statistically significant (\(P < 0.05\)). In this case, 14 variables were used, at the minimum cost function point (Fig 7E): 861, 899, 920, 921, 971, 1,049, 1,100, 1,204, 1,206, 1,261, 1,365, 1,447, 1,496 and 1,596 cm\(^{-1}\) (Fig. 7D; see ESI Table S14).
**Immunochemistry**

To further evaluate potential epigenetic changes contributing to trans-generational variability we performed immunohistochemistry in the form of methylation studies. The classes 1983-1984 (n=10) and 2012-2013 (n=10) were compared blindly. Methylation was graded according to the intensity of staining from 0 to 3 (0=no staining, 1=weak staining, 2=moderate staining, 3=strong staining). The percentage of cells exhibiting the particular grade within different cellular compartments (epithelial, basal, stromal and vascular cells) was also recorded (see ESI, Table S15). The class of 1983-1984 exhibited global methylation with 8 samples displaying strong (3) and 2 moderate staining (2) in nearly 100% of cells for all cellular compartments. 6 samples from class of 2012-2013 exhibited strong staining a 4 moderate staining.


Discussion

This study aimed to identify spectral differences between benign prostate tissues acquired following TURP procedures carried out over the last 30 y on similarly-aged men. Such spectral differences could be the first evidence of phenotypic alterations from one generation to the next. A total of \( n = 156 \) tissues were analysed using ATR-FTIR and Raman spectroscopy. The specific prostatic histological area examined was the transition zone as this is the tissue region excised at TURP (McNeal, 1981). About 75% of prostate cancers originate in the peripheral zone, which is located postero-laterally to the urethra (Fig. 1) (McNeal, 1981). Some 25% of prostate cancers also arise in the transition zone and behave differently to peripheral zone cancers, both morphologically and functionally (Augustin et al., 2003). Micro-environmental cellular communication plays a significant role in cancer initiation and progression; (Hanahan & Weinberg, 2011) therefore examination of any part of the prostate may provide information that may lead to better understanding disease pre-disposing alterations, e.g., prostatic intraepithelial neoplasia (Bostwick, 2000).

IR spectra were obtained from the mid-IR region from 900 to 1,800 cm\(^{-1}\) as most bio-molecular spectral signatures reside within this area (Movasaghi et al., 2008). Raman spectra used contained wavenumbers from 750 to 1,500 cm\(^{-1}\) for the same reason (Movasaghi et al., 2007). Computational analysis allowed discrimination of prostatic tissue according to the year of surgery. The rationale for this approach was to determine if a trans-generational change in the spectral phenotype of this tissue might be detectable. There was apparent separation between the clusters of different categories that became more pronounced as the period between sample collections became larger.
These three computational methods applied to the spectra obtained by both Raman or ATR-FTIR spectroscopy had varying degrees of success in correctly classifying the specimens into categories. For the ATR-FTIR spectral dataset, the weakest approach for classification was PCA-LDA with 49.9% of the population data correctly classified. Six PCs were used as they provided enough variance (>90%) without introducing unwanted noise and therefore arbitrary separation. The related scores plot (Fig. 2B) shows co-clustering between some of the categories, but also separation between the classes of 1983-1984 and 2012-2013.

GA-LDA was the best method for classification of the ATR-FTIR dataset with 92.3% of the sample correctly classified. SPA-LDA ranked second for classification proficiency (84.2%). Both approaches revealed segregation and a temporal progression between the different categories. Interestingly, both chemometric approaches identified a shift where the “1983-1984” category cluster is completely segregated from the “2012-2013” one.

The Raman spectral data analysis also revealed significant segregation between the different categories. The different chemometric methods had varying success rates in correctly classifying the data. PCA-LDA and SPA-LDA correctly classified 35.8%, while GA-LDA correctly classified 38.6% of the sample population. Despite its weaker classification attainment, Raman spectroscopy pointed to spectral regions representing similar biochemical entities to ATR-FTIR; for example Amide I, Amide III, collagen and more importantly changes involving DNA/ RNA nucleotide bases and backbone. The markedly reduced variability exhibited by Raman spectroscopy in comparison to ATR-FTIR may be due to the area of tissue interrogated for the acquisition of each spectrum with each technique. The larger surface area sampled by the ATR probe (≈250 µm × 250 µm) has an averaging effect
which in this case may be advantageous as it delivers information on the biochemical signature over multiples of cells within the same histological region. Raman on the other hand acquires spectra from a much smaller area and therefore is affected more by micro topographical variations. Nevertheless, the two techniques are potentially complementary, highlighting variability within similar biomolecular regions.

The hypothesis that the chemo-molecular make-up of the prostate gland has changed within one generation is supported by the biospectroscopic techniques employed in this study. The prostate tissues used originated from procedures to treat BPH, which is influenced by nutritional variations including alcohol, vegetables and red meat (Kristal et al., 2008; Lagiou et al., 1999). BPH also has potential causal relationships with features of metabolic syndromes like diabetes, hypertension, obesity, high insulin and low HDL-cholesterol (Giovannucci et al., 1994; Hammarsten et al., 1998; Ozden et al., 2007). These relationships may be determined by genetic or epigenetic events that develop due to hormone-driven events or chemical exposures causing the formation of DNA adducts (John et al., 2009; F. L. Martin, 2013). Both ATR-FTIR and Raman spectral analysis highlighted marked trans-generational variation in the spectral regions containing DNA and RNA bands (≈ 1,000-1,490 cm⁻¹) involving nucleic acids, phosphate and deoxyribose modifications. This may point towards alterations that have occurred through chemical genotoxicity or through epigenetic modification of chromatin structure (Lu & Thompson, 2012). Also interesting is that SPA and GA algorithms identified wavenumbers indicating variability within the protein region involving amino acid conformational changes in C-O, C-H and N-H. This could be due to post-translational modifications related to genetic and/or epigenetic changes evident within the DNA/RNA spectral regions. Interestingly, the featured spectral areas may point
towards a genetic or epigenetic alterations with the variation becoming more pronounced as the period between sample acquisitions increases. Although the small population analysed by immunohistochemistry does not allow statistically significant results more samples from the 1983-1984 class showed significant methylation than from the 2012-2013 class. Global demethylation of the genome in parallel with CGI hypermethylation of particular genes with tumour-suppressor function associated with progression to prostatic cancer (Hoque, 2009).

This study was performed using prostate tissues taken from TURP procedures. Although H&E parallel sections of the tissue blocks used for spectroscopy did not show any complicating diathermy effect, this might also need to be taken into account. We tried to select a homogeneous population for our sampling. All men were between 60 and 69 y old. Age is the most important predictor of prostate cancer and its incidence rate increases sharply from 144/100,000 to 500/100,000 for men over the age of 65 (Kirby et al., 2010). We sampled a population that varied by 10 y in age in order to increase our sample size. The related confounding variability may have affected our results.

All samples in our study were free from prostate cancer. Approximately 10 to 20% of TURP procedures result in the incidental detection of invasive disease (Ornstein et al., 1997). Therefore a big portion of individuals with “silent” prostate cancer may have been excluded from the tested sample. The main limitation of the study is the lack of information regarding the actual lifestyle of our cohort. We unfortunately could not retrieve information on body mass index, weight, diet and alcohol consumption for all individuals. Also, we could not retrieve from their notes, relative comorbidities, for example diabetes or hypertension. What we wanted to test though was if there is any variability within prostate tissue with time of tissue
collection independent of other variables; therefore, knowing associated risk factors may have caused the introduction of unwanted bias to our study.

**Conclusion**

Prostate-related population diversity has not been significantly addressed to date. With this study we attempted to discover spectroscopic alterations that would classify prostate tissue from TURP procedures for BPH according to the year the operation was undertaken. We endeavoured to identify prostate variability that may be related to lifestyle changes that have happened within one generation\textsuperscript{13-15}.

Utilising two spectroscopic technologies coupled with three chemometric techniques, we observed significant discrimination of the prostate samples according to their year of collection. Also evident was complete segregation of the prostate tissues collected from two different generations nearly 30 y apart as well as progression through the years. Lifestyle changes during the studied generation have been extensively documented. Their association with changes in prostate tissue from individuals suffering from BPH is indicated by our study.

More extensive research in this field is required to assess the ability of vibrational spectroscopy to identify the existence of variations in prostate tissue with time. A study that extends over several generations, say from the 1920s to the present, may unearth further alterations in the biochemical composition of the prostate gland. These alterations may harbour biomarkers associated with the increase in prostatic cancer incidence linked to a Westernised lifestyle adaptation. This in turn may assist the identification of lifestyle adjustments for the prevention of prostate cancer.
Acknowledgments

Kássio M.G. Lima thanks the CNPq (The National Council for Scientific and Technological Development, Brazil) for his Postdoctoral Fellowship (Ref. 246742/2012-7). Guilherme Post Sabin (independent scholar, Cachoeirinha, Rio Grande do Sul, Brazil) is acknowledged for coding the matlab tool to remove cosmic rays for Raman spectra

References


Figures

Figure 1. Prostate anatomy, sample preparation and ATR-FTIR or Raman spectroscopy. A) Prostate anatomy illustrating the different histological zones. TURP removes part of the transition zone. B) Low-E slide containing a prepared sample. C) Micrograph of a prostate sample as visualised during Raman spectroscopy. D) H&E stained section for histological comparison and to ensure no diathermy artefacts contaminate the sample. E) Unprocessed ATR-FTIR spectral dataset (x-axis: wavenumbers (cm\(^{-1}\)), y-axis: absorbance) F) Unprocessed Raman spectral dataset. (x-axis: wavenumbers (cm\(^{-1}\)), y-axis: absorbance)
Figure 2. Processing of the ATR-FTIR and Raman derived spectral datasets for all categories by PCA-LDA. A) Pre-processed ATR-FTIR spectral dataset. B) Scores (DF1 × DF2 × DF3) plot calculated by PCA-LDA. C) Loadings plot derived from PCA-LDA. D) Prepossessed Raman spectral dataset. E) Scores (DF1 × DF2 × DF3) plot calculated by PCA-LDA. F) Loadings plot derived PCA-LDA.
Figure 3. Processing of the ATR-FTIR and Raman spectral datasets for all categories by SPA-LDA. A) Twenty three wavenumbers selected by the SPA-LDA model for the ATR-FTIR spectral dataset. B) Graph representing the power calculation used to identify the optimum number of wavelengths used for SPA. C) Scores (DF1 × DF2 × DF3) plot calculated by SPA-LDA. D) Seventeen wavenumbers selected by SPA-LDA model for the Raman spectral dataset. E) Graph representing the power calculation used to identify the optimum number of wavelengths to be used for SPA. F) Scores (DF1 × DF2 × DF3) plot calculated by SPA-LDA.
Figure 4. Processing of the ATR-FTIR and Raman datasets for all classes by GA-LDA A) Thirty two wavenumbers selected by GA-LDA model for the ATR-FTIR dataset. B) Graph representing the power calculation used to identify the optimum number of wavelengths to be used for SPA. C) Scores (DF1 × DF2 × DF3) plot calculated by using the variables selected by GA-LDA from ATR-FTIR spectra obtained from prostate tissues segregated into seven categories. D) Forty nine wavenumbers selected by GA-LDA model for the Raman spectral dataset. E) Graph representing the power calculation used to identify the optimum number of wavelengths to be used for SPA. F) Scores (DF1 × DF2 × DF3) plot calculated by using the variables selected by GA-LDA.
Figure 5: Processing of the ATR-FTIR- and Raman-derived spectral datasets for categories: 1983-1984 and 2012-2013 by PCA-LDA. A) Pre-processed ATR-FTIR spectral dataset. B) Scores (DF1 × DF2) plot calculated by PCA-LDA. C) Loadings plot derived from PCA-LDA. D) Prepossessed Raman spectral dataset. E) Scores (DF1 × DF2) plot calculated by PCA-LDA. F) Loadings plot derived PCA-LDA.
**Figure 6:** Processing of the ATR-FTIR and Raman spectral datasets for categories: 1983-1984 and 2012-2013 by SPA-LDA.  
A) Four wavenumbers selected by the SPA-LDA model for the ATR-FTIR dataset.  
B) Graph representing the power calculation used to identify the optimum number of wavelengths used for SPA.  
C) Scores (DF1 × DF2) plot calculated by SPA-LDA.  
D) Three wavenumbers selected by SPA-LDA model for the Raman spectral dataset.  
E) Graph representing the power calculation used to identify the optimum number of wavelengths to be used for SPA.  
F) Scores (DF1 × DF2) plot calculated by SPA-LDA.
Figure 7: Processing of the ATR-FTIR and Raman spectral datasets for categories: 1983-1984 and 2012-2013 by GA-LDA A) Seventeen wavenumbers selected by GA-LDA model for the ATR-FTIR spectral dataset. B) Graph representing the power calculation used to identify the optimum number of wavelengths to be used for SPA. C) Scores (DF1 × DF2) plot calculated by using the variables selected by GA-LDA from ATR-FTIR spectra obtained from prostate tissues segregated into seven classes. D) Fourteen wavenumbers selected by GA-LDA model for the Raman spectral dataset. E) Graph representing the power calculation used to identify the optimum number of wavelengths to be used for SPA. F) Scores (DF1 × DF2) plot calculated by using the variables selected by GA-LDA.
**Figure 8:** Immunohistochemistry of samples form the 1983-1984 and 2013-2014 classes. **A)** 8 out of 10 samples from the 1983-1984 class showed intense global methylation. **B)** 6 out of 10 samples from the 2012-2013 class showed intense global methylation. The rest of the samples showed moderate methylation in all cellular types.
A biospectroscopic analysis of human prostate tissue obtained from different time periods points to a trans-generational alteration in spectral phenotype

Georgios Theophilou\textsuperscript{1,2}, Kássio M.G. Lima\textsuperscript{1,3}, Matthew Briggs\textsuperscript{2}, Pierre L. Martin Hirsch\textsuperscript{1,2}, Helen F. Stringfellow\textsuperscript{2}, Francis L. Martin\textsuperscript{1*}

\textsuperscript{1}Centre for Biophotonics, LEC, Lancaster University, Lancaster LA1 4YQ, UK
\textsuperscript{2}Department of Obstetrics and Gynaecology, Central Lancashire Teaching Hospitals NHS Foundation Trust, Preston, UK
\textsuperscript{3}Institute of Chemistry, Biological Chemistry and Chemometrics, Federal University of Rio Grande do Norte, Natal 59072-970, RN-Brazil

\textsuperscript{*}Corresponding Author: Prof Francis L Martin PhD, Centre for Biophotonics, LEC, Lancaster University, Lancaster LA1, 4YQ, UK; Tel.: +44(0) 1524 510206; Email: f.martin@lancaster.ac.uk
Table S1: Number of training, validation and prediction specimens (or spectra) in each year of collections for prostatic tissues from the FTIR spectral data.

<table>
<thead>
<tr>
<th>Year of collection</th>
<th>Set Training</th>
<th>Validation</th>
<th>Prediction</th>
</tr>
</thead>
<tbody>
<tr>
<td>1983-1984</td>
<td>140</td>
<td>30</td>
<td>30</td>
</tr>
<tr>
<td>1988-1989</td>
<td>170</td>
<td>40</td>
<td>40</td>
</tr>
<tr>
<td>1993-1994</td>
<td>141</td>
<td>30</td>
<td>30</td>
</tr>
<tr>
<td>1998-1999</td>
<td>120</td>
<td>30</td>
<td>30</td>
</tr>
<tr>
<td>2003-2004</td>
<td>150</td>
<td>30</td>
<td>30</td>
</tr>
<tr>
<td>2008-2009</td>
<td>140</td>
<td>30</td>
<td>30</td>
</tr>
<tr>
<td>2012-2013</td>
<td>144</td>
<td>32</td>
<td>32</td>
</tr>
</tbody>
</table>
**Table S2:** Number of training, validation and prediction specimens (or spectra) in each year of collections for prostatic tissues from the Raman spectral data.

<table>
<thead>
<tr>
<th>Year of collection</th>
<th>Set Training</th>
<th>Validation</th>
<th>Prediction</th>
</tr>
</thead>
<tbody>
<tr>
<td>1983-1984</td>
<td>130</td>
<td>30</td>
<td>30</td>
</tr>
<tr>
<td>1988-1989</td>
<td>170</td>
<td>40</td>
<td>40</td>
</tr>
<tr>
<td>1993-1994</td>
<td>140</td>
<td>30</td>
<td>30</td>
</tr>
<tr>
<td>1998-1999</td>
<td>129</td>
<td>25</td>
<td>25</td>
</tr>
<tr>
<td>2003-2004</td>
<td>148</td>
<td>30</td>
<td>30</td>
</tr>
<tr>
<td>2008-2009</td>
<td>140</td>
<td>30</td>
<td>30</td>
</tr>
<tr>
<td>2012-2013</td>
<td>150</td>
<td>30</td>
<td>30</td>
</tr>
</tbody>
</table>
Table S3: Principal segregating wavenumbers for all categories derived from the loadings (LD1) curve associated with PCA-LDA of the ATR-FTIR spectral dataset.

<table>
<thead>
<tr>
<th>Wavelength (cm$^{-1}$)</th>
<th>Biological fingerprint</th>
</tr>
</thead>
<tbody>
<tr>
<td>1227</td>
<td>PO$_2^-$ asymmetric (phosphate I)</td>
</tr>
<tr>
<td>1400</td>
<td>Symmetric stretching vibration of COO$^-$ group of fatty acids and amino acids</td>
</tr>
<tr>
<td>1574</td>
<td>C=N adenine</td>
</tr>
<tr>
<td>1624</td>
<td>Peak of nucleic acids due to the base carbonyl stretching and ring breathing mode</td>
</tr>
<tr>
<td>1674</td>
<td>Unassigned band</td>
</tr>
<tr>
<td>1720</td>
<td>C=O</td>
</tr>
</tbody>
</table>
**Table S4:** Selected wavelengths for SPA-LDA analysis for the ATR-FTIR dataset for all 7 classes. The model calculated that 23 wavelengths were needed for correct segregation without the introduction of noise.

<table>
<thead>
<tr>
<th>Wavelength (cm$^{-1}$)</th>
<th>Biological fingerprint</th>
</tr>
</thead>
<tbody>
<tr>
<td>968</td>
<td>Symmetric stretching mode of dianionic phosphate monoesters of phosphorylated proteins or cellular nucleic acids DNA</td>
</tr>
<tr>
<td>1018</td>
<td>$\nu$(CO), $\nu$(CC), $\delta$(OCH), ring (polysaccharides, pectin)</td>
</tr>
<tr>
<td>1053</td>
<td>$\nu_s$ CO-O-C O stretching coupled with C-O bending of the C-OH of carbohydrates Glycogen</td>
</tr>
<tr>
<td>1153</td>
<td>Stretching vibrations of hydrogen-bonding C-OH groups</td>
</tr>
<tr>
<td>1234</td>
<td>Composed of amide III as well as phosphate vibration of nucleic acids CH$_6$, 20 a,a’ rock</td>
</tr>
<tr>
<td>1315</td>
<td>Amide III band components of Collagen</td>
</tr>
<tr>
<td>1392</td>
<td>Carbon particle</td>
</tr>
<tr>
<td>1415</td>
<td>Deformation C-H, N-H, stretching C-N</td>
</tr>
<tr>
<td>1446</td>
<td>$\delta$(CH2), lipids, fatty acids $\delta$(CH) (polysaccharides, pectin)</td>
</tr>
<tr>
<td>1462</td>
<td>Paraffin</td>
</tr>
<tr>
<td>1489</td>
<td>In-plane CH bending vibration</td>
</tr>
<tr>
<td>1512</td>
<td>In-plane CH bending vibration from the phenyl rings CH in-plane bend</td>
</tr>
<tr>
<td>1539</td>
<td>Protein amide II absorption- predominately $\beta$-sheet of amide II</td>
</tr>
<tr>
<td>1562</td>
<td>Ring base</td>
</tr>
<tr>
<td>1593</td>
<td>C=N, NH$_2$ adenine</td>
</tr>
<tr>
<td>1620</td>
<td>Peak of nucleic acids due to the base carbonyl stretching and ring breathing mode</td>
</tr>
<tr>
<td>1631</td>
<td>Amide I region</td>
</tr>
<tr>
<td>1651</td>
<td>Amide I region</td>
</tr>
<tr>
<td>1666</td>
<td>C5O stretching vibration of pyrimidine base</td>
</tr>
<tr>
<td>1693</td>
<td>A high frequency vibration of an antiparallel b-sheet of amide I (the amide I band is due to in-plane stretching of the C=O band weakly coupled to stretching of the C-N and in-plane bending of the N-H bond)</td>
</tr>
<tr>
<td>1716</td>
<td>C=O thymine Amide I (arises from C=O stretching vibration) C=O stretching vibration of DNA and RNA C=O stretching vibration of purine base</td>
</tr>
<tr>
<td>1735</td>
<td>C=O stretching (lipids)</td>
</tr>
<tr>
<td>1797</td>
<td>Lipids</td>
</tr>
</tbody>
</table>
Table S5: Selected wavelengths for GA-LDA analysis for the ATR-FTIR dataset for all 7 classes. The model calculated that 32 wavelengths were needed for correct segregation without the introduction of noise.

<table>
<thead>
<tr>
<th>Wavelength (cm(^{-1}))</th>
<th>Biological fingerprint</th>
</tr>
</thead>
<tbody>
<tr>
<td>987</td>
<td>OCH(_3) (polysaccharides-cellulose)</td>
</tr>
<tr>
<td>999</td>
<td>Ring stretching vibrations mixed strongly with CH in-plane bending</td>
</tr>
<tr>
<td>1002</td>
<td>Unassigned band</td>
</tr>
<tr>
<td>1026</td>
<td>Carbohydrates peak for solutions Vibrational frequency of CH(_2)OH groups of carbohydrates (including glucose, fructose, glycogen, etc.) Glycogen</td>
</tr>
<tr>
<td>1029</td>
<td>O-CH(_3) stretching of methoxy groups</td>
</tr>
<tr>
<td>1072</td>
<td>Phosphate I band for two different C-O vibrations of Deoxyribose in DNA in disordering structure</td>
</tr>
<tr>
<td>1191</td>
<td>Deoxyribose</td>
</tr>
<tr>
<td>1199</td>
<td>Collagen Phosphate (P=O) band</td>
</tr>
<tr>
<td>1299</td>
<td>Deformation N-H cytosine</td>
</tr>
<tr>
<td>1303</td>
<td>Unassigned band</td>
</tr>
<tr>
<td>1350</td>
<td>Unassigned band</td>
</tr>
<tr>
<td>1353</td>
<td>Unassigned band</td>
</tr>
<tr>
<td>1365</td>
<td>Stretching C-O, deformation C-H, deformation N-H</td>
</tr>
<tr>
<td>1373</td>
<td>Stretching C-N cytosine, guanine</td>
</tr>
<tr>
<td>1381</td>
<td>δCH(_3) Stretching C-O, deformation C-H, deformation N-H</td>
</tr>
<tr>
<td>1388</td>
<td>Carbon particle</td>
</tr>
<tr>
<td>1392</td>
<td>Unassigned band</td>
</tr>
<tr>
<td>1404</td>
<td>CH(_3) asymmetric deformation</td>
</tr>
<tr>
<td>1415</td>
<td>Deformation C-H, N-H, stretching C-N</td>
</tr>
<tr>
<td>1458</td>
<td>δ(_{as})CH(_3) of collagen</td>
</tr>
<tr>
<td>1496</td>
<td>C=C, deformation C-H</td>
</tr>
<tr>
<td>1504</td>
<td>In-plane CH bending vibration from the phenyl rings</td>
</tr>
<tr>
<td>1512</td>
<td>In-plane CH bending vibration from the phenyl rings</td>
</tr>
<tr>
<td>1543</td>
<td>Amide II</td>
</tr>
<tr>
<td>1554</td>
<td>Ring base</td>
</tr>
<tr>
<td>1562</td>
<td>Ring base</td>
</tr>
<tr>
<td>1589</td>
<td>Ring C-C stretch of phenyl</td>
</tr>
<tr>
<td>1600</td>
<td>C=N cytosine, N-H adenine</td>
</tr>
<tr>
<td>1647</td>
<td>Amide I in normal tissues-for cancer is in lower frequencies</td>
</tr>
<tr>
<td>1708</td>
<td>C=O thymine</td>
</tr>
<tr>
<td>1720</td>
<td>C=O</td>
</tr>
<tr>
<td>1751</td>
<td>ν(C=C) lipids, fatty acids</td>
</tr>
</tbody>
</table>
Table S6: Principal segregating wavenumbers for categories: 1983-1984 and 2012-1013 derived from the loadings (LD1) curve associated with PCA-LDA of the ATR-FTIR spectral dataset.

