APPLICATION OF WHOLE GENOME AMPLIFICATION FOR THE INVESTIGATION OF GENOMIC MUTATIONS IN HODGKIN'S LYMPHOMA

by

PRADEEP RAMAGIRI

This project is submitted in partial fulfilment of the requirements for the award of the MRes in Biomedical Research

School of Cancer Sciences

College of Medical and Dental Sciences

University of Birmingham

August 2012

Abstract

Tumour cells of classical Hodgkin's lymphoma (cHL) comprise only a fraction (1% or less) of the total cellular infiltrate. The low tumour cell numbers pose a great difficulty in the investigation of underlying genetic events in cHL development. This hurdle can be overcome by exome sequencing of a small number (~50) of isolated HL cells. This study was aimed at establishing 'whole genome amplification (WGA) of the HL cell lines and patients' samples, which can be potentially used to for the exome sequencing. This study used 6 HL cell lines (HDLM-2, KM-H2, L-428, L540, L-591, and L-1236) and 4 HL patients' samples (1.5, 6.4, 7.28, and 7.8) to generate PCR products for exons 2 to 9 of the *TNFAIP3* gene. We also aimed to demonstrate whether microdissection and whole genome amplification methods will introduce any changes to the sequences, by analysing the mutation spectra for the whole genome 'amplified', and 'unamplified' (KM-H2 and L-591) cell lines. We obtained good quality WGA product for all the microdissected cell line samples and patient samples. However, only patient sample 6.4 amplified all of these exons except the exon 6 and 7.1 in the following PCR experiments. The quality of the sequences obtained for the amplified cell line samples (76.5%) was as good as the quality for the unamplified cell lines (79.7%). We observed same nucleotide changes in the cell lines KM-H2 (deletion of intron 2– exon 6 region), and L-1236 (G491A) as reported in a previous study (**16**).

Acknowledgement

I sincerely thank Dr. Wenbin Wei, Prof. Paul Murray, and Dr. John Arrand for their kind support and guidance.

Table of contents

Table of figures

1) Introduction

Hodgkin's lymphoma (HL) is one of the most prevalent types of malignant lymphomas in the western countries, with an incidence rate of 3 in 100,000 persons each year. In the United Kingdom (UK), 1,852 people diagnosed with HL, and 319 HL related deaths were registered between the year 2009 and 2010. Although HL accounts for less than 1% of all cancers in the UK (**3**), about 4-5% of all childhood cancers in the UK (around 64 children per year) are registered as HL. HL is predominant in older children (between 10 to 14 years old) and no infant cases with HL are registered so far (**12**). HL can be successfully treated in the developed countries like the UK. However, the mortality rate is high in under-developed countries (one in two registered cases), and developing countries (one in four cases) (**3**).

This disease was first described in 1832 by Thomas Hodgkin **(15)**. A majority of HL cases (95%) belong to the classical form of the disease and the remaining 5% are of nodular lymphocyte predominant HL (NLPHL). Based on the histology and cellular composition, classical HL is subdivided into nodular sclerosis (NS), mixed cellularity (MC), lymphocyte-depleted (LD), and lymphocyte-rich (LR) forms. The tumour cells in classical Hodgkin's lymphoma (cHL) are known as the Hodgkin and Reed-Sternberg (HRS) cells, and in nodular lymphocyte predominant lymphoma (NLPHL) they are known as lymphocyte predominant (LP) cells. Irrespective of the subtype, the tumour cells account for only 1-10% of the total cellular infiltrate. This low tumour cell content along with poor chromosome morphology has made it very difficult for researchers to investigate the molecular events in the development of HL.

1.1) Molecular biology of Hodgkin's lymphoma

Germinal centre (GC) B cells that escape apoptosis after acquiring fatal immunoglobulin V gene mutations are believed to give rise to HRS cells. In rare cases (~ 2%), HRS cells originate from T cells. On the other hand, LP cells derive from antigen-selected GC B cells (**2**).

Comparative genomic hybridisation (CGH) was used to detect copy number aberrations in HRS cells. Chromosomal arms 1p, 6q, 7q, 11q, 12q, and 14q have shown recurrent multiple chromosomal break points. Frequent recurrent gains have been observed on chromosomes 2, 5, 9, and 12, whereas, frequent losses are reported on chromosomes 13, 21 and Y (**9**).The genomic instability of the HRS and LP cells will depend on the complexity of genetic alterations and poses difficulty in identifying the genetic aberrations related to pathogenesis (**2**).

1.2) Deregulation of multiple signalling pathways in HRS cells

In recent years, deregulated expression of several proteins and aberrant activation of a number of signalling pathways that are associated with the pathogenesis of HRS cells have been identified, namely, the transcription factors (NF-κB, STAT, AP-1, and Notch1), multiple receptor tyrosine kinase (RTKs), PI3K, and MEK/ERK pathways. These pathways are vital for the proliferation and survival of the HRS cell phenotype (**2**).

1.2.1) Anti-apoptotic mechanisms in HRS cells

As the HRS cells are derived from pre-apoptotic GC B cells, evading apoptosis is vital in the transformation. As mentioned above several signalling pathways are aberrantly activated to facilitate the generation of HRS cells, most importantly NF-κB.

1.2.2) NF-κB pathway activation

The NF-κB transcription factor family consists of five members, Rel A, Rel B, c-Rel, NF-κB1, and NF-κB2, which can act as homo-and/or heterodimers. In the classical NF-κB pathway, the p50/p65 dimer was kept in the cytoplasm by binding to IκBα, a NF-κB inhibitor. As a result of activating the pathway, IKK (IkB kinase complex) kinases phosphorylate IκBα and thereby induce its proteosomal degradation resulting in the translocation of NF-κB dimers into the nucleus and which subsequently activate multiple target genes. This activation can be partially mediated through receptor signalling, such as the TNF family member CD40. Other receptors involved in similar NF-κB activation include CD30, TACI, BCMA, RANK and Notch1 (**15**).

However, these signalling pathways are not sufficient for the strong constitutive NF-κB activation. Several genetic aberrations have been identified in HRS cells that affect the NF-κB pathway. The REL gene show copy number gains or amplification in 40% to 50% of HL cases (**15**). The *NFKBA* gene (which encodes IκBα and IκBε) has also been shown to have mutations. Recent studies have shown *TNFAIP3* (Tumour Necrosis Factor, Alpha-induced Protein-3) as a novel tumour suppressor gene that is involved in the regulation of the NF-κB pathway. The *TNFAIP3* gene encodes A20 protein that has ubiquitinase and deubiquitinase functions. A20 is a negative regulator of NF-κB signalling and acts upstream of the IKK kinases (**15**). Some of the previous studies on HL described mutations/deletions in this gene in 30%- 40% of the cases (**15-16**).

1.3) TNFAIP3

TNFAIP3 gene expression is induced by tumour necrosis factor (TNF). *TNFAIP3* encodes a zinc-finger protein, A20, which is known to inhibit NF-κB activation and TNF-mediated apoptosis (**18**). The *TNFAIP3* gene is located on Chromosome 6q23

and has 9 exons. Coding sequence of this gene is 2373 nucleotides long, and encodes 790 amino acids (**20**).

TNFAIP3 mutation seems to have an inverse correlation with the presence of Epstein-Barr virus (EBV) in HRS cells. Around 70% of EBV^{-ve} cases showed *TNFAIP3* mutations (mostly deleterious) where as only 12% of the EBV infected patients showed mis-sense mutations in*TNFAIP3* gene. This might indicate an alternative NF-κB activation mechanism involving A20 inactivation and EBV infection (**16**). Functional studies involving re-expression of A20 in A20 deficient HL cell lines resulted in reduced expression of NF-κB target genes and negatively affected cell survival, demonstrating that A20 is a tumour suppressor (**16**). The role of *TNFAIP3* as a tumour suppressor gene has also been demonstrated in other lymphomas, such as primary mediastinal B cell lymphoma (PMBL), which are also characterized by constitutive activation of the NF-κB pathway, as mentioned above (**16**).

For the better understanding of the genetic alterations and their role in the HL development, we need to isolate the tumour cells and study the variations across the whole genome.

1.4) Microdissection method

Microdissection techniques can be divided into three main categories: manual extraction, selective ablation, and laser capture microdissection (LCM). In this study we used LCM method to isolate the tumour cells from the fixed cell line samples. LCM was first designed by 'Lance Lotta' and co-workers in the mid 1990 at the NIH, Maryland, USA. There are 3 main types of LCMs, Infrared (IR) LCM, Ultra violet (UV) LCM, and a combined version (IR/UV LCM). In the IR LCM method, the surrounding tissue of the target specimen were filled with a melted ethylene vinyl acetate (EVA) by laser activation and then the target specimen selectively adhere to the thermoplastic membrane by low energy laser pulse. The targets are then extracted by removing the polymer from the tissue surface. This method uses low energy beams so low photo chemical effects on the specimen, although visualization in this method is fuzzy. UV LCM method, in contrary, selects and cuts the target specimen with a fine laser beam. This method avoids the unwanted debris surrounding the specimen and aided with better visualization. We used UV LCM method to microdissect the HL cells in our study (**31**).

1.5) Whole genome amplification (WGA)

A single cell usually contains ~6-7 picograms (pg) of genomic DNA (**30**). Most of the sequencing based methods such as whole genome sequencing or exome sequencing methods require ~3-5 micrograms (µg) of a good quality starting material (DNA) per reaction. Hence, we need to amplify the genome by more than 1000 fold, if we are working on a sample material which contains only few (1-100) cells.

Various commercial WGA kits are available to meet the above requirements, such as QIAgen's 'REPLI-g', GE Healthcare's 'GenomiPhi', and Sigma's 'GenomePlex'. These kits work on the either 'PCR' or 'multiple displacement amplification' (MDA) method. PCR-based methods use DNA polymerase from a thermophilic bacterium (*Thermus aquaticus*) and amplify the DNA by repeated 'denaturation-elongation' steps at various temperatures. MDA method employs 'bacteriophage (phi 29) DNA polymerase' for the rolling circle amplification of a DNA template at a 'constant' temperature (isothermal). MDA method is much slower (~18 hours) than the PCR based method (~3 hours) (**29**).

The above commercial kits facilitate a straight forward application of WGA method to generate the necessary quality and quantity of the product. However, the choice of the kit will depend on the research question and the sample material (old, degraded, fixed, paraffin embedded and frozen etc.) as the reliability (percentage of samples met the requirement), fidelity, and coverage will vary between these technologies. For genotyping applications, the longer MDA method ('REPLI-g') found to have better 'accuracy' and coverage than the faster PCR method (Genomeplex) (**29**).

1.6) Aim

In this study, we aimed 1) to establish 'whole genome amplification method' that can be potentially used in genome sequencing of HL cell lines and HL patient samples; 2) to confirm genetic changes in the *TNFAIP3* gene of HL cell lines discovered by a previous study; 3) to investigate whether there are genetic changes in the *TNFAIP3* gene of HL cell line L540; 4) to investigate whether microdissection and whole genome amplification methods will introduce any changes to the genome sequences using the KM-H2 and L-591 cell lines by comparing WGA amplified and unamplified DNA; and 5) to investigate the applicability of the whole genome amplification method to the HL patient samples.

2) Materials and Methods

2.1) COSMIC (Catalogue of Somatic Mutations in Cancer) search for the frequently mutated genes in Hodgkin's lymphoma

COSMIC is a comprehensive database designed to gather, curate, organise, and present the information on somatic mutations in cancer and put that information in the public domain (**23**). This database allows researchers to investigate the key oncogenes, gene fusions and structural rearrangement annotations across numerous cancer samples. COSMIC integrates the manually curated cancer mutation data from scientific literature with the output from the Cancer Genome Project (CGP) (**5**). I searched for the genes that are heavily mutated in the Hodgkin's lymphoma in this data base by selecting the type of cancer and the histology subtype using COSMIC web interface.

2.2) Patient samples and cell lines

This study was part of an ongoing research that has been approved by the National Children's Cancer and Leukaemia Group (CCLG) Biological Studies Steering Group. These samples were collected by the tissue bank of the National Children's Cancer and Leukaemia Group (CCLG). Four patient samples that were used in this study (labelled as 1.5, 6.4, 7.21 and 7.8) were already fixed and microdissected into 'PALM membrane adhesive cap' (PALM tubes) (50 cells in each tube) and stored at -20^0 C.

Human Hodgkin's lymphoma cell lines HDLM-2 (DSMZ No: ACC 17), KM-H2 (DSMZ No: ACC 8), L-428 (DSMZ No: ACC 197), L-540 (DSMZ No: ACC 72), L-591(DSMZ No: ACC 602), and L-1236 (DSMZ No: ACC 530) were obtained from *Leibniz Institute DSMZ-German Collection of Microorganisms and Cell Cultures,* Inhoffenstraße 7B, 38124 Braunschweig, GERMANY (**11**).

All the 6 cultured HL cell lines, HDLM-2, KM-H2, L-428, L-540, L-591, and L-1236, were cultured and kindly provided by Eszter Nagy, a PhD student at Prof. Paul Murray's lab, School of Cancer Sciences, College of Medical and Dental Sciences, University of Birmingham, Birmingham, UK.

2.3) Laser microdissection and pressure catapulting of cells

Appropriate volumes of cultured KM-H2 (6.25 mL at $8X10^4$ cells/mL concentration) and L-591 (1.47 mL at $17X10^4$ cells/mL) were centrifuged at 3000 r.p.m for 5 minutes and the pellet was re-suspended with 500 µL Phosphate Buffered Saline (PBS) and 100 µL of this spun on to membrane-covered slides (PALM) using a Cytospin at 1000 r.p.m. for 5 minutes. Immediately, the slides were air dried and were fixed using cold ethanol and stained with Haematoxylin. This entire procedure was performed on ice.

Then the slides were air dried and stored at -20°C until the microdissection was performed. The slides were then taken to the Human Biomaterials Resource Centre, College of Medical and Dental Sciences, University of Birmingham, UK (**10**), samples were microdissected using the PALM microbeam HBO100/AX10 Laser Capture Microdissection (LCM) apparatus (Carl Zeiss PALM Microbeam, Carl Zeiss Ltd, Cambridge, UK) (**27**) After calibrating the 'power and focus' (Energy = 40, Focus= 69, LPC= -20, and Delta = 2), the cells were selected for cutting and capturing on to the caps of 'PALM membrane adhesive cap' tubes (PALM tubes).

Figure 1 show how a single cell on the KM-H2 PALM membrane slide was selected and captured. The slides were micro-dissected in different batches of a single cell and 50 cells onto the cap of PALM tubes. The caps of these tubes were checked under the LCM for the presence of cell sections adhered to them. The 'PALM membrane adhesive cap' tubes with cells were stored at -20°C until they were used in the whole genome amplification method.

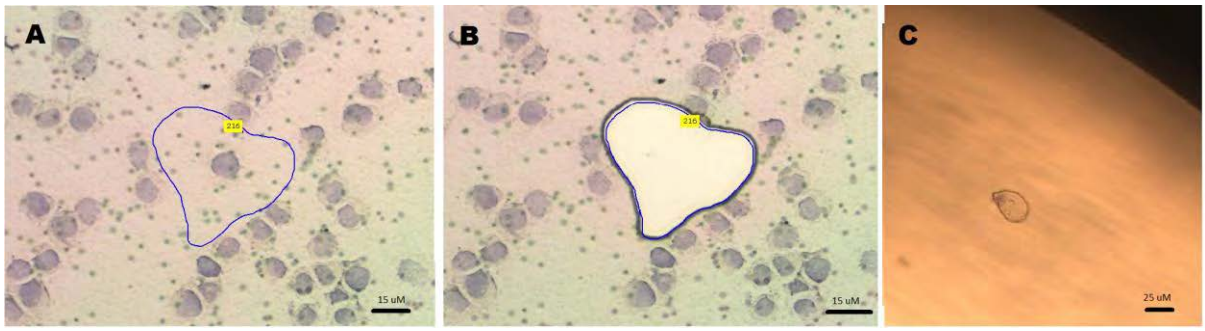

Figure 1. Microdissection of a single KM-H2 cell. A: A single KM-H2 cell was selected; B: cut was made; C; image of the membrane section containing the single KM-H2 cell captured on the PALM tube cap. The bar in the figures A, B, and C represent scale 15 uM,15 uM, and 25uM respectively.

