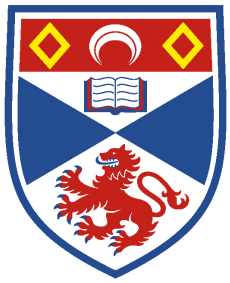
# Temporal dynamic expression of genes identified using Bayesian Network Analysis is Prognostic of Overall Survival in Ovarian Cancer

120012670
Supervisor: Dr V. A. Smith



University of
St Andrews

This is submitted in partial fulfilment for the degree of BSc(Hons)
Molecular Biology
at the
University of St Andrews

(6369 words)
18/04/16

***Declaration of Authorship and Acknowledgements***

*Except where duly acknowledged, the work reported in this paper is my own.*

*My supervisor, Dr V. A. Smith provided typical input, advice, discussion and support as well as the provision of some template scripts. Her time, guidance and encouragement were infinitely helpful and greatly appreciated.*

*Mr T. Vogogias (Collaborator) provided the use of his cluster visualisation tool COUNTERPOINT, currently in development, along with advice and information about the workings of the tool.*

*Professor W. Cresswell and Mr R. Fallon provided advice and help troubleshooting R coding issues.*

*Cameron Lockhart and Susan McGill provided advice, encouragement and unbounding friendship that I will always be grateful for.*

*Matriculation number:* 120012670 *Date:* 18/04/16

# CONTENTS

# ABSTRACT (193 words)

The treatment of ovarian cancer still faces several challenges: it is often diagnosed at a late stage, and many patients relapse due to acquisition of resistance to chemotherapeutic agents. We aimed to identify differentially expressed genes, whose expression depends on conditions such as drug resistance and treatment type, which can predict prognosis in patients.

Gene expression data obtained from either carboplatin sensitive or carboplatin resistant cell lines, and treated with either carboplatin alone, or in combination with paclitaxel, was analysed to look for differentially expressed genes. Such genes were then clustered and used to build Bayesian networks to assess causal dependency on treatment and cell line variables. It was hypothesised that directly dependent gene clusters would be prognostic in an independent clinical dataset.

3 gene clusters appeared to have gene expression profiles dependent on the presence or absence of treatment. All three gene groups allowed clinical data sets to be clustered into groups that had significantly different overall survival and thus are prognostic. *TUBA4A* alone is able to predict prognosis at a level of statistical significance. The prognostic genes identified provide avenues of further research into the development of a clinically viable prognostic tool.

# INTRODUCTION

During 2015, 21,290 women in the US were diagnosed with ovarian cancer and 14,180 were killed by the disease (Siegel *et al.*, 2015) making it a major cause of cancer fatality in women. Despite the existence of relatively effective chemotherapeutic treatments (du Bois, *et al*., 2003), which are used in combination with surgery (Raja, Chopra and Ledermann, 2012), late diagnosis - due to few early symptoms - means that long-term survival remains relatively poor (Shapira, *et al.,* 2014). Therefore, a major aim of the current ovarian cancer research field is to identify new biomarkers, especially those expressed early in the cancers progression, that could be used to improve diagnosis and hence patient survival (Kobayashi, *et al.,* 2012). Ideally, a biomarker will be found with sufficient sensitivity and specificity to be used in a screening program, but thus far an adequate candidate is not available (Nguyen, *et al*., 2013).

Another issue facing the field is that, despite the majority of patients initially responding to the current available treatment regimens - carboplatin, and carboplatin in combination with paclitaxel (Luvero, Milani and Ledermann, 2014) - many patients develop resistance to the medication after a period of time (Agarwal and Kaye, 2003). Carboplatin is a platinum based alkylating agent which functions by creating guanine crosslinks in the DNA, disrupting replication and preventing cell growth (Knox, *et al.,* 1986). However resistance often develops, increasing DNA repair and drug inactivation (Eckstein, 2011). Paclitaxel is a non-platinum based agent that acts by inhibiting microtubules during mitosis and meiosis and thus causing cell cycle arrest and apoptosis (Kumar, *et al.,* 2010). However, even in combination, the drugs are imperfect. Therefore, current research is also aiming to elucidate more of the mechanisms underlying the development of resistance whilst discovering ways to

predict its development and looking for new drug targets that may help overcome the problem of resistance (Konstantinopoulos, Spentzos and Cannistra, 2008). The discipline is increasingly looking towards technology and computational biology to aid such research.

The recent growth in the biotechnology field has lowered the cost of techniques such as sequencing and gene expression assays, making them more viable for use in the clinical setting. Therefore, there has been an increase in novel approaches incorporating gene expression analysis into research methods and clinical tools when investigating and treating ovarian cancer (Lisowska, *et al*., 2014). Gene expression analysis methods have been used to find both novel markers (Nolen and Lokshin, 2012; Beer, *et al.,* 2013) and to aid prognosis (Spentzos, *et al.,* 2004, Cai *et al.,* 2015).

In particular, network analysis has been used to identify prognostic signatures of ovarian cancer, but using only a static gene expression array (Coveney, *et al.,* 2015). Bayesian network analysis of cancer gene expression data has also previously been implemented in the study of breast cancer (Hedenalk, *et al*., 2001; van de Vijver, *et al.,* 2002; Gevaert, *et al*., 2006) and thyroid cancer (Polanski, *et al.,* 2007) amongst others. Bayesian network analysis may provide a beneficial method of machine learning when using patient data to make prognostic decision as the methodology allows for "noisy data", a common feature of clinical data (Friedman, *et al.,* 2000).