<table>
<thead>
<tr>
<th>Wavelength (cm(^{-1}))</th>
<th>Biological fingerprint</th>
</tr>
</thead>
<tbody>
<tr>
<td>1231</td>
<td>Overlapping of the protein Amide III and the nucleic acid phosphate vibration</td>
</tr>
<tr>
<td>1400</td>
<td>Symmetric stretching vibration of COO(^-) group of fatty acids and amino acids</td>
</tr>
<tr>
<td>1447</td>
<td>Asymmetric CH(_3) bending of the methyl groups of proteins</td>
</tr>
<tr>
<td>1578</td>
<td>Ring C-C stretch of phenyl</td>
</tr>
<tr>
<td>1624</td>
<td>Peak of nucleic acids due to the base carbonyl stretching and ring breathing mode</td>
</tr>
<tr>
<td>1674</td>
<td>Unassigned band</td>
</tr>
</tbody>
</table>
Table S7: Selected wavelengths for SPA-LDA analysis for the ATR-FTIR dataset for classes 1983-1984 and 2012-2013. The model calculated that 4 wavelengths were needed for correct segregation without the introduction of noise.

<table>
<thead>
<tr>
<th>Wavelength (cm(^{-1}))</th>
<th>Biochemical fingerprint</th>
</tr>
</thead>
<tbody>
<tr>
<td>1504</td>
<td>In-plane CH bending vibration from the phenyl rings</td>
</tr>
<tr>
<td>1620</td>
<td>Peak of nucleic acids due to the base carbonyl stretching and ring breathing mode</td>
</tr>
<tr>
<td>1647</td>
<td>Amide I in normal tissues-for cancer is in lower frequencies</td>
</tr>
<tr>
<td>1728</td>
<td>C=O band</td>
</tr>
</tbody>
</table>


Table S8: Selected wavelengths for GA-LDA analysis for the ATR-FTIR dataset for classes 1983-1984 and 2012-2013. The model calculated that 17 wavelengths were needed for correct segregation without the introduction of noise.

<table>
<thead>
<tr>
<th>Wavelength (cm(^{-1}))</th>
<th>Biochemical fingerprint</th>
</tr>
</thead>
<tbody>
<tr>
<td>1049</td>
<td>C-O stretching coupled with C-O bending of the C-OH of carbohydrates, Glycogen</td>
</tr>
<tr>
<td>1053</td>
<td>νC-O &amp; δC-O of carbohydrates Shoulder of 1121cm(^{-1}) band, due to DNA</td>
</tr>
<tr>
<td>1253</td>
<td>Not assigned</td>
</tr>
<tr>
<td>1415</td>
<td>Deformation C-H, N-H, stretching C-N</td>
</tr>
<tr>
<td>1423</td>
<td>Not assigned</td>
</tr>
<tr>
<td>1500</td>
<td>In-plane CH bending vibration from the phenyl rings</td>
</tr>
<tr>
<td>1504</td>
<td>In-plane CH bending vibration from the phenyl rings</td>
</tr>
<tr>
<td>1512</td>
<td>Not assigned</td>
</tr>
<tr>
<td>1516</td>
<td>Amide II</td>
</tr>
<tr>
<td>1519</td>
<td>Not assigned</td>
</tr>
<tr>
<td>1527</td>
<td>Stretching C=N, C=C</td>
</tr>
<tr>
<td>1531</td>
<td>Modified guanine</td>
</tr>
<tr>
<td>1535</td>
<td>Stretching C=N, C=C</td>
</tr>
<tr>
<td>1539</td>
<td>Not assigned</td>
</tr>
<tr>
<td>1543</td>
<td>Amide II</td>
</tr>
<tr>
<td>1546</td>
<td>Amide II (δN-H, νC-N)</td>
</tr>
</tbody>
</table>
Table S9: Principal segregating wavenumbers for all categories derived from the loadings (LD1) curve associated with **PCA-LDA** of the **Raman** spectral dataset.

<table>
<thead>
<tr>
<th>Wavelength (cm(^{-1}))</th>
<th>Biological fingerprint</th>
</tr>
</thead>
<tbody>
<tr>
<td>1418</td>
<td>CH(_2) scissoring vibration (lipid band)</td>
</tr>
<tr>
<td>1457</td>
<td>Deoxyribose</td>
</tr>
<tr>
<td>1576</td>
<td>Nucleic acid mode</td>
</tr>
<tr>
<td>1657</td>
<td>Triglycerides (fatty acids)</td>
</tr>
<tr>
<td>1704</td>
<td>C=O stretching vibrations of cortisone</td>
</tr>
<tr>
<td>1739</td>
<td>Ester group</td>
</tr>
</tbody>
</table>
**Table S10**: Selected wavelengths for SPA-LDA analysis for the Raman dataset for all 7 classes. The model calculated that 17 wavelengths were needed for correct segregation without the introduction of noise.

<table>
<thead>
<tr>
<th>Wavelength (cm(^{-1}))</th>
<th>Biochemical fingerprint</th>
</tr>
</thead>
<tbody>
<tr>
<td>1000</td>
<td>Phenylalanine, Bound &amp; free NADH</td>
</tr>
<tr>
<td>1001</td>
<td>Phenylalanine</td>
</tr>
<tr>
<td>1004</td>
<td>Phenylalanine (of collagen) (v_s(C-C)), symmetric ring breathing, phenylalanine (protein assignment)</td>
</tr>
<tr>
<td>1062</td>
<td>C-C skeletal stretch random conformation</td>
</tr>
<tr>
<td>1109</td>
<td>Benzoid ring deformation</td>
</tr>
<tr>
<td>1244</td>
<td>Amide III</td>
</tr>
<tr>
<td>1294</td>
<td>Methylene twisting</td>
</tr>
<tr>
<td>1295</td>
<td>Methylene twisting</td>
</tr>
<tr>
<td>1306</td>
<td>CH(_3)/CH(_2) twisting or bending mode of lipid/ collagen CH(_3)/CH(_2) twisting, wagging &amp;/or bending mode of collagens &amp; lipids</td>
</tr>
<tr>
<td>1336</td>
<td>Polynucleotide chain (DNA purine bases) (\delta(CH3)) (\delta(CH2))twisting, collagen (protein assignment)</td>
</tr>
<tr>
<td>1373</td>
<td>T, A, G (ring breathing modes of the DNA/RNA bases)</td>
</tr>
<tr>
<td>1376</td>
<td>Unassigned band</td>
</tr>
<tr>
<td>1436</td>
<td>CH(_2) scissoring</td>
</tr>
<tr>
<td>1437</td>
<td>CH(_2) deformation</td>
</tr>
<tr>
<td>14551</td>
<td>CH(_2)CH(_3) deformation</td>
</tr>
<tr>
<td>1655</td>
<td>Amide I</td>
</tr>
<tr>
<td>1671</td>
<td>Amide I</td>
</tr>
</tbody>
</table>
Table S11: Selected wavelengths for GA-LDA analysis for the Raman dataset for all 7 classes. The model calculated that 49 wavelengths were needed for correct segregation without the introduction of noise.

<table>
<thead>
<tr>
<th>Wavelength (cm(^{-1}))</th>
<th>Biochemical fingerprint</th>
</tr>
</thead>
<tbody>
<tr>
<td>842</td>
<td>Glucose</td>
</tr>
<tr>
<td>845</td>
<td>Unassigned band</td>
</tr>
<tr>
<td>874</td>
<td>C-C stretching, hypro (collagen assignment)</td>
</tr>
<tr>
<td>892</td>
<td>Backbone, C-C skeletal</td>
</tr>
<tr>
<td>920</td>
<td>C-C stretch of proline ring/glucose/lactic acid C-C, praline ring (collagen assignment)</td>
</tr>
<tr>
<td>946</td>
<td>Unassigned band</td>
</tr>
<tr>
<td>965</td>
<td>Hydroxyapatite</td>
</tr>
<tr>
<td>967</td>
<td>Lipids</td>
</tr>
<tr>
<td>971</td>
<td>v(C-C) wagging</td>
</tr>
<tr>
<td>997</td>
<td>C-O ribose, C-C</td>
</tr>
<tr>
<td>998</td>
<td>v45(CC), observed in the spectra of single human RBC</td>
</tr>
<tr>
<td>1010</td>
<td>Unassigned band</td>
</tr>
<tr>
<td>1022</td>
<td>Glycogen</td>
</tr>
<tr>
<td>1067</td>
<td>Proline (collagen assignment)</td>
</tr>
<tr>
<td>1087</td>
<td>v1CO232, n3PO342, v(C-C) skeletal of acyl back-bone in lipid (gauche conformation)</td>
</tr>
<tr>
<td>1168</td>
<td>Lipids v(C=)δ(COH) (lipid assignment) n(C-C), carotenoid</td>
</tr>
<tr>
<td>1182</td>
<td>Cytosine, guanine, adenine</td>
</tr>
<tr>
<td>1185</td>
<td>Anti-symmetric phosphate vibrations</td>
</tr>
<tr>
<td>1201</td>
<td>Nucleic acids and phosphates Aromatic C-O and C-N</td>
</tr>
<tr>
<td>1251</td>
<td>Guanine, cytosine (NH2)</td>
</tr>
<tr>
<td>1265</td>
<td>Amide III</td>
</tr>
<tr>
<td>1271</td>
<td>Amide III</td>
</tr>
<tr>
<td>1310</td>
<td>CH3/CH2 twisting, wagging &amp;/or bending mode of collagens &amp; lipids</td>
</tr>
<tr>
<td>1342</td>
<td>G (DNA/RNA) CH deformation (proteins and carbohydrates)</td>
</tr>
<tr>
<td>1373</td>
<td>T, A, G (ring breathing modes of the DNA/RNA bases)</td>
</tr>
<tr>
<td>1405</td>
<td>ν,COO2 (IgG)</td>
</tr>
<tr>
<td>1421</td>
<td>A, G (ring breathing modes of the DNA/RNA bases)</td>
</tr>
<tr>
<td>1423</td>
<td>NH in-plane deformation</td>
</tr>
<tr>
<td>1457</td>
<td>Deoxyribose</td>
</tr>
<tr>
<td>1483</td>
<td>Unassigned band</td>
</tr>
<tr>
<td>1496</td>
<td>Unassigned band</td>
</tr>
<tr>
<td>1499</td>
<td>C=C stretching in benzenoid ring</td>
</tr>
<tr>
<td>1507</td>
<td>Cytosine</td>
</tr>
<tr>
<td>1518</td>
<td>ν(C=), porphyrin,</td>
</tr>
<tr>
<td></td>
<td>Carotenoid peaks due to C-C &amp; conjugated C5C band stretch</td>
</tr>
<tr>
<td>1560</td>
<td>Tryptophan</td>
</tr>
<tr>
<td>1575</td>
<td>Ring breathing modes in the DNA bases G, A (ring breathing modes of the DNA/RNA bases)</td>
</tr>
<tr>
<td>1629</td>
<td>Ca=Ca stretch Amide C=O stretching absorption for the b-form polypeptide films</td>
</tr>
<tr>
<td>1652</td>
<td>Lipid (C=C stretch)</td>
</tr>
<tr>
<td>1657</td>
<td>Fatty acids Amide I (collagen assignment) Triglycerides (fatty acids)</td>
</tr>
<tr>
<td>1600</td>
<td>Amide I</td>
</tr>
<tr>
<td>1666</td>
<td>Collagen</td>
</tr>
<tr>
<td>1673</td>
<td>Amide I</td>
</tr>
<tr>
<td>1700</td>
<td>ν(C=O)OH (amino acids aspartic &amp; glutamic acid)</td>
</tr>
<tr>
<td>1710</td>
<td>One of absorption positions for the C=O stretching vibrations of cortisone</td>
</tr>
<tr>
<td>1729</td>
<td>Ester group</td>
</tr>
<tr>
<td>1733</td>
<td>One of absorption positions for the C=O stretching vibrations</td>
</tr>
<tr>
<td>1741</td>
<td>Ester group</td>
</tr>
<tr>
<td>1745</td>
<td>ν(C=O), phospholipids Triglycerides (fatty acids) ν(C=O) (polysaccharides, pectin)</td>
</tr>
</tbody>
</table>
Table S12: Principal segregating wavenumbers for categories: 1983-1984 and 2012-1013 derived from the loadings (LD1) curve associated with PCA-LDA of the Raman spectral dataset.

<table>
<thead>
<tr>
<th>Wavenumber (cm(^{-1}))</th>
<th>Biological fingerprint</th>
</tr>
</thead>
<tbody>
<tr>
<td>1419</td>
<td>Ester group</td>
</tr>
<tr>
<td>1459</td>
<td>Deoxyribose (\delta(CH_2))</td>
</tr>
<tr>
<td>1567</td>
<td>Unassigned band</td>
</tr>
<tr>
<td>1654</td>
<td>C=C stretch &amp; the Amide I bands, Amide I</td>
</tr>
<tr>
<td>1709</td>
<td>C=O stretching vibrations of cortisone</td>
</tr>
<tr>
<td>1742</td>
<td>Ester group</td>
</tr>
</tbody>
</table>
**Table S13:** Selected wavelengths for **SPA-LDA** analysis for the **Raman** dataset for classes 1983-1984 and 2012-2013. The model calculated that 3 wavelengths were needed for correct segregation without the introduction of noise.

<table>
<thead>
<tr>
<th>Wavelength (cm(^{-1}))</th>
<th>Biochemical fingerprint</th>
</tr>
</thead>
<tbody>
<tr>
<td>891</td>
<td>Saccharide band (overlaps with acyl band)</td>
</tr>
<tr>
<td>1001</td>
<td>Symmetric ring breathing mode of phenylalanine</td>
</tr>
<tr>
<td>1295</td>
<td>Not assigned</td>
</tr>
</tbody>
</table>

**Table S14:** Selected wavelengths for **GA-LDA** analysis for the **Raman** dataset for classes 1983-1984 and 2012-2013. The model calculated that 14 wavelengths were needed for correct segregation without the introduction of noise.

<table>
<thead>
<tr>
<th>Wavelength (cm(^{-1}))</th>
<th>Biochemical fingerprint</th>
</tr>
</thead>
<tbody>
<tr>
<td>861</td>
<td>Phosphate group</td>
</tr>
<tr>
<td>899</td>
<td>Monosaccharides (β-glucose), (C-O-C) skeletal mode</td>
</tr>
<tr>
<td>920</td>
<td>C-C stretch of proline ring/glucose/lactic acid C-C, praline ring (collagen assignment)</td>
</tr>
<tr>
<td>921</td>
<td>Not assigned</td>
</tr>
<tr>
<td>971</td>
<td>ν (C-C) wagging</td>
</tr>
<tr>
<td>1049</td>
<td>Glycogen</td>
</tr>
<tr>
<td>1100</td>
<td>C-C vibration mode of the gauche-bonded chain</td>
</tr>
<tr>
<td>1204</td>
<td>Amide III &amp; CH(_2) wagging vibrations from glycine backbone &amp; proline side chains, collagen</td>
</tr>
<tr>
<td>1206</td>
<td>Hydroxyproline, tyrosine (collagen assignment)</td>
</tr>
<tr>
<td>1261</td>
<td>Not assigned</td>
</tr>
<tr>
<td>1365</td>
<td>Tryptophan</td>
</tr>
<tr>
<td>1447</td>
<td>CH(_2) bending mode of proteins &amp; lipids</td>
</tr>
<tr>
<td>1496</td>
<td>Not assigned</td>
</tr>
<tr>
<td>1596</td>
<td>Not assigned</td>
</tr>
</tbody>
</table>
Table S15. Methylation studies for classes of 1983-1984 and 2012-2013. An intensity score from 0 to 3 was given to each cellular type along with the percentage of cells showing that intensity.

<table>
<thead>
<tr>
<th>Class/ Sample</th>
<th>Epithelial cells</th>
<th>Basal cells</th>
<th>Stromal cells</th>
<th>Vascular cells</th>
</tr>
</thead>
<tbody>
<tr>
<td>1983-1984</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>Percentage %</td>
<td>100</td>
<td>100</td>
<td>100</td>
</tr>
<tr>
<td>Intensity score</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>1983-1984</td>
<td>Percentage %</td>
<td>100</td>
<td>100</td>
<td>100</td>
</tr>
<tr>
<td>Intensity score</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>1983-1984</td>
<td>Percentage %</td>
<td>100</td>
<td>100</td>
<td>100</td>
</tr>
<tr>
<td>Intensity score</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>1983-1984</td>
<td>Percentage %</td>
<td>100</td>
<td>100</td>
<td>100</td>
</tr>
<tr>
<td>Intensity score</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>1983-1984</td>
<td>Percentage %</td>
<td>100</td>
<td>100</td>
<td>100</td>
</tr>
<tr>
<td>Intensity score</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>1983-1984</td>
<td>Percentage %</td>
<td>100</td>
<td>100</td>
<td>100</td>
</tr>
<tr>
<td>Intensity score</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>1983-1984</td>
<td>Percentage %</td>
<td>100</td>
<td>100</td>
<td>100</td>
</tr>
<tr>
<td>Intensity score</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>1983-1984</td>
<td>Percentage %</td>
<td>95</td>
<td>95</td>
<td>100</td>
</tr>
<tr>
<td>Intensity score</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>2012-2013</td>
<td>Percentage %</td>
<td>100</td>
<td>100</td>
<td>100</td>
</tr>
<tr>
<td>Intensity score</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>2012-2013</td>
<td>Percentage %</td>
<td>100</td>
<td>100</td>
<td>100</td>
</tr>
<tr>
<td>Intensity score</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>2012-2013</td>
<td>Percentage %</td>
<td>100</td>
<td>100</td>
<td>100</td>
</tr>
<tr>
<td>Intensity score</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>Intensity score</td>
<td>2012-2013 Percentage %</td>
<td>2012-2013 Intensity score</td>
<td>2012-2013 Percentage %</td>
</tr>
<tr>
<td>---</td>
<td>----------------</td>
<td>------------------------</td>
<td>---------------------------</td>
<td>------------------------</td>
</tr>
<tr>
<td>3</td>
<td>2</td>
<td>100</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>4</td>
<td>2</td>
<td>100</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>5</td>
<td>2</td>
<td>100</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>6</td>
<td>1</td>
<td>100</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>7</td>
<td>2</td>
<td>100</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>8</td>
<td>2</td>
<td>100</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>9</td>
<td>3</td>
<td>100</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>10</td>
<td>3</td>
<td>100</td>
<td>3</td>
<td>3</td>
</tr>
</tbody>
</table>
References


Chapter 5

Spatial and temporal age-related spectral alterations in benign human breast tissue

Georgios Theophilou, Simon W. Fogarty, Júlio Trevisan, Rebecca Strong, Kelly Heys, Imran I Patel, Helen F. Stringfellow, Pierre L. Martin-Hirsch, Francis L Martin*
Centre for Biophotonics, LEC, Lancaster University, Lancaster LA1 4YQ, UK

Contribution:
I retrieved the synchrotron spectral datasets.
I performed the data analysis.
I have written the first draft of the manuscript,

Professor F. L Martin ........................................
G. Theophilou ........................................
Spatial and temporal age-related spectral alterations in benign human breast tissue

Georgios Theophilou, Simon W. Fogarty, Júlio Trevisan, Rebecca Strong, Kelly Heys, Imran I Patel, Helen F. Stringfellow, Pierre L. Martin-Hirsch, Francis L Martin*

Centre for Biophotonics, LEC, Lancaster University, Lancaster LA1 4YQ, UK

Correspondence to: Prof Francis L Martin, Centre for Biophotonics, LEC, Lancaster University, Lancaster LA1 4YQ, UK; Email: f.martin@lancaster.ac.uk; Tel.: +44(0)1524 510206
Summary

Epidemiological evidence suggests that cancer with a pathogenesis attributable to exogenous carcinogenic agents may appear decades after initial exposure. Environmental factors including lifestyle and diet have been implicated in the aetiology of cancers of the breast cancer. Breast tissue undergoes continuous molecular and histological changes from the time of thelarche to menopause and thereafter. These alterations are both cyclical and longitudinal, and can be influenced by several environmental factors including exposure to oestrogen through pregnancy or breastfeeding. Available research of latent stages of breast carcinogenesis has been limited to when hyperplastic lesions are present. Investigations to identify a biomarker of commitment to disease in normal breast tissue are hindered by the molecular and histological diversity of disease free-breast tissue. Benign tissue from reduction mammoplasties provides an opportunity to study biochemical differences between women of similar ages as well as alterations with advancing age. Herein, synchrotron radiation-based Fourier-transform infrared microspectroscopy (SR-FTIR) was used to examine the terminal ductal lobular epithelium (TDLU), intra- and inter-lobular epithelium to identify spatial and temporal changes within these areas. Principal component analysis (PCA) followed by linear discriminant analysis of mid-infrared spectra revealed unambiguous inter-individual as well as age-related differences in each histological compartment interrogated. Moreover, exploratory PCA of luminal and myoepithelial cells within the TDLU indicated the presence of specific cells, potentially stem cells. Understanding alterations within benign tissue may assist in the identification of alterations within the latent pre-clinical stage of breast cancer.
Introduction

The clinical manifestation of breast cancer may be considered to be the final expression of a complex sequential process that begins with exposure to a causative agent (Grover & Martin, 2002). Tumour formation involves temporal alterations in genetic morphology or expression, which directly or indirectly disturbs normal cellular regulation of proliferation and growth inhibition, leading to malignancy (Pedraza-Farina, 2006). The period from the initiating event to tumour formation is termed the “latency period” (Rothman, 1981). This definition implies that cancers in which environmental exposures play a role arise several years after exposure (Grover & Martin, 2002; Land et al., 2003). This latency period may be of different length depending on the type, timing and length of exposure as well as inherent predisposition to the particular type of cancer (Armenian, 1987).

Exposure to carcinogens will certainly vary significantly between individuals as will their response to such an exposure (Johnson et al., 2011; Kahlenborn et al., 2006). Factors that predispose women to a risk of breast cancer include early menarche, late menopause, nulliparity or delayed parity and, use of contraception and hormone replacement therapy (Broeders & Verbeek, 1997; Wohlfahrt et al., 1999). All these characteristics are associated with increased exposure to oestrogen. It is now accepted that “Westernized” lifestyle either through immigration or adoption of Western diet are likely causative factors for breast or other hormone-dependent cancers (Grover & Martin, 2002; Yager & Davidson, 2006).