2.4) DNA extraction of the HL cell lines

All the 6 HL cell lines, HDLM-2, KM-H2, L-428, L-540, L-591 and L-1236 were cultured and appropriate volumes of the cell cultures ($3X10^6$ cells) spun down and their pellet was re-suspended in PBS and sent to the Genomics lab, at the University of Birmingham for the DNA extraction (**7**).

2.5) Whole genome amplification (WGA) of the HL cell lines

Whole genome amplification of the microdissected cells was initially carried out using QIAGEN's REPLI-g Mini Kit (25 X, Catalogue no.150023, **19**) according to the manufacturer's instructions. REPLI-g kit works on a principle known as Multiple Displacement Amplification (MDA) method where random primers (hexamers) bind to the template and generate fragments at a constant temperature with the help of a high fidelity DNA polymerase enzyme, usually 'Φ29 DNA polymerase'. The resulting fragments of DNA are larger than conventional PCR products and with lower error frequency.

However, the REPLI-g Mini Kit didn't give consistent amplification of the microdissected cell material. As an alternative, GenomePlex® Single Cell Whole Genome Amplification Kit (WGA4) from Sigma-Aldrich, UK (**7**) was used for whole genome amplification from the microdissected cells. WGA4 works on a principle of 'non-enzymatic random fragmentation of the genomic DNA' (**28**). The WGA was performed according to the manufacturer's instructions. After the lysis procedure tubes that contained cell(s) were again checked under the microscope for the presence of cells on their caps to ensure the cells were lysed and 'off' their caps.

2.6) PCR

PCR primers designs were taken from a study by Roland Schmitz et al. (**16**) and used for the PCR and sequencing reactions (Supplementary Table 1). These primers have a melting temperature (Tm) ranging between 55°C and 74°C. Primers were ordered from Sigma-Aldrich, at 100 µM concentration (**24**). The primer stocks were diluted with nuclease free water as per the requirement (100 picomoles/µL for PCR, and 3.2 picomoles/ µL for sequencing reactions). The diluted primers and the stocks were stored at -20°C.

PCR was set up using QIAGEN's multiplex PCR Kit for a total of 20 µL reaction mixture according to manufacturer's instructions. Hot start PCR was programmed as follows: 95°C for 5 minutes (Denaturing of DNA) followed by 35 cycles of 94°C for 45 seconds, 65°C for 1 minute (Annealing) and 72°C for 1 minute (Extension) and then the final extension at 72°C for 10 minutes.

The PCR product was then validated by running 6.0 µL of the product and 3.0 µL of 100bp ladder from Invitrogen, UK (Catalogue.No.15628-019) on a 1.5% Agarose gel. The bands were checked for approximate length of the product for each set of reaction (Supplementary Table 2).

2.7) PCR product purification

ExoSAP-IT for PCR Product Clean-Up (catalogue no. 78200 200 UL), Affymatrix, UK, was used to clean up the PCR products according to the manufacturer's instructions. The concentration of the clean PCR product was measured by loading 1.0 µL on the Nanodrop 1000 spectrophotometer (Thermo scientific, USA).

2.8) Sequencing

Clean PCR products were then diluted to the appropriate concentration (as requested by the Genomics lab) in a total of 10 µL reaction volume (Supplementary table 3) and sent to the sequencing service facility at Genomics lab, University of Birmingham, UK. These samples were later loaded on to the ABI 3730 Sequencing machine at the Genomics lab.

2.9) Sequence Analysis

Sequences retrieved from the Genomics lab, University of Birmingham, for the cell lines and patient sample were in the .ab1 format (format for the raw DNA data taken out from a scientific instrument and output from Applied Biosystem's Sequence Analysis software). All these files were aligned and edited with the genomic reference sequence, (NCBI Reference Sequence: NC_000006.11, gi|224589818:138188581-138204449, **21**) using '**Sequencher 5.0 Demo**' software(**17**).

3) Results

3.1) COSMIC search of frequently mutated genes in Hodgkin's lymphoma

COSMIC database was searched for the key cancer genes associated with Hodgkin's lymphoma. The search yielded the top 20 genes most frequently mutated in the HL patient samples, namely SOCS1, TNFAIP3, NRAS, TP53, CYLD, HRAS, KDM6A, CDKN2A, CDKN2a(p14), BCR, ETV6, PCM1, NPM1, KRAS, JAK2, ATM, BRAF, PIK3CA, PTEN, and WT1. Of these 20 genes 3 genes were found to be mutated in higher percentage of patient population in which they were tested; namely, *TNFAIP3, SOCS1, and NRAS* (Figure 2). A literature review on the importance of these genes in HL was carried out to select a single gene for our study. All of these genes are important in the HL disease development. However, TNFAIP3 gene was selected on the basis of its central role in HL development and its location on 6q23 (this region is frequently deleted in B cell lymphomas). *SOCS1* is located on chromosome 16p13.13 and the genomic instability in this region is not implicated in the HL development, so far.

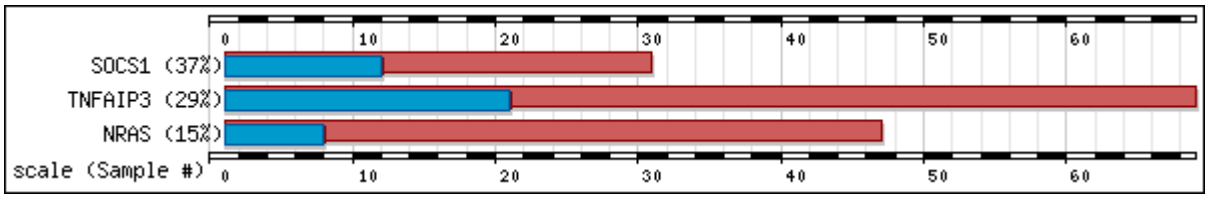

Figure 2. Genes frequently mutated in Hodgkin's lymphoma. The red bar represents total number of samples and the blue bar represents the number of samples with mutations (Source: 23).

3.2) PCR of TNFAIP3 exons using unamplified genomic DNA

Test PCR reactions were set up with A20 E2F/E2R primers on all the six cell line DNA (Figure 3_1), and with all the *TNFAIP3* gene primers on unamplified L-591 cell line to obtain optimal PCR conditions (Figure 3_2). Another PCR reaction was set up with all the primer sets for unamplified KM-H2 cell line sample. The PCR products were run on a 1.5% agarose and checked for the bands at the appropriate size (Supplementary Table 2).

As mentioned in some of the previous studies (**16**), a deletion of a region between intron 2 and exon 6 in the *TNFAIP3* gene was observed in the unamplified sample of the KM-H2 cell line in this study (Figure 3_3).

This PCR method was repeated for the rest of the cell line and for the whole genome amplified KM-H2 and L-591, with all the *TNFAIP3* gene primer sets until we obtained PCR product with appropriate band size.

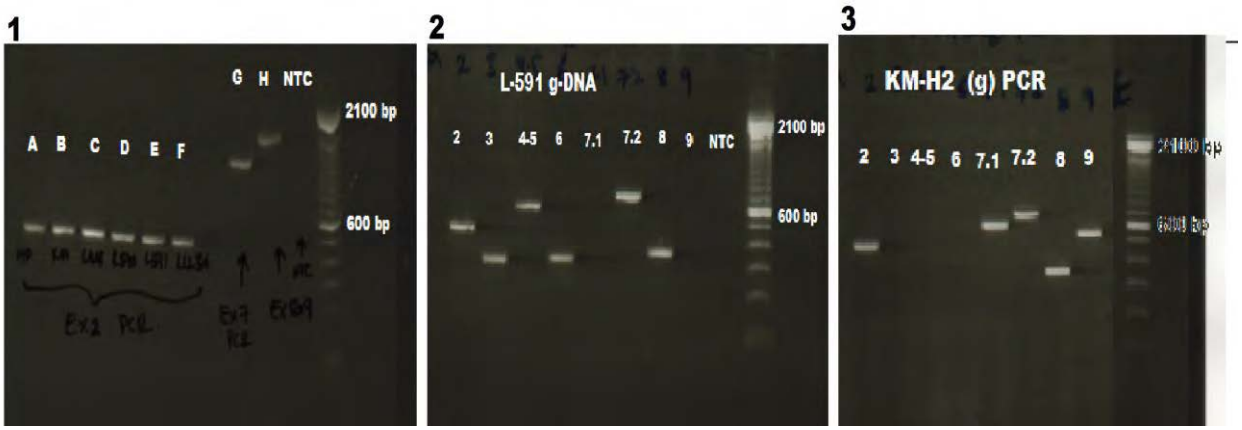

Figure 3. Gel images of the PCR experiments. 1. A-F: exon 2 PCR products (462 bp) for HDLM-2, KM-H2, L-428, L-540, L-591, and L-1236 cell lines respectively; G,H: exon 7 and 9 PCR products for HDLM-2; NTC: non-template control; 2. L-591 genomic DNA PCR. Numbers 2, 3, 4/5, 6, 7.1, 7.2, 8, 9 represent the corresponding *TNFAIP3* exon PCR products for L-591 genomic DNA with band sizes, 462, 306, 588, 311, 568, 653, 328, and 527 respectively (Supplementary table 2) ; 3. The PCR using the KM-H2 genomic DNA shows exon 3 to 6 are missing.

3.3) Whole genome amplification

Microdissected sections on the 'PALM tubes' caps were checked before and after the whole genome amplification procedure, to ensure the capture and lysis of the cells. Whole genome amplification of microdisected samples containing a single KM-H2 cell section, and 50 L-591 cells section was carried out using REPLI-g mini kit. Products of the WGA were run on a 2% agarose gel (Figure 4).

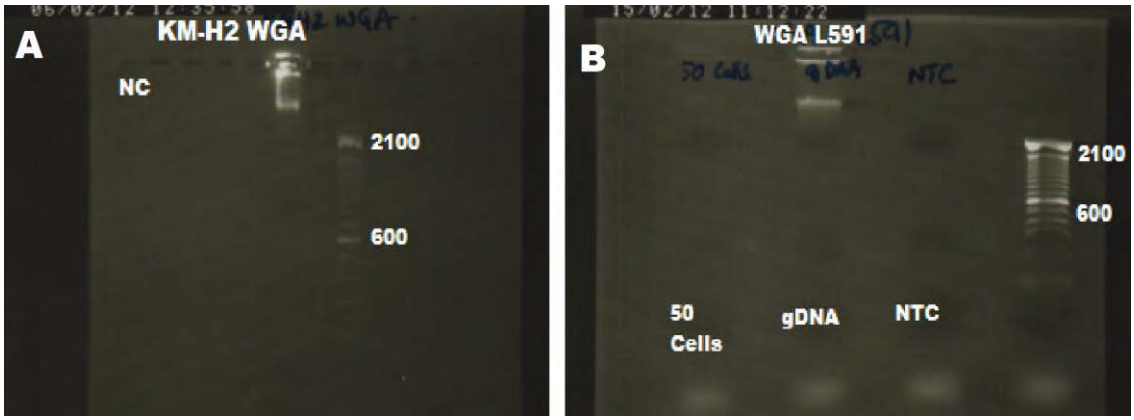

Figure 4. Images of a single KM-H2 cell, and 50 L-591 cells WGA products is showed. A: REPLIi-g kit successfully amplified KM-H2 cell genome, NC: non template control; B: REPLI-g failed to amplify 50 L-591 cells. However, the gDNA (positive control) was amplified. NTC: non-template control.

REPLI-g kit was inconsistent in whole genome amplification process of other cell samples with a single cell and 50 cells (Figure 4), and often showing positive results for the non template control (NTC) in the reactions.

As an alternative, SIGMA's GenomePlex Single Cell Whole Genome Amplification kit (WGA4) was used in the whole genome amplification of microdissected cell line (KM-H2 and L-591) and patient (1.5, 6.4, 7.21, and 7.8) samples (Figure 5) according to the manufacturer's instructions. The amplified products were run on a 2% agarose gel. WGA4 kit amplified all the samples.

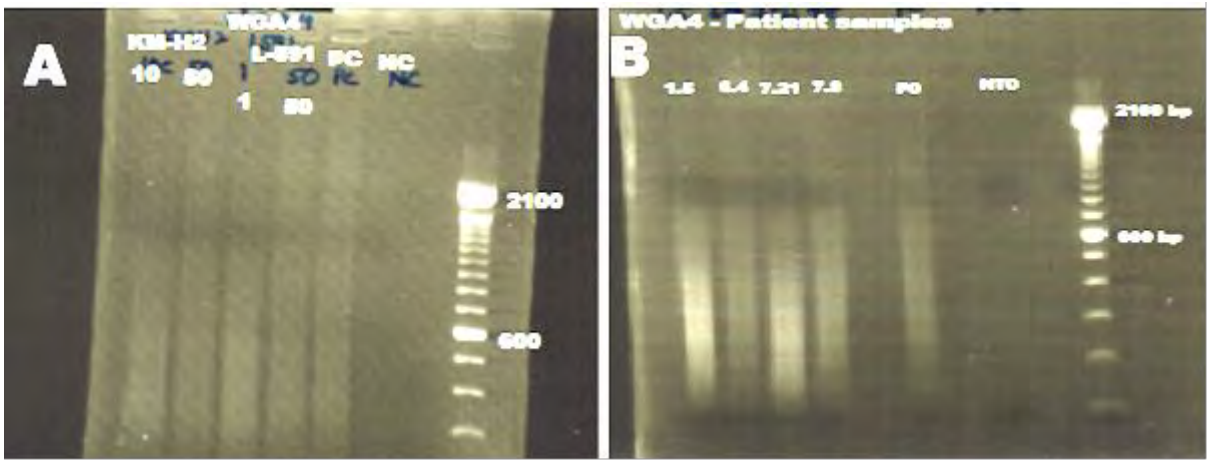

Figure 5. KM-H2, L-591 and patient samples amplification using WGA4 kit. A: 10 cells and 50 cells of KM-H2; a single cell and 50 cells of L-59; PC: positive control; NC: non-template control; B: patient samples 1.5, 6.4, 7.21, and 7.8; PC: positive control; NTC: non-template control.

3.4) Whole genome amplification – PCR

A test PCR with the WGA4 products of KM-H2 (10 cells, 50 cells) and L-591 (a single cell and 50 cells) microdissected samples was set up using the exon 8 primers. Only the PCR for WGA4 product of L-591 microdissected sample with 50 cells has worked (Figure 6_1).

To rule out the possibility, that, this PCR failure could have been due to the lower amount of initial DNA template concentration, another PCR test was set up with various template concentrations (0.5, 1.0, 2.0, 3.0, and 4.0 µL) of WGA4 products with the exon 8 primers. In this experiment WGA4 products of KM-H2 (10 cells, 50 cells) and 50 L-591 cells microdissection sample have worked. The L-591 single cell WGA4 product again failed to amplify this region (Figure 6_2).

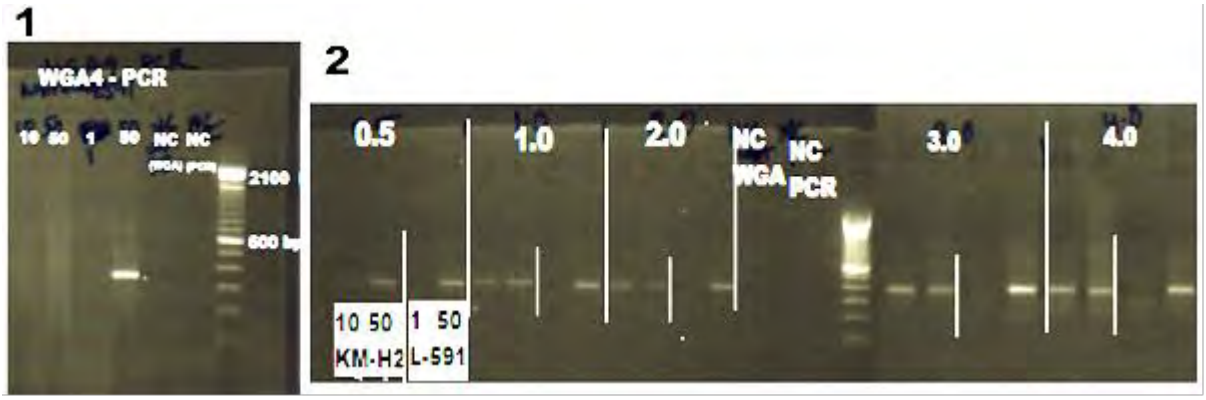

Figure 6. Test PCR on WGA4 products of KM-H2 and L-591 samples. 1. A test PCR with the WGA4 products of KM-H2 (10 cells, 50 cells) and L-591 (a single cell and 50 cells) microdissected samples was set up using the exon 8 primers. NC (WGA): non-template control from WGA4 experiment; NC (PCR): non-template control for this PCR experiment; 2. A test PCR with various template concentrations (0.5, 1.0, 2.0, 3.0, and 4.0 µL) of KM-H2 (10 cells and 50 cells) and L-591 (a single cell and 50 cells) WGA4 products; NC (WGA): non-template control from WGA4 experiment; NC (PCR): non-template control for this PCR experiment.

PCR on WGA4 product of L-591 for all of the *TNFAIP3* exons were successful (Figure7_1:A) and the PCR experiments on WGA4 product of KM-H2 yielded the same results as those of cell lines, such as the deletion in the region between intron 2 and exon 6 (Figure7_1:B).

WGA4 was used to amplify some Hodgkin's lymphoma patient samples, namely, 1.5, 6.4, 7.21 and 7.8 (50 cells each). The amplified products were run on a 2% agarose gel (Figure 7_2). PCR experiments on patient samples 1.5, 7.4 and 7.8 were failed (Figure7_2:A). However, patient sample 6.4 successfully PCR amplified all the exons except exon 7.1(Figure7_2:B).

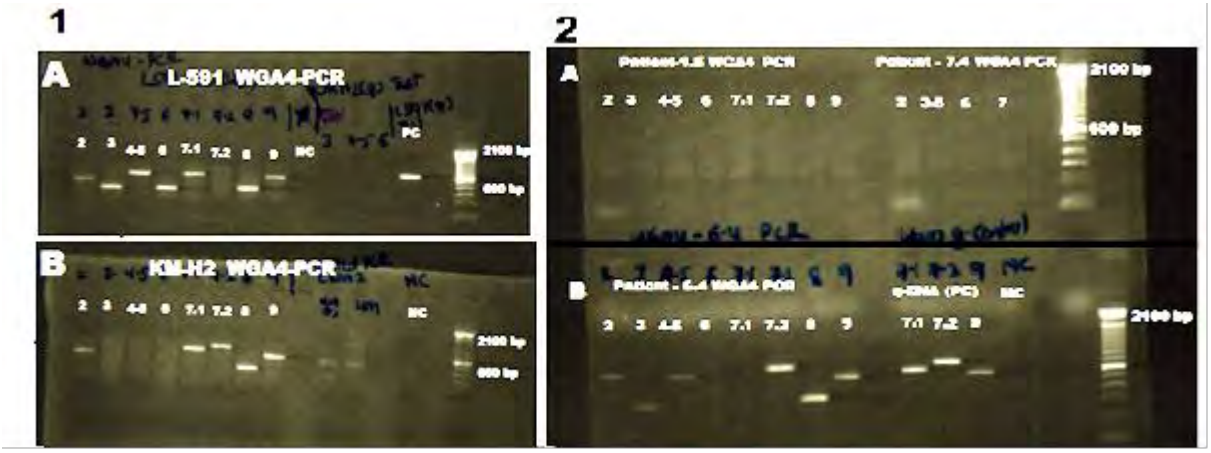

Figure 7. PCR on WGA4 products of KM-H2, L-591, and patient samples. 1. PCR on the WGA4 products of KM-H2 (50 cells) and L-591 (50 cells) microdissected samples was set up using all the exon primers. NC (WGA): non-template control from WGA4 experiment; NC (PCR): non-template control for this PCR experiment. 2. PCR on patient sample WGA4 product. A: WGA4 products for patient sample 1.5 and 7.21 have failed in the PCR; B: Patient sample 6.4 has picked up bands for all the exons except 7.1. NC: negative control; g-DNA: positive control.

3.5) Quality of DNA sequences from unamplified cell lines and amplified cell lines and patient samples

The quality of the sequences was assessed based on the percentage of the alignment (in Sequencher 5.0 software) between sequence traces obtained and the reference sequence. The quality of sequences for the unamplified cell lines was better (79.7%) than those of WGA sequences (76.5%) (Table1). Good quality sequences obtained for the rest of the unamplified cell lines (HDLM-2, L-428, L-540,

and L-1236) cell lines. Interestingly, the quality of the sequences obtained from the patient sample 6.4 were as good as the unamplified cell line samples (Table 2).

Table 1. Showing the quality of sequences for the unamplified and WGA KM-H2 and L-591

		KM-H2				L591			
	PCR product (bp)	**WGA (bp)**	**PASS (%)**	**UA (bp)**	**PASS (%)**	**WGA (bp)**	**PASS (%)**	**UA (bp)**	**PASS (%)**
Exon 2	462	473	63	1270[a]	75	387	79.8	389	73.5
Exon 3	306	del	del	del	del	F	F	1215[c]	71.4
Exon 4-5	588	del	del	del	del	600	81.2	559	85.5
Exon 6	311	del	del	del	del	301	80.7	301	81.7
Exon 7.1	568	561	83	1099[b]	82	545	91.6	543	91.9
Exon 7.2	653	F	F	F	F	F	F	F	F
Exon 8	328	333	59.5	317	81	301	79.4	301	71.5
Exon 9	528	515	87.6	533	79.4	508	54.7	513	83.8

bp:base pairs; **UA:**unamplified; **del:**deletion; **F:**sequencing failure/poor quality sequences**; a**: exon7.1 IntR to exon 2; **b:**Int ex7; **c:** exon 3 IntF to exon 5

Table 2. Showing the quality of sequences for the unamplified and patient sample

	PCR product (bp)	L-428		L-540		L-1236		HDLM-2		Patient (6.4) sample	
		UA (bp)	PASS (%)	UA (bp)	PASS (%)	UA (bp)	PASS (%)	UA (bp)	PASS (%)	UA (bp)	PASS (%)
Exon 2	462	477	73	477	71.9	477	71.7	481	70.7	F	F
Exon 3	306	278	87	283	86.2	282	85	282	79.8	278	88.8
Exon4-5	588	566	91.5	569	90.5	569	91.2	598	88.3	561	86.8
Exon 6	311	290	84	300	81.7	293	83.3	291	82	FP	FP
Exon 7.1	568	547	91.6	545	94	544	93.8	549	92	FP	FP
Exon 7.2	653	637	84	639	85	636	90.4	642	89.1	629	93.5
Exon 8	328	303	88.8	302	90.4	301	87	300	88.7	299	91
Exon 9	528	512	88.7	512	88.9	513	89.9	512	88.9	500	94.6

bp:base pairs; **UA:**unamplified; **del:**deletion; **F:**sequencing failure/poor quality sequences**; FP:**PCR failure

3.6) Comparison of WGA amplified and unamplified DNA sequence

To investigate whether microdissection and whole genome amplification methods will introduce any changes to the genome sequences, exon sequence of TNFAIP3 gene of WGA amplified and unamplified DNA from the KM-H2 and L-591 cell lines were compared. No sequence differences were found.

3.7) Analysis of the TNFAIP3 mutations in HL cell lines and patient sample

All of the seven nucleotide changes that were observed in this study were illustrated in the Table 3. Nucleotide T deletion was observed at the base position 8751 (intron 5-6) (gi| reference sequence) in L-540 and L-591 (both unamplified and WGA amplified) (Table 3).

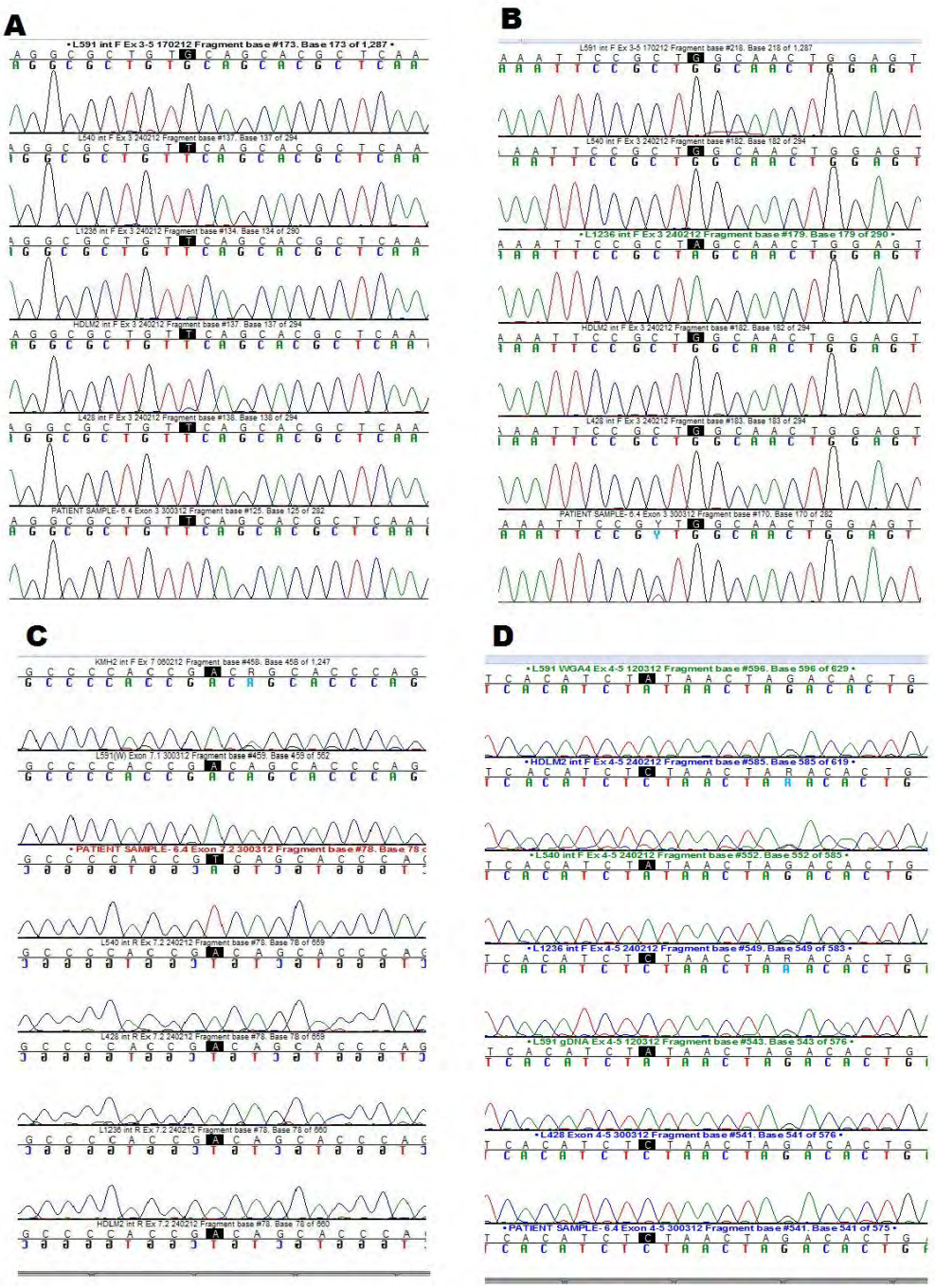

Figure 8. Chromatograms showing nucleotide changes observed in the cell lines and patient samples. The chromatogram shows A. T446G transversion for the L-591 (unamplified) cell line, B. G491A transition for the L-1236 cell line, C. A1434T transversion for the patient 6.4; and D. C8751A transversion for the L-591 and L-540 cell lines.

Some of the nucleotide changes observed in this study match with the previous studies on the HL cell lines (**16**), such as the deletion of the region between intron2 and exon 6 in the KM-H2 cell line, and mutation in L-591 at coding base 491 (G491A) which has changed the amino acid Tryptophan to STOP codon. A1434T transversion and T3003A transitions were observed in the patient sample 6.4.

Mutations observed in L-1236 (G491A) and patient sample 6.4 (A1434T) were searched (http://siftdna.org) for potential implication on the protein function. The results show that both of these mutations are 'novel' with tolerated (A1434T) and damaging (G491A) effect. We could not obtain population frequency for any of the genetic changes observed in this study.

Table 3. Mutations found in cell lines and patient sample

Sample	EBV	Subtype	Nucleotide change [a]	Amino acid change [b]
KM-H2 (B)	−	NS	Δ intron 2–exon 6	Frameshift aa 99
L-540 (g)	−	NS	C8751A[c]	Intron 5-6
L-1236 (g)	+	NS	G491A	W142STOP
L-591 (B)	+	NS	C8751A[c]	Intron 5-6
Patient 6.4	−	−	A1434T	T463S
Patient 6.4	−	−	T3003A	3' UTR

NS: Nodular Sclerosis; **MC:** Mixed Cellular; **Δ:** deletion; **g:** unamplified; **B:** found in both amplified and unamplified; **a:** position on the coding sequence **(22); b:** corresponding to PDB accession no.NP_006381**; c:** corresponding to the genomic reference sequence, (NCBI Reference Sequence: NC_000006.11, gi|224589818:138188581-138204449 (**21**).

4) Discussion

In this study, we initially used REPLI-g kit to whole genome amplify the cell line samples. However, due to inconsistencies in the amplification results and unspecific amplification of samples, we replaced it with the WGA4 kit. WGA4 kit amplified DNA of all the microdissected cell lines samples and patient samples. However, WGA products for patient samples 1.5, 7.28, and 7.8 failed to amplify all of the *TNFAIP3* exons in the following PCR experiments. Patient sample 6.4 amplified all of these exons except the exon 6 and 7.1.

Better quality sequences were obtained for the cell line (>76.5%) and patient sample 6.4 (90.9%). WGA method gave good coverage for the *TNFAIP3* gene in these samples. However, sequencing reactions for exon 3 (L-591) and exon 7.2 (KM-H2 and L-591) had consistently failed.

After analysing the mutation spectra for the whole genome 'amplified', and 'unamplified', KM-H2 and L-591 cell lines, we came to a conclusion that microdissection and whole genome amplification methods didn't introduce any changes to the sequences. However, differences in the quality of the sequences obtained and the failure to generate sequences for all the exons for these amplified samples had hampered our investigation.

Our study for the first time revealed C8751A transversion in the *TNFAIP3* gene for the L-540 cell line. The same mutation was observed in the L-591 cell lines in this study.

Our results were compared to a previous study (**16**) that looked into the mutation spectra for the *TNFAIP3* gene in the same cell lines. We observed same nucleotide changes in the cell lines KM-H2 (deletion of intron 2– exon 6 region), and L-1236 (G491A). However, improvement in the quality of the sequence and the coverage for the region of interest in the WGA products will result in a better whole genome or exome sequencing of the target sample.

The functional role of the genetic changes observed in our study was investigated in http://siftdna.org. The results show that the mutations observed in L-1236 (G491A) and patient sample 6.4 were 'novel' with tolerated (A1434T) and damaging (G491A) effect. We could not obtain population frequency for any of the genetic changes observed in this study.

This study enabled us to amplify DNA from very few numbers of cells (~50 cells) which will be sufficient enough to use in the future exome/ whole genome sequencing applications to understand the role of genetic alterations in Hodgkin's lymphoma.

References

1. Baylin, S.B., Stem cells, cancer, and epigenetics (October 31, 2009), StemBook, ed, The Stem Cell Research Community, StemBook, doi/10.3824/stembook.1.50.1(http://www.stembook.org).

2. Brauninger, A., R. Schmitz, et al. (2006). "Molecular biology of Hodgkin's and Reed/Sternberg cells in Hodgkin's lymphoma." Int J Cancer **118**(8): 1853-1861.

3. Cancer Research UK (http://www.cancerresearchuk.org/).

4. Farrell, K. and R. F. Jarrett (2011). "The molecular pathogenesis of Hodgkin lymphoma." Histopathology **58**(1): 15-25.