Little is known regarding the molecular changes that may develop before the appearance of pre-clinical and clinical breast cancer (Allred et al., 2001). Changes that appear at the initiation stage or during the latency period may provide useful biomarkers for early identification of women at risk of developing breast cancer.
Also, these changes may be temporary, regressive, permanent or progressive (Allred et al., 2001; Arpino et al., 2005). Biomarkers that could identify lesions with a low risk of progression towards malignancy or even the chance of regression would be advantageous (Buyse et al., 2006; Esserman et al., 2011; O’donoghue & Esserman, 2013). Women exhibiting high-risk alterations might be encouraged to make appropriate alterations in their lifestyle to try and “rectify” these changes (Blackburn & Wang, 2007).

In order to enable understanding of pathological processes involved in carcinogenesis, we first need to be able to identify physiological differences within breast tissue between similarly aged women as well as alterations that occur with increasing age. The areas wherein these variations are most interesting are within the terminal ductal lobular unit (TDLU) along with the supporting intra- and inter-lobular stroma (Figure 1a). These areas are thought to be responsible for cancer initiation process (Adriance et al., 2005; Allred et al., 2001; Ronnov-Jessen et al., 1996). The TDLU consists of terminal ductules ending in acini, bounded by luminal epithelial cells, which are surrounded by myoepithelial cells (Figure 1b). TDLUs have different compositions depending on their developmental stage from pre-puberty to menopause (Figure 1c). The pre-pubertal “simple” TDLU consists of one central ductule with three or four branches. After menarche, the TDLU’s morphology depends on the stage of the menstrual cycle with luminal cells growing in size as the cycle progresses from the follicular to the luteal phase. During pregnancy and lactation, the TDLU hypertrophies and remains in a similar state to the luteal phase.
Figure 1  (a) Diagram representing a lactiferous duct with an aggregation of TDLUs. The space between acini within the TDLU is occupied by intra-lobular stroma while the space between different TDLUs is occupied by inter-lobular stroma. (b) Diagrammatic interpretation of the terminal ductal lobular unit (TDLU), illustrating the types of cells that surround an acinus. (c) Diagrammatic description of the developmental progression of the TDLU with advancing age. (i) pre-pubertal, (ii) pubertal, (iii) Mature, (iv) lactating, (v) post-menopausal. (d) Example of principal component analysis and linear discriminant analysis (PCA-LDA) of TDLU, intra-lobular stroma and inter-lobular stroma. The tissue section was selected from a subject within the 40-49 age group. (1) Parallel sections were stained with H&E for histological representation. (2) Numbered gridiron overlays were added to micrographs of the sections to aid spectral selection. (3) Image maps were produced from which spectra were extracted. (4) Class means representing spectral differences between different cell types. (5) PCA-LDA scores plots of different cells where each spectral point is derived from the average of 5 IR spectra.
Postmenopausally, the lobule has fewer ductules and a denser intralobular stroma. With advancing age, the TDLU undergoes complete atrophy but the branching duct tree remains forever. In cancer, the micro-architecture of the TDLU is disturbed.

Breast tissue from reduction mammoplasties provides an opportunity to study spatial and temporal variation that may exist within the TDLU and surrounding areas of the mammary gland when there is no evidence of malignant or pre-malignant changes. The cancer risk in this population is comparable or marginally reduced relative to the general population (M. H. Brown et al., 1999; Hassan & Pacifico, 2012).

In the pursuit for the discovery of these differences, biospectroscopy may play an important role as it can identify structural alterations of molecules within cells according to their chemical bonds (H. Y. Holman et al., 2000; Mourant et al., 2003). This technique has been used to detect molecular alterations associated with cancer in various tissues (Gajjar et al., 2012; Gajjar et al., 2013; German et al., 2006; Pichardo-Molina et al., 2007; Tobin et al., 2004). Recent studies have also examined its potential in identifying biomarkers that can be used in screening for cancer (Harris et al., 2009; Walsh et al., 2007).

Synchrotron radiation-based Fourier-transform infrared (SR-FTIR) microspectroscopy is superior to conventional FTIR spectroscopy for the detection of molecular biomarkers. The reason for this is that a synchrotron emits a collimated light beam that is more brilliant than that of a bench-top spectrometer. This provides an excellent signal-to-noise ratio (SNR) that is 1000 times greater to that of conventional IR sources and allows spatial resolutions as small as 10µm (L. M. Miller & Dumas, 2006; Tobin et al., 2004).
Interrogation of biological tissues by IR spectroscopy results in thousands of spectra due to the complex chemical composition of cells. The vast amount of data obtained by a typical spectroscopic experiment may be analysed using multivariate analysis. This aims to simplify this information for logical visual representation. Two of the commonly utilised multivariate analysis approaches are: principal component analysis (PCA) and linear discriminant analysis (LDA) or a combination of both (Martin F L et al., 2007).

This study aims to identify spectral differences in breast tissue of women of similar ages as well as changes with time. This could be the first step towards the recognition of the origins of breast cancer in the path that leads to its prevention.

Materials and Methods

Sample preparation

Human breast tissue was obtained from eleven patients undergoing reduction mammoplasty for indications other than breast-related pathology. Consent was taken with ethical approval according to the Declaration of Helsinki. Five individuals were aged 20 to 29 years old, three were in their thirties and three in their forties. The breast tissue samples obtained were formalin-fixed and paraffin-embedded. Ten-µm-thick tissue sections were floated onto 1cm × 1cm BaF₂ slides (Photox Optical Systems). These were de-waxed by serial immersion in sequential fresh xylene baths (×3) for five minutes and washed in an acetone bath for another five minutes. Resulting samples were allowed to air dry and then placed in a desiccator until processing.
4-µm-thick parallel tissue sections were stained with H&E. These assisted with correct identification of the different cell types when overlaying mapping grids on the micrographs visualized through the SR-FTIR microscope (Figure 1d).

**Synchrotron radiation-based FTIR micro-spectroscopy**

Spectral images were acquired using a Bruker Vertex 80v spectrometer coupled to a Bruker Hyperion 3000 microscope containing a mercury cadmium telluride detector cooled with liquid nitrogen, on the 22 IR beam-line at the Diamond Light Source Ltd, UK ([www.diamond.ac.uk](http://www.diamond.ac.uk)). Spectra were collected in transmission mode via a 36× objective lens employing an aperture of 10 µm × 10 µm with a step size of 10 µm intervals, 256 co-additions were acquired; it was ensured that maps were generated within an acquisition time of ~6 h. Background spectra were taken every 10 spectra to compensate for beam and atmospheric alterations. Spectra were then converted to absorbance using OPUS 8 software from Bruker Optics.

**Spectral pre-processing**

Absorbance spectral images were converted to suitable digital files (.txt) for input to Matallab software. Computational analysis was carried out using in-house written scripts for Matlab (Trevisan et al., 2013). The wavenumber regions entered were between 4,000cm⁻¹ and 600cm⁻¹. Spectra were then cut to include the regions between 1,800-900cm⁻¹ as this is the spectral region associated with biologically active molecules (Movasaghi et al., 2008). They were smoothed using the first derivative Savitzky-Golay filter, rubber-band-like baseline corrected, vector normalised and normalized to Amide I (1650cm⁻¹). This did not affect the original spectral resolution.
**Computational analysis**

The pre-processed spectral data were explored using PCA. This was carried out as an unsupervised technique using the first 10 principal components (PCs). Generally, the first 10 PCs account for approximately 99% of the variance within a sample population, without introducing excessive noise (Martin F L et al., 2007). The output from PCA was inputted into LDA. LDA is a supervised technique that maximizes the inter-category variance. The software analyses 90% of the data while using 10% of the data to train itself. This process is repeated 10 times in a cyclical fashion, so that all data are used for both analysis and system supervision. The statistical significance of each PC and LDA analysis contributing to class segregation was determined by the *ANOVA* test in Graphpad 7 when more than 2 classes were present. For visualization purposes, scores plots and loadings curves were generated.

Scores plots, derived from PCA-LDA allow visualization of a spectrum as a single point, whose coordinates are its scores on a number of axes. This aims to simplify visualization of potential differences between the particular classes as well as identify co-clustering of similar spectral signatures. Loading curves allow identification of distinguishing wavenumbers when comparing classes with each other. The *x*-axis represents wavenumbers from 900 to 1800 cm\(^{-1}\). The *y*-axis represents the absorbance coefficient. The highest peaks and troughs on this axis identify the wavenumbers that are most responsible for separation between selected classes. The 6 greatest absorbance coefficient deviations were selected. These wavenumbers were then compiled onto tables alongside tentative assignments. The resulting tables point to biochemical entities responsible for class segregation.
Results

Using SR-FTIR microspectroscopy, samples from 11 patients were interrogated. A raster scan approach was applied to include terminal ductal lobular epithelial (TDLU) regions within the specimens. This allowed a high SNR with apertures close to the diffraction limit. The spatial resolution was $10\mu m \times 10\mu m$. The high resolution allowed separation of histologically different layers of the TDLU and surrounding regions. A micrograph of the involved areas with overlaid markers was used for identification of the specific regions of which spectra were acquired. Following interrogation, spectral differences were apparent between the location-derived spectra (Figure 1d). These differences allowed classification into inter-lobular stroma, intra-lobular stroma and TDLU, which was further classified to myoepithelial cellular layer and luminal cellular layer. Correlation with parallel H&E tissue sections ensured correct selection of different cellular classes.

In total there were $n=539$ spectra within the inter-lobular stroma class, $n=442$ spectra within the intra-lobular stroma class and $n=591$ spectra within the TDLU. Within the TDLU there were $n=155$ spectra from the luminal cellular layer and $n=436$ from the myoepithelial layer. The very large number of individual spectra would impede easy visualization on scores plots. Therefore, for most classes every two, three or five spectra in chronological order were averaged (Table 1). The different areas were interrogated individually to identify putative spatial and temporal differences within our population. Furthermore, the myoepithelial and luminal cellular layers within the TDLU were separately examined to identify cells within these layers with divergent spectral signatures that may represent potential biomarkers of disease (Petersen et al., 2003).
**Inter-individual variations**

The infrared spectra derived from the synchrotron maps were extracted and assigned to their pre-defined histological location (TDLU, inter-lobular stroma and intra-lobular stroma). IR spectra from each individual woman represented a separate class. These classes were allocated to their designated age group: 3\textsuperscript{rd}, 4\textsuperscript{th} and 5\textsuperscript{th} decade of life. Each location was analysed separately for each individual within every age group to investigate the existence inter-individual variation and identify the responsible wavenumbers. PCA-LDA cascade analysis was used to reduce each spectrum to a single point at the same time maximizing intra-category homogeneity and inter-category heterogeneity. Derived scores plots identified clustering of spectra taken from the same individual independent of location (Figure 2). Moreover they identified clear separation between individuals of similar ages for all histological classes. This separation was highly significant and associated with different wavenumbers for different ages and histological locations.
Figure 2. PCA-LDA scores plots showing separation in every histological compartment tested between individuals in all age groups. The X-axis represents LD1 and the Y-axis LD2. Each spectral point is derived from PCA-LDA of the average of 5 spectra. TDLU, Intra-lobular stroma and Inter-lobular stroma were examined separately. Each individual was processed as an independent category.

Figure 3 shows the loadings curves containing the 6 principal discriminating wavenumbers for each category. Table 2 illustrates the discriminating wavenumbers for each location and age category alongside their corresponding molecules. Different spectral signatures were responsible for maximum segregation between individuals for the different histological classes. It was noted that 2 molecules were responsible for inter-individual variation in all age groups: RHS Amide I (1,630 cm⁻¹) was responsible for segregation between the TDLU’s and DNA/ RNA (1,080 cm⁻¹) was responsible for segregation between intra-lobular stromata of individuals of all ages.
Figure 3. Loadings (LD1) plots derived from PCA-LDA comparing individuals from each age group for differences within each histological compartment. The X-axis is cm$^{-1}$ and the Y-axis represents absorbance coefficient. The five wavenumbers contributing to the most segregation were derived from the points furthest away from the X-axis.
Temporal variations

Following the observation that the wavenumbers segregating TDLUs and intra-LS between individuals carried over all age groups, IR spectra were examined to determine the existence of variation between ages for the selected histological areas. This was performed using the same parameters and method of analysis (PCA-LDA using the first 10 PCs). The pre-processed spectral data was classified according to their allocated histological locations and age groups (instead of individuals within a group). Resulting scores plots (Figure 4) revealed clustering within age groups but not as much segregation as was expected. Despite the overlap, the classes were significantly segregated ($P < 0.0001$).

**Figure 4.** (A) PCA-LDA scores plots showing separation between age groups in every histological compartment. The X-axis represents LD1 and the Y-axis LD2. (B) Loadings plots (LD1) showing the principal discriminating wavenumbers in graphical form. The X-axis is cm$^{-1}$ and the Y-axis represents absorbance coefficient.
Corresponding loadings plots (Figure 4) identified the 6 wavenumbers responsible for maximum segregation between age groups for each histological location. Another spectral signature corresponding to RHS Amide I \((1,456\text{cm}^{-1})\) was one of the segregating wavenumbers for the TDLU area. Therefore, RHS Amide I was responsible for segregation between individuals as well as between age groups within the TDLU location. In the case of Intra-LS, DNA/RNA \((1,080 \text{ cm}^{-1})\) alterations were not found to be responsible for segregation between age groups. Instead the principal segregating wavenumber was associated with ring base \((1,554\text{cm}^{-1})\).

All age groups were segregated in every histological class by glycogen \((1040\text{cm}^{-1})\).

**Inter-individual and temporal variations within the TDLU**

Concentrating on the TDLU, spectra derived from point maps acquired \textit{via} the 10 \(\mu\)m \(\times\) 10 \(\mu\)m beam aperture were extracted and assigned to myoepithelial or luminal cells. Spectra from these histological areas were classified initially according to individuals within age groups and were analysed using a similar technique to the previous analyses to identify wavenumbers responsible for segregation between these more specific areas. There was some segregation between the different cellular classes, although with considerable overlap. This segregation was nevertheless significant \((P < 0.0002)\) in all age groups (Figure 5).
Figure 5. (A) PCA-LDA 1D scores plots showing some separation between luminal and myoepithelial cells within the TDLU for the different age groups. The X-axis represents LD1 (B) loadings plots showing the principal discriminating numbers. The X-axis is cm\(^{-1}\) and the Y-axis represents absorbance coefficient. (C) Tables of discriminating wavenumbers with their corresponding biochemical markers.

After the confirmation that luminal and myoepithelial cells are segregated, each location was separately analysed to identify wavenumbers responsible for segregation between age groups within them (Figure 6). Amide I again featured as one of the significant segregating wavenumbers in both luminal and myoepithelial cell layers. Interestingly, within the myoepithelial layer one of the principal discriminating wavenumber was for PO\(_2^-\) (1,094 cm\(^{-1}\)). This wavenumber has been found to be associated with stem cells in a variety of tissues including TDLUs in a previous study performed by this group and other studies (Brown et al., 1999; Patel et al., 2014).
Figure 6. (A) PCA-LDA scores plots showing separation between different age groups for the TDLU compartment with a table of the principal discriminating wavenumbers and corresponding biochemical markers. (B) PCA-LDA scores plots showing separation between different age groups for the luminal and myoepithelial cell regions. The X-axis represents LD1 and the Y-axis LD2. (C) Tables show the 5 major segregating wavenumbers with their corresponding biochemical markers.
**Exploratory PCA within the myoepithelial and luminal layers**

To further investigate the existence of spectral discriminating factors within the myoepithelial and luminal layers, IR spectra taken via the 10 µm × 10 µm aperture were extracted from the image maps of all individuals. These spectra were not averaged. They underwent exploratory PCA which is an unsupervised analytic technique using the first 10 PCs, which account for about 99% of variance. Three-dimensional scores plots were extracted representing the 3 first PCs (Figure 7a). Each point on the resulting scores plots represented a single point on the image map. Both plots identified spectral points that segregated from the clustered spectra. These “outliers” were particularly obvious on the PC3 axis in the case of myoepithelial cells. Loadings plots for PC3 identified the wavenumbers responsible for the separation of “outliers” from the clustered spectra (Figure 7b). In a similar way PC1 was most responsible for spectral points segregating from clustered spectra on PCA scores plots for luminal cells. Loadings plots for PC1 were used to extract the first 6 segregating wavenumbers. The major discriminating factor in both cases was Amide I.
Figure 7. (A) 3-D scores plots of exploratory PCA using 10 first PCs examining luminal and myoepithelial cells separately to identify outliers which may represent residing stem cells. (B) Corresponding loadings plots for PC1 for luminal cells and PC3 for myoepithelial cells identify the major segregating wavenumbers. (C) Tables show the 6 major segregating wavenumbers with their corresponding biochemical markers.

Discussion

The purpose of this study was to identify spectral differences within normal breast tissue of women of similar ages as well as differences that occur with age. Using the IR radiation beam of a synchrotron facility, normal breast tissues from eleven healthy women were examined. Specifically the areas interrogated were terminal ductal lobular units (TDLU) and surrounding intra-lobular and inter-lobular stroma.
Morphological and molecular alterations within these areas have been shown to be directly associated with carcinogenesis in breast (Allred et al., 2001; Yang et al., 2013). Spectroscopic analysis of these areas in healthy individuals may reveal specific molecular causes of the vast heterogeneity that exists within breast tissues. It may also reveal alterations that predate carcinogenesis as defined by primary pre-cancerous changes including hyperplasia, premalignant changes and carcinoma in situ (Yang et al., 2013). The SR-FTIR spectra obtained were from the mid-IR region of the spectrum from 900 to 1800 cm\(^{-1}\) within which the spectral signatures of most biomolecules can be identified (Movasaghi et al., 2008).

Computational analysis of the obtained spectra identified significant differences between individuals of similar ages. Some of the discriminating wavenumbers responsible for this variation were responsible for similar variation in all age groups. Namely, Amide I was responsible for inter-individual variability within the TDLU while DNA/RNA (O-P-O stretching) was responsible for separation of INTRA-LS in all age categories. Other spectral bio-molecular signatures were only associated with separation of specific histological locations only in one age group (Table 2). Also certain wavenumbers could identify inter-individual variations within age groups while others could identify inter-individual variation in all age groups, illustrating the vast heterogeneity that exists. Many factors contribute to this heterogeneity. They include: previous history of breast cancer, positive family history with or without BRCA mutations, nulliparity, late parity, high body mass index, use of hormonal contraception or hormone replacement therapy and menopausal status (Helmrich et al., 1983; Kelsey et al., 1981; Kelsey et al., 1993).

When analysing spectral signatures of the same histological areas for temporal variations between the three defined age groups it was noted that there was co-
clustering of spectra from the same age groups and segregation between groups for all areas (Figure 4). Histologically breast tissue undergoes several changes with age. These changes start at thelarche with the branching of the lactiferous ducts and reach maturity in puberty with the formation of the adult TDLU. Changes continue in a cyclical fashion with menstrual cycles. Pregnancy and lactation cause hyperplasia of the TDLU. After menopause the TDLU involutes but the pattern of involution is different for nulliparous and parous women. The morphological and functional differentiation of the mammary epithelium is directly dependent on systemic hormones (mainly oestrogen and progesterone) but also by local signalling by the adjacent stroma (Howlett & Bissell, 1993). Our study was able to identify potential spectral alterations that may be associated with age-related histological appearances. Some of these alterations were unique to particular histological areas (TDLU, Inter-LS or Intra-LS) while some were responsible for separation between ages in all areas (Table 3). Spectral alterations within the TDLU and surrounding stroma, may provide evidence to support age related changes in the functional interaction between these areas. Furthermore these interaction alterations may be associated with the initial steps in breast carcinogenesis (Howlett & Bissell, 1993; Hu & Polyak, 2008; Ronnov-Jessen & Bissell, 2009; Ronnov-Jessen et al., 1996).

It is widely accepted that that the first morphological changes associated with cancer occur in the bi-layered TDLU epithelium (Vargo-Gogola & Rosen, 2007). It has also been hypothesized that micro-anatomical changes predating pre-cancerous changes reside in the same areas (Jonine D. Figueroa, 2012). In order to further examine the role of IR spectroscopy in identifying such changes within each layer of the TDLU, IR spectra taken from this location were reclassified into luminal and myoepithelial cell categories. These spectra were analysed using multivariate analysis.
as before. When investigating inter-individual variations, the resultant 1-D scores plots revealed significant separation between the two layers in all age groups. Related loadings curves identified the responsible wavenumbers and their corresponding molecules are presented in Figure 6. These spectral variations may be associated with morphological differences that are specific to a particular cellular layer of the TDLU rather than the whole TDLU structure. They may be used to pinpoint the cells associated with the increase in breast cancer risk.

The same two layers were examined for spectral variation between age groups. 2-D scores plots identified some segregation between age groups for both cellular layers. Amide I featured as a major discriminatory molecule for both cell types as well as for the whole TDLU. Unsupervised exploratory PCA of the luminal and myoepithelial cells identified aberrant spectral signatures in both layers. These signatures may represent multi-potent or uni-potent stem cells responsive are either age related or hormone dependent alterations. Indeed, there is expanding evidence that FTIR is capable of identifying stem cells in several tissues including cornea, epidermis and intestine (Fogarty et al., 2013; Kelly et al., 2010; Patel et al., 2012; Walsh et al., 2008; Walsh et al., 2009). Similarly stem cells within the mammary gland may undergo continuous differentiation under hormonal or micro-environmental influences and account on the diversity of breast tissue (Howlett & Bissell, 1993). They may also represent pluripotent progenitor cells whose abnormal differentiation under oxidative stress in adjacent stroma can lead to carcinogenesis (Howlett & Bissell, 1993; Hu & Polyak, 2008; Medema & Vermeulen, 2011; Vargo-Gogola & Rosen, 2007).
Conclusion

With this study we demonstrated that ST-FTIR micro-spectroscopy coupled with multivariate computational analysis might be to identify discriminating biomarkers for both inter-individual and temporal variation within breast tissue. We also demonstrated the histological locations where this variation potentially occurs. This is particularly important as it demonstrates the potential interplay between external environmental influences, endogenous hormonal control and micro-environmental communication with the epithelial cells of the TDLU similarly to other tissues (Cunha & Ricke, 2011; Hu & Polyak, 2008; Purandare et al., 2013; Ricke et al., 2012; Ronnov-Jessen et al., 1996). Although specific molecular changes associated with the vast variability encountered in the mammary gland remains elusive spectral imaging is able to identify classes of molecules that may be used in the search for biomarkers associated with the initiation of breast disease. In the future, FTIR spectroscopy may be able to track molecular changes within particular cell layers involved in disease to produce a database of related biomarkers (Trevisan et al., 2012). Moreover spectroscopy involves non-destructive procedures that do not produce oxidative radiation, as is the case with mammography (Yaffe & Mainprize, 2011). Therefore it can be used to obtain molecular profiles of cell populations in situ (Patel et al., 2012). Identification of such biomarkers may be followed by the application of biospectroscopic techniques in clinical practice. Spectral alterations associated with increased risk of breast cancer in a healthy population, may be used as biomarkers in potential population-screening programs without the need to identify high-risk individuals for inclusion to such a program. Further research in the field is required to assess spectroscopic applications in the search for biomarkers for screening for breast
disease. Once specific biomarkers are established similar techniques may translate into clinical practice for the evaluation biomarkers within live breast tissue.