5. Forbes, S. A., G. Tang, et al. (2010). "COSMIC (the Catalogue of Somatic Mutations in Cancer): a resource to investigate acquired mutations in human cancer." Nucleic Acids Res **38**(Database issue): D652-657.

6. Genomics lab, University of Birmingham, Birmingham, UK. (http://www.genomics.bham.ac.uk/sequencing.htm).

7. GenomePlex® Single Cell Whole Genome Amplification Kit (WGA4) from Sigma-Aldrich,UK (http://www.sigmaaldrich.com/catalog/product/sigma/wga4?lang=en&region=GB).

8. Gundry, M., W. Li, et al. (2012). "Direct, genome-wide assessment of DNA mutations in single cells." Nucleic Acids Res **40**(5): 2032-2040.

9. Hartmann, S., J. I. Martin-Subero, et al. (2008). "Detection of genomic imbalances in microdissected Hodgkin and Reed-Sternberg cells of classical Hodgkin's lymphoma by array-based comparative genomic hybridization." Haematologica **93**(9): 1318-1326.

10. Human Biomaterials Resource Centre, College of Medical and Dental sciences, University of Birmingham, UK (http://www.birmingham.ac.uk/research/support/facilities/biorepository.aspx).

11. *Leibniz Institute DSMZ-German Collection of Microorganisms and Cell Cultures,* Inhoffenstraße 7B, 38124 Braunschweig, GERMANY (http://www.dsmz.de/catalogues/catalogue-human-and-animal-cell-lines.html).

12. Macmillan Cancer Support factsheet - *Hodgkin lymphoma in children (2010), (*www.childrenwithcancer.co.uk).

13. NCBI Reference Sequence: NC_000006.11, gi|224589818:138188581-138204449, (http://www.ncbi.nlm.nih.gov/nuccore/NC_000006.11?report=fasta&from=138188581&to=138204449).

14. Navin, N. and J. Hicks (2011). "Future medical applications of single-cell sequencing in cancer." Genome Med **3**(5): 31.

15. Kuppers, R. (2009). "Molecular biology of Hodgkin lymphoma." Hematology Am Soc Hematol Educ Program: 491-496.

16. Schmitz, R., M. L. Hansmann, et al. (2009). "TNFAIP3 (A20) is a tumor suppressor gene in Hodgkin lymphoma and primary mediastinal B cell lymphoma." J Exp Med **206**(5): 981-989.

17. Sequencher 5.0 (http://genecodes.com/).

18. Verstrepen, L., K. Verhelst, et al. (2010). "Expression, biological activities and mechanisms of action of A20 (TNFAIP3)." Biochem Pharmacol **80**(12): 2009-2020.

19. QIAGEN's REPLI-g Mini Kit, QIAGEN, UK (25 X, Catalogue no.150023,) (www.qiagen.com/wga).

20. TNFAIP3 protein sequence at NCBI. (http://www.ncbi.nlm.nih.gov/protein/5454132).

21. TNFAIP3 gene information at NCBI. (http://www.ncbi.nlm.nih.gov/gene/7128).

22. TNFAIP3 coding sequence information at NCBI. (http://www.ncbi.nlm.nih.gov/CCDS/CcdsBrowse.cgi?REQUEST=GENEID&DATA=7128).

23. COSMIC (Catalogue of Somatic Mutations in Cancer) (http://www.sanger.ac.uk/cosmic).

24. Sigma-Aldrich, UK. (http://www.sigmaaldrich.com/life-science/custom-oligos/custom-dna.html).

25. QIAGEN's multiplex PCR Kit, QIAGEN, UK (100X, Catalogue no.206143) (http://www.qiagen.com/products/pcr/multiplexpcrsystem/multiplexpcr.aspx).

26. Zhang, Y., K. Weber-Matthiesen, et al. (1997). "Frequent deletions of 6q23-24 in B-cell non-Hodgkin's lymphomas detected by fluorescence in situ hybridization." Genes Chromosomes Cancer **18**(4): 310-313.

27. Carl Zeiss PALM Microbeam, Carl Zeiss Ltd, Cambridge, UK (http://microscopy.zeiss.com/microscopy/en_de/products/laser-microdissection/microbeam.html).

28. Arneson, N., S. Hughes, et al. (2008). "GenomePlex Whole-Genome Amplification." CSH Protoc **2008**: pdb prot4920.

29. Treff, N. R., J. Su, et al. (2011). "Single-cell whole-genome amplification technique impacts the accuracy of SNP microarray-based genotyping and copy number analyses." Mol Hum Reprod **17**(6): 335-343.

30. Dolezel J. et al. Nuclear DNA content and genome size of trout and human. Cytometry 2003;51:127–128.

31. Barbara D. et al. Laser Capture Microdissection in the Genomic and Proteomic Era: Targeting the Genetic Basis of Cancer. Int J Clin Exp Pathol.2008; 1:475-488

6) SUPPLEMENTARY NOTES

Table 1. Primer sequences used for genomic DNA amplification of TNFAIP3

Primer name	Sequence	Amplification of
A20E2exF	TGCCTACAGATCAGGGTAATGACAAG	exon 2
A20E2NewexF	GGAGTCGTATTAAAGTCAGGCTAA	exon 2
A20E2NewexR	GGCAAAAGAAACACAACAGAAC	exon 2
A20E2intF	GTTTCCTGCAGGCAGCTATAGAGG	exon 2
A20E2R	AGCTTCATGAATGGGGATCCAGCAG	exon 2
A20E3exF	ACCATTCAGTCCCCTAGAATAGCAG	exon 3
A20E3intF	ACCTTTGCTGGGTCTTACATGCAG	exon 3
A20E3R	TATGCCCACCATGGAGCTCTGTTAG	exon 3
A20E4-5exF	TGAATAATTGTAGAGTGATGTCAGAATGAC	exon 4/5
A20E4-5intF	TACAGGGAGTACAGGATACATTCAAGC	exon 4/5
A20E4-5R	GGAAAACCCTGATGTTTCAGTGTCTAG	exon 4/5
A20E6exR	AATCACTCTACTGTTGAGCTTCAGG	exon 6
A20E6F	TGAGATCTACTTACCTATGGCCTTG	exon 6
A20E6intR	TCAGGTGGCTGAGGTTAAAGACAG	exon 6
A20E7.1exF	GGTTCTACAATTCTTGCCATAATCCAC	exon 7
A20E7.1intF	GAGCTAATGATGTAAAATCTTGTGTGTG	exon 7
A20E7.1R	CAAAATCCGTTGTGCTGCACATTCAG	exon 7
A20E7.2exR	CAGTTCTGCCTGACTGCCTACATG	exon 7
A20E7.2F	CTCTCGGGGAGAAGCCTATGAGC	exon 7
A20E7.2intR	GAACAAAACCCCTTCTGGACAGCAG	exon 7
A20E8exR	ATGAGGAGACAGAACCTGGCAGAG	exon 8
A20E8F	ACTGTCAGCATCTCTGTATCGGTG	exon 8
A20E8intR	TGTCACTGTCGGTAGAAAACGCTC	exon 8
A20E9exF	GTAGACTCCACACTCTCCAATGAG	exon 9
A20E9intF	GTGCTCTCCCTAAGAAATGTGAGC	exon 9
A20E9R	GGGTTACCAAACCTGAGCATCGTGC	exon 9
A20E9Rnew	CGGGTTACCAAACCTGAGCATCGTG	exon 9

E2-E9:exon2-exon9; **F:**forward primer; **R:** reverse primer.

Table 2. List of the primers used and their PCR products

Primer name	Position on gDNA	Product (bp)
A20 E2intF->E2R	3690->4152	462
A20 E2NewF->E2NewR	3712->4110	398
A20 E3intF->E3R	7336->7642	306
A20 E4-5intF->E4-5R	8193->8781	588
A20 E6intF->E6R	9875->9914	311
A20 E7.1intF->E7.1R	10918->11486	568
A20 E7.2intF->E7.2R	11966->12012	653
A20 E8intF->E8R	12888->12925	328
A20 E9intF->E9R	13522->14049	527
A20 E9intF ->E9Rnew	13522->14050	528

E2-E9:exon2-exon9; **F:**forward primer; **R:** reverse primer; **gDNA:** genomic DNA.

Table 3. Template requirements for sequencing at the Genomics lab

Template DNA	DNA Quantity (ng)
single-stranded DNA	25-50 (*)
double-stranded DNA	150-300 (*)
PCR : 100-200 bp	1-3 (**)
PCR : 200-500 bp	3-10
PCR : 500-1000 bp	5-20
PCR : 1000-2000 bp	10-40
PCR : >2000 bp	20-50 (*)

(*) : for BigDye® Terminator versions 1.0, 2.0 and 3.0 higher quantities necessary

(**) : rule of thumb: PCR product length (in bp) / 50 = amount in ng.

Comparing and contrasting different approaches to identifying predictive biomarker combinations

BY

PRADEEP RAMAGIRI

THIS PROJECT IS SUBMITTED IN PARTIAL FULFILMENT OF THE REQUIREMENTS FOR THE AWARD OF THE MRES IN BIOMEDICAL RESEARCH

SCHOOL OF CANCER SCIENCES

COLLEGE OF MEDICAL AND DENTAL SCIENCES

UNIVERSITY OF BIRMINGHAM

AUGUST 2012

Abstract

Adrenocortical carcinoma (ACC) is a heterogeneous malignancy with an incident rate of about 0.7-2.0 cases per million people per year. However, it accounts for more than 15% of the total adrenal incidentalomas registered every year. Distinguishing malignant ACC from benign Adrencortical Adenoma (ACA) has been a major problem. Hence, in this study we attempt to statistically design a biomarker tool based on steroid metabolic excretion data from ACA and ACC patients. Data in question is of 32 distinct adrenal derived steroid secretions obtained from 102 ACA patients and 45 ACC patients. Variable selection was performed by forward selection, backward search, stepwise selection, and all subset combination methods using logistic regression. An investigation into prediction model building was carried out and an assessment was made on the model fitness using various criteria, such as log-likelihood, deviance, AIC, and Wald's statistics. Sensitivity and specificity for some of the best models was evaluated using ROC curve. Some of the best prediction models from all subset combination models were chosen to inspect all the patient samples in the data. Models with 6, 7, and 8 variable subsets were shown to have best prediction capability with perfect sensitivity and specificity. However, investigation into these variable subsets revealed 'overfitting'. Models with 4 and 5-variable subset combinations have relatively high prediction capability with a significant sensitivity and specificity (AUC=99.3% and AUC=99.8%, respectively). The strength of the estimates in these models was relatively stable when compared to 6-8 variable subset models.

Acknowledgement

I sincerely thank Prof. Jon Deeks and Dr. Wenbin Wei for their kind support and guidance.

Table of contents

Table of figures

1) Introduction

Adrenal tumours are one of the most prevalent forms of cancer with an incident rate of 2% in general population and about 7% in the older generation (70 years or above age groups). Incidental discovery of Adrenal tumour (Adrenal incidentalomas) during computed tomography (CT) and autopsy resulted in a rapid increase in their prevalence rate (1, 8, and 12). Adrenal tumours cause serious health problems in patients, such as Cushing's syndrome, Conn's syndrome (hyperaldosteronism), virilisation in females, feminisation in males, and multiple endocrine neoplasia (MEN) (1, 3, 8, 12-13).

Adrenal tumours can be divided into two groups based on their origin, 'Adrenocortical adenoma' (ACA) and 'Adrenocortical carcinoma' (ACC) of Adrenal Cortex origin, and 'Neuroblastoma' and 'Pheochromocytoma' of Adrenal Medulla origin. The majority of adrenal tumours are benign, but still cause some serious health problems. Only 10% of the Pheochromocytomas and all ACC are malignant. ACC is an aggressive form of cancer which can occur in all age groups (1, 8, 12).

Most often ACC are not diagnosed, until they have grown very large and/or metastasize to other organs, due to their location deep in the retro peritoneum. Surgical removal of ACC is carried out as part of the main treatment but is not feasible with many patients, due to complications associated with the hormonal imbalances after surgical removal of the tumour. It is very difficult to distinguish ACC from ACA based on histopatholgical and biochemical methods. Some of the imaging techniques, such as [^{18} F] Fluorodeoxyglucose positron emission tomography, offer better diagnostic capability with higher specificity (91%) and sensitivity (97%) (13). But this technology is very expensive and sparsely available. Hence, it is pivotal to differentiate ACC from ACA by some other means.

1.1) Evidence based medicine

Traditionally, medicine has been very much subjective. This trend has been changing in recent years, where 'evidence based' medicine is gaining popularity among physicians and health policy makers. Evidence based medicine can be best explained as 'a conscientious, explicit and judicious use of current best evidence in making decisions about the care of individual patients'. Clinical prediction models may provide the evidence based input by providing estimates of the individual probabilities of risks and benefits. Clinical prediction models consider a number of characteristics related to the disease, patient and the treatment (7).

Prediction models may help design preventive interventions for patients with high risk of having or developing a disease. From a clinical point of view, prediction models may provide valuable information to patients and their physicians on the probability of a diagnosis or a prognostic outcome.

1.2) Statistical modelling for prediction

In data analysis, statistical models summarise patterns in the data. Prediction mainly estimates risks of disease development and survival rates for the patients. Prediction also tests hypotheses, such as the importance of certain predictor variables in a disease and the correlation between variables (14).

Statistical model for Prediction can be distinguished into 3 main classes: regression, classification, and neural networks according to Ewout W. Steyerberg (10). However, majority of other statisticians discern the prediction modelling methods into two classes: Regression analysis (for continuous variables) and classification (for non-continouos variables). Regression analysis investigates the relationship between a 'dependent/outcome' variable and one or more 'independent/predictor' variable(s) by employing several techniques for modelling and analysis. Dependent/outcome variable can be defined as the observable result, such as the presence or absence of cancer, when any one of the independent variable is manipulated (5). Independent variables are usually the measurements of an effect, such as response to treatment/drug, and effect of temperature on durability of materials. Regression helps understand how the value of a dependent variable changes with changes to any one of the independent variables (10).

Classification method identifies a set of categories/sub-populations to which a new observation belongs, based on 'training set' of data containing observations whose category membership is known. The individual observations are analysed into a set of quantifiable properties, known as explanatory variables, features etc. These properties may be of categorical, ordinal, integer valued, or real valued. Neural networks can be considered as the non-linear extensions of linear logistic models. The generalised non-linear models (GNLM) are implemented as neural networks. In this method, outcome is related to the non-linear combinations of the predictor variables, in contrast to the methods where the outcome is related to simple linear combination of estimated regression coefficients and predictor values (14).

For binary outcome such as the diagnostic outcome (presence or absence of a disease) there are several methods available for prediction modelling. Logistic regression facilitates as a quite flexible model to derive predictions. In this approach, interactions and nonlinearity can be incorporated. Other models such as Naive Bayes can be seen as a simplified version of logistic regression, which ignores

correlations between predictors. The choice of the model depends on various factors, such as the sample size and research question(s) (14).

This study used logistic regression models to distinguish ACC from ACA patients in the data.

1.3) Logistic regression

Logistic regression is a flexible method as it can incorporate categorical and continuous variables, non-linear transformations, and interaction terms. Logistic regression is a form of 'generalised linear models'. Similar to linear regression model, the binary outcome 'Y' is linked to a linear combination of a set of predictors and regression coefficients 'β' (4). Logistic 'link' function is used to restrict predictions between '0' and '1'.

Logistic regression works by relating the log odds (logit) to a linear combination of the explanatory variables. Logistic regression was applied to predict the transformation of the outcome (otherwise we could predict unrealistic values of the probability (i.e. outside 0 to 1 range). Log odds value lies between minus infinity and plus infinitive. Logistic regression can be interpreted using odds and odds ratio as below,

$$\log(\text{odds}) = \text{logit}(P) = \ln\left(\frac{P}{1-P}\right)$$

Therefore, $\log_e(\text{odds}) = a+bx$

Where, 'a' is an intercept where predictor variable x=0, 'b' regression coefficient that relate to 'x', and 'e' is the base of natural logarithm.

In the case of one predictor variable 'P' the logistic regression equation from which the probability of outcome is predicted is given by:

$$P = \frac{e^{a+bX}}{1+e^{a+bX}}$$

In logistic regression, estimates of the coefficients are obtained in an iterative procedure known as 'maximum-likelihood'.

Maximum likelihood is generally used to estimate β_i regression coefficients in a standard logistic regression model. β_i regression coefficients can be interpreted as the effect of a 1-unit increase in $\times_i$ while keeping the other predictors in the model constant. In logistic regression model, β_i is unadjusted or univariate for a single predictor, but with multiple predictors, β_i is adjusted or conditional on the values of other predictors (14).

1.4) Building prediction models

In the development of prediction models some important considerations have to be made. The very first and the most important one is the 'research question'. Some of the other important questions that need to be addressed are: What do we know about the predictors? How are the predictors in the study defined? What is the outcome? How to deal with the missing values? The other important considerations involve coding of the predictors, specifying a model, 'overfitting', and validation of the model (14).

1.4.1) Coding

As in many cases the raw data is often not in an appropriate form for entering in the regression models. Hence, they need to be manipulated. This process is called 'coding'. Data analysis usually starts with obtaining an impression of the data. This includes finding the missing values and the distribution of the predictors. Descriptive analyses, such as frequency distribution and Inter-quartile (IQ) range are quite useful to find the distribution of predictor variables in the data (14).

1.4.2) Model selection

Selection of the predictor variables can also be made based on the subject knowledge. The list of the number of variables in the study can be reduced by a literature review. Variables can be selected based on their distributions. Variables with relative importance can be deleted if they have a large number of missing values (14).

Variables are often strongly correlated with each other. This 'collinearity' is one of the major 'variance inflation factors', which degrades the precision of estimate coefficients (11). The number of variables in a model must not depend only on the statistical significance but also on the number of questions it will answer. A parsimonious model is not necessarily a better model, as it might not answer all the biological research questions. However, it can be the case when pre-specified

models were compared. A model with smaller number of variables is easier to interpret and practice.

Validation of the prediction models on new subjects, outside the subjects on which they were built, is the main aim of outcome prediction. A major threat to this validation is 'Overfitting', where predictions are not valid for the new subjects even though the data under study are well described. These models might have been built by capitalising on the specifics and idiosyncrasies of the sample (9). 'Overfitting', can also be defined as the 'curse of dimensionality', or as fitting a statistical model that has too many degrees of freedom (10). As a result of overfitting, too 'optimistic' impression of a model will be achieved in new subjects from the underlying population. Optimism = True performance (underlying population) - Apparent performance (estimated performance in the sample) (14).

1.5) Challenges involved in statistical prediction modelling

Statistical prediction modelling faces various challenges, such as model uncertainty, and limited sample size.

Model uncertainty arises when a model is not fully pre-specified, before fitting it to a data set. Iterative model checking and model modifications are often followed in this case. On the other hand, standard statistical methods assume that a model was pre-specified. Hence, parameter estimates such as regression coefficients, their corresponding standard errors, 95% confidence intervals, and *p*-values are largely unbiased. When structure of the model was at least partly based on findings in the data, bias may occur, and usually the uncertainty of conclusions drawn from the model is underestimated (2, 6).

A large sample size is necessary to address most of the scientific questions with empirical data. In the majority of the prediction studies 'effective sample size' is much smaller than indicated by the total number of subjects in a study (9). For example, a study containing 1000 patients with an event incidence rate of 1% will have an effective sample size of 10. We may not be able to derive a reliable prediction model from a small sample size. A large sample size is desirable to study several aspects of prediction modelling, such as gene-disease associations and multivariable prognostic modelling. A small size renders in making strong modelling assumptions. With a small samples size we may have to assume the linearity of the predictor variable and no/less correlation between the predictor variables. Hence, we may have a limited power to test the deviations from these model assumptions. Sample size may restrict our scientific approach and may dictate on what we can achieve as more complex questions, such as 'what are the most important predictors in the model?' can be solved with a large sample data (14, 15).

1.6) Aim of the study

This study was aimed at designing prediction models by using the differences in the steroid and steroid metabolite excretion levels between the ACC and ACA patients. In this study we investigated and compared different methods for selecting variables. We also aimed to investigate statistical significance of the estimates in the prediction models and examine the estimates for these models on all of the patients in the data.

2) Material and Methods

2.1) Patient data

The data used in this study was provided by Wiebke Arlt group from 'Centre of Endocrinology, Diabetes, and Metabolism, University of Birmingham, Birmingham, UK'. The data contains urinary steroid metabolomes, analysed in a single run by gas chromatography/mass spectrometry (GC/MS), of 24 hour urine samples from adrenal tumour patients. These samples were collected, with informed consent, in six specialist referral centres participating in the European Network for the study of Adrenal Tumours (ENS@T; www.ensat.org) between 2006 and 2009. Appropriate ethical approval was obtained from the local ethical boards. These samples were identified as either benign ACA or malignant ACC with the help of histological, biochemical, clinical examination as well as imaging.

Wiebke Arlt group have used the same data to design a biomarker tool by applying 'Generalized Matrix Learning Vector Quantization (GMLVQ)' method to differentiate ACC from ACA. They have published their research findings in 'Journal of Clinical Endocrinol Metab', in 2011(17). This group have also used logistic regression method, but concluded GMLVQ was a better method for building prediction models with this data set. In this study, I have made an attempt to learn and apply logistic regression method to build alternative prediction models.

The data in the present study contains 32 predictor variables. These variables were grouped into Androgen metabolites, Androgen precursor metabolites, Mineralocorticoid metabolites, Mineraloccorticoid precursor metabolites, Glucocorticoid precursor metabolites, and Glucocorticoid metabolites, based on their biochemical nature. The detailed description of these 32 variables is available in the Supplementary Table 2.

2.2) Coding

The data taken from the Wiebke Arlt group's study (17) was labelled as 'raw data' and the predictor variables in the raw data were labelled from P1 to P32. Data was manipulated by replacing any '0' values with '0.5', so that when log transformed, the patient measurements with '0' values were not denoted as 'NA'. Log transformation expands the smaller values and squeezes the bigger values thus making the distribution more symmetric. Predictor variables in the log-transformed changed data were labelled from s1 to s32. The mean values observed for the log-transformed

changed data were less than that of the log transformed raw data. Best variables from both forms of the data set were taken to build prediction models.

2.3) The Logistic regression model for binary responses

A logistic regression model helps us to predict the probability of a particular outcome in relation to a list of predictor (independent) variables. In our study, the outcome has two categories. The predicted value lies between '0' and '1' (in this study, '0' represents the presence of Adrenocortical adenoma (ACA) and 1 represents presence of Adrenocortical carcinoma (ACC). Hence, Binary logistic regression was applied. The outcome variable will be the proportion (probability) of individuals with the characteristic (i.e., ACC or ACA). This approach is a form of 'discriminant' analysis.

2.4) Data was processed using R language

The majority of statistical problems in this study were performed using R language on R platform (R version 2.13.1 and R Commander version 1.7-3). All of the functions run in R were obtained from the 'Comprehensive R Archive Network (CRAN) (16).The scripts that were used to perform the statistical analysis were detailed in the Supplementary **Table3**.

2.5) SPSS

IBM SPSS software was used to generate prediction models using 'stepwise forward selection' method and 'stepwise backward search' method.

2.6) Stata

Stata V11.0 was used to generate ROC curves using the 'lroc' command following fitting the logistic models.

2.7) Bluebear

'Bluebear' cluster is a parallel computing service provided by the 'IT services' at the University of Birmingham, Birmingham, UK. It consists of 1500 processing cores and approximately 150 terabytes of user disk space with a sophisticated cluster computing system. Computationally intensive, all subset combinations method was performed using logistic regression on Bluebear platform. This method generated

subset combinations that had significant values for assessing measure of fit, such as 'deviance' and 'AIC'.

2.8) Building prediction models

In the previous sections, I have estimated and interpreted the coefficients in the logistic regression model. To build a 'best prediction model' for distinguishing ACA from ACC using the dataset, I needed to devise a strategy to a) select predictor variables for the model , and b) employ a set of methods to assess 'fit' of the models by analysing the estimates in the prediction models.

2.8.1) Variable selection

One of the aims in this study is to build the most parsimonious prediction model that still can describe the data. Prediction models with less number of variables can be numerically more stable and more easily generalized. Addition of more variables to the prediction model will increase the estimated standard errors and as a result the prediction model becomes dependent on the observed data. This approach can lead to 'over fitting' of the prediction model, where idiosyncrasies in the data are fitted rather than the 'generalizable' patterns. As a result, this model is not applicable to new patients.

It is important to include the relevant variables in the prediction model, to gain as much control of confounding as possible within the given data set. But, care must be taken for not 'over fitting' the model. This can be very real in our case as the number of variables is relatively large compared to that of patients.

In this study, potentially important variables were selected based on their estimated coefficients, estimated standard error, the likelihood ratio test for significance of the coefficient, and Wald statistic. There are various techniques available for choosing variables for a regression model. One of the popular methods that can be used to build prediction models is generalized linear regression, which employs 'forward selection', 'backward elimination', 'stepwise regression ', and 'best subsets selection'(Supplementary Table3).

a) Forward selection

Forward selection is one of the simplest data-driven prediction models building approach, where variables are added one at a time. The 'significance' of the variable is a measure of the statistical significance of the coefficients for that variable. In logistic regression, the significance is assessed with the help of likelihood ratio chi-square test. Hence, at any single step in the procedure, the most 'significant' variable

will produce the greatest change in the log-likelihood as compared to a model not containing the variable. The 'significance' of the variables is usually tested based on a pre-set P-value. The conventional P-value set up is at 0.05, but can choose a customary P-value level at say 0.10 or 0.15, to explore the nature of this method (4).

The model building starts with adding the most 'significant' variable to an empty model. Then at each step, each variable that is not already in the model is tested for the inclusion in the model. The most 'significant' of these variables is added to the model, so long as its P-value is below the pre-set level. This method is iterated until none of remaining variables are 'significant' when added to the model. This multiple use of hypothesis testing nature can lead to higher 'type 1 error rate' for a variable, where an 'unnecessary' chance is given for a variable inclusion. Because of this error prone nature, forward selection is usually a very good exploratory method (4). In forward selection, each addition of a variable to a model can also render one or more of the variables in the model 'insignificant' due to their 'correlation' with each other.

b) Backward elimination

In this method, an initial screening of the variables for their 'significance' is done based on the pre-set P-value. Then a model is fitted with all the variables of interest. Then at each step, the least 'significant' variable is excluded based on the pre-set P-value. This procedure is repeated until we end up with a model containing all the 'significant' variables.

But, this model building approach can omit a variable in the initial steps which can be a very 'significant' variable to include in the final reduced model. This flaw can be avoided by using the 'stepwise regression' method.

c) Stepwise regression

Stepwise regression method is a bidirectional selection method, which uses both forward selection and backward elimination. The model either starts with an empty model or a full model and the predictor variables are selected either for inclusion or exclusion from the model in each step based entirely on statistical criteria. Missing values restrict the number of available cases to this method, especially, when a full model is employed.

d) All subset combinations

An alternate method is to fit all possible regression models, and to evaluate these models according to some criterion, such as deviance or AIC. In this method a

number of best regression models can be selected. However, the fitting of all possible regression models is very computer intensive and time consuming.

2.9) Criteria for model fitting

Log-likelihood, Deviance, AIC, and Wald's statistics were used to assess the fit of the prediction model.

2.9.1) Log likelihood ratio

The log likelihood ratio (LLR) corresponds to a difference in log likelihoods. LLR is maximum at the maximum-likelihood estimate (MLE) and equals zero. MLE is the value that corresponds to the largest possible likelihood of the event (π)

Log (LR) = L(π)-L(MLE)

Log-likelihood indicates how much unexplained information is there after the model has been fitted. The deviance statistic -2LL is a goodness of fit indicator. Large values of log-likelihood indicates poor fitting. Log-likelihood can be used to compare the state of logistic regression model against baseline state. A baseline state is when there is only the constant in the prediction model.

$$\chi^2 = -2LL_R - (-2LL_F) = -2\ln\left(\frac{likelihood_R}{likelihood_F}\right)$$

Where, 'R' is simple model and 'F' is a complex model (usually has an extra variable). The parameters in the simple model must be a proper subset of the parameters in the complex model. This model follows χ^2 distribution with k degrees of freedom where k is the number of predictors in the new model. χ^2 was used to test whether addition of a variable will reduces the goodness-of-fit measure.

2.9.2) Akaike Information Criteria (AIC)

Akaike Information Criteria (AIC) is used to measure the relative 'goodness of fit' of a prediction model. This criteria work on the basis of trade off between 'accuracy' and 'complexity' of the model.

The general equation for AIC can be given as below,

$$AIC = 2k - 2\ln(L)$$

Where, ‘k’ is the number of parameters in the model and ‘L’ is the maximum value of likelihood function for the estimated model.

The model with the least AIC value is considered as a better model. AIC includes a penalty for increasing the number of estimated parameters. This would discourage ‘over fitting’ (increase in the parameters improves goodness of fit, regardless of the parameters used in generating the data).

2.9.3) Assessing the contribution of predictors (Wald’s statistics)

The contribution of a predictor can be assessed by Wald statistics (z). Wald statistics is defined as the coefficient of the predictor (b) divided by the standard error (SE) of the coefficient, b/SE(b).

2.10) Area under the curve (AUC)

Area under the curve, otherwise known as the area under the receiver operating characteristic (ROC) curve, is one of the traditional measures for binary and survival outcomes. It’s a visualisation method for discriminant analysis. ROC curve plots the sensitivity (true positive rate) against the 1-specificity (false positive rate) for consecutive cut-offs for the probability of an outcome. A model with AUC=1 is a perfect model with greater specificity and sensitivity. AUC=0.5 is as good as guessing (15).

2.11) Correlation test

Correlation is a method of analysis for association between two continuous variables. The degree of ‘association’ is measured by ‘correlation coefficient (r)’.

Pearson product-moment correlation coefficient

This is a standard method of measuring the association between variables. This method leads to a quantity called ‘r’ which can take any value from -1 to +1, r=0 means no correlation, and r=1 means perfect correlation. This correlation coefficient ‘r’ measures the degree of ‘straight line’ association between the values of the two variables. Pearson’s correlation coefficient between two variables is defined as the covariance of the two variables divided by the product of their standard deviations.

3) Results

3.1) Data analysis

Initial interrogation of the data revealed that it is positively skewed with some missing values (total 57, all of them in ACC patients), 0 (total 13, mostly in ACC patients), 1(total 16, mostly in the ACA patients) with a highest measured value of 812807 (for variable P3 in patient 10) (Table1). We employed logistic regression method in our study which makes no 'assumptions' on the distribution of the data.

Table1. Raw data containing 'zeroes' and 'missing values'

Variable	Missing	'Zero' value
P3		1(139)
P5	1(10)	
P6	1(1)	
P9	2(11,15)	1(30)
P10		1(30)
P11	11(1-4,10-16)	2(34-35)
P12	7(1-4,10,11,15)	4(45,48,59,73)
P13	10(1-4,10-15)	
P15	11(1-4,10-16)	2(24,34)
P18	3(12,13,14)	1(96)
P21	10(1-4,10-15)	1(34)
P32	1(1)	

Predictor variables in the raw data with missing values and zeroes (the corresponding patient number in parentheses).

To study the distribution of the predictor variables in the data, the median and quartiles were calculated (Table2). The median and IQR values for these predictor variables were found to be larger in the ACC patients. Figure1 shows the difference in distribution between the ACC and ACA patients. IQR values for ACC were found to be significantly larger than in ACA. This will further strengthen our theory, that steroid and steroid metabolite excretions vary significantly between ACC and ACA patients.