**ACKNOWLEDGEMENTS** We gratefully acknowledge Lancashire Teaching Hospitals NHS Trust for facilitating these studies. The Sciences and Technologies Facilities Council is thanked for grant support to access the Diamond synchrotron facility.
**Tables**

**Table 1.** Spectra for each histological compartment for most patients were averaged by a factor of 2, 3 or 5 to aid visualisation in scores plots.

<table>
<thead>
<tr>
<th>Patient/Averaging factor</th>
<th>TDLU</th>
<th>Inter-LS</th>
<th>Intra-LS</th>
<th>Myo-epithelial</th>
<th>Luminal</th>
</tr>
</thead>
<tbody>
<tr>
<td>P1</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>P2</td>
<td>3</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>P3</td>
<td>3</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>P4</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>P5</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>P6</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>P7</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>P8</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>P9</td>
<td>2</td>
<td>1</td>
<td>5</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>P10</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>P11</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>
Table 2. Principal segregating wavenumbers between individuals of each age group for each histological compartment. All wavenumbers were tested for significance using the ANOVA test with $P < 0.0001$

<table>
<thead>
<tr>
<th>Wavenumber/Area</th>
<th>TDLU Intra-LS</th>
<th>Inter-LS</th>
<th>TDLU Intra-LS</th>
<th>Inter-LS</th>
<th>TDLU Intra-LS</th>
<th>Inter-LS</th>
</tr>
</thead>
<tbody>
<tr>
<td>1,780*</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1,709: C=O</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1,693: Amisde I (C=O)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1,671: Amide I (C=C)</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1,650: Amide I (C=O stretch and C-N stretch)</td>
<td></td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1,630: RHS Amide I (C=O stretch and C-N stretch)</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1,550: Amide II (N-H bend and C-N stretch)</td>
<td></td>
<td>X</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1,470-1,473: (CH2 bend)</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1,451: (CH3)</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1,375: (C-N: cytocine,</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Value</td>
<td>guanine</td>
<td>1,232: (PO2)</td>
<td>X</td>
<td>X</td>
<td>1,225: DNA (O-P-O asymmetric stretch)</td>
<td>X</td>
</tr>
</tbody>
</table>
Table 3. Principal segregating wavenumbers between age groups for each histological compartment. All wavenumbers were tested for significance using the Anova test with $P < 0.0001$.

<table>
<thead>
<tr>
<th>Wavenumber</th>
<th>TDLU</th>
<th>Intra-LS</th>
<th>Inter-LS</th>
</tr>
</thead>
<tbody>
<tr>
<td>1,680: Amide I</td>
<td></td>
<td></td>
<td>X</td>
</tr>
<tr>
<td>1,650: Amide I (C=O stretch and C-N stretch)</td>
<td>X</td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>1,550: Amide II (N-H bend and C-N stretch)</td>
<td>X</td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>1,495: C=C</td>
<td></td>
<td></td>
<td>X</td>
</tr>
<tr>
<td>1,456-1,460: CH₃ methyl groups</td>
<td>X</td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>1,375: C-N cytosine guanine</td>
<td></td>
<td></td>
<td>X</td>
</tr>
<tr>
<td>1,238-1,242: ν(PO₂)</td>
<td>X</td>
<td></td>
<td>X</td>
</tr>
<tr>
<td>1,225: DNA (O-P-O asymmetric stretch)</td>
<td>X</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1,219: (PO₂)</td>
<td></td>
<td></td>
<td>X</td>
</tr>
<tr>
<td>1,200: RNA Ribose (C-O stretching)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1,080: DNA/RNA (O-P-O stretching)</td>
<td>X</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1,061: C-O deoxyribose</td>
<td></td>
<td></td>
<td>X</td>
</tr>
<tr>
<td>1,040: Glycogen (C-O-H bond)</td>
<td>X</td>
<td>X</td>
<td></td>
</tr>
</tbody>
</table>
References


Chapter 6

Discussion
Discussion:

Overview:

In recent years powerful instrumentation and analytic techniques have allowed the utilisation of biospectroscopic methods in cancer research. The associated techniques have been assessed for their potential in identifying cancer in tissues and bio-fluids. They have also been used to classify cancer subtypes and locate stem cells that may be associated with cancer initiation.

This thesis investigates ATR-FTIR, synchrotron based FTIR and Raman spectroscopy for their potential in identifying differences within seemingly healthy populations. These novel methods may be able to extract biomarkers that may classify a section of a population that has a “high” risk of developing specific types of cancer.

Biospectroscopy is shown to be a valid adjunct or even an alternative to well established techniques like histopathology and immunochemistry and has been used in vitro and in vivo (Buschman et al., 2000). It’s potential to be used in the clinical setting assisting in both diagnostics and therapeutics is slowly coming to light. Further research will inevitably improve the quality of related hardware and software and may provide a simple alternative to routine investigations and procedures being performed today.

Carcinogenesis is a complex multifactorial process and no single research methodology is sufficient in identifying events that are associated with commitment to cancer development or progression. This thesis explores the role of biospectroscopy in carcinogenesis research. “Systems biology” is the combination of data from several high throughput technologies aiming to examine the functional progression to tumorigenesis, delving into genetics, epigenomics, transcriptomics,
proteomics, metabolomics, toponomics (Hood et al., 2004). Biospectroscopy may play a central role in systems biology as it can search for related chemical biomarkers independent on the functional compartment where they may reside.

This thesis also investigates spectral similarities and differences that may exist within the functional histological areas of benign breast tissue. Women of similar age groups were examined for inter-individual spectral variations, as did women of differing age groups. Despite the vast variability between breast tissues of women of even similar ages, spectral biomarkers were identified that remain unchanged with advancing age but distinguish between women within the same age group. These biomarkers may potentially harbour alterations that may be utilised for research for the very early stages of commitment to cancer development. The same project attempted the localisation of stem cells from which carcinogenesis may generate within the TDLU.

It is well documented that the risk of prostatic cancer varies significantly amongst men of different ethnicities and also depends on geographical position (Ferlay et al., 2015). As with any hormone dependent cancers it is believed that the increased risk of prostatic cancer noticed in the Western hemisphere is associated with the adoption of a “Westernised lifestyle” (Leitzmann & Rohrmann, 2012). Within this thesis we set out to determine whether benign prostatic tissue has altered spectroscopically within the past 30 years. Prostatic tissue of similarly aged men was interrogated by ATR-FTIR and Raman spectroscopy, which combined with chemometric analysis identified transgenerational alterations within this tissue. Areas associated with epigenetic variation which itself is dependent on environmental influences were most prominently related to these alterations. These findings were further validated by the use of immunohistochemical techniques that identified differences in the methylation status of prostate tissues harvested one generation apart.
The conventional approach to the diagnosis of ovarian cancer has inherent limitations due to the low sensitivity and specificity of non-invasive tests such as radiological and biochemical tests. Therefore histopathological diagnosis is usually required at the time of surgery to distinguish between benign, borderline and malignant ovarian tumours. The characterisation of these tumours is essential for planning the type of operation performed and adjuvant therapy. ATR-FTIR has the potential to be utilised within surgical theatres to either examine excised ovarian tumours in real time or even \textit{in vivo}. This thesis describes a pilot study examining the value of ATR-FTIR in discriminating benign, borderline and malignant ovarian tissues and also differentiating between the different subtypes of ovarian carcinomas \textit{in vitro}. The application of ATR-FTIR in conjunction with a classification machine was able to correctly classify the different ovarian tumours opening the road for more extensive research in fresh tissue or \textit{in vivo} using handheld ATR devises.

This thesis provides evidence that biospectroscopy has great potential to be used in research to identify biomarkers associated with early carcinogenesis. It has identified minute differences in benign mammary tissue that persist with age and where biomarkers associated with increase risk for future cancer may reside. It has also shown that benign prostatic tissue has undergone trans-generational alterations that may be influenced by lifestyle changes within the past thirty years. It was also able to prove that biospectroscopic techniques are able to classify ovarian tumours.

Although the work done encourages the progression of the field of biospectroscopy it is not without limitations. For example, as a proof of principle, diagnostic segregation of ovarian cancer from control as well as separation of carcinoma subtypes by ATR-FTIR is possible in laboratory-based experiments. However, its application in the clinical setting using fresh tissue or tissue \textit{in-vivo} will require further research. Also,
while both ATR-FTIR and Raman techniques have successfully identified trans-generational differences within prostatic tissue the results will need validation in larger and more diverse population groups. More-over this research has identified inter-individual and temporal variation within benign mammary tissue but has not directly associated this variation with an increased risk of breast cancer. The specific spectral regions contributing to the identified variability will need to be further examined comparing breast cancer with controls to pinpoint relevant biomarkers.

Overall, this thesis addresses four general topics regarding biospectroscopic methodology. Firstly, it explores the role of biospectroscopy within a systems biology approach towards identification of carcinogenic events. Secondly, it investigates variability in benign breast tissue of similarly aged women and groups of women of different ages identifying specific spectral regions responsible for that variability. At the same time it suggests the specific location of stem cells within the TDLU. Thirdly this work investigates trans-generational differences within benign prostatic tissue suggesting epigenetic variability. Lastly it provides evidence that biospectorscopy can correctly classify ovarian tissue distinguishing between benign, borderline malignant and cancerous tumours.

The next sections describe each of the studies carried out towards the completion of this thesis. They include a summary of the hypotheses, findings and conclusions and the rationale for inclusion in this work. Lastly a generalized conclusion discusses suggestions for future work that may explore further the translatory potential of biospectroscopy from a laboratory-based research tool to a technology that can be applied in clinical research and ultimately used in front-line medical practice.
Extracting biomarkers of commitment to cancer development: potential role of vibrational spectroscopy in systems biology.

Systems biology is a new approach to understanding physiological and pathophysiological processes (Kitano, 2002). Systems biology can bring together several high throughput methodologies for identifying elements of cellular function that may be associated with tumour formation. It has an important role to play in research concerning carcinogenesis due to the complexity and diversity of the mechanisms that govern initiation and progression to malignancy (Sarasin, 2003; Vineis et al., 2010). This approach collects data from the sciences of genomics, epigenomics, transcriptomics, proteomics, metabonomics and toponomics to answer questions about the pathways that may be followed by cells individually and as a group for the initiating steps for tumourigenesis to occur. New technologies have allowed rapid and precise mapping of cellular processes associated with tumourigenesis identifying several biomarkers on the way. The mechanisms governing alterations that lead to cancer are multifactorial and do not only involve genetics or epigenetics, for example. Therefore relying on a technology that is capable to only explore one of the functional elements of the cellular pathophysiology would always be incomplete. This can be explained by the needle in the haystack allegory, where each of the sciences mentioned above is looking for a needle in a different haystack in a field of haystacks. The place of biospectroscopy within such an allegory is that this technology has the potential to point towards the haystack or haystacks where the needle is probably located. Biospectroscopy methodology at this time is not accurate enough to identify specific biomarkers for carcinogenesis but can point towards particular cellular compartments where they may reside, for example within
the proteomic or metabolic make-up of the cells. This would allow more directed data analysis by other techniques for extracting such biomarkers. The datasets produced by biospectroscopic instrumentation can be made compatible with current online datastores and search machines that are used for systems biology type research. The resulting data combinations along with powerful analytic techniques that can explore the data and validate their results internally and externally have the potential to answer specific questions regarding the mechanisms behind malignant phenotypes and may find common pathways for the development of this group of diseases.

**Biospectroscopy for the diagnostic segregation of human ovarian tumours.**

Current methods for the diagnosis of ovarian cancer have significant limitations stemming from the heterogeneity of ovarian tumours and the inability to define accurately their malignant potential without surgical excision and histolopathological evaluation (Brun et al., 2000). Although biochemical tumour markers in combination with radiological investigations have met some success in identifying individuals whose ovarian tumours have an increased risk of being malignant, clinical intra-operative judgement and histopathological diagnosis inform planning the radicality of surgery and subsequent chemotherapeutic management. Previous research suggested that ATR-FTIR on blood and serum, coupled with variable selection techniques has the ability to segregate ovarian cancer stages (Lima et al., 2015). We wanted to evaluate the same technique for its potential use on ovarian tissue rather than biofluids. Our aim was to explore the diagnostic potential of biospectroscopic methods on ovarian tissue and characterise specific molecular alterations associated with ovarian tumour types. We utilised several analytical methods including PCA-LDA, SPA-LDA,
GA-LDA, multivariate control charts, spectral ratios and classification machines to achieve segregation of these tumours with excellent results.

Ovarian tissues were collected from a biobank with appropriate ethics approval and were de-waxed and dehydrated before ATR-FTIR spectral acquisition. The resulting spectral datasets were classified according to their histopathological diagnosis.

Different chemometric techniques achieved relatively accurate classification of ovarian tumour types into benign, borderline malignant and carcinomas. Interestingly lipid to protein ratio, phosphate to carbohydrate ratio and RNA to DNA ratios segregated ovarian carcinomas from borderline and benign tumours. Multivariate control charts are routinely used in quality control but their successful application in the classification of ovarian tumours may signify their value in cancer research.

An assortment of relevant variable selections techniques was combined to form a classification machine, which successfully categorised different ovarian carcinoma subtypes. This illustrated that for certain applications a combination of analytical methods is better than the utilisation of a single one and not exceedingly difficult to employ.

This trial provided evidence for the potential that biospectroscopic methods possess in correctly classifying ovarian tumours. ATR-FTIR spectroscopy should now be applied on fresh tissue in real time (intra-operatively) and compared with concurrent histopathology for validation of the accuracy of this method. With time handheld ATR-FTIR instrumentation may be used to attempt classification of ovarian tumours in vivo.
Identification of trans-generational variation in human benign prostatic tissue by biospectroscopic techniques.

Prostatic cancer incidence is consistently increasing with time (Parkin et al., 1999). This fact may be associated with environmental and lifestyle adaptations that have occurred over decades (Alberti, 2010). “Westernised lifestyle” which is associated with an increase in obesity, diabetes, metabolic syndrome and others may play a role in the increase in the incidence of hormone dependent cancers. We set out to identify the molecular markers of potential variation between benign prostatic tissues that were harvested from similarly aged men classified according to the year these tissues were harvested by TURP procedures.

Paraffin embedded tissue blocks \( n=156 \) were de-waxed, washed and processed by ATR-FTIR and Raman techniques. The resulting spectral datasets were distributed according to the year of collection and analyzed by PCA-LDA, SPA-LDA and GA-LDA. ATR-FTIR was better than Raman at correctly classifying these tissues according to the year of collection, possibly due to the larger area sampled from each specimen owing to the size of the diamond in contact with each sample. There was distinct segregation between prostatic tissue harvested more than 30 years apart, but also a progressive movement from one class to the other when viewing the associated scores plots. The spectral regions that contributed to this classification included areas involved with epigenetic variation, namely methylation. Immunohistochemistry was therefore performed to evaluate methylation differences between tissues separated by one generation with positive results.

This study points to environmental and lifestyle factors contributing to an increased incidence of prostatic cancer, albeit indirectly. Further research will be needed to
perform more specific epigenetic studies to identify specific genomic regions affected
by these alterations. The population used for this study extended from 1984 to 2013.
A population that extended several generations in the past would allow the
identification of potential alterations that have occurred at times when lifestyle was
more drastically different.

**Spatial and temporal age related spectral alterations in benign human breast
tissue: visualizing pre-initiation events in disease free tissue.**

Breast tissue varies significantly amongst women according to age, ethnicity, parity,
pregnancy, exogenous hormones, previous breast cancer, positive family history with
or without *BRCA* mutations and menopausal status (Broeders & Verbeek, 1997;
Yager & Davidson, 2006). This variability makes it extremely difficult to identify
biomarkers that may be indicative of an increased risk of developing cancer in the
future. Identifying spectral similarities and differences between benign breast tissues
in target histological areas, where breast cancer is believed to arise may be central in
identifying putative screening biomarkers for cancer. These areas involve the TDLU
and the surrounding stroma as histologically these areas are implicated in ductal
carcinoma in situ, which precedes overt carcinoma (Gudjonsson et al., 2005).
Identifying the molecules that account for the vast variability of the TDLU and
surrounding inter- and intra-lobular stroma, both between individuals and between
groups of individuals of different ages could allow the identification of biomarkers
that could be used in future screening for breast cancer.

Paraffin embedded tissue blocks (*n*=11) were de-waxed and washed and processed
using the synchrotron facilities at the Diamond Synchrotron (Oxford). Point spectra
were obtained that included the luminal, and myoepithelial layers of the TDLU and surrounding stroma. This allowed comparison between the sample population for the specific histological areas. It also allowed the exploration of these areas for the localization of putative stem cells. Abnormalities within these stem cells may contain the initiating steps towards breast carcinogenesis.

The resulting spectral datasets were analyzed using PCA-LDA. The analyses identified differences but also similarities that existed between individuals of similar ages. Some of this spectral similarity was carried through all age groups and may represent a constant molecular structure. This constant may warrant more research comparing benign breast tissue with cancerous concentrating on these spectral areas. Any identified variability here may be associated with cancer development. Therefore FTIR spectroscopy may be able to track molecular changes within particular cell layers involved in carcinogenesis to produce a database of related biomarkers that can be used in healthy population screening (Trevisan et al., 2012).
Conclusions:

Vibrational spectroscopy techniques for biological research have gained a lot of momentum in recent years. Their application has been assisted greatly by the creation of powerful chemometric analysis software. The fact that relatively simple processing techniques are required to examine both tissues and bio-fluids has allowed their use in laboratory-based research. Moreover spectroscopy involves non-destructive procedures and can be used on live subjects in some instances.

Biospectroscopy is able to identify specific molecules by detecting the precise chemical bond structure within them. Therefore different proteins, carbohydrates, lipids and nucleic acids can be described depending on their individual “spectral finger-print”. The spectral properties of cellular components can thus be examined for putative biomarkers for disease. Identification of such biomarkers may be followed by the application of bio-spectroscopical techniques in clinical practice.

The application of biospectroscopy in clinical practice is twofold; one function is to extract biomarkers associated with disease or increased risk of developing a disease. Another function is to classify populations according to their spectral differences. This allows classification of tissues and bio-fluids according to their histopathological and cytopathological phenotypes. Therefore, this group of technologies has the potential to assist in the identification and classification of disease including cancer.

This thesis has provided information on the value of biospectroscopy in cancer research. It has explored the use of associated techniques in the identification of cancer and the classification of cancer subtypes. It has also investigated the potential of biospectroscopy to extract biomarkers of variability in healthy individuals indirectly relating them to their risk for cancer development.
References


Complete List of references


http://www.nature.com/nprot/journal/v9/n8/abs/nprot.2014.110.html - supplementary-information


of different types of prostate cell. *Biophys J*, 90(10), 3783-3795. doi: 10.1529/biophysj.105.077255


304


Patel, II, Shearer D A, Fogarty S W, Fullwood N J, Quaroni L, Martin F L & Weisz J. (2014). Infrared microspectroscopy identifies biomolecular changes associated with chronic oxidative stress in mammary epithelium and stroma of breast tissues from


rapid linking of molecular changes to clinical endpoints. *Am J Pathol, 159*(6), 2249-2256. doi: 10.1016/S0002-9440(10)63075-1


advanced human prostate cancer through loss of expression. *Proc Natl Acad Sci USA*, 95(9), 5246-5250.


341

Appendix 1

Ethical approvals, consent forms and patient information leaflets
18 October 2010

Dr Helen Stringfellow
Lancashire Teaching Hospitals NHS Trust
Pathology, Royal Preston Hospital
Sharoe Green Lane
PR2 9HT

Dear Dr Stringfellow

Title of the Research Tissue Bank: Archival genito-urinary tissue, blood, urine and saliva collection
REC reference: 10/H0308/75
Designated Individual: Dr Timothy P Dawson

Thank you for your letter of 30 September 2010, responding to the Committee’s request for further information on the above research tissue bank and submitting revised documentation.

The further information was considered at the meeting of the Sub-Committee of the REC held on 15 October 2010. A list of the members who were present at the meeting is attached.

Confirmation of ethical opinion

On behalf of the Committee, I am pleased to confirm a favourable ethical opinion of the above research tissue bank on the basis described in the application form and supporting documentation as revised.

Duration of ethical opinion

The favourable opinion is given for a period of five years from the date of this letter and provided that you comply with the conditions set out in the attached document. You are advised to study the conditions carefully. The opinion may be renewed for a further period of up to five years on receipt of a fresh application. It is suggested that the fresh application is made 3-6 months before the 5 years expires, to ensure continuous approval for the research tissue bank.

Approved documents

The documents reviewed and approved at the meeting were:

<table>
<thead>
<tr>
<th>Document</th>
<th>Version</th>
<th>Date</th>
</tr>
</thead>
<tbody>
<tr>
<td>REC application</td>
<td>58126/139495/3/344</td>
<td>30 July 2010</td>
</tr>
<tr>
<td>Covering Letter</td>
<td>Dr Helen Stringfellow</td>
<td>30 July 2010</td>
</tr>
</tbody>
</table>

This Research Ethics Committee is an advisory committee to East of England Strategic Health Authority. The National Research Ethics Service (NRES) represents the NRES Directorate within the National Patient Safety Agency and Research Ethics Committees in England.
Licence from the Human Tissue Authority
Thank you for providing a copy of the above licence.

Research governance

A copy of this letter is being sent to the R&D office responsible for Royal Preston Hospital. You are advised to check their requirements for approval of the research tissue bank.

Under the Research Governance Framework (RGF), there is no requirement for NHS research permission for the establishment of research tissue banks in the NHS. Applications to NHS R&D offices through IRAS are not required as all NHS organisations are expected to have included management review in the process of establishing the research tissue bank.

Research permission is also not required by collaborators at tissue collection centres (TCCs) who provide tissue or data under the terms of a supply agreement between the organisation and the research tissue bank. TCCs are not research sites for the purposes of the RGF.

Research tissue bank managers are advised to provide R&D offices at all TCCs with a copy of the REC application for information, together with a copy of the favourable opinion letter when available. All TCCs should be listed in Part C of the REC application.

NHS researchers undertaking specific research projects using tissue or data supplied by a research tissue bank must apply for permission to R&D offices at all organisations where the research is conducted, whether or not the research tissue bank has ethical approval.

Site-specific assessment (SSA) is not a requirement for ethical review of research tissue banks. There is no need to inform Local Research Ethics Committees.

Statement of compliance

The Committee is constituted in accordance with the Governance Arrangements for Research Ethics Committees (July 2001) and complies fully with the Standard Operating Procedures for Research Ethics Committees in the UK.

After ethical review

Now that you have completed the application process please visit the National Research Ethics Service website > After Review

Here you will find links to the following:

a) Providing feedback. You are invited to give your view of the service that you have received from the National Research Ethics Service and the application procedure. If you wish to make your views known please use the feedback form available on the website.

b) Annual Reports. Please refer to the attached conditions of approval.

c) Amendments. Please refer to the attached conditions of approval.