Table2. IQR and median values for the predictor variables

		MEDIAN		IQR	
Variable	**Common name**	**ACC**	**ACA**	**ACC**	**ACA**
P1	An	1130	632	1898	802.75
P2	Etio	3671	803.5	6029	791
P3	DHEA	612	58	15725	111.75
P4	16α-OHDHEA	653	201.5	2566	278.25
P5	5-PT	1900.5	121	7292.5	91.25
P6	5-PD	3412	257	13460.5	231.5
P7	THA	112	93.5	137	110.5
P8	5α-THA	76	88	114	66.75
P9	THB	147	105	308.5	101.25
P10	5α-THB	155	221.5	275	182.25
P11	3α5β-THALDO	24	22	36.25	20
P12	THDOC	102.5	15.5	184.75	16.5
P13	5α-THDOC	22	4	53.5	4.75
P14	PD	839	137.5	2262	120
P15	3α5α-17HP	18.5	9.5	31.25	16.75
P16	17HP	511	120	658	172.25
P17	PT	1484	372.5	3215	513.25
P18	PTONE	32	18.5	86.25	24.75
P19	THS	2151	122	4346	103.25
P20	F	245	85	534	73.75
P21	6β-OH-F	356	133	1176	127.75
P22	THF	2836	1811	3791	1143
P23	5α-THF	852	1264.5	1174	1412.5
P24	α-cortol	557	355.5	1289	272.25
P25	β-cortol	740	536	1018	414.5
P26	11β-OH-An	653	552.5	1520	512.75
P27	11β-OH-Et	366	265	1479	316.75
P28	E	164	126	250	98.5
P29	THE	3701	3478.5	4638	2577.25
P30	α-cortolone	1840	1340	1899	783.25
P31	β-cortolone	677	665.5	813	516
P32	11-oxo-Et	483.5	401	1623.25	455.75

Common names of the variables in the study and their distribution in ACC and ACA patients are shown here. Inter quartile range (IQR) in ACC patients are showing higher values.

Steroid profiling of the samples revealed the differences between ACA and ACC patients (Table2, Figure1). Higher excretion of active androgens (An and Etio) and androgen precursor metabolites (5-PD, DHEA, and 16α-OHDHEA) were observed in ACC. In mineralocorticoid precursor metabolites, only THDOC and 5α-THDOC show elevated levels of excretion in ACC. ACC also showed increased levels of Glucocorticoid precursor metabolites (PD, PT, and 17-HP), and Glucocorticoid metabolites (THS, THE, 6β-OH-F, and α-cortol) excretion when compared to ACA (Table2, Figure1).

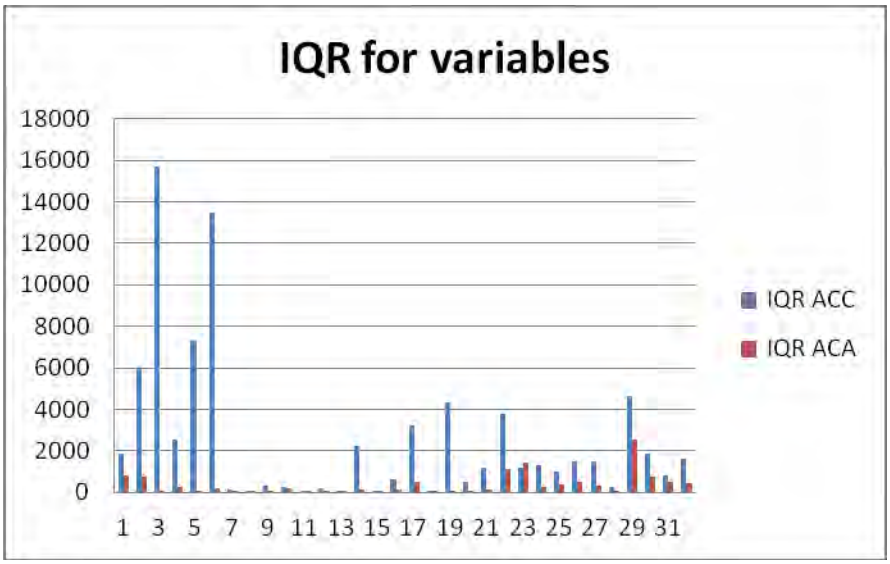

Figure 1. Inter quartile range (IQR) for predictor variables in ACC and ACA patients. The Blue and Red columns represent ACC and ACA patients in the study respectively.

Analysis of the sum of steroid subclass metabolites had revealed elevated levels of glucocorticoid metabolites excretion in 82% of ACC, compared to only 29% of ACA. An excess of androgen precursor, either combined with elevated levels of glucocorticoid precursors (36%) or both glucocorticoid and mineralocorticoid precursor (33%) was seen in the majority (67%) of ACC patients. In 9% of the ACC and 1% ACA patients, isolated glucocorticoid precursor excess was observed. In 2% of ACC and ACA patients elevated levels of isolated androgen precursor metabolites was observed. No excessive levels of isolated androgen metabolite were found.

There was no effect of age, sex, tumour size, or presence of metastasis on the level of metabolic excretion observed, in these ACC and ACA patients (Weibke Arlt et al).

3.2) Coefficients of the predictive variables

3.2.1) Raw data

Coefficients of the predictive variables in the raw data were analysed using R (Table3). Estimations for coefficients were made on both raw data and log transformed changed data based on log-likelihood ratio, Akaike Information Criteria (AIC), Deviance, p-value and Wald's statistics. Based on log-likelihood ratio of the variables, P19, P5, P6, P13, P2, P21, P12, P17, P14, and P3 have significant predictive capabilities. However, the order of significance based on the p-value and deviance was P5, P6, P19, P2, P13, P17, P14, P3, P16, and P4 (Table3).

Based on AIC criteria, the order of significance obtained was P5, P6, P19, P13, P2, P21, P12, P17, P14 and P3. These four selection methods yielded the same five highly significant predictive variables at the top of the table. However, Wald's statistics [Pr(>|z|)] yielded P2, P5, P13, P16, P17, P6, P12, P14, P24, and P19 variables with significant predictive capability (Table3). The differences in the output can be due to the differences in the way they trade off statistical significance with the number of parameters (as in case of AIC).

3.2.2) Log form of changed data

Raw data was manipulated by replacing '0' value with '0.5' so that when log transformed the patient measurements with '0' values were not denoted as 'NA'. The analysis of log transformed changed data revealed that the most significant variables for log-likelihood, deviance, p-value, and AIC were the same, namely, s19, s5, s14, s6, s16, s17, s2, s3, s4, and s20 in the order of most significance first. However, for Wald's statistics the most significant variables, although identical to the other methods, were yielded in a different order: s14, s19, s5, s6, s2, s16, s17, s3, s4, and s20 (Table3).

3.2.3) Comparing the coefficients of raw and log transformed data

The estimates of the coefficients from raw data and log transformed changed data were then analysed to check which scale of the data has most significant predictor variables. The investigation revealed estimates for coefficients for predictor variables, s10, s12, s14, s16, s17, s20, s21, and s23 from the log-scale were more significant than their counterparts in the raw data. A new predictor data set (Best data set) was designed which is a concoction of the most significant predictor variables from both scales (Table3). Best data set was used in the prediction modelling.

Table 3. Coefficients for Raw data and log transformed data

Raw data				Log transformed changed data			
Var	Wald's p-value	AIC	Change in Deviance	Var	Wald's p-value	AIC	Change in Deviance
P1	< 0.0001	167.1	17.99	s1	< 0.001	173.32	11.78
P2	< 0.0001	117.64	67.45	s2	< 0.0001	128.76	56.33
P3	< 0.0001	138.71	46.38	s3	< 0.0001	139.67	45.42
P4	< 0.0001	142.39	42.7	s4	< 0.0001	148.52	36.57
P5	< 0.0001	81.89	100.82	s5	< 0.0001	110.58	74.52
P6	< 0.0001	90.58	92.13	s6	< 0.0001	121.06	64.03
P7	0.0548	181.41	3.69	s7	0.2634	183.84	1.25
P8	0.2951	184	1.1	s8	0.3862	184.34	0.75
P9	< 0.001	169.46	10.83	s9	0.9875	185.09	0
P10	0.1669	183.18	1.91	s10	0.0507	181.27	3.82
P11	0.1115	154.42	2.53	s11	< 0.0001	168.19	16.91
P12	< 0.0001	126.39	41.32	s12	< 0.01	176.99	8.1
P13	< 0.0001	99.83	59.88	s13	< 0.01	176.05	7.57
P14	< 0.0001	137.11	47.99	s14	< 0.0001	116.33	68.76
P15	< 0.001	143.78	13.17	s15	0.3367	184.17	0.92
P16	< 0.0001	140.54	44.55	s16	< 0.0001	123.16	61.93
P17	< 0.0001	129.42	55.67	s17	< 0.0001	123.74	61.35
P18	0.019	172.34	5.5	s18	0.4318	184.48	0.62
P19	< 0.0001	98.27	86.83	s19	< 0.0001	99.978	85.12
P20	< 0.0001	165.88	19.21	s20	< 0.0001	159.03	26.06
P21	< 0.0001	124.57	35.14	s21	0.5764	184.78	0.31
P22	< 0.0001	168.45	16.64	s22	< 0.001	171.63	13.46
P23	0.571	184.77	0.32	s23	0.0115	178.7	6.39
P24	< 0.0001	161.1	23.99	s24	< 0.0001	164.63	20.46
P25	< 0.001	171.96	13.13	s25	0.0247	180.05	5.05
P26	< 0.0001	167.53	17.56	s26	0.0156	179.25	5.84
P27	< 0.0001	164.36	20.73	s27	0.0156	179.24	5.85
P28	< 0.01	176.15	8.94	s28	0.0102	178.49	6.6
P29	0.0165	179.35	5.74	s29	0.4242	184.45	0.64
P30	< 0.01	177.11	7.98	s30	0.2222	183.6	1.49
P31	< 0.01	176.95	8.14	s31	0.4004	184.39	0.71
P32	< 0.0001	163.71	19	s32	0.3732	184.3	0.79

Estimates for the predictive variables from the raw data and log transformed changed data was shown in the table. **Var:** predictor variables

3.3) Model fitting

The predictor variables were selected by employing logistic regression using various statistical approaches, such as forward selection, backward search, stepwise selection, and all subset selection methods. These methods were performed on R, SPSS, and Stata software platforms as mentioned in the materials and methods section.

3.3.1) Forward selection

Forward selection method was performed in the SPSS software as I was having algorithm convergence problems in R. Stepwise forward selection had fitted s12, s17, s20, and s23 variables into a model (Table4).

Table 4. Variable equation in forward selection

Variables in the Equation

							OR		CI	
		B	S.E.	Wald	Sig.	OR	Lower limit	Upper limit	Lower limit	Upper limit
Step 1[a]	s12	1.786	0.323	30.515	0	5.968	3.17	11.24	1.153	2.419
	Constant	-7.572	1.229	37.99	0	0.001				
Step 2[b]	s12	2.06	0.351	34.491	0	7.847	3.94	15.61	1.372	2.748
	s23	-1.047	0.306	11.684	0.001	0.351	0.19	0.64	-1.647	-0.447
	Constant	-1.398	1.965	0.506	0.477	0.247				
Step 3[c]	s12	1.574	0.369	18.224	0	4.825	2.34	9.95	0.851	2.297
	s17	1.193	0.468	6.487	0.011	3.297	1.32	8.25	0.276	2.11
	s23	-1.312	0.35	14.028	0	0.269	0.14	0.53	-1.998	-0.626
	Constant	-5.513	2.699	4.174	0.041	0.004				
Step 4[d]	s12	1.242	0.38	10.675	0.001	3.461	1.64	7.29	0.497	1.987
	s17	1.412	0.505	7.806	0.005	4.104	1.53	11.04	0.422	2.402
	s20	0.623	0.309	4.074	0.044	1.865	1.02	3.42	0.017	1.229
	s23	-1.549	0.396	15.329	0	0.212	0.1	0.46	-2.325	-0.773
	Constant	-7.223	2.956	5.97	0.015	0.001				

a. Variable(s) entered on step 1: s12; b. Variable(s) entered on step 2: s23;

c. Variable(s) entered on step 3: s17; and d. Variable(s) entered on step 4: s20.

3.3.2) Backward search and bidirectional search

Backward and bidirectional search in R yielded the same predictor variables in the model and failed to fit full 32 variable models in all the software I used in the study. Algorithms in these programs software did not converge. This can be blamed on the

smaller number of events (45) compared to the number of predictor variables. Hence, this method cannot be applied in our study (Table 5). Backward and bidirectional search yielded a model with 18 variables (P2 + P4 + P5 + P8 + P11 + P13 + P15 + P18 + P19 + P22 + P26 + P27 + P28 + P31 + s10 + s12 + s20 + s21) with an AIC value of 38.

Table 5. Bidirectional search method

Step		Deviance	AIC
	Intercept	7.76	36.45
+	s12	9.03	55.09
+	s10	8.91	53.23
+	P31	8.76	50.95
+	P2	8.65	49.33
+	P22	8.56	47.79
+	P5	8.44	45.86
+	P27	8.42	45.57
+	P4	8.37	44.82
+	P19	8.28	43.36
+	P13	8.28	43.28
+	P28	8.28	43.26
+	P18	8.23	42.46
+	P8	8.15	41.15
+	P15	8.14	41.01
+	s21	8.13	40.91
-	P6	7.76	38.45
-	P30	7.75	38.39
-	P25	7.75	38.39
-	s23	7.75	38.38
-	P29	7.75	38.35
-	P3	7.75	38.34
-	P24	7.75	38.34
-	P9	7.75	38.32
-	P7	7.75	38.23
-	P32	7.74	38.22
-	P1	7.72	37.87
-	s16	7.72	37.86
-	s14	7.72	37.85
+	P26	7.94	37.60
-	s17	7.71	37.58
+	s20	7.92	37.28
+	P11	7.90	36.97

‘+’addition of the variable to the model;’ -‘ is deletion

3.3.3) All subset combinations (from best data set)

This was a CPU intensive approach, where all the possible subset combinations were calculated from the 32 variables. Logistic regression analysis was carried out on the best data set raw data using 'Generalized Linear Model' (GLM) method in R. A script was written to obtain subsets containing 2 to 8 variable set combinations using the 'Bluebear' (Supplementary Table 3). This method resulted in the subset combinations which had significant values for the assessing measure of fit such as 'deviance' and/or 'AIC'.

The time scales given below were for running the 'glm' (generalized linear model) script that was on a regular desktop. The script for 2-variable subset models yielded 496 combinations in less than 10 minutes. Best of these combinations was (**A**): P5 + P19, with deviance=57.5 and AIC=63.5.The script for 3-variable subset yielded 4960 different combinations in less than 30 minutes. The most significant model derived from the 3-variable subset combination was (**B**): P5 + P19+P31, with deviance=39.8 and AIC=47.8(Table 6).

35,960 4-variable subset combinations were calculated, and the output was produced in less than 3 hours. The predictive significance in these models had improved significantly as compared to 2, 3-variable subset models, with the best model (**C**): P2+P19+ s21+P31. There were 201,376 different 5-variable subset models that can be obtained from 32 variables in 10 hours. The best model in these combinations was (**D**): P2+P6+P11+P19+P31 (Table 6).

The 'glm' script for 6-variable combination subset took more than 40 hours to run. So, for the variations in this script and for the large numbered (7 and above) variable subsets, the scripts were run in R on the 'Bluebear' platform. The script for 6-variable subset models returned 906,192 with the most significant model (**E**): P2+P6+P19+s21+P25+P30. The predictive capability has significantly increased for these models (Table 6).

The scripts for 7 and 8-variable subset model returned 3,365,856 and 10,518,300 combinations respectively. The best models for these combinations were (**F**):P2+P6+P19+s21+P22+P25+P30,and (**G**): P2+P3+P6+P15+P19+P22+P25+P30, respectively. These scripts took around 40 hours each to return the results on the Bluebear.

The significance of the model subsets increased with the increase in the number of variables in the model in all the subset combinations with deviance. However, the significance measure 'AIC' increased up to 6 variable subset model and then started declined on addition of further variables to the model (Table 6).