This Research Ethics Committee is an advisory committee to East of England Strategic Health Authority
The National Research Ethics Service (NRES) represents the NRES Directorate within the National Patient Safety Agency and Research Ethics Committees in England
We would also like to inform you that we consult regularly with stakeholders to improve our service. If you would like to join our Reference Group please email referencegroup@nres.npsa.nhs.uk

10/H0308/75  Please quote this number on all correspondence

Yours sincerely

[Signature]

Dr Rowan Burnstein
Chair

E-mail: Nicky.Storey@eoe.nhs.uk

Enclosures:

List of names and professions of members who were present at the meeting and those who submitted written comments

Standard approval conditions

Copy to:

Dr Timothy Dawson
Lancashire Teaching Hospitals NHS Trust
Pathology, Royal Preston Hospital
Sharoe Green Lane
Preston
Lancashire
PR2 9HT

This Research Ethics Committee is an advisory committee to East of England Strategic Health Authority
The National Research Ethics Service (NRES) represents the NRES Directorate within the National Patient Safety Agency and Research Ethics Committees in England
Cambridgeshire 2 Research Ethics Committee

Attendance at Sub-Committee of the REC meeting on 15 October 2010

Committee Members:

<table>
<thead>
<tr>
<th>Name</th>
<th>Profession</th>
<th>Present</th>
</tr>
</thead>
<tbody>
<tr>
<td>Revd Dr Derek Fraser</td>
<td>Chaplain</td>
<td>Yes</td>
</tr>
<tr>
<td>Mrs Rebekah Ley</td>
<td>Assistant Director Medico-Legal and Patient Experience</td>
<td>Yes</td>
</tr>
<tr>
<td>Dr Ian Nimmo-Smith</td>
<td>Statistician</td>
<td>Yes</td>
</tr>
</tbody>
</table>

This Research Ethics Committee is an advisory committee to East of England Strategic Health Authority

The National Research Ethics Service (NRES) represents the NRES Directorate within the National Patient Safety Agency and Research Ethics Committees in England
Patient Information Sheet

Title of Project: Collection of Genito-Urinary Tissue and/or Blood / Urine / Saliva for Future Research Use - Cancer Research Tissue Bank

Name of Researchers: Dr H. Stringfellow, Dr C. Nicholson, Mr P. Martin-Hirsch, Mr P. Keating, Mr N. Wood, Mr S. Matanheila, Miss R. Blades, Mr P. Javie

We would like to invite you to take part in our research study. Before you decide we would like you to understand why the research is being done and what it would involve for you. One of our team will go through the information sheet with you and answer any questions you have. We'd suggest this should take about 30 minutes. Talk to others about the study if you wish. Ask us if there is anything that is not clear.

Thank you for reading this.

What is the purpose of this project?
We wish to collect and store ("bank") tissue samples (blood, urine, saliva and organ tissue) from patients with womb, ovary and prostate disorders to support future research into the cause, diagnosis, treatment and outcome of such disorders. Some of these studies may include genetic research aimed at understanding the way in which genes (molecules instructing cell division and growth) influence the behaviour of these disorders. Researchers in Lancashire Teaching Hospitals and other medical research organisations will be able to access the tissue, urine, saliva and blood collection subject to ethical approval.

Why have I been chosen?
You are being asked to take part in this project because you are being investigated for a disorder of ovary, womb or prostate. Your management includes surgery during which tissue is removed routinely for access, diagnosis or treatment. We would like to ask you whether you would be willing to allow this tissue and any urine, saliva and blood samples to be included in the research collection (tissue bank).

Do I have to take part?
NO, it is up to you to decide whether or not to take part. If you do decide to take part you will be given this information sheet to keep and be asked to read, understand and sign a consent form. If you decide to take part you are still free to withdraw at any time and without giving a reason. Your management and treatment will not be influenced in any way whether you wish to take part or not.
What will happen to me if I take part?

- Participating in this study by donating samples to the tissue bank will not affect your treatment in any way. The length of your operation and stay in hospital will not be affected and no additional surgery, procedures or appointments will be necessary.
- We will ask you to give us permission (signed consent) to include samples removed as part of your diagnostic or treatment procedures in our studies. It may involve taking saliva samples which is additional to tissue routinely removed for diagnostic or treatment purposes.
- Tissue will only be stored for research use once all diagnostic needs have been met.
- You may be asked to give blood or urine samples at various times during your treatment in order to check how your treatment is affecting you. If you give permission for a blood sample to be stored for research extra blood will be taken on some of these occasions, but no additional venepuncture (blood testing) will be required.
- We will ask you for permission to consult your medical records at Lancashire Teaching Hospitals (or other relevant medical records elsewhere) for some information relevant to your illness.
- This information will include your age, gender, type and site of surgery, pathology diagnosis, radiological (X-ray) features, your medical treatment and the response to treatment. This data will be stored on an NHS database designed to store research data. Access to this database will be restricted to those who need access and the Genito-urinary Tissue Bank Management Team will decide who has access.

If you agree to take part in this study, we will have all that we need for our research and will not need to contact you again. Blood, urine, saliva and archival tissue samples will be treated as gifts to research within the usual diagnostic archive, in accordance with Human Tissue Authority regulation.

What are the possible disadvantages and risks of taking part?
There are no disadvantages or risks to taking part.

What are the possible benefits of taking part?
There are no specific benefits to you directly, but the results of investigations using this tissue may help others with similar disorders. Research results obtained from your specimens are not going to provide any information that is clinically relevant to you directly.

What happens to tissue in the bank?
Tissue will be securely stored until allocated to an ethically approved project. During analysis some will be "used up". This allocation will go on until the supply is exhausted. Tissue is stored in locked freezers in the pathology department of Royal Preston Hospital. The person responsible for overseeing the tissue bank is Dr Helen Stringfellow, Consultant Pathologist.

Will my taking part in this study be kept confidential?
YES. If you agree to take part in this study, only the relevant information mentioned above will be extracted from your records. Any information about you released by the Tissue Bank will have removed personal details such as your name and address. The same will apply to the tissue samples used in any laboratory studies. The information and tissues will only be known by a research number, which will prevent researchers from knowing your identity.

What will happen if I don't want to carry on with the study?
You can contact your consultant and withdraw consent at any time. Any unused tissue or blood stored in the Tissue Bank will be transferred to the diagnostic archive and stored or disposed according to departmental diagnostic protocols. No further allocations will be made to research
projects. Data from previously allocated tissue, blood, urine or saliva may already exist and will remain associated with those projects anonymous to the researcher.

Will any genetic tests be done?
DNA derived from tissue may be examined for abnormalities, which may give information on the cause of a disorder. It will not produce results with a direct influence on you or your relatives.

What happens if something goes wrong?
The planned research will have no influence on your treatment. The banking of your tissue, urine, saliva or blood for research carries no risk. Regardless of this, if you wish to complain about any aspect of the way you have been approached or treated, the normal National Health Service complaints mechanisms will be available to you.

What will happen to the results of the research study?
Results will be presented at conferences and published as scientific papers, but you will not be identified in any report or publication. Results obtained from your specimens are unlikely to include information of immediate clinical relevance to you, but should anything helpful be found, this will be conveyed to your treatment team.

Who is organising and funding the research?
The genito-urinary tissue bank is organised by pathologists, surgeons, oncologists and researchers at Lancashire Teaching Hospitals. Funding for tissue banking can be obtained from NHS research funds, Charitable Funds and from ongoing project funding. None of the participants of the genito-urinary tissue banking project are being paid for their participation in the project. Future research studies utilising banked tissues will be funded by a variety of funds/charities.

Who has reviewed the study?
All research in the NHS is looked at by an independent group of people, called a Research Ethics Committee, to protect your interests. This study has been reviewed and given a favourable opinion by Cambridgeshire 2 Research Ethics Committee.

You will be given a copy of the information sheet and a signed consent form to keep.

Thank you very much for reading this information sheet.

Contacts for Further Information

Mr Pierre Martin-Hirsch, Consultant Surgeon, Royal Preston Hospital
e-mail: Pierre.martin-Hirsch@lthtr.nhs.uk
01772 524214

Dr Helen Stringfellow, Consultant Pathologist, Royal Preston Hospital
e-mail: helen.stringfellow@lthtr.nhs.uk
01772 522149

Mr Shyam Matanhelia, Consultant Urologist, Royal Preston Hospital
e-mail: shyam.matanhelia@LTHTR.nhs.uk
01772 522468

Dr Caroline Nicholson, Consultant Pathologist, Royal Preston Hospital
e-mail: caroline.nicholson@lthtr.nhs.uk
01772522140

Patient Information for use in Lancashire Teaching Hospitals: Gynaecological / Urological tumour tissue bank
Version 2, 30th September 2010  Page 3 of 3
PATIENT CONSENT FORM

Title of Project: Collection of Genito-Urinary Tissue and/or Blood/Urine/Saliva for Future Research Use – Cancer Research Tissue Bank

Name of Researchers: Dr H. Stringfellow, Dr C Nicholson, Mr P Martin-Hirsch, Mr P Keating, Mr N. Wood, Mr S Matanhelda, Miss R Blades, Mr P Javie

1. I ……………………….. confirm that I have read and understand the information sheet (Version 2, 30th September 2010) for the above tissue collection and have had the opportunity to ask questions.

2. I understand that my participation is voluntary and that I am free to withdraw at any time without giving any reason, without my medical care or legal rights being affected.

3. I understand that samples will be collected during surgery, at post operative appointments and during any subsequent routine appointments.

4. I understand that sections of any of my medical notes from Lancashire Teaching Hospitals or elsewhere may be looked at and information taken from them to be analysed in strict confidence by responsible individuals from the research team or from regulatory authorities where it is relevant to my taking part in research. I give permission for these individuals to have access to my records.

5. I agree to take part in the above study.

6. Consent for storage and use in possible future research projects: I agree that the samples I have given and the information gathered about me can be stored by the Pathology Laboratory at Lancashire Teaching Hospitals for possible use in future projects, subject to additional project specific ethical approval.

7. Genetic research: I understand that future approved projects utilising the sample(s) I have given, may include genetic research aimed at understanding the genetic influences on tumours, but that the results of these investigations are unlikely to have any implications for me personally.

Please sign and date

Name of Patient __________________________ Date __________________________ Signature __________________________

Name of Person taking declaration (if different from researcher) __________________________ Date __________________________ Signature __________________________

Researcher __________________________ Date __________________________ Signature __________________________

Thank you for agreeing to participate in this research

1 copy for patient, 1 copy for Tissue Bank, 1 copy to be kept with hospital notes

Consent Form for use in Lancashire Teaching Hospitals: Gynaecological / Urological tumour tissue bank

Version 2, 30th September 2010
Patient Information Collection Sheet for Research

Study Code

Patient Initials

Age

Post code

Date of Surgery

Type of Specimen  Uterus: Case/ Control
                  Ovary: Case/ Control

Tamoxifen exposed  Y  N  Current  Past

Duration of exposure  How long ago stopped?

Smoking status  current  previous  never  not known

Breast cancer history  Y  N

BRCA carrier  Y  N  not known

Menopause:  Y  N  LMP:  Cycle: ___/____

BMI

Parity

Ethnicity

Previous hormonal manipulation:

Ca 125

Cancer:  Y  N  Endometrial/Ovarian

Pathology – subtype  grade

Additional information (previous cancer diagnoses, other co-morbidities etc)

________________________________________________________________________

________________________________________________________________________

________________________________________________________________________
Appendix 2

Other related articles published during the MD
Vibrational spectroscopy of biofluids for disease screening or diagnosis: translation from the laboratory to a clinical setting

Alana L. Mitchell1, Ketan B. Gajjar2, Georgios Theophilou1, Francis L. Martin1, and Pierre L. Martin-Hirsch1,3

1 Centre for Biophotonics, Lancaster Environment Centre, Lancaster University, Lancaster LA1 4YQ, UK
2 Addenbrooke’s Hospital, University of Cambridge, UK
3 Department of Obstetrics and Gynaecology, Lancashire Teaching Hospitals NHS Foundation Trust, Preston, UK

Received 12 February 2014, revised 21 February 2014, accepted 22 February 2014
Published online 20 March 2014

Key words: ATR-FTIR spectroscopy, biofluid, biomarkers, microfluidic platform, point-of-care, Raman spectroscopy, screening

There remains a need for objective and cost-effective approaches capable of diagnosing early-stage disease in point-of-care clinical settings. Given an increasingly ageing population resulting in a rising prevalence of chronic diseases, the need for screening to facilitate the personalising of therapies to prevent or slow down pathology development will increase. Such a tool needs to be robust but simple enough to be implemented into clinical practice. There is interest in extracting biomarkers from biofluids (e.g., plasma or serum); techniques based on vibrational spectroscopy provide an option. Sample preparation is minimal, techniques involved are relatively low-cost, and data frameworks are available. This review explores the evidence supporting the applicability of vibrational spectroscopy to generate spectral biomarkers of disease in biofluids. We extend the inter-disciplinary nature of this approach to hypothesise a microfluidic platform that could allow such measurements. With an appropriate lightsource, such engineering could revolutionise screening in the 21st century.

1. Introduction

The need for simple, non-invasive methods to diagnose or screen for important medical conditions has never been more relevant. For example, every year thousands of patients are referred onto two-week-wait cancer pathways in the UK’s NHS service for investigation of a multitude of symptoms that may or may not lead to a cancer diagnosis [1]. In an ageing population, this is economically unsustainable as
Women insert a cotton bud into the vagina and the vaginal secretions acquired are subsequently tested for oncogenic HPV subtypes. Despite its limitations, the use of blood serum or plasma has an important role in routine diagnostics, e.g., creatinine levels for the diagnosis of renal impairment, or, bilirubin, alanine transaminase or alkaline phosphatase for liver impairments (Table 1) [8, 9]. A major issue with expanding the use of such biofluids to the screening process revolves around issues of specificity and sensitivity [10]. An excellent example of this is the commonly used marker of ovarian cancer, carbohydrate antigen (CA) 125 (CA-125, also known as cancer antigen 125), which has long been proposed as a possible screening biomarker. Although the sensitivity and specificity of serum CA-125 is 81% and 75% respectively, it is only raised in ≈50% of early cases of ovarian cancer [11], with levels rising in a variety of other conditions ranging from other gynaecological complaints such as endometriosis to non-gynaecological conditions such as liver impairment. Given that pre-symptomatic ovarian cancer is an insidious disease, this is a major problem. Provisional results of its use as a screening tool suggest it might only be useful in high-risk populations [12] as specificity is too low for mass population screening.

Biofluids such as serum contain a high protein concentration [13]. Levels and profiles of such proteins and other biomolecules have been ascertained in healthy subjects and any values outside references ranges might indicate disease, e.g., elevated serum prostate specific antigen (PSA) levels for prostate cancer. However, biomarker development remains a challenging objective because conventional methods often rely on one knowing the exact molecular entity to be determined and there being a label (e.g., an antibody) to a particular conformation of this molecule; as Table 1 shows, despite extensive investment over the last three or four decades there remain relatively few biofluid biomarkers. The problem with chronic diseases is that their development generally entails multi-stage processes over decades in a heterogeneous human population that is genetically outbred and exposed to enormously disparate environmental influences [14]. Thus, although cancer or neurodegenerative diseases, even when arising from the same tissue site, may be included under a particular blanket diagnosis, a particular pathology such as breast cancer may in fact consist of very differing pathologies, for instance based on mutational profiles [15]; these are likely to be susceptible to different treatment regimens. In fact, there is increasing evidence that this heterogeneity may exist at the individual level, giving rise to the notion of a need for personalised medicine [16]. If this is the case, it is not far-fetched to conclude that a biomarker or profile of such molecular entities in a biofluid could reflect a particular pathology; however, within a heterogeneous population pool these molecules will likely exhibit different conformational or isomeric forms. In this scenario, employing a specific labelling approach is unlikely to have the required generic applicability; one needs a molecular pathology tool capable of generating a fingerprint of all the constituents within this surrogate tissue, i.e., biofluid. One such approach is the application of biospectroscopy tools, including Fourier-transform infrared (FTIR) spectroscopy, attenuated total reflection FTIR (ATR-FTIR) spectroscopy or Raman spectroscopy [17]. These approaches generate fingerprint spectra of biological material. Within this review, we examine their applicability as a screening or diagnosis tool and also, we speculate on how these laboratory-based techniques might in the future be translated to point-of-care settings in clinical practice.

2. Biospectroscopy methods: translatable tools towards clinical practice

Vibrational spectroscopy has been proposed as a reagent-free, non-destructive approach towards tack-
ling the traditional problems associated with bio-
fluids in diagnosis and screening (Figure 1). These
methods primarily exploit the fact that chemical
bonds constituent within biomolecules absorb in
the mid-infrared (IR) region (wavelength \( \lambda = 2.5-25 \mu m \)); from this can be derived fingerprint spectra.
As such, spectra represent an omics fingerprint of
the underlying biological chemistry of an analysed
biological sample; it can allow one to track the cell
lineage within a tissue [18] or determine cellular ac-
tivity [19]. However, with instrument developments
the consequent ability for data acquisition has grown
markedly and this has led to requirements for suit-
able computational frameworks [20] and a need for
robust biomarker extraction approaches [21]. Ulti-
mately, what will be needed in a point-of-care cli-
cal setting is an easy-to-use analytical tool using a
validated classification or biomarker extraction sys-
tem, which could for instance be based on a green
(normal), amber (early warning) and red (disease
present) light system (Figure 2). The attractiveness
of this is that it would readily allow for on-the-spot
repeat measurements at minimal cost.

2.1. Infrared (IR) spectroscopy

Within cellular structures or biomolecules, constitu-
tent chemical bonds with an electric dipole moment
are IR active. In the mid-IR spectral region, the
most important spectral regions for biospectroscopy
analyses are generally the fingerprint region (1,500-
600 cm\(^{-1}\), which identifies skeletal and bending
fingerprint vibrations) and the double-bond region
(2,000–1,500 cm\(^{-1}\) associated with C=C, C=O and
C=N). The 1,800–900 cm\(^{-1}\) spectral region can be
tered the ‘bio-fingerprint’ region associated with
the structure and function of the tissue, cellular or
biofluid samples analysed, all based on chemical
bond vibrations [20]. These fingerprint spectra con-
sist of wavenumber-absorbance intensities and can

![Figure 2 A hypothesised point-of-care, may be personal, continuous (self-)monitoring instrument based on a microfluidic platform. A lancet or small needle as used for diabetic testing can be used draw a sufficient blood volume (=50 µl) which is drawn up and divided along two channels: one portion navigates through a microcentrifuge to release the blood plasma fraction, the other via a separator to allow clotting and forward conduit of serum. Along parallel channels the plasma or serum is directed to an internal sensor such as an ATR diamond. A lightsource (a globar or even a quantum cascade laser) facilitates generation of fingerprint spectra. Generated spectra are transmitted to a centralised database for automatic pre-processing and classification; based on this, a return signal triggers a green (everything normal), amber (early warning signs) or red (disease diagnosis) light. Not only would such a system allow for ready and repeat screening for a range of diseases negating the costs of transport of samples and laboratory analysis, one could monitor the emergence of pathological states and inform treatment, i.e., towards personalised medicine.

With thanks to Matthew Briggs, Medical Illustrator at Lancashire Teaching Hospitals NHS Trust.
be arbitrarily split according to different biomolecular constituents [22]. Proteins mainly contribute to the Amide I peak at ~1650 cm⁻¹ (80% C=O stretching, 10% C–N stretching and 10% N–H bending) and the Amide II peak at ~1540 cm⁻¹ (60% N–H bending, 40% C–N stretching). Lipids contribute absorptions at ~1740 cm⁻¹ (C=O stretching) and the PO₄²⁻ stretching vibrations of DNA are located at ~1080 cm⁻¹ (symmetric; νₛPO₄²⁻) and ~1225 cm⁻¹ (asymmetric; νₐPO₄²⁻). Position and intensity of absorption bands are determined by the molecular or conformational structure of these constituents; intensity may reflect underlying cellular activity with elevated levels being associated with higher turnover whereas a horizontal shift in the Amide I band may point to α-helix to β-pleated sheet protein conformational changes [23, 24]. FTIR microspectroscopy measurements can be conducted in either point or imaging modes; the former can be conducted via small sampling apertures (5 μm × 5 μm to 150 μm × 150 μm), allowing areas of interest to be interrogated following identification [18]. In single-point mode FTIR microspectroscopy with a small sampling aperture, dispersion effects may occur if the area under analysis contains IR opaque spherical objects, such as condensed nuclei [25].

One such variation of this technology that we believe has the potential to have enormous impact in this area is ATR-FTIR spectroscopy; within this simple and robust method, an IR beam is directed through a crystal, typically a diamond [26]. This beam subsequently undergoes total internal reflection within the crystal, generating a evanescent wave form that penetrates just beyond the crystal (the penetration depth depends on a number of factors including the refractive index of the crystal and the incidence angle of the beam) to interact with a tissue sample placed in contact with the crystal; the attenuation of the beam is exploited to generate an absorbance spectrum. A variety of crystal types can be used including diamond, germanium or zinc. ATR-FTIR spectroscopy may cover a relatively large sampling area (for instance, the diamond used in our studies for single-point measurements is ~250 μm × 250 μm [26]) which makes it an ideal technique for cytology as large numbers of cells may be studied at any given time, increasing the likelihood of determining the presence of a small number of atypical cells. Dispersions effects are also weaker in ATR-FTIR spectroscopy; therefore, fewer spectral artefacts attributed to light scattering are generated during the acquisition of spectral measurements from biological samples [27]. It is our contention that as a robust method, ATR-FTIR spectroscopy is very applicable to the routine monitoring of biofluids [28]; consequently, Figure 2 puts forward a point-of-care tool for either the General Practitioner’s surgery or even as a self-monitoring device. Such a microfluidic device would only require a small aliquot of blood, which could be separated into plasma or serum. Directed to independent ATR diamonds, a spectral reading could then be remotely classified giving an answer within minutes or seconds. This would allow for repeat measurements and the monitoring of emerging disease. It would also allow one to personalise care either through the identification of a profile of spectral biomarkers or which therapeutic strategy was most appropriate.

2.2. Raman spectroscopy

As a complimentary tool to IR spectroscopy, Raman spectroscopy also generates bio-fingerprint spectra within the 2,000–400 cm⁻¹ region; however, Raman is based on inelastic light scattering of monochromatic light from a laser in the visible, near-IR or near-ultraviolet regions. As it is a scattering technique, Raman microspectroscopy is unaffected by aqueous, potentially permitting in vivo and live-cell imaging [29]. Chemical bonds are excited to a virtual state through excitation with photons; these bonds then relax into a different vibrational state causing inelastic photon scattering and frequency shift. These scattered photons cause a shift in frequency indicative of specific vibrational modes; from this light scattering effect, a fingerprint spectrum can be derived [29]. It is also non-destructive, non-contact and reagent-free [30, 31]. Because Raman spectroscopy is not perturbed by aqueous, the spectral region is dominated by relevant protein features; it has recently been shown that by using centrifugal filter devices to concentrate analytes, an enhanced signal-to-noise ratio (SNR) can be achieved [13]. Further developments resulting in spatially-offset Raman spectroscopy (SORS) that allow deep non-invasive characterisation of biological tissues, which can be combined with SERS to give the technique known as SESORS, has been shown to allow one to derive quantitative, in vivo, transcantage measurements [22]. Further applications of Raman spectroscopy have shown that a SERS multiplex system has the potential to facilitate the simultaneous detection of three bacterial meningitis pathogens potentially in biofluids such as CSF, eliminating the need for time-consuming culture-based techniques [33].