Table 6. Best 3 models from each subset combination

V1	V2	V3	V4	V5	V6	V7	V8	Deviance	AIC
2	6	19	21	28	31			0	14
2	6	11	19	25	30			0	14
2	6	19	21	25	30			0	14
2	6	11	19	22	25	29		0	16
2	6	15	19	22	25	30		0	16
2	6	19	21	22	25	30		0	16
2	6	17	19	21	22	25	30	0	18
2	3	6	15	19	22	25	30	0	18
2	3	6	19	21	22	25	30	0	18
2	6	11	19	31				13.256	25.256
2	6	13	19	31				14.179	26.179
2	6	19	21	31				14.247	26.247
2	19	21	31					24.281	34.281
6	19	21	30					24.625	34.625
2	13	19	31					25.073	35.073
5	19	31						39.832	47.832
5	19	30						39.977	47.977
6	19	30						40.089	48.089
5	19							57.521	63.521
5	23							59.007	65.007
1	5							63.268	69.268

The coefficients (deviance and AIC) of the best 3 models from each subset combination was shown here. V1 to V8 are variables 1 to 8 in the model. The change in deviance has improved significantly with the addition of an extra variable. However, AIC value degraded with the addition of an extra variable to the 6-variable subset model.

Investigations on the significance of estimates for each predictor variable were carried using the best prediction model of each subset combination.

3.4) Area under the ROC curve

The area under the receiver operating characteristic (ROC) curve was used to measure the discrimination and visualize the sensitivity and specificity of the prediction models in the study. The ROC curves for the best models from 2-8 variable subsets (A to G) and forward selection model were drawn. The sensitivity and specificity for 2-variable subset was found to be 94%. A similar sensitivity and specificity (AUC= 94.8%) was found for the stepwise forward selection model

(s12+s17+s20+s23). The measurements for the sensitivity and specificity has increased with the addition of each variable, 3-variable subset (97.8%), 4-variable subset (99.3%), 5-variable subset (99.8%), and attained a perfect 100% sensitivity and specificity from 6-variable subsets onwards (Figure2).

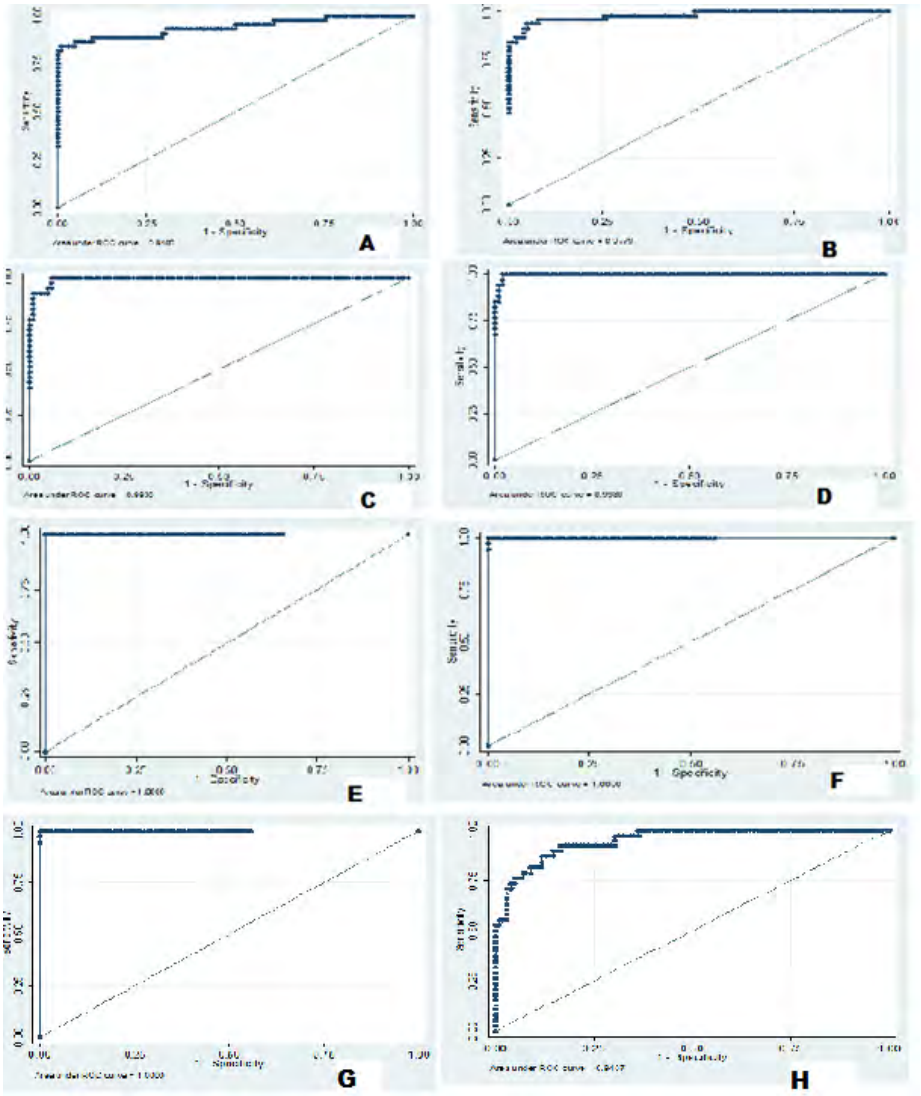

Figure 2. Area under the ROC curve values for 2-variable (A) through to 7-variable (G) subset models (as mentioned in the Section 3.3.3) and forward selection model (H).

3.5) Correlation among the best predictive variables

3.5.1) Best predictor variables from the subset models

Best predictor variables were investigated by checking the most frequent variables in the best subset models. An investigation into the 12 best models from the 2 to 8 variable subsets was carried out to find the most frequent variables in these best models subsets. As depicted in Figure 3, the most frequent variables in these best subset models were, P2, P5, P6, P11, P15, P19, s21, P22, P25, P30, and P31.

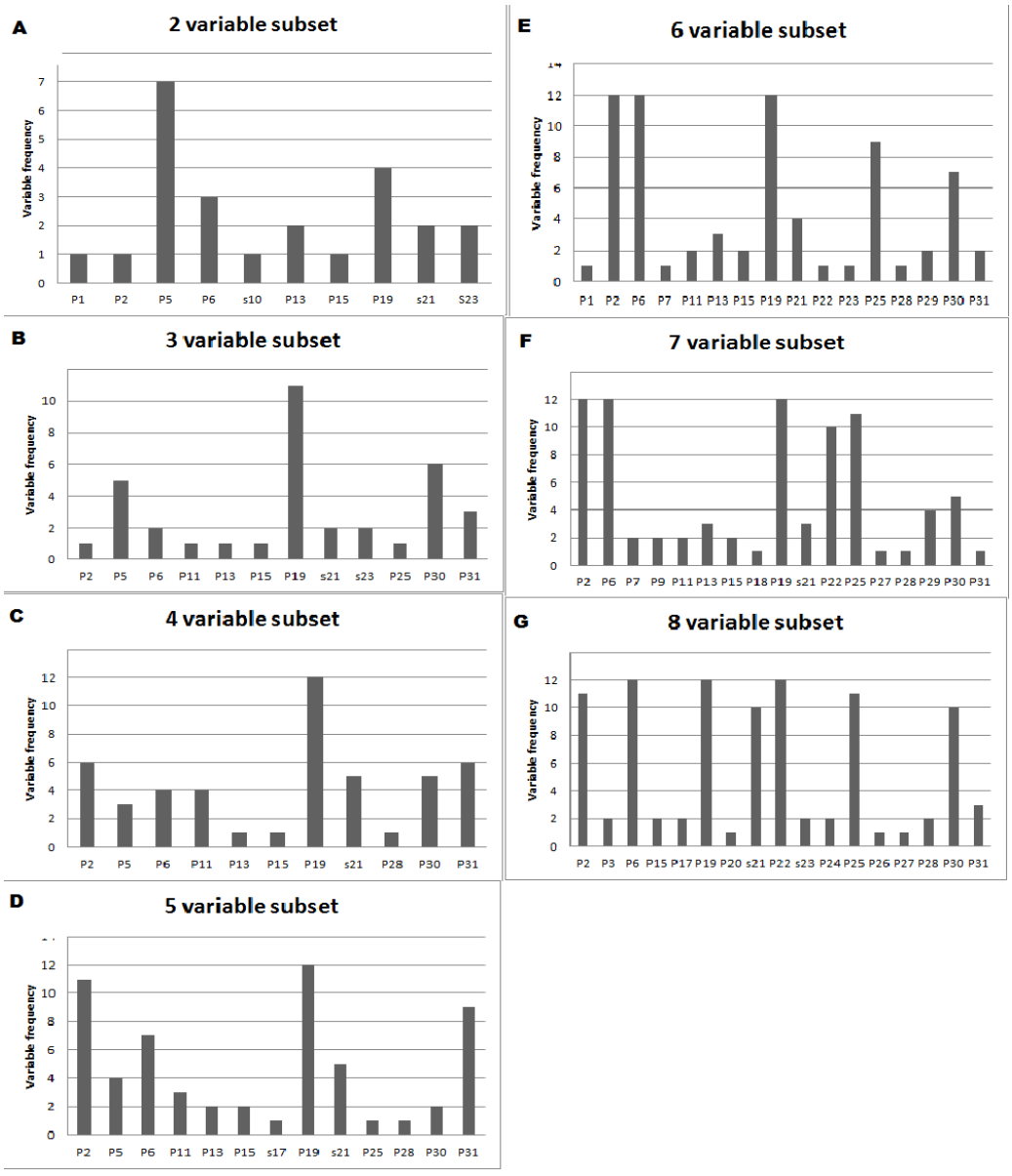

Figure 3. Frequencies of the variables in 2-8 variable subset models were shown here. Here, bar diagrams A-G are showing the variable frequencies in 2-8 variable subsets.

3.5.2) Correlation

An investigation on the correlation among the best predictive variables was carried out. Strong association between the predictor variables will undermine the significance for the estimates and also result in the inflation of variance. Analysis revealed high correlation among variables from the same steroid group as compared to the other steroid types. Predictor variable P2 has higher correlation with P6, P22, and s21. Variables P5, P3 and P6 have high correlation with each other. P6, P2, P5, and P3 were highly correlated. Similarly, predictor variables P22, P25, P31, P30, s21, and P2 had high correlation with each other (Table 7&8).

Table 7. Pearson product-moment

	P2	P3	P5	P6	P11	P15	P19	P22	P25	P30	P31
P2											
P3	0.44										
P5	0.4	0.71									
P6	0.57	0.51	0.54								
P11	0.28	0.19	0.32	0.18							
P15	0.37	0.02	0.19	0.36	0.1						
P19	0.32	0.31	0.2	0.39	0	0.15					
P22	0.56	0.07	0.16	0.27	0.15	0.16	0.14				
P25	0.56	0.12	0.18	0.33	0.23	0.2	0.22	0.91			
P30	0.36	0.05	0.18	0.3	0.26	0.27	0.25	0.78	0.77		
P31	0.49	0.01	0.13	0.33	0.18	0.2	0.09	0.87	0.88	0.83	
s21	0.55	0.17	0.29	0.44	0.27	0.32	0.23	0.69	0.65	0.72	0.63

Significant correlations were highlighted in yellow (high) and red (low) colour. The identical half of the table was trimmed to make it look more aesthetically appealing

Table 8. Pair-wise p-values

	P2	P3	P5	P6	P11	P15	P19	P22	P25	P30	P31
P2											
P3	< 0.00001										
P5	< 0.00001	< 0.00001									
P6	< 0.00001	< 0.00001	< 0.00001								
P11	0.0011	0.0249	< 0.0001	0.0373							
P15	< 0.00001	0.8369	0.0308	< 0.00001	0.2281						
P19	0.0002	0.0002	0.0206	< 0.00001	0.9954	0.0843					
P22	< 0.00001	0.3933	0.0583	< 0.0016	0.0852	0.066	0.099				
P25	< 0.00001	0.1604	0.0365	< 0.00001	< 0.01	0.0184	0.0108	< 0.00001			
P30	< 0.00001	0.597	0.0403	0.0003	< 0.01	< 0.01	< 0.01	< 0.00001	< 0.00001		
P31	< 0.00001	0.9094	0.1344	< 0.00001	0.0366	0.0224	0.2774	< 0.00001	< 0.00001	< 0.00001	
s21	< 0.00001	0.0549	0.0007	< 0.00001	< 0.0013	< 0.0001	< 0.01	<0.00001	< 0.00001	< 0.00001	< 0.00001

P2 to s21 were predictor variables from the best models. The identical half of the table was trimmed to make it look more aesthetically appealing.

3.5.3) Inspection of predictor variables from each subset

The prediction capability of the best model from each subset was analysed with the help of 'scatter plot' graph (Figure4). When these models were validated on each of the patients in the data set, a clear variation in the predictive capabilities between these models was found. Variable subsets 2, 3, 4 and 5 showed inconsistency in the predictive capability. However, variable subsets 6, 7, and 8 showed a perfect prediction capability.

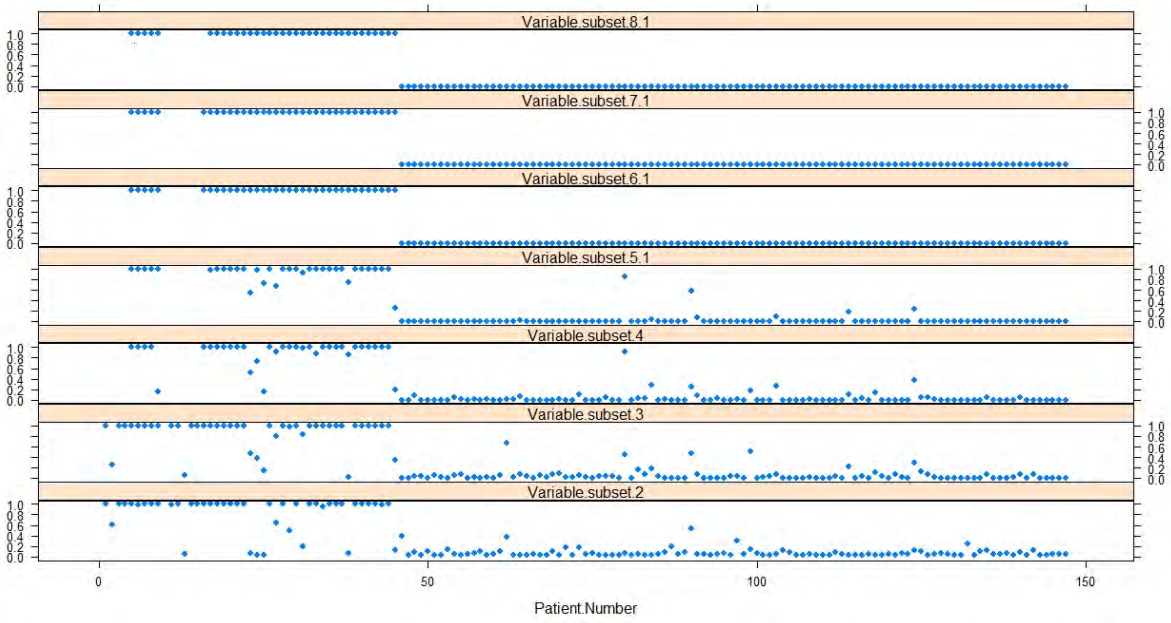

Figure 4. Scatter plot depicting the model fit for each patient. Each scatter plot show fitness of a subset model on each the patient measurements in the data (number 1-45 are ACC patients, P=1, and patient number 46-147 are ACA patients P=0). The subset model represented here is the best model (A-G) from each variable subset.

The estimates for the best predictor variables from all the subset models were obtained by logistic regression. These estimates were then standardized by dividing the predictor variable measurements with their respective Inter Quartile Range (IQR) values. The strength of the predictor variable estimates was examined starting from their own estimates and throughout the model building procedure. A good model will have its estimates gaining/losing gradually in this process. A sudden multi-fold hike or drop can be a result of 'overfitting'.

The strength of the estimates was found to be unstable with the addition of a variable to 5-variable subset model, as it is evident in the case of P19, where the significance of the estimate increased gradually up to 5 variable subset, a sudden 19 fold

increase in the 6 variable subset model, and a 75% drop in the strength of estimate in the 7 variable subset. The strength of the estimates for individual predictor variable in the majority of subset models was random as well. This phenomenon of 'overfitting' can be blamed on the small sample size, interaction, and collinearity (Table 9).

Table 9. Inspection of predictor variable's estimates in each subset model

	Intercept	Estimate	IQR	A	B	C	D	E	F	G
P5	-3.11	2.17	414.00	1.70	2.83					
P19	-2.47	1.20	381.00	0.93	2.99	5.93	11.44	197.09	47.36	
P31	-1.45	0.40	560.00		-3.63	-6.96	-22.02			
P2	-2.76	1.57	1339.00			4.21	8.84	128.99	66.09	83.21
s21	-7.20	1.29	1.14			0.87		28.83	26.31	
P6	-2.90	1.61	641.00				2.71	66.22	63.55	64.07
P11	-1.34	0.14	22.00				-2.03			
P25	-1.49	0.44	543.00					-187.28	-179.79	-196.57
P30	-1.58	0.47	1089.00					-170.32	-201.90	-120.34
P22	-1.44	0.38	1967.00						130.39	141.01
P3	-1.52	0.36	433.00							-6.22
P15	-1.72	0.57	19.50							-24.08

The estimates for each predictor variables in the best model from each subset were analysed. Here, IQR, Inter Quartile Range, A is 2-variable subset: P5+P19, B is 3-variable subset: P5+P19+P31, C is 4-variable subset: P2+P19+s21+P31, D is 5-variablesubset:P2+P6+P11+P19+P31, E is 6-variable subset: 2+P6+P19+s21+P25+P30, F is 7-variable subset: P2+P6+P19+s21+P22+P25+P30, and G is 8-variable subset: P2+P3+P6+P15+P19+P22+P25+P30.

4) Discussion

Analysis of urinary steroids and steroid metabolite data from adrenal tumour patients revealed that some of the steroids and steroid metabolites have a distinct secretion pattern between ACC and ACA patients. In ACC patients, a significantly higher excretion of An, Etio (active androgens), 5-PD, DHEA, 16α-OHDHEA (androgen precursor metabolites), THDOC, 5α-THDOC (mineralocorticoid precursor metabolites) PD, PT, 17-HP, THS (Glucocorticoid precursor metabolites), THE, 6β-OH-F, and α-cortol (Glucocorticoid metabolites) was observed when compared to ACA (Table1&Figure1).

This study was aimed at designing a biomarker tool by exploiting the differences in steroid and steroid metabolites excretion levels between ACC and ACA patients. The choice of the statistical method, regression analysis, was made basing on some criteria, such as the performance of the method, distribution of the data, missing values, sample size, and expertise.

Logistic regression analysis revealed that some of the variables have higher predictive capability. However, the order of significance was slightly different depending on the measurement criteria, for example, log-likelihood, deviance, p-value, and AIC criteria gave s19 (THS), s5 (5-PT), s14 (PD), s6 (5-PD), s16 (17HP), s17 (PT), s2 (Etio), s3 (DHEA), s4 (16α-OH DHEA), and s20 (F) in the order of most significance first. However, for Wald's statistics the most significant variables, although identical to the other methods, were yielded in a different order: s14, s19, s5, s6, s2, s16, s17, s3, s4, and s20 (Table3). The differences in the output can be attributed to the random presence of large measurement values for the predictor variables in the data, which renders Wald's statistics insignificant (type2 error) (13).

This study used stepwise regression methods and all subset combination for fitting the models. Among stepwise regression methods, only forward selection method has worked with this data. Forward selection method fitted s12, s17, s20, and s23 variables into a model. This model has a significant sensitivity and specificity (AUC=94.8%). However, a thorough investigation into the strength and stability of model estimates, interaction and collinearity between variables, and validation on a new data set is required. Backward selection method failed to fit any models as the number of events in the study is only 45 (ACC=45) and there were 32 predictor variables to fit. A large sample would have been ideal for the application of other testing methods.

Among the methods employed, all subset combination methods yielded most parsimonious models with better statistical significance (Table5). However, this method is computationally intensive and will require more time and high performance computing, such as cluster computing. Inspection of subset models on each of the

patient sample in the data revealed an increase in the predictive capability with the increase in the number of variables (Figure2&4, Supplementary table 3). The sensitivity and specificity measure for these models using ROC curve showed a perfect predictive capability in the larger subset models (from 6-variable subset and above) (Figure 2). However, an investigation into the significance of estimates in each model revealed a possible 'overfitting' in these models.

Among the subset models 4 and 5-variable subset models showed a significant prediction capability with a sensitivity and specificity (AUC) measure of 99.3 % and 99.8%. The strength of the variable estimates in these models remained stable throughout the process. However, a closer examination into the 'collinearity' and 'interaction' between the predictor variables in the model is necessary.

Backward and bidirectional search yielded a model with 18 variables (P2 + P4 + P5 + P8 + P11 + P13 + P15 + P18 + P19 + P22 + P26 + P27 + P28 + P31 + s10 + s12 + s20 + s21) with an AIC value of 38. Although this model could describe the variables in the model, this method might degrade the precision of the estimates of predictor variables in the model.

This study has been an exploratory one. But, future studies can benefit from employing a larger sample population, a better strategy to deal with the missing values, a literature review and an expert advice in reducing the number of predictor variables, and by comparing with the alternative mathematical models such as Trees and neural networks at the same time (2, 6, 9, and 10).

References

1. Barzon, L., N. Sonino, et al. (2003). "Prevalence and natural history of adrenal incidentalomas." Eur J Endocrinol **149**(4): 273-285.
2. Chatfield C. 1995 Model uncertainty, data mining and statistical inference. J R Stat Soc Ser A 158:419-466.
3. Cancer research UK, http://www.cancerresearchuk.org/.
4. David WH. and Stanley Lemeshow. Applied logistic regression, New York, Chichester: Wiley, c1989.
5. Douglas GA. Practical statistics for medical research. New York: Chapman & Hall/CRC, 1999.
6. Draper D. 1995 Assessment and propagation of model uncertainty.
7. Guyatt GH. et al. 2000 Users' Guides to the Medical Literature:XXV. Evidence-based medicine:principles for applying the Users' Guide to patient care. Evidence-based Medicine Working Group, Jama 284:217-226.
8. Grumbach, M. M., B. M. Biller, et al. (2003). "Management of the clinically inapparent adrenal mass ("incidentaloma")." Ann Intern Med **138**(5): 424-429.
9. Harrell FE. Regression modelling strategies: with applications to lineal models, logistic regression, and survival analysis. New York: Springer, 2001.
10. Hastie T. et al. The elements of statistical learning: data mining, inference, and prediction. New York: Springer, 2001. J R Stat Soc Ser B 57:45-97.
11. Kirkwood BR. and Jonathan AC Sterne. Essential Medical Statistics. 2nd Ed. Malden, Mass:Blackwell Science, 2003.
12. Mansmann, G., J. Lau, et al. (2004). "The clinically inapparent adrenal mass: update in diagnosis and management." Endocr Rev **25**(2): 309-340.
13. Aspinall, S. R., A. H. Imisairi, et al. (2009). "How is adrenocortical cancer being managed in the UK?" Ann R Coll Surg Engl **91**(6): 489-493

14. Steyerberg EW. Clinical Predictction Models: A practical Approach to Development, Validation, and Updating. New York: Springer,2010.

15. Steyerberg EW. et al. 2010 Assessing the Performance of Prediction Models: A Frame work for Traditional and Novel Measures.Epidemilogy 21:128-138.

16. The Comprehensive R Archive Network: http://cran.r-project.org/.

17. Arlt, W., M. Biehl, et al. (2011). "Urine steroid metabolomics as a biomarker tool for detecting malignancy in adrenal tumors." J Clin Endocrinol Metab **96**(12): 3775-3784

Supplementary material

Supplementary Table 1. Demographic and clinical characteristics of adrenal tumour (AT) patients

	ACA (n=102)	**ACC (n=45)**
Median age (range) at time of urine collection (yr)	60 (19-84)	55(20-80)
Sex (male, female)	39, 63	24, 21
Tumour load at the time of collection	AT (n=102)	AT, no metastasis (n=9) AT+metastasis (n=26) ACC metastasis after removal of primary tumour (n=10)
Maximum diameter of AT at the time of urine collection (median and range)	26 (9-78) mm	90 (14-230)mm
Surgical removal of AT	24/102 (24%)	30/45 (67%)
Median Weiss score [a]	1 (0-2) (n= 15)	6 (3-9) (n=21)
Duration of follow-up (median and range)	52 (26-201) months since diagnosis (n=102)	14 (1–187) months from diagnosis to death due to metastatic ACC in deceased patients (n=35) 45 (30–100) months since diagnosis in alive patients (n=10)

a The Weiss system scores the presence or absence of nine histopathological features (Weiss score range 0–9); scores under 3 are indicative of a benign adrenal tumour, scores of 3 are borderline, and scores of 4 and above are indicative of malignancy (14).

b Seven of the 10 surviving patients suffer from metastatic disease. The three remaining patients have not shown evidence of recurrence yet (all three initially presented with early-stage disease [ENS@T II (13); primary tumour diameters 80, 89, and 160 mm; histology indicative of ACC with Weiss scores of 5, 7, and 4, respectively; current survival times 45, 51, and 42 months, respectively].

** Taken from Wiebke et al. 2011(Reference 17).

Supplementary Table2. Description of the steroid and steroid metabolites used in the study

No.	Abbreviation	Common name	Chemical name	Metabolite of
Androgen metabolites				
1	An	Androsterone	5α-androstan-3α-ol-17-one	Androstenedione,testosterone, 5α-dihydro testosterone
2	Etio	Etiocholanolone	5β-androstan-3α-ol-17-one	Androstenedione,testosterone
Androgen precursor metabolites				
3	DHEA	Dehydroepi-androsterone	5-androsten-3β-ol-17-one	DHEA + DHEA sulphate (DHEAS)
4	16α-OHDHEA	16α-hydroxy-DHEA	5-androstene-3β,16α-diol-17-one	DHEA + DHEAS
5	5-PT	Pregnenetriol	5-pregnene-3β,17, 20α-triol	17-hydroxy pregnenolone
6	5-PD	Pregnenediol	5-pregnene-3β, 20α-dioland 5, 17, (20)-pregnadien-3β-ol	pregnenolone
Mineralocorticoid metabolites				
7	THA	Tetrahydro-11-dehydro corticosterone	5β-pregnane-3α, 21-diol,11, 20-dione	corticosterone, 11-dehydro corticosterone
8	5α-THA	5α-tetrahydro-11-dehydro corticosterone	5α-pregnane-3α, 21-diol-11,20-dione	corticosterone, 11-dehydro corticosterone
9	THB	Tetrahydro corticosterone	5β-pregnane-3α, 11β,21-triol-20-one	corticosterone
10	5α-THB	5α-tetrahydro corticosterone	5α-pregnane-3α, 11β,21-triol-20-one	corticosterone
11	3α5β-THALDO	Tetrahydro aldosterone	5β-pregnane-3α, 11β,21-triol-20-one-18-al	aldosterone
Mineralocorticoid precursor metabolites				
12	THDOC	Tetrahydro-11-deoxy corticosterone	5β-pregnane-3α, 21-diol-20-one	11-deoxy corticosterone

13	5α-THDOC	5α-tetrahydro-11-deoxy corticosterone	5α-pregnane-3α, 21-diol-20-one	11-deoxy corticosterone
Glucocorticoid precursor metabolites				
14	PD	Pregnanediol	5β-pregnane-3α, 20adiol	progesterone
15	3α5α-17HP	3α, 5α-17-hydroxypregnanolone	5α-pregnane-3α, 17α-diol-20-one	17-hydroxy progesterone
16	17HP	17-hydroxy pregnanolone	5β-pregnane-3α, 17α,-diol-20-one	17-hydroxy progesterone
17	PT	Pregnanetriol	5β-pregnane-3α, 17α,20α-triol	17-hydroxy progesterone
18	PTONE	Pregnanetriolone	5β-pregnane-3α, 17,20α-triol-11-one	21-deoxycortisol
19	THS	Tetrahydro-11-deoxycortisol	5β-pregnane-3α, 17, 21-triol-20-one	11-deoxycortisol
Glucocorticoid metabolites				
20	F	Cortisol	4-pregnene-11β, 17, 21-triol-3, 20-dione	cortisol
21	6β-OH-F	6β-hydroxy-cortisol	4-pregnene-6β, 11β, 17,21-tetrol-3, 20-dione	cortisol
22	THF	Tetrahydrocortisol	5β-pregnane-3α, 11β,17, 21-tetrol-20-one	cortisol
23	5α-THF	5α-tetrahydrocortisol	5α-pregnane-3α, 11β,17, 21-tetrol-20-one	cortisol
24	α-cortol	α-cortol	5β-pregnan-3α, 11β, 17,20α, 21-pentol	cortisol
25	β-cortol	β-cortol	5β-pregnan-3α, 11β, 17,20β, 21-pentol	cortisol
26	11β-OH-An	11β-hydroxy androsterone	5α-androstane-3α, 11β-diol-17-one	cortisol (+androgens)
27	11β-OH-Et	11β-hydroxy etiocholanolone	5β-androstane-3α, 11β-diol-17-one	cortisol (+androgens)

28	E	Cortisone	4-pregnene-17α, 21-diol-3, 11, 20-trione	cortisone
29	THE	Tetrahydro cortisone	5β-pregnene-3α, 17, 21-triol-11, 20-dione	cortisone
30	α-cortolone	α-cortolone	5β-pregnane-3α, 17,20α, 21-tetrol-11-one	cortisone
31	β-cortolone	β-cortolone	5β-pregnane-3α, 17,20β, 21-tetrol-11-one	cortisone
32	11-oxo-Et	11-oxo etiocholanolone	5β-androstan-3α-ol-11,17-dione	cortisone (+androgens)

** Taken from Wiebke et al. 2011(Reference 17).

Supplementary Table3. Scripts used in model selection methods

Stepwise regression

```
> fit <-
glm(cancer~P1+P2+P3+P4+P5+P6+P7+P8+P9+P11+P13+P15+P18+P19+P22+P24+P25+P26+P27
+P28+P29+P30+P31+P32+s10+s12+s14+s16+s17+s20+s21+s23,data=AT)

> step <- stepAIC(fit, direction="both")

Start:  AIC=60.3

cancer ~ P1 + P2 + P3 + P4 + P5 + P6 + P7 + P8 + P9 + P11 + P13 +
    P15 + P18 + P19 + P22 + P24 + P25 + P26 + P27 + P28 + P29 +
    P30 + P31 + P32 + s10 + s12 + s14 + s16 + s17 + s20 + s21 +
    s23
```

Stepwise backward elimination

```
> fit <-
glm(cancer~P1+P2+P3+P4+P5+P6+P7+P8+P9+P11+P13+P15+P18+P19+P22+P24+P25+P26+P27
+P28+P29+P30+P31+P32+s10+s12+s14+s16+s17+s20+s21+s23,data=AT)

> step <- stepAIC(fit, direction="backward")

Start:  AIC=60.3

cancer ~ P1 + P2 + P3 + P4 + P5 + P6 + P7 + P8 + P9 + P11 + P13 +
    P15 + P18 + P19 + P22 + P24 + P25 + P26 + P27 + P28 + P29 +
    P30 + P31 + P32 + s10 + s12 + s14 + s16 + s17 + s20 + s21 + s23
```

Script used to generate all subset combinations in the blue bear

```
####data
data<-(read.csv(file.choose(), header = T, sep = ","))

####Number of predictors
preds<-32

####Number chosen
choose<-4

library(gtools)

####Number of combinations
#library(combinat)
#library(R.basic)
#n_choose<-nChooseK(preds,choose)
n_choose<-choose(preds,choose)

####Time count

time.start<-Sys.time()

####Combinations

pred_combs<-combinations(n=preds,r=choose)

####Fitting models

glm_output<-rep(0, n_choose)
for (i in 1:n_choose)
{
glm_m_output<-glm(data$cancer~data[, pred_combs[i,1]]+data[, pred_combs[i,2]]+data[,
pred_combs[i,3]]+data[, pred_combs[i,4]], family="binomial")
glm_output[i]<-glm_m_output$deviance
}

####deviance output and ordering

deviance_mod<-cbind(pred_combs, glm_output)
deviance_mod[order(deviance_mod[,5]),]
deviance_mod<-as.data.frame(deviance_mod)

time.end<-Sys.time()

print(difftime(time.end, time.start))

write.table(deviance_mod, file="c://Prj2//Models_deviance//AT_chgv_logdeviance_mod3.csv",
sep=",", row.names=FALSE, quote = FALSE)
```