3. Biospectroscopy to identify disease

The use of biofluids in biospectroscopy for disease screening and diagnosis is possible in a wide range of conditions, including cancer, diabetes and neurodegeneration. There have been a plethora of studies...
ATR-FTIR spectroscopy has been used to demonstrate a high level of diagnostic accuracy in the detection and classification of gliomas, malignant brain tumours that are responsible for 50% of intracranial lesions [48, 49]. Current diagnostic options for these types of tumours include CT and magnetic resonance imaging (MRI) scans but ultimately these rely on an inpatient diagnostic brain biopsy, a highly invasive procedure. Biomarkers include cytokines (which modulate the immune response of tumours as well as their growth) and angiogenesis factors (towards the proliferation of new blood vessels from ones already formed, also directly influenced by cytokines). Bio-Plex Pro™ immunoassays provide cytokine and angiogenesis factor levels that differ between serum from glioma vs non-cancer patients; specifically angiopetin, follistatin, hepatocyte growth factor (HGF), interleukin 8 (IL-8), leptin, platelet-derived growth factor (PDGF)-BB [composed of two B chains] and platelet endothelial cell adhesion molecule-1 (PECAM-1); on its own, this test gave sensitivities and specificities as high as 88% and 81%, respectively. With ATR-FTIR spectroscopy, sensitivities and specificities of 87.5% and 100% were achieved, respectively; this suggests that combination of these techniques in an orthogonal diagnostic regime could provide enhanced diagnostic capability [48].

The analysis of biofluids by Raman spectroscopy towards cancer diagnostics has been demonstrated in the fields of gynaecological cancer [50] and gastroenterology; not only was diagnosis achievable but also an ability to distinguish between high- and low-grade lesions, suggesting a potential role in screening. For example, alcohol use or infection with hepatitis C are risk factors for cirrhosis of the liver, which in itself increases the likelihood of developing hepatocellular carcinoma (HCC), which accounts for 80% of all liver tumours [51]. Current screening recommendations for patients with cirrhosis include six-monthly ultrasound scanning, which is operator dependent. A number of serological markers, primarily AFP, are used as biomarkers of HCC; however, whilst its role in diagnostics is well-established, its application towards routine screening is of much more limited value because high AFP levels are often observed in patients with just cirrhosis, in the absence of HCC. This lack of disease specificity is a major problem with the majority of conventional biomarkers. Preliminary studies have suggested that Raman spectroscopy can be used to distinguish patients with HCC from those without the disease amongst a cohort of patients all suffering from cirrhosis; this was when spectrochemical analysis was coupled with chemometric methods to detect subtle spectral differences and gave an overall accuracy of 91.9% [52].

Various approaches have been developed to increase the accuracy and usability of Raman spectroscopy in cancer diagnostics, including surface-enhanced Raman spectroscopy (SERS). In general, SERS involves using immuno-labels with an affinity towards particular biomarkers (for instance, cell-surface epitopes) coupled with secondary antibodies conjugated to SERS substrates. By exploiting nanoparticles (NPs) such as silver or gold (Au), enhanced spectra are generated to allow for better characterization, detection and identification of biomolecules in a shorter timeframe [53–55]. This has led to a marked increase in interest in SERS as a diagnostic approach in biomedical research. Research areas include the preparation and discovery of new SERS substrates such as NPs, surface functionalization [56], imaging [57] and clinical translation [58]. An example of work in this area includes the use of SERS in detection of decreased Raman peaks in the saliva of lung cancer patients compared to those who did not have the disease; this was attributed to differences in levels of proteins and nucleic acids in saliva [59].

Recent advances in the preparation of SERS immunoassays include suspension of magnetic beads with antibody-labelled AuNPs. Magnetic beads are used to support the antibodies so a solution of the immunocomplexes can be formed, followed by immobilization with a magnetic bar of these complexes on microtubes. Theoretically, this technique should overcome some limitations of immunoassays on solid substrates (i.e., the need for repeat washings and long incubation times) and it has been suggested that this leads to more reproducible results; robust reproducibility remains a common problem in conventional SERS [60]. SERS can be applied to samples of plasma, serum or saliva from cancer patients and in some circumstances, it outperforms conventional immunoassays, e.g., ELISA [61].

3.2. Diabetes

An increasingly unhealthy Westernized lifestyle appears responsible for diabetes becoming a major medical concern. There are three main pathophysiological subtypes of this condition: Type 1, which is primarily autoimmune leading to destruction of pancreatic islet cells; Type 2, which often correlates with obesity leading to peripheral decreases in insulin resistance; and, gestational/pregnancy-induced, which can lead to complications for both mother and child [62]. Monitoring of blood glucose levels is performed by applying a drop of capillary blood from the tip of a finger to a chemical test strip, and recent technological advances allow continuous surveillance [63]. Reagent-based, continuously-measuring sensors rely strongly on the stability of required reagents, which are often enzymes. In addition, this approach con-
in this area, but whether laboratory-based investigations within which control vs. atypical cohorts are often balanced in terms of sample sizes are translatable to a point-of-care clinical setting where abnormal cases are rare events within a well-population remains to be established. Prospective clinical trials are required within a well-population screening service; this would prove whether biospectroscopy approaches have the capability to identify the small proportion of at-risk individuals amongst the large numbers that require no follow-up. It would also require correlation with gold standard endpoints such as histology for cancer diagnosis. To demonstrate its applicability towards disease screening or diagnosis, biospectroscopy analyses would likely need to be initially incorporated into an existing screening programme in addition to other routine analyses. Long-term follow-up and correlation with gold standard endpoints is vital in the application of new techniques for cancer diagnosis; to be implemented in practice, sensitivity and specificity for disease diagnosis needs to be exceptional as does the ability to determine emerging or progressive disease.

3.1. Cancer

The discovery of new biomolecules for cancer diagnosis is a global research priority with blood-based diagnostics showing promise in areas of disease subtype classification, response to treatment and diagnosis of disease relapse [34]. For instance, dried blood spot sampling is re-emerging as a screening tool given the simplicity of the technique for collection, storage and transport [35, 36]. There has been major interest in identifying biomarkers of malignancy of gynaecological origin, brain, prostate and the gastrointestinal tract (Table 1). Such biomarkers are molecules that arise from the presence of abnormal cells within a tumour; theoretically, these may then appear in blood, urine or saliva. The detection of nucleic acids in the blood of oncology patients was first recognized in the 1970's [37, 38]. High levels of nucleic acid fragments can be found in the blood of these patients; could biospectroscopy techniques determine these entities as the presence of cancer [39, 40]? Various explanations for the presence of nucleic acids detected within blood include elevated apoptosis or necrosis within cancerous tissue and consequent release into the general circulation. Small numbers of cells may also be released from a tumour and may circulate in the peripheral blood of oncology patients; these may also be detectable [41, 42]

Ovarian cancer has often been termed the “silent killer”. The origin of this term refers to the initial insidious presentation of the disease with often vague and unremarkable symptoms that may be attributed to a wide range of benign conditions such as indigestion, reflux, “middle-age spread” and “water work problems”. As a result, by the time of diagnosis, the disease is often in its advanced stages. Medical professionals are currently being encouraged to adopt a low threshold for suspicion in middle-aged women presenting with abdominal pain, bloating or changes in bowel or bladder function following guidelines produced by the National Institute for Clinical Excellence (NICE) [43]. Current methods for diagnosis include clinical examination, various scanning modalities including ultrasound and computed tomography (CT) as well as the measurement of a serum molecule, CA-125. In addition to serum CA-125, other blood-based biomarkers including macrophage-colony stimulating factor (M-CSF), ovarian cancer antigen X1 (OVX1), human epididymis protein 4 (HEP-4), lipoprotein-associated sialic acid, CA-72-4, CA-15-3 and CA-54-61 have been tried and tested with poor results due to unacceptable levels of sensitivity and specificity. β-Human chorionic gonadotropin (β-HCG) and α-fetoprotein (AFP) levels have a role in the diagnosis of germ-cell tumours that occur primarily in younger patients [44]. Currently, a combination of the clinical, serum and radiological findings indicate probable diagnosis and from these results, further treatment including surgery and chemotherapy can be planned. However, the lack of specificity of CA-125 limits its role as an effective screening test which, given the late presentation of the disease, is so badly needed. In a UK-based prevalence trial for ovarian cancer screening (UKCTOCS) in postmenopausal women, annual CA-125 followed by TVS as a second-line test showed a sensitivity of 75% when borderline ovarian tumours (the tumours least likely to metastasise or re-occur) were excluded from the results [45]. In addition, similar research carried out in the United States did not show any survival benefits using the same modalities [46].

ATR-FTIR spectroscopy has been shown to be a promising approach for the diagnosis of ovarian cancer using blood serum, offering a promising screening tool. A UK pilot study has demonstrated that blood plasma- or serum-based ATR-FTIR spectroscopy with a classification ensemble has the potential to discriminate ovarian cancers from controls [47]. The most notable results came from the analysis of blood plasma of patients with ovarian cancer where ≈97% patients were correctly diagnosed when compared to the gold standard of histological diagnosis by a pathologist. Ovarian cancer was also detectable using serum, with ≈95% of samples from patients with disease correctly classified. With 30 patients in each group that is in itself a bias, clearly larger datasets with different grades and sub-types of ovarian cancer are needed to validate this study but these findings suggest what this area can offer in the screening and diagnosis of this disease.
sumes glucose, whereas reagent-free approaches such as spectrochemical ones do not. Studies have shown that a wide variety of optical methods may be suitable for a reagent-free observation of glucose [64], such as polarimetry or refractometry. Spectroscopy of the molecular vibrations of glucose by IR or Raman spectroscopy [65–67] has been shown to be a sensitive and specific method in vitro [68, 69] and ex vivo [70]. Raman spectroscopy (employing a 785 nm laser) detected minute amounts of glucose in diluted (10-fold) urine, with a 92% accuracy to classify abnormal (8 mg/dL) and normal urine samples according to their glucose concentrations [71]. Studies on rodent models employing such spectrochemical methods with cluster analysis successfully demonstrated differences between diabetic and control subjects using a variety of substrates including muscle samples, kidney plasma membrane and liver microsomal membrane [72].

As a first step towards the continuous monitoring of glucose, reagent-free transmission spectroscopy in the mid-IR region has been carried out in vitro using a quantum cascade laser (QCL) and an optical silver halide fibre [73]. QCL-based systems, although still quite specialized, are finding applications in medical diagnostics [74]. Because they require relatively low power and are so small, QCL-based systems may eventually replace larger and slower FTIR and other spectroscopy systems for laboratory work and may thus, with further development, have a more applicable role in a clinical-based setting. Such clinical application requires extremely fast sampling times, relatively small size (i.e., volume or mass), and accurate results in order to avoid misdiagnosis [75]. QCL mechanisms are based on intersubband-transitions of electrons inside a quantum-well structure. Therefore, unlike other semiconductor light sources, the emitted wavelength is not determined by the band gap of the used material but on the total thickness of the substrate under analysis [76]. The use of QCL systems in continuous glucose monitoring demonstrates the possibility of a reagent-free, fibre-based, and most importantly, minimally invasive in vitro glucose sensor. The sensor could potentially be used as an implant in order to continuously monitor the body's glucose levels. With the development of an effective, reagent-free continuous glucose sensor, the connection to an insulin pump may eventually lead to the development of a completely closed-loop system for glucose control. Recently, a tunable (1030 cm⁻¹–1230 cm⁻¹) external-cavity QCL allowed simultaneous determination of glucose, lactate and triglycerides in blood serum [77]; miniaturization of such set-ups make point-of-care applications a genuine possibility.

In addition to the day-to-day complications of strict glucose level monitoring, the major implications in the development of diabetes are the long-term consequences of end-organ damage, especially renal impairment and hypertension. Prevention of these complications is linked to strict glycaemic control; the development of non-invasive monitoring could enhance compliance to the dietary and lifestyle routines that diabetics have to adhere to. Currently, the glomerular filtration rate (GFR) [78] is used to calculate protein excretion in urine over a 24 h period. eGFR levels determined from single blood samples are not as accurate as 24 h urine collection and therefore is not a reliable substitute in optimal patient management. Thus, the main drawback is the need to collect urine samples for over a long timeframe.

The advantages of vibrational spectroscopy towards detecting changes in urica and creatinine concentrations when compared to standard immunoassays rely on the fact that urine can be evaluated with such approaches in real-time, quickly and without the need for reagents. This enables analysis to be done in each sample collected during the day, increasing the case of analysis and patient comfort both of which may lead to more accurate results [79]. Raman spectroscopy has been shown to be applicable to this with studies now available suggesting a role in the detection of biomarkers in the urine of diabetic patients with renal impairment that may lead to earlier diagnosis of the complication and better management [80]. Urine samples were collected from control subjects vs. patients with diabetes and hypertension in the absence of complications, a high degree of complications and those requiring renal dialysis. It was found that levels of urica and creatinine significantly decreased (mirrored in the corresponding Raman peaks) as disease progressed from control to low risk/high risk to patients undergoing dialysis, indicating results that could generate diagnostic information regarding possible complications and allowing one to tailor therapy to improve disease prognosis.

3.3. Neurodegeneration

With advances in medical treatments/care, the average human life span is getting longer; however, an increasingly elderly population is giving rise to an increase in the complexity of health problems and a population ever more reliant on healthcare. Arguably, Alzheimer’s disease (AD) is one of the most distressing medical conditions of the elderly with significant social, psychological and financial burdens in addition to medical complications, which affect not just the patient but the wider family unit [81]. Clinically-used psychological tools that have some predictive value in determining the degree of memory impairment include the min-mental state examination.
(MMSE) in conjunction with formal assessment by a specialist in old age psychiatry [82].

A number of mechanisms have been proposed to explain the pathogenesis of AD including amyloid metabolism, oxidative stress, inflammation, and lipid dysregulation [83]. Oxidative stress and free radical damage are critical factors in the pathology of aging and many age-associated degenerative diseases, such as Parkinson's disease (PD) and AD [84]. Brain tissue is susceptible to destruction by free radicals due to its requirement for high oxygen consumption, high levels of polyunsaturated fatty acids in neuronal membranes, and low antioxidant levels 85. Moreover, amyloid β-protein is known to play an important role in oxygen radical generation [86, 87]. It can directly induce the formation of free radicals and a correlation with neurotoxicity has been noted [87].

Studies from the last twenty years looking at differences in the brains of controls, multiple s Loners (MS) or AD patients using conventional FTIR spectroscopy methods had little difficulty in determining differences between white and grey matter due to spectral alterations attributed to lipid content in white matter; it was more difficult to determine differences between control vs. AD grey matter with only subtle and inconsistent differences noted in Amide I peaks [88, 89]. There is a need to improve the characteristics of spectra obtained from amyloid plaques, especially at a single-cell level [90]. Since these earlier studies, there has been an enormous research effort in the search for specific biomarkers of AD. For example, decreased levels of Ap43-42 and increased levels of either 'total tau' proteins or phosphorylated tau in cerebrospinal fluid (CSF) are now becoming accepted as early diagnostic biomarkers for AD [91]. In addition, a novel enzyme-linked immunosorbent assay (ELISA) system allows for the detection of Aβ oligomers in which an anti-Aβ monoclonal antibody, BAN50, is used for both capture and detection of the oligomers, in a single antibody sandwich ELISA (SAS-ELISA) system. The BAN50 SAS-ELISA cannot detect Aβ monomers. This is because the capture antibody fills the only available binding site the antibody has available to it. However, it can detect Aβ oligomers because they have multiple binding sites. The BAN50 SAS-ELISA has been shown to specifically detect high molecular weight Aβ oligomers in CSF samples from patients with AD or mild cognitive impairment; these signals are significantly higher than those obtained from age-matched controls, as well as correlating negatively with scores obtained in the MMSE. These results indicate a possible role for the SAS-ELISA as a useful molecular diagnostic marker for AD, and perhaps a surrogate marker for disease severity [92]. Again, there is a real possibility to combine such endpoints with reagent-free spectrochemical end-points; in fact, these latter technologies could allow for real-time measurements of emerging pathology, a major challenge in neurodegenerative disease.

From a clinical perspective, CSF sampling is an invasive and uncomfortable procedure that would not be feasible on a widely-introduced scale and would be unlikely to be acceptable to most patients, many of whom will be elderly with co-existing medical problems [76]. Research has been conducted into the use of vibrational spectroscopy of blood samples that would likely be a more acceptable method of obtaining biofluids. In light of the work carried out on CSF, it would be a fair assumption that by-products of oxidative substrate modification would be present in the peripheral circulation. Studies on human plasma [95] suggest that an increase in oxidative stress leads to an increase in R-OH bonds and a decrease in both CH and NH bonds, detectable by IR spectroscopy. A metabolomic signature was detected in AD plasma consistent with altered protein and oxidative states although further work is required to determine the exact chemical components responsible.

4. Computational issues

The acquisition of spectra is only the first step in developing vibrational spectroscopy as an established clinical tool. In the early days of spectroscopy, research data analysis was limited to the rudimentary computer science available. Now, however, there are a number of analytical options available as the processing capacity of modern computational systems have increased [94]. Traditionally, data analysis for vibrational spectroscopy was largely carried out by commercial software companies such as CytoSpec, Neurodeveloper, Unscrambah, Piroette, OPUS or Wrire. Commercially-used MATLAB toolboxes include the Neural Network and the Bioinformatics Toolbox. Others are free to use, but closed-source leaving researchers dependent on the software manufacturers. In addition, open-source software can be largely restricted to a given task [95]. Our group addresses these technical complications with the development of IRootLab, a free and open-source MATLAB toolbox for vibrational biospectroscopy data analysis. This online programme offers a class library containing a comprehensive range of methods, concepts and visualizations for the analysis and simplification of the data analysis process [96]. Regardless of which analytical programme is used, there needs to be clear steps taken in order to generate a comprehensive dataset, including quality control, pre-processing, feature extraction, clustering, classification and biomarker extraction commonly involving principal component analysis loadings [97].
Further advances in methodological-computational development may lead to the development of classifications using a compound dataset consisting simultaneously of blood plasma and serum as well as further exploration of classification ensemble architectures [21, 47]. A future point-of-care device would remotely send spectra to a central classification database; the outcome would then trigger a simple response for the user (Figure 2).

5. Conclusions

The potential advantages of using IR or Raman spectroscopy of biofluids for disease detection are: no reagents are required, a profile of spectral alterations can be determined and the methods are suitable for automation. In the handling of some biofluids, such as blood, there remain challenges to be overcome [98]. Whilst there is a large body of compelling evidence to support the application of vibrational spectroscopy to biofluids such as blood to triage patients for particular conditions such as myocardial infarction [99, 100], the integration of these techniques into standard clinical practice remains some way off [101] and definitive clinical trials are needed to define their true diagnostic accuracy [102]. The enormous advantage of a real-time analytical monitoring device is that in the face of emerging disease such as neurodegeneration, one could take simple medications to slow down the progress of the pathology [103]. An obvious place to initially implement such a screening/diagnostic approach would be in the developing world where new inexpensive technologies are most likely to make a dramatic impact on population screening [17]. In addition, further validations in large multi-centred randomised control studies incorporating the gold standard of current diagnostic methods are required. Many current studies have used relatively small sample numbers and thus may be biased; a real-world scenario involves a large weighting towards negative outcomes with only a very small number generating a positive result requiring intervention.

The sensitivity of vibrational spectroscopy, especially ATR-FTR spectroscopy, to identify low-level effects [104, 105] that may lead to transformation (a precursor of malignancy) [106] has been demonstrated. Not only do computational algorithms allow one to classify according to cell type (or phenotype; also applicable to biofluids), but they also allow one to extract discriminating features that may give rise to a new concept of spectral biomarkers. This interdisciplinary approach will require a major mindset change in translational research, which traditionally understands alterations in gene expression or protein levels as biomarker endpoints; harnessing these methods as in SERS [107] to more conventional approaches may facilitate this transition. Once these methods are established within tertiary care, there may be scope for moving into general practice and primary care with the design of hand-held devices that could be used to identify patients requiring referral to tertiary care for more definitive investigations and diagnosis. Although this concept may seem very far removed from current clinical practice, there have been various protocols and designs that may provide a starting point for such an idea to become a reality. A 3-D mammalian cell separator biochip has been designed that introduces a femtosecond laser written device capable of sorting cells and collecting viable cell populations [108]. This could be the first step in such devices being introduced as high-throughput optofluidic instruments with a wide range of applications in medical technology both within and outside vibrational spectroscopy, and a valuable tool in the quest for effective diagnostic and screening measures in everyday clinical practice.

Acknowledgements. Thanks to Lancashire Teaching Hospitals NHS Trust and Rosemere Cancer Foundation for funding the Clinical Fellows conducting these studies at Lancaster University. Special thank to Medical Illustrator Matthew Briggs (Lancashire Teaching Hospitals NHS Trust) for generating figure images.

Abbreviations AD, Alzheimer’s Disease; AFP, α-fetoprotein; β₂GPO₄, asymmetric phosphate stretching vibrations; ATR, attenuated total reflection; Au, gold; BDGF, bone-derived growth factor; BHCQ, β-Human chorionic gonadotropin; CA, carbohydrate antigen; CSF, cerebrospinal fluid; CT, computed tomography; ELISA, novel enzyme-linked immunosorbent assay; FTIR, Fourier transform infrared; GFR, glomerular filtration rate; HCC, hepatocellular carcinoma; HE4, human epididymis protein 4; HGF, hepatocyte growth factor; HPV, human papilloma virus; IL-8, interleukin 8; IR, infrared; MCF, macrophage colony stimulating factor; MMSE, mini-mental state examination; MRI, magnetic resonance imaging; MS, multiple sclerosis; NICE, National Institute for Clinical Excellence; NP, nanoparticle; O VX1, ovarian cancer antigen X1; PD, Parkinson’s disease; PECAM-1, platelet endothelial cell adhesion molecule-1; PSA, prostate specific antigen; QCL, quantum cascade laser; SAS, single antibody sandwich; SERS, surface-enhanced Raman spectroscopy; SERS, surface-enhanced Raman spectroscopy; v₂PO₄, symmetric phosphate stretching vibrations; TVS, transvaginal scan.

References


Exploiting biospectroscopy as a novel screening tool for cervical cancer: towards a framework to validate its accuracy in a routine clinical setting

Biospectroscopy is an emerging field that harnesses the platform of physical sciences with computational analysis in order to shed novel insights on biological questions. An area where this approach seems to have potential is in screening or diagnostic clinical settings, where there is an urgent need for new approaches to objectively interrogate large numbers of samples in an objective fashion with acceptable levels of sensitivity and specificity. This review outlines the benefits of biospectroscopy in screening for precancer lesions of the cervix due to its ability to separate different grades of dysplasia. It evaluates the feasibility of introducing this technique into cervical screening programs on the basis of its ability to identify biomarkers of progression within derived spectra (‘biochemical-cell fingerprints’).

Cervical cancer was responsible for 275,000 deaths worldwide in 2008, with 530,000 new cases being diagnosed that year. It is the third most common cancer in women worldwide and the most common in Eastern Africa and South Central Asia (330). Current estimations suggest that almost all cervical cancers are attributable to the human papilloma virus (HPV) (3). The majority of women are infected with HPV soon after the onset of sexual activity (4), but only a few will develop invasive disease (4). HPV infection primarily occurs in the differentiated epithelium and is usually transient. This viral infection results in the production of IgG1 and IgA antibodies. Persistent long-term infection superimposed on events compromising the host’s immune response contributes towards the development of cervical intraepithelial neoplasia (CIN), that is, the grade of disease (5). Ever since HPV infection was suggested as a necessary cause of cervical cancer, testing for oncogenic virus genotypes has proven useful in triage of women when cytology is not definitive; it can also be an alternative primary screening tool (6).

Currently, the most common screening method for precancer of the cervix is cervical cytology. Cervical cytology is the technique that involves using a spatula or a cyto-brush; cellular material is obtained by moving the spatula across the cervical os (i.e., the neck of the womb) while rotating it 360°. The cellular material is fixed with alcohol and then stained by the Papanicolaou technique or it is prepared in liquid-based cytology processes; the latter improve adequacy rates and quality of resultant slides. Cervical cytology is graded depending on the cellular morphology and most cytological classifications resemble the American Bethesda classification system. Abnormal squamous cells are graded into (7):

- Atypical squamous cells;
- Atypical squamous cells of undetermined significance (ASCUS);
- Atypical squamous cells – cannot exclude high-grade squamous intraepithelial lesion (if present, termed ASCUS-H);
- Low-grade squamous intraepithelial lesion;
- High-grade squamous intraepithelial lesion;
- Squamous cell carcinoma;
- Atypical glandular cells not otherwise specified;
- Atypical glandular cells, suspicious for adenocarcinoma in situ or cancer;
- Adenocarcinoma in situ.

The term ‘atypical squamous cells – cannot exclude high grade squamous intraepithelial lesion’ refers to a smear where a high-grade lesion
Key Terms
Cervical cytology: A speculum or a cyber-brush is used to obtain cellular material across the cervix. This is then fixed with alcohol and stained by the Papanicolau technique to allow visual examination of the cellular morphology.

Sensitivity: True positive rate – percentage of abnormal cytology specimens correctly classified.

Biospectroscopy: Application of spectroscopy techniques to address biological questions.

Biochemical-cell fingerprint: Derived IR spectrum that is unique to a biological sample under observation.

Specificity: True negative rate – percentage of normal cytology specimens correctly classified.

Current cervical cancer screening programs
Considerable attention has been devoted to the organization of cervical cancer screening programs, which minimize the adverse effects and maximize the benefits of screening. In many resource-poor countries, screening is still opportunistic because of lack of resources and the lack of organized screening. Consequently, the introduction of cervical cancer screening in the UK has been a significant decline in the number of deaths from this disease [10]. Although successful, cytology-based cervical cancer screening programs are not without problems. Cervical cytology is associated with poor sensitivity (i.e., probability that an affected patient will be correctly tested as positive – true positives) and a poor positive predictive value (PPV, i.e., probability that a patient is affected given that she has been tested as such) [11]. Furthermore, excisional treatment for disease is associated with an increased risk of preterm delivery in subsequent pregnancies [12]. Smears remain inadequate in 1–2% of cases, even with modern methods towards using liquid-based cytology, which has led to a marked improvement from the 9% inadequate smear rates of older approaches [13]. There are also major manpower pressures on healthcare organizations with a huge cost implication; in England alone, the cost was estimated at £137 million in 2008 or 0.2% of the total health budget per annum [14]. There is an urgent need for a cheaper and more robust screening technology.

With the advent of HPV testing and vaccination, cervical cancer screening is en route to a whole new phase in its implementation. Primary prevention of cervical cancer by prophylactic vaccination against "high-risk" HPV subtypes appears to have been advocated as the way forward in many European and Western countries; however, uptake has been variable [15]. Implementation of a vaccination program is unrealistic in resource-poor countries. It is also important to remember that cervical cancer screening will need to be continued even if countries have high vaccination compliance, as current vaccines are not multivalent and they do not prevent all oncogenic infections. Screening women between the ages of 20 and 24 years has been demonstrated to be less effective at preventing invasive disease compared with older age groups. Hence, it is reasonable to screen women from the age of 25 years in both the developed and developing world [16]. Molecular biomarkers have the potential to improve diagnostic testing for CIN3. One such biomarker is p16-INK4a, a cell cycle inhibitor, whose expression is increased as a response to high-risk HPV oncoproteins E6 and E7; it was found to have marginally better sensitivity than HPV testing in high-grade lesions [17–23]. In spite of these advances, there remains a need for an objective screening tool that is cheap and robust. Ideally, such a screening tool would also differentiate at an early stage between women committed to progressing to higher grade disease from those presenting with lesions that will regress in the absence of intervention, that is, give a dichotomous marker.

Biospectroscopy
The application of biospectroscopy techniques is one such approach that has been tested as applicable to the objective diagnosis of precancer of the cervix. The term 'biospectroscopy' generally refers to the application of spectroscopic techniques (e.g., MS, Raman spectroscopy, IR spectroscopy, etc.) to the study of biological material, often in a biomedical context. In this review, we will concentrate on FTIR spectroscopy. The principle of the technique involves exposing cellular material to broadband mid-IR electromagnetic radiation, thereby deriving absorption spectra that are unique to each biological sample under observation. A large number of chemical bonds present in biomolecules exhibit periodic vibration patterns (i.e., bending, stretching, rocking, wagging, and scissoring) that occur at frequencies within the IR spectral range (4000–400 cm⁻¹). Employing an interferometer to encode optical frequencies in time, the application of FTIR spectroscopy allows for recorded data to be Fourier transformed, in order to generate an absorbance (vibration) spectrum. By determining the fraction of incident light absorbed, IR spectroscopy devices generate sample absorption profiles, giving rise to what may be coined a ‘biochemical-cell fingerprint’ of the material under analysis [24–26]. FTIR spectroscopy is an advance on IR spectroscopy.
instruments dating back to the 1960s, when it finally became possible to calculate Fourier transform with the use of a computer. FTIR instruments represent a substantial improvement in throughput and sensitivity over other IR techniques (e.g., dispersive IR spectroscopy) [27]. Biospectroscopy has immense potential within biomedical sciences, ranging from detection of toxins and pollutants in the human body [28,29] to identifying stem cell regions within human tissues [30,31]. Cervical cytology cells can thus be analyzed to objectively segregate normal samples from abnormal ones [32-34]. It is suggested that IR spectroscopy applied to cervical cytology has the ability to pick up subtle changes before the cytological changes are visible using optical microscopy [35]. Molecular markers have the ability to identify low-grade lesions that are likely to progress to high-grade disease and those that are likely to regress [36].

Screening for cervical cancer may be the primary target in order to reduce both disease incidence and mortality. However, another important public health issue is further management following a positive screening result, whether it is cytologic or viral. In other words, with regard to current triaging of low-grade cytological abnormalities (or following a positive HPV DNA test) or CIN2+, and future more ambitious prediction of high-grade cytological abnormalities for invasive cervical cancer progression: how can one effectively manage such patients? In addition, what is needed is an absolute 'cure test' that may enable very early recognition (ideally from the sixth post-operative month) of not only those going on to have residual intraepithelial disease post-treatment, but also those going on to develop invasive cervical cancer post-treatment. An attempt in that direction is the concept of a 'Scoring System', that is in a form that is individualized (i.e., under the umbrella of personalized medicine), a combination of several HPV-related biomarkers (HPV typing, mRNA E6 and E7 expression, etc.) and other epidemiological data, available from a woman's history, such as demographic characteristics (age, parity, etc.), sexual behavior history and potential cofactor information (e.g., condom use and smoking). Such as system, in addition to cytology and colposcopy, including biospectroscopy, could allow a 'Nomogram of Risk' estimation in order to identify individual risk for CIN2+, CIN3+ or even invasion, regardless of the cytologic and colposcopic phenotype [37-39].

**Aims & objectives**

Cervical cancer remains one of the leading causes of female mortality in the developing world and there is an urgent need to develop inexpensive and robust platforms for screening for this disease. These programs must be cost effective and cover as large a population as possible. The screening test must be objective and robust. The aim of this review is to highlight the benefits of biospectroscopy as an objective screening tool for cervical dysplasia [40]. It also aims to design a hypothesized screening program using FTIR spectroscopy, and to validate its sensitivity and specificity in the setting of a developing country.

**Infrastructure of a screening program: call & recall**

Most developing countries will need to implement a cervical cancer screening program employing cytological evaluation, as they will be unable to cover the entire adolescent population by vaccination. Even in developing countries where cervical smears have been offered free of charge since 1995, uptake has been approximately equal to 47% [41]. The major problem with cancer screening in developing countries such as Malaysia or India is that there is no recall system and most healthcare facilities are located in urban areas generally leaving rural communities unscreened. Even if smear tests are provided free of charge, the education and/or ability among the general population to avail of such facilities is lacking. Hospitals that provide services for free are generally overcrowded with long waiting queues. In the event that there is a significant increase in cervical cytology coverage in the population, the lack of laboratory services, screening facilities and colposcopy referral centers will immediately arise as a limiting factor towards these programs' success. Attitudes towards cervical cancer, religious as well as cultural beliefs, also greatly impact on whether women will attend, even if the services are provided. Prior to implementing universal screening, women and their families need to be educated about cervical cancer screening within a health education program, in a culturally sensitive manner [42] through community-based and outreach programs [42-45]. The next step is policy-making and improving the infrastructure for screening and making it more accessible. Organizing a recall system in very populous countries will be extremely challenging. Cervical cancer screening embedded within other screening
programs, such as those for anemia, might make this more achievable [44], by reducing the cost of infrastructure development. As outreach screening programs may already be in place with significant numbers of people benefiting, one can predict good attendance and acceptance rates for a parallel cervical screening program.

Significant infrastructure is required to set up a call–recall system. For instance, it is estimated that some 6.4 million women undergo annual screening in Italy. This costs the government €158.5 million/year. Of these, 2.4% will have an abnormal smear. It is estimated that approximately 28,000-42,000 women will be treated for abnormal smears and this costs the government €22.9 million/year [45]. Despite estimating the economic burden of cervical dysplasia, it is difficult to extrapolate the cost from one country to another. The next question that must be asked is: how often should a country with limited resources screen women so as to achieve maximal benefit with available finances? 5-yearly screening in Finland with intensified screening in high-risk populations, showed women with an abnormal smear to have a ninefold elevated risk of cervical cancer if aged <40 years and a sixfold risk if aged >40 years as compared with asymptomatic women with a normal smear [47]. The NHS cervical screening program advises 3-yearly smears between the ages of 25 and 49 years and then 5-yearly smears from 50 to 64 years [48,49]. Creighton et al. demonstrated that the 2-yearly screening programs in Australia could just as readily be a 3-yearly program with significant savings and without an increase in cancer burden [50]. The benefit of 3-yearly screening (87%) over a 5-yearly program (83%) is only marginal in the 55-year age group, but in under-55s, three-yearly (84%) offers markedly greater protection than 5-yearly screening (73%) [49]. Resource-poor countries need to tailor their cervical cancer screening program to what is financially viable and feasible, and a 5-yearly program with intensified screening in high-risk populations may be the way forward. Yet, screening is only useful if abnormal smears are acted upon. Audits from Columbia highlight the fact that adequate coverage will not reduce mortality from cervical cancer unless adequate treatment and follow-up are provided [51].

**Screening technologies**

**HPV screening**

Liquid-based cytology combined with HPV testing over two rounds of screening showed no benefit over liquid-based cytology alone for the detection of high-grade lesions [52]. Furthermore, HPV testing is not useful in determining the need for colposcopy referral in younger women (<40 years) presenting with low-grade abnormalities [53,54]. There is some evidence to suggest that HPV testing in women who are over 55 years old is more specific than conventional screening and decreases colposcopy referrals and follow-up [55,56]. HPV testing at the time of colposcopy is also useful for women with a normal colposcopy, but with a smear suggestive of glandular disease [57]; it also increases the accuracy of colposcopy for high-grade disease [58]. HPV screening may be useful as an ancillary diagnostic tool. An obvious drawback of utilizing HPV testing as a screening tool in a low-resource setting is the cost, despite the fact that inexpensive innovative techniques for HPV testing are becoming readily available. However, these low-cost HPV tests are unlikely to have the same array of biomarkers that are likely to be available in developed world screening programs; this could result in the lack of capture of certain oncogenic HPV genotypes in at-risk populations. A suitable alternative could be an objective biophysical method such as biospectroscopy, which has been shown to possess the ability to differentiate between various HPV types [59] and could potentially distinguish between patients carrying high-risk HPV from those with low-risk HPV or even those who are HPV negative. Other than the initial cost of instrument acquisition, biospectroscopy techniques require minimal long-term running costs, in stark contrast to consumable-hungry molecular approaches. This makes such reagent-free analytical techniques, in regions where there is often a rich legacy in their application, an ideal alternative in screening conundrums. Whether or not it is cost effective as a primary screening tool will need to be decided by large-scale randomized trials [60].

**Adjunctive colposcopy techniques**

Women identified as having an abnormal cytology by screening are generally referred to secondary care for a definitive diagnostic test. This is generally done by colposcopy, which involves examination of the cervix under low power magnification. Identification of precancerous / cancerous lesions is aided by the application of acetic acid and Lugol’s iodine. Colposcopy is subjective and, despite morphology classification systems, the diagnostic accuracy is imperfect with the potential of disease being missed or over-called. New technologies have been devised to augment colposcopic accuracy.
While originally evaluated as a triage to colposcopy, the LuViva® multimodal hyperspectroscopy device is also being investigated as a primary screening tool for moderate and high-grade cervical cytology. The LuViva Advanced Cervical Scan detects changes in cervical cells by shining a combination of white and UV light on the cervix and measuring the patterns of reflected light. There are other newer imaging techniques currently being evaluated such as Optical Coherence Tomography (Niris® Imaging System; Imalux®, OH, USA), the Epitheliometer, and the Dynamic Spectral Imaging System (DySiS; DySiS medical, Edinburgh, UK), all of which can be used as an adjunct to colposcopy. DySiS produces a map to rate extent and duration of aceto-whitening (where a 3–5% acetic acid solution is placed on the cervix and dysplastic tissue turns visibly white). Optical coherence tomography uses IR light similar to ultrasound pulse-echo imaging. Image resolution is optimal in the 1–3 mm range. Optical coherence tomography is used as an adjunct to colposcopy and may improve the specificity of this approach slightly [6].

The principle of applying the epitheliometer is that electrical impedance is different between normal and dysplastic tissue. This device is a pencil probe that detects dysplasia/neoplasia by coming into contact with abnormal tissue using the principle of electrical impedance spectroscopy. It does have the potential to be used as an adjunct to colposcopy in the diagnosis of high-grade CIN [92]. This device is still in the trial phase and although it may be useful in reducing the number of biopsies taken in a resource-poor setting, in the future, it is unlikely to play a role in primary screening of women in its current state.

Visual inspection with acetic acid

Visual inspection with acetic acid is a practically feasible option for resource-poor countries with a PPV of around 58% [63]. It is associated with a high number of false positives and, consequently, generates a significant cost to the system [92]. Nevertheless, when performed by trained personnel it is an effective method to prevent cervical cancer in the developing world [63], at least until a cheaper and more robust, objective method of screening is available.

FTIR spectroscopy

Historically, IR spectroscopy was first used by William Herschel who separated the electromagnetic spectrum using a prism in 1799; this demonstrated that there is an increase in temperature beyond the red part of the visible spectrum. The basis for FTIR spectroscopy was laid in the 1880s with the invention of the Michelson interferometer. The interferometer is an optical device that allows IR light to be electronically sampled, and therefore computationally manipulated. The IR region as a whole lies between visible light and microwaves. It can be divided into three subcategories: far-, mid- and near-IR. All three have and can be used for biospectroscopy studies. The far-IR has the lowest energy and is not routinely applied to analyze biological material. In contrast, the near-IR has the most energy and is widely used in a wide range of spectroscopic studies (mostly chemical and mineral studies) due to its ability to identify several types of chemical bonds within this range. Mid-IR is the region of the electromagnetic spectrum between 4000 and 400 cm⁻¹ (the physical unit expressed in cm⁻¹, or ‘waves per centimeter’, is the inverse of the wavelength). It is the most useful for biological studies because it contains the fundamental vibrational modes of a large number of biomolecules, especially in the region known as the ‘biochemical-cell fingerprint region’ (1800 to 900 cm⁻¹).

Methods

IR instruments can be operated in three different ‘modes’: transmission, transfection and attenuated total reflection (ATR) [66,67]. In transmission mode, the IR beam passes through the sample to an IR detector – for this, the sample substrate must be IR-transparent. In contrast to this, in transfection mode the sample sits on an IR-reflective substrate (e.g., ‘Low-E’ slides from Keveley Technologies Inc., OH, USA), which reflects the IR beam back to be detected in a location near the IR source (i.e., the IR beam passes through the sample twice); recent discovery of an optical artifact known as the electric-field standing wave casts doubt on the reliability of measurements in transfectance mode [66]. ATR is a technique that uses an ‘internal reflection element’ (IRE), which is an IR-transparent body with high refractive index (typically a crystal, e.g., diamond or ZnSe). The IRE stays in direct contact with the sample. ATR is based on the ‘evanescent wave’ phenomenon. When the IR beam is internally reflected in the IRE, a fraction of the wavelength actually penetrates (proportional to the wavelength) beyond the IRE into the sample, which absorbs the IR beam.
The resultant spectrum is the result of the absorbance of IR by cellular components, hence the notion of a spectral ‘biochemical-cell fingerprint’ [24,26]. Following total internal reflection, the beam is then reflected back into the spectrometer to give absorption spectra, which can then be analyzed. One application of this technique is to differentiate between normal, dysplastic and cancer cells. The hypothesis herein being that the biochemical-cell fingerprint of a normal cell will differ from that of a dysplastic cell, which will again differ from that of a cancer cell. There are still questions as to the best bio-spectroscopy technique to employ; transmission measurements require material to be free of any IR opacity (there are questions as to whether condensed chromatin might be IR opaque), whereas transmission measurements are often associated with optical effects that require corrective algorithms (this is associated with large memory and processing requirements that are likely not to be available in resource-poor countries for the foreseeable future). However, ATR spectroscopy appears to be the most applicable of these techniques and spectral analysis is less prone to optical interferences. Cell-by-cell analysis has also been touted, but this currently requires a higher level of instrumentation and again significant processing power/capacity, especially considering some $5 \times 10^9$ cells need to be examined per individual smear. Harnessed to powerful computational algorithms, bulk spectral analysis of cellular material using ATR spectroscopy has been proven capable of extracting underlying alterations associated with CIN in a sub-population of atypical cells [32,33,45].

Chemical bonds within a cell have specific frequencies within the IR region at which they vibrate. Each of these bonds corresponds to various energy levels, which can be excited when they are ‘hit’ by light at corresponding frequencies. For example, glycogen exhibits a triad of peaks between 1250 and 1000 cm$^{-1}$; if pure, these occur at 1151, 1078 and 1028 cm$^{-1}$. Varying levels of glycogen in a woman’s menstrual cycle can be measured by IR spectroscopy, highlighting its ability to detect subtle changes in biochemical constitution [71]. The spectral region between 1740 cm$^{-1}$ and 1740 cm$^{-1}$ contains two protein peaks known as Amide I (1650 cm$^{-1}$; 89% C=O stretching vibration of the amide group coupled to the bending of the N-H bond and the stretching of the C-N bond) and Amide II (1590 cm$^{-1}$; 40% C-N stretching and 60% N-H deformation). The region between 1480 cm$^{-1}$ and 1200 cm$^{-1}$ contains spectral bands related to other biochemical components, including DNA, RNA, phosphates and phospholipids. When comparing spectra derived from adenocarcinoma of the cervix with those from normal tissues, Neelapanni et al. found significant changes at 1025 cm$^{-1}$ (gycogen), 1080 cm$^{-1}$ (gycogen and nucleic acids), 1155 cm$^{-1}$ (C-OH groups of serine, threonine and tyrosine, and C-O group of carbohydrates), 1240 cm$^{-1}$ (PO$_2^-$ groups of nucleic acids), 1400 cm$^{-1}$ (methyl group of lipids and proteins) and 1450 cm$^{-1}$ (methylene group of lipids and proteins) [72]. Figure 1 shows a sample spectrum with wave numbers of significance. Walsh et al. employed ATR-FITR spectroscopy with multivariate analysis and identified possible biomarkers of cervical cancer progression [32]. There are confounding factors that alter cellular spectra, for example, age, smoking and HPV status. Benign variations such as inflammation, metaplasia and blood will also alter the IR spectra collected [71,73,74]. However, with the implementation of liquid-based cytology, a cellular pellet can now be obtained devoid of blood and other contaminants (Figure 2).

Protocol for preparation of sample for bio-spectroscopy measurements

Cervical cytoplast samples need to be prepared prior to spectroscopic analysis. Samples must be devoid of all fixative and desiccated. The following is a technique to prepare a cervical cytoplast sample for spectroscopic analysis — sample preparation requires only a few minutes and requires no staining procedure. Figure 2 outlines the essential steps in sample processing. The procedure begins with the taking of a smear into ThinPrep Solution (PreservCyt$^\text{TM}$ Solution, Cytyc Corp., MA, USA) in the exact same manner a specimen is routinely taken for liquid-based cytology. This sample (3 ml, depending on cell density; this is only an example) is required for ATR-FITR spectroscopy. The solution containing the cellular material is collected in a tube and centrifuged at 1500 rpm for a period of 5 min. The supernatant fluid is removed with a pipette. The remaining pellet is washed with 3 ml of autoclaved distilled water (dH$_2$O) and centrifuged. This procedure is repeated twice more (i.e., a total of three washes with dH$_2$O) and the supernatant fluid is removed to give a cellular pellet devoid of methanol from the ThinPrep solution [66,75]. This pellet is then diluted with 0.5 ml of autoclaved dH$_2$O and re-suspended at room temperature onto a 1 cm x 1 cm glass
slide; the thickness of the sample on the slide should ideally be ≥3 μm to avoid any interfering effects from the glass substrate. The slides are allowed to air-dry prior to being placed in a desiccator until analysis. A major advantage is that sample acquisition and preparation fit within current practice for obtaining cervical smears and additional steps such as centrifuging are routine; they are also not time-consuming and would allow for high throughput of individual samples from a large population.

Example protocol for spectral acquisition
Spectra can be acquired using a range of different instruments (for instance, the authors use a Bruker TENSOR 27 FTIR spectrometer with a Helios ATR attachment containing a diamond IRE; incidence angle of the IR beam: 45°, in conjunction with Bruker OPUS computer software). The proprietary software is used to configure the spectrometer, carry out spectral acquisition, and perform spectral data manipulations. Prior to acquiring any spectra, the ATR diamond is cleaned with autoclaved dH₂O and observed to be clean, after which a background spectrum is acquired. The slide with cellular material is then placed on a stand and the diamond crystal is brought into contact with it; spectral acquisition can be concluded within 1 min. One possible preprocessing sequence (performed within OPUS software) is rubberband baseline correction followed by normalization to the Amide I (approximately equal to 1650 cm⁻¹) absorbance band. It is recommended that several IR spectra be acquired per sample; a number of ten spectra per sample may suffice in the majority of cases. Spectra may be then subsequently averaged per sample, or used in bulk, on a case-by-case basis. However, this judgment is based on the fact that the methodology is still in its infancy and needs to ensure robustness, but as it is refined such requirements will be reduced, thus speeding up the technique for clinical usage even more. Given that a well-mixed sample retrieved in liquid-based cytology is placed on a slide and that multiple spectral acquisitions can be taken from random points, heterogeneity in terms of spatial distribution does not present as a problem; indeed, one must remember that a typical cervical cytology specimen will contain multiple very different cell types including endocervical and ectocervical cells – correcting for this latter biological heterogeneity can be more challenging in order to isolate within a mixture of differing cell types, a small proportion (say <2%) that might be atypical.

FTIR spectroscopy data analysis
The complexity challenges of handling FTIR spectroscopy data are not dissimilar to applications for other methods, including NMR and MS, and therefore, the principles underlying the computational techniques employed tend to be similar. The major challenge is that increasingly biospectroscopy datasets tend to involve enormous numbers of spectral acquisitions (>10⁶ is not uncommon), each having a large number of variables, but only a few ‘features’ carry most of the relevant information; for instance, distinguishing biomarkers. Biospectroscopy data analysis leans towards searching for ‘features’ that will help address questions, such as:
- How can one distinguish between different physical groups (i.e., differentiating normal from dysplastic cells)?
- What are the intracellular mechanisms involved?

Computational analysis is carried out by statistical, pattern recognition and signal processing methods. Methods that allow for biomarker identification are usually preferred over ‘black-box’ methods, even if the latter perform slightly better. Data processing involves a sequence of steps, each one with its own purpose (Figure 3):

- Preprocessing: to correct for any noise, baseline alterations and other undesirable factors in the absorption spectra;
- Feature extraction: searching for relevant 'features' for data visualization or biomarker identification (e.g., between-category distinguishing wavenumbers);
- Classification: this phase typically has training and test stages. The former uses training data to train a classifier, which will be subsequently used in the latter stage for screening or diagnostic purposes.

Sensitivity & Specificity

As applied to biospectroscopy studies, various studies have employed biospectroscopy techniques to classify cervical cytology data, with varying degrees of sensitivity (i.e., ‘true positives’) and specificity (i.e., the probability that a patient is normal given that she was tested as normal – ‘true negatives’). Raman spectroscopy is another biospectroscopy technique; it is a powerful analytical tool that can be employed to identify the internal structure of molecules, and it does this by detecting the inelastic scattering of light from chemical bonds; this also generates a biochemical-cell fingerprint (80–81). Raman spectroscopy in vivo (using histology as the gold standard for diagnosis), while differentiating between high-grade disease and benign tissue, was found to have 89% sensitivity and 81% specificity (82). DeSantis et al. suggested the sensitivity of cervical spectroscopy was 95% with a specificity of 55% for benign lesions (employing an in vivo fluorescence variant of biospectroscopy); thus, the approach exhibited the potential to accurately detect moderate and high-grade dysplasia while reducing false-positive rates (83). However, employing ATR-FTIR spectroscopy, FTIR or Raman spectroscopy is, in our opinion, more applicable because one derives a
The prospect of applying biospectroscopy as a screening tool
Biospectroscopy has tremendous potential in the field of cervical cancer screening. Its capability to differentiate between grades of cytology, age groups and HPV-infected subtypes has been demonstrated [99]. However, this technology has not yet been trialed in a routine clinical setting. Most studies have compared the ability of biospectroscopy to distinguish different grades based on prior classification with visual screening of cytology. It is unfortunate that visual screening of cytology is a poor test, as most studies use it as the gold standard to compare new screening tests with. The major advantage of biospectroscopy lies in its potential application to identify biomarkers for disease status at many organ sites and not just cervical precancerous cells, without requiring staining or isotope-labeling. The potential of spectroscopy in categorizing cancer and intraepithelial neoplasia in tissues such as endometrium, prostate and various parts of the gastrointestinal tract has already been demonstrated [82-85]. The clinical implications of adopting spectroscopy as a screening or diagnostic tool are vast; in particular, the recent advances enabling spectra to be acquired in vivo and potentially in the operating theatre suggests that it could potentially revolutionize clinical practice [96]. Biospectroscopy techniques could also have a multipurpose functionality and utilization for mass screening for many other cancers, or their precancerous lesions, thereby increasing its cost-effectiveness in low-resource settings.

A hypothesized clinical setting
There is a need to validate the sensitivity and specificity of spectroscopy in a clinical setting. However, prior to this, it is necessary to map out a plan of a hypothesized clinical setting. At the moment, spectroscopy techniques remain an interdisciplinary research tool. The sample preparation and spectral acquisition process is inexpensive, but multiple steps are required. If done manually, it is potentially prone to errors due to sample misplacement or lack of attention in spectral acquisition; such problems are not unfamiliar when repetition leads to tiredness in a busy testing environment. Therefore, there is a need for the engineering of robotic instruments to increase throughput. For example, an instrument that would automatically search and acquire spectra on different parts of a cellular pellet would substantially reduce the labor of data collection and input the acquired spectral data into a classification algorithm. Figure 2 shows the steps in sample processing and data acquisition. The clinical setting designed to validate spectroscopy would be incorporated into a well-population screening program, where the current method employed is cervical cytology with follow-up colposcopy facilities. Figure 4 demonstrates the stepwise approach to validate FTIR spectroscopy in a clinical setting:

- Once funding is available and ethical approval obtained, each patient must be screened with liquid-based cytology;
- All smears that are negative with a negative HPV will be used as the gold standard, designated normal;
- Abnormal smears will then be classified as per the Bethesda system into ASCUS, ASCUS-H, low-grade squamous intraepithelial lesion, high-grade squamous intraepithelial lesion, squamous cell carcinoma, abnormal glandular cells of undetermined significance, atypical endocervical cells, atypical endometrial cells and adenocarcinoma;
- All abnormal spectra will be typed as per the cytological classification prior to computational data analysis;
- Patients with an abnormal smear will be referred for colposcopy and all abnormal lesions will be biopsied. Two biopsies over one at colposcopy increases the pick-up rate of CIN [97], so it is reasonable to advise routine biopsy of an abnormal lesion and the two biopsies necessary. Figure 3 shows the route from patient referral to spectral acquisition and data analysis;
- Separate studies will determine if there is a better diagnosis when biospectroscopy classifications are based on different histological grades as opposed to cytology.

There is a theoretical possibility that since biospectroscopy assesses the severity of nuclear atypia, through analysis of cellular components, it might be able to identify 'true' higher-grade disease that has not yet become visible on cytology. In clinical practice, CIN2 in adolescents is often managed conservatively. CIN2 may regress in up to 23% of patients per year and in up to 55% over 4 years [99]. To avoid over treating CIN2, it is essential to identify biomarkers of progression. Spectra can be obtained
from the cytology in this patient group, looking to differentiate retrospectively between the women whose disease progressed, from those in whom it remained static or regressed. This predictive application of spectroscopy needs further evaluation [100-104]. In the analysis, one can also examine the spectral classification of high-risk HPV-positive versus high-risk HPV-negative smears. Everything comes at a cost. It is important that in the clinical setting proposed above, the cost of smear acquisition, cytology processing and spectroscopy can be evaluated individually. The sensitivity and specificity can also be obtained individually, comparing it with cytological diagnosis first and then with histological diagnosis.

Future perspective
Starting a screening program that incorporates cytology and FTIR spectroscopy seems most feasible for a developing country, due to reduced labor costs with a significant benefit to the target population, yet the costs of initiating such a program may still be significant. If a new screening program based on biospectroscopy is to be initiated, the laboratory facilities need to be adequate with laboratory technicians replacing cytologists. The cost of training such technicians to perform spectroscopy methodologies must also be calculated. Colposcopy facilities with trained colposcopists, colposcopy nurses and histopathology facilities must also be in place. Following objective classification of cytology employing biospectroscopy [105-107], patients testing positive would go to colposcopy. Referral pathways for patients with a diagnosis of cancer as well as guidelines for management of precancer are essential. However, the fact remains that to successfully implement such a program in a developing country, awareness and political motivation is what appears to be most essential.

Acknowledgements
We thank the hospital staff who have facilitated many of our studies over the years.

Financial & competing interests disclosure
Funding from Rosemere Cancer Foundation is gratefully acknowledged. Role of funding source provided on equipment grants to allow preliminary studies in FL Martin’s laboratory at Lancaster University. The authors have no other relevant affiliations or financial involvement with any organization or entity with a financial interest in or financial conflict with the subject matter or materials discussed in the manuscript apart from those disclosed.

No writing assistance was utilized in the production of this manuscript.
## Executive summary

- A cost-effective and more robust screening tool for precancer of the cervix is needed.
- In parallel, effective and efficient screening programs based on current evidence are also needed; these will need to reach out to the socially deprived and ethnic minorities.
- The role of human papilloma virus (HPV) testing along with cytology, ‘cytological triage’, has been evaluated in many studies, where HPV-positive, smear-negative women get referred to colposcopy and HPV-positive, smear-negative women continue to get screened with HPV tests. Although HPV testing is 63% more sensitive than conventional cytology, it all comes at a much greater cost.
- Vaccination will alter the time intervals of screening in years to come, and although it does provide some cross-protection against other strains, long-term data on the durability of protection and efficiency are lacking.
- Nevertheless, we will not be able to do away with a screening program, especially since mass vaccination in resource-poor countries is unlikely to be applicable in the foreseeable future.
- Given this scenario, biospectroscopy along with computational data analysis offers the potential of an automated screening technique that is cheap and robust.
- Fourier-transform IR spectroscopy has the ability to detect subtle but important biochemical differences between cells, which allow it to differentiate between various grades of cellular abnormality or atypia.
- It also lends the potential to differentiate between specimens that are infected with one or multiple subtypes of the HPV virus. However, further research is required to verify whether spectroscopy can be used to successfully identify different grades of cytology and biomarkers of progression.
- The potential of biospectroscopy for detection of precancer of the cervix and assessment of cancer progression has been demonstrated in a large number of studies.
- This provides a strong argument for the allocation of resources and for technology transfer to the clinical setting (including machine and software engineering) towards the implementation of a new clinical screening sample incorporating biospectroscopy techniques.

## References


- Suggests a spectral phenotype in exfoliative cervical cytology, associated with transient versus persistent human papilloma virus infection.


- Describes sample preparation (a tissue section or cytology specimen), the application of IR spectroscopy tools, and computational analysis.


- Examines the remaining challenges in biospectroscopy studies towards sample preparation, instrumentation and data handling.


91. Schubert JM, Bird B, Papamarkakis K, et al. Spectral cytopathology of cervical samples: detecting cellular abnormalities in...


### Websites


Appendix 3

Abstracts and Posters
Abstract

Epidemiological evidence suggests that cancer with a pathogenesis attributable to exogenous carcinogenic agents may appear decades after initial exposure.

Hormone sensitive cancers including breast cancer implicate such environmental factors as lifestyle and diet for their aetiology.

Breast tissue undergoes continuous molecular and histological changes from the time of thelarche to menopause and thereafter. These alterations are both cyclical and longitudinal and depend on can be influenced by several environmental factors such as exposure to estrogen through pregnancy and breastfeeding, use of contraception and diet to name but a few.

Available research of latent stages of breast carcinogenesis has been limited to the time when hyperplastic lesions are present. Investigations aiming to identify a biomarker of commitment to disease in normal breast tissue are hindered by the molecular and histological diversity of disease free breast tissue.

Benign breast tissue from reduction mammoplasties provides an opportunity to study cellular biochemical differences between women of similar ages as well as differences appearing with advancing age. In this study, synchrotron radiation Fourier-transform infrared micro-spectroscopy (SR-FTIR) has been...
used to examine the terminal ductal lobular epithelium (TDLU), intra- and inter-lobular epithelium aiming to identify spatial and temporal changes within these areas. Principal component analysis followed by linear discriminant analysis of the mid-infrared spectra obtained by these cellular areas revealed unambiguous inter-individual as well as age related differences in each histological compartment interrogated.

Moreover, exploratory principal component analysis of luminal and myoepithelial cells within the TDLU indicated the presence of specific cells, which may potentially represent stem cells. Understanding of alterations within benign tissue may assist in the identification of a potential biomarker of commitment to disease within the latent pre-clinical stage of breast cancer.
Spatial and temporal age-related spectral alterations in benign human breast tissue: visualising pre-initiation events in disease-free tissue

Introduction

- Epidemiological evidence suggests that cancer with a pathogenesis attributable to exogenous carcinogenic agents may appear decades after initial exposure.
- Hormone-sensitive cancers (including breast cancer) are implicated in such environmental factors as lifestyle and diet for their etiology.
- Breast tissue undergoes continuous molecular and histological changes from the time of adolescence to the postmenopausal period.
- In this study, synchrotron radiation Fourier-transform infrared micro-spectroscopy (SR-FTIR) has been used to examine the terminal ductal lobular unit (TDLU), intra-lobular (intra-LU) and inter-lobular (driven-B) systems, aiming to identify spatial and temporal changes within these areas.

Method

- Human breast tissue from reduction mammoplasties of 3 age groups were analysed: 20-29, 30-39, and 40-49.

1. Tissue sections were prepared for spectroscopy with parallel H&E sections for histological identification.
2. Spectral image maps were obtained.
3. Spectra were cut to the region 1,030-900cm⁻¹, averaged every 5, normalized against Amide I peak.
4. Computational analysis using PCA followed by LDA was performed and results visualized as 2D scores plots and major discriminatory wavenumbers identified on loadings plots.

Spatial differences: between Individuals of similar ages

<table>
<thead>
<tr>
<th>Age Group</th>
<th>30-39 yrs</th>
<th>40-49 yrs</th>
</tr>
</thead>
<tbody>
<tr>
<td>TDLU</td>
<td>Wave number</td>
<td>Wave number</td>
</tr>
<tr>
<td>LDU</td>
<td>Wave number</td>
<td>Wave number</td>
</tr>
<tr>
<td>MDI</td>
<td>Wave number</td>
<td>Wave number</td>
</tr>
</tbody>
</table>

- Scores plots show significant separation between individuals for all examined regions using LDU and LDI.
- Loadings plots (LDI) identify wavenumbers responsible for separation between groups (highest and lowest peaks).
- Table 1 shows the principal segregating wavenumbers and their corresponding chemical assignments.
- There are significant inter-individual spectral variations within age groups.
- Segregation of TDLU among individuals of all age groups is obvious in the Amide I spectral region.
- Also, DNA/RNA (C=O stretching) spectral signature alterations demonstrate variation within the intra-LU areas for individuals of all age groups.

Table 1: Spatial changes: discriminating wavenumbers

<table>
<thead>
<tr>
<th>Wavenumber</th>
<th>TDLU</th>
<th>LDU</th>
<th>MDI</th>
</tr>
</thead>
<tbody>
<tr>
<td>1,030-900</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Conclusion

- Principal component analysis followed by linear discriminant analysis of the mid-infrared spectra obtained from breast tissue samples revealed unambiguous inter-individual as well as age-related differences in each histological compartment investigated.
- Amide I spectral region alterations were responsible for segregation between individuals of all age groups as well as segregation between age groups.
- Alterations within the DNA/RNA regions identified variation between individuals of all age groups. They were also responsible for segregation between age groups.
- Individual variation amongst individuals of similar age groups is likely caused by differences in the likelihood of diet, use of estrogen containing medication as well as genetic inheritance.
- In the search for potential spectral biomarkers of disease, it is important to understand normal spectral variation between individuals of similar age as well as variation between different age groups.

Table 2: Temporal changes: discriminating wavenumbers

<table>
<thead>
<tr>
<th>Wavenumber</th>
<th>TDLU</th>
<th>LDU</th>
<th>MDI</th>
</tr>
</thead>
<tbody>
<tr>
<td>1,030-900</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Surgical management of ovarian tumours largely depends on their histopathological diagnosis. Currently, screening for ovarian malignancy with tumour markers in conjunction with radiological investigations has a low specificity for discriminating benign from malignant tumours. Also, pre-operative biopsy of ovarian masses increases the risk of intra-peritoneal dissemination of malignancy. Intra-operative frozen section, although sufficiently accurate in differentiating tumours according to their histological type, increases operation times. This results in increased surgery-related risks to the patient and additional burden to resource allocation.

We set out to determine whether attenuated total reflection Fourier-transform infrared (ATR-FTIR) spectroscopy, combined with multivariate analysis can be applied to discriminate between normal, borderline and malignant ovarian tumours and classify ovarian tumour subtypes according to the unique spectral signatures of their molecular composition.

Formalin fixed, paraffin-embedded ovarian tissue blocks were de-waxed, mounted on Low-E slides and desiccated before being analysed using ATR-FTIR spectroscopy. Multivariate analysis in the form of principal component analysis (PCA), sequential progressive algorithm (SPA) and genetic algorithm (GA), followed by linear discriminant analysis (LDA) of the obtained spectra revealed clear segregation between benign versus borderline versus malignant tumours as well as segregation between different histological tumour subtypes, when these approaches are used in combination.

ATR-FTIR spectroscopy coupled with chemometric analysis has the potential to provide a novel diagnostic approach in the accurate intra-operative diagnosis of ovarian tumours assisting surgical decision making to avoid under-treatment or over-treatment, with minimal impact to the patient.
Attenuated total reflection Fourier-transform infrared spectroscopy coupled with variable selection techniques discriminates subtypes of human ovarian cancer

Georgios Theophilou, Kaelio M.G. Lima, Alan Mitchell, Helen F. Stringfellow, Pierre L. Martin-Hirsch and Francis L. Martin

Introduction

- Surgical management of ovarian masses largely depends on their histopathological diagnosis.
- Screening for ovarian malignancy with tumour markers and radiological investigations has a low specificity.
- Pre-operative biopsy of ovarian masses increases the risk of intraoperative dissemination of malignancy.
- Intra-operative frozen section, although sufficiently accurate in differentiating tumours according to their histological type, increases operation times.
- Attenuated total reflection Fourier-transform infrared (ATR-FTIR) spectroscopy, combined with multivariate analysis (PCA-LDA) and variable selection techniques (PCA-LDA and GA-LDA) were applied to examine if this technique can discriminate between different ovarian tumour types according to the unique spectral signatures of their molecular composition.

Method

1. Human ovarian tissue from 109 cases receiving surgery without neoadjuvant chemotherapy.
2. Formalin-fixed, paraffin-embedded tissue blocks of different ovarian tumour types were de-waxed, mounted on Low-E slides and desiccated before being analysed using ATR-FTIR spectroscopy.
3. Spectra were cut to the region 1,650-900cm⁻¹ and normalized against Amide I peak.
4. Multivariate analysis using PCA-LDA and variable selection techniques (SPA-LDA, GA-LDA) were performed and results visualized as 2D and 3D scores plots.

Results

Analysis of all classes together

- 7 ovarian cancer types were classified as follows:
  0. High grade serous
  1. Low grade serous
  2. Endometrioid
  3. Mixed
  4. Mucinous
  5. Clear cell
  6. Coronaromas

Analysis of classes in pairs

- 70% for training set
- 15% for internal validation set
- 15% for external validation set

Classification machine

- A classification machine was constructed using the three analytical techniques to discriminate classes in pairs.
- The table below represents the classification algorithms for each class pair.
- The algorithms are in order of accuracy (%): Red > Orange > Green

<table>
<thead>
<tr>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
</tr>
</thead>
<tbody>
<tr>
<td>PCA-LDA</td>
<td>SPA-LDA</td>
<td>GA-LDA</td>
<td>PCA-LDA</td>
<td>SPA-LDA</td>
<td>GA-LDA</td>
<td>PCA-LDA</td>
</tr>
<tr>
<td>85.66</td>
<td>85.66</td>
<td>85.66</td>
<td>72.50</td>
<td>72.50</td>
<td>72.50</td>
<td>72.50</td>
</tr>
<tr>
<td>85.66</td>
<td>85.66</td>
<td>85.66</td>
<td>72.50</td>
<td>72.50</td>
<td>72.50</td>
<td>72.50</td>
</tr>
<tr>
<td>85.66</td>
<td>85.66</td>
<td>85.66</td>
<td>72.50</td>
<td>72.50</td>
<td>72.50</td>
<td>72.50</td>
</tr>
<tr>
<td>85.66</td>
<td>85.66</td>
<td>85.66</td>
<td>72.50</td>
<td>72.50</td>
<td>72.50</td>
<td>72.50</td>
</tr>
<tr>
<td>85.66</td>
<td>85.66</td>
<td>85.66</td>
<td>72.50</td>
<td>72.50</td>
<td>72.50</td>
<td>72.50</td>
</tr>
<tr>
<td>85.66</td>
<td>85.66</td>
<td>85.66</td>
<td>72.50</td>
<td>72.50</td>
<td>72.50</td>
<td>72.50</td>
</tr>
<tr>
<td>85.66</td>
<td>85.66</td>
<td>85.66</td>
<td>72.50</td>
<td>72.50</td>
<td>72.50</td>
<td>72.50</td>
</tr>
</tbody>
</table>

Conclusion:

We show that ATR-FTIR spectroscopy coupled with variable selection techniques has the potential to provide a novel diagnostic approach in the accurate diagnosis of ovarian tumours assisting intra-operative surgical decision making to avoid undertreatment or overtreatment, with minimal impact to the patient.

Acknowledgements:

For reference and queries please contact g.theophilou@lancaster.ac.uk
A biospectroscopic analysis of human prostate tissue obtained from different years points to a trans-generational alteration in spectral phenotype

Georgios Theophilou¹, ², Kássio M.G. Lima¹, ³, Matthew Briggs², Guilherme Post Sabin³, Pierre L. Martin Hirsch¹, ², Helen F. Stringfellow², Francis L. Martin¹*

¹Centre for Biophotonics, LEC, Lancaster University, Lancaster LA1 4YQ, UK
²Department of Obstetrics and Gynaecology, Central Lancashire Teaching Hospitals NHS Foundation Trust, Preston, UK
³Institute of Chemistry, Biological Chemistry and Chemometrics, Federal University of Rio Grande do Norte, Natal 59072-970, RN-Brazil

Abstract

Prostate cancer is the most commonly diagnosed male malignancy in the world; however, there is marked geographic variation in its incidence. This may be associated with adoption of a Westernised lifestyle. We set out to determine whether attenuated total reflection Fourier-transform infrared (ATR-FTIR) spectroscopy or Raman spectroscopy combined with principal component analysis-linear discriminant analysis (PCA-LDA) or variable selection techniques employing genetic algorithm (GA) or successive projection algorithm (SPA) could be utilised to explore differences between prostate tissues obtained from differing years. In total, 156 prostate tissue samples from transurethral resection of the prostate (TURP) procedures for benign prostatic hyperplasia (BPH) were collected from 1983 to 2013. These were distributed according to the year of collection to form seven categories: 1983-1984 (n=20), 1988-1989 (n=25), 1993-1994 (n=21), 1998-1999 (n=21), 2003-2004 (n=21), 2008-2009 (n=20) and 2012-2013 (n=21). Ten-μm-thick tissue sections were floated onto Low-E (IR-reflective) slides for ATR-FTIR or Raman spectroscopy. Resulting scores plots for PCA-LDA, SPA-LDA or GA-LDA from ATR-FTIR data revealed marked segregation between the seven categories. In fact, there was a chronological development of prostate tissue spectroscopic alterations with successive categories. This classification was less evident following Raman spectroscopy but here also, a significant separation between categories was identified. Moreover, examination of the two categories that are at least one generation (30 years) apart indicated highly significant segregation, especially at spectral regions containing DNA and RNA bands (≈ 1,000-1,490 cm⁻¹), involving nucleic acids, phosphate and deoxyribose modifications. This may point towards alterations that have occurred through chemical genotoxicity or through epigenetic modification of chromatin structure. Immunohistochemical studies for DNA methylation and hypomethylation supported the results obtained by vibrational spectroscopy. This study points to a trans-generational phenotypic change in human prostate as a function of spectral alterations.