In keeping with these aims, we analyse here a temporal gene expression data series, previously used by Koussounadis, *et al.,* (2014) looking at cancer cell's response to the two major medication regimens in clinical use. By comparing a normal, platinum sensitive ovarian cancer cell line, OV1002, to a platinum resistant ovarian cancer cell line, HOX424, we hope to identify differentially expressed genes

and pathways. We aim to examine whether genes identified in this *in vitro* data are prognostic *in vivo.*

This is done by building Bayesian networks, modelling the statistical relationships between the cancer involved in genes and the different conditions - such as cell line and treatment type - which are typically interconnected and co-regulated (Segal, *et al.,* 2003). Genes identified as important in tumour progression are then investigated for use as prognostic signatures in a clinical data set. It is hypothesised that the differentially expressed genes with the closest statistical interaction with treatment condition and cell line will be prognostic of survival in ovarian cancer.

Here we identified several genes from our clustering and Bayesian network analysis with interesting expression patterns that may be of further interest as potential biomarkers or drug targets. We also selected several gene clusters using Bayesian network analysis that were predictive of survival in independent clinical datasets of ovarian cancer patients.

# DATA AND METHODS

## Gene Expression Data

Data analysed was taken from Koussounadis, *et al.* (2014). Briefly, xenografts were produced by implanting two different cell lines into female mice. The two cell lines produced by Faratian, *et al.* (2011) were OV1002, a platinum sensitive carcinoma, and HOX424, a carcinoma with reduced platinum sensitivity. Following implantation of the tissue, the tumour was allowed to develop until day 0 of the experiment, on which the mice were administered with either carboplatin, carboplatin + paclitaxel, or nothing (as a control) via injection. Xenograft samples from treated animals were collected on days 1, 2, 4, 7 and 14, and xenograft samples from control animals were collected on days 0, 1, 7 and 14. In total, 101 xenograft samples were taken, with each condition having 2-4 biological repeats. The totalRNA was prepared from each frozen sample and the quality of the totalRNA checked. Samples were then hybridised to Illumina hT-12 Beadchips – a previously validated method (Sims *et al,* 2012) – and the assay performed. The microarray data is available from Gene Expression Omnibus (GEO) with accession number GSE49577. The series matrix, phenodata and RAW data were downloaded, and the series matrix manually edited before being read into R as a data frame and then converted into an expressions set.

## Differential Gene Expression

Using the Bioconductor package *limma*, lists of differentially expressed genes for each condition were calculated comparing the daily expression values to the pooled control (DE_geneexpr_byday.R, Appendix A). This was done by first subsetting the data into groups separating condition and day. Contrast matrices were built to differentiate between control and treated groups and these contrast matrices

created top tables of differentially expressed genes, detailing their associated p value. Genes with a false discovery rate (FDR) adjusted p value ≤0.05 were considered, and these lists of differentially expressed genes were compared to the lists of differentially expressed genes determined by Koussandis, *et al.* (2014) (data not shown). Due to timing restraints, the differentially expressed genes found in the supplementary material of Koussandis*, et al.* (2014) were used in the further analysis. As very few differentially expressed genes in the condition HOX424 treated with carboplatin were identified as having a FDR adjusted p value ≤0.05, this condition was omitted from further analysis.

**Clustering**

The significant differentially expressed genes (FDR adjusted p≤0.05) for each condition were formatted and entered into the clustering tool COUNTERPOINT (Vogogias, 2016). This tool clusters the genes using an algomerative approach along a Euclidian distance metric. It aims to optimise clustering by providing real-time visual feedback as the two parameters are changed to alter the clustering. The two parameters are similarity and distinctiveness. To yield our clusters, a similarity score of 0.527 and a distinctiveness score of 0.5 were used. These scores were chosen based on visual estimation of the most appropriate balance between cluster number and homogeneity. This yielded 8 clusters within the OV1002 carboplatin treated condition, 6 clusters within the OV1002 carboplatin and paclitaxel treated condition and 10 clusters for the HOX424 carboplatin and paclitaxel treated condition. Group content ranged from 10s to 1000s of genes per cluster. Means were made across the genes for each cluster, producing average temporal gene expression profiles for each

cluster that represented the general expression pattern across the days for use in Bayesian network analysis.

**Bayesian network inference**

The software Banjo was used to build Bayesian networks from the differentially expressed genes (Smith, *et al.,* 2006). Banjo is a java based program that infers statistical dependencies between variables based on Bayes theory. Bayes theory states:

$$p(A)=p(A|B)p(B)/p(A)$$

This allows a numerical value to be calculated which defines the probabilistic relationship between any two variables. This relationship can be calculated for a number of variables and these relationships are organised into directed acyclic graphs, with nodes representing variables and the links between them representing the statistical relationships. The links, or arcs as technically termed, between the nodes represent direct dependencies, with direction of the arc indicating causality or diagnostic uses (Jensen and Nielsen, 2007). Banjo version 2.20 (available at http://users.cs.duke.edu/~amink/software/banjo/) was applied to all genes, with all 24 clusters from each condition acting as nodes alongside custom nodes representing variables. Greedy searches were performed using a RandomLocalMove proposer. The data was discretised into three groups within Banjo using a quantitative method (q3), which divides the data into 3 equal groups along an Euclidian distance. Further settings were ESS=1 and MaxParentCount=5. During the network search, a maximum of 10,000,000 restarts were permissible, with a maximum search time of 3 hours.

The first network analysed contained the 24 gene cluster nodes as well as one custom node representing the cell line (OV1002, platinum sensitive, and HOX424,

platinum resistant) and treatment condition (control, carboplatin, or carboplatin and paclitaxel). The same network analysis was then performed using all 24 cluster nodes but omitting treatment as a node, leaving only cell line as a condition node. This was repeated using the 24 cluster nodes but with no condition node for cell line, and instead one node representing type of treatment (treatment) and another representing the presence or absence of treatment, although not specifying the treatment combination (treatmentyn). Because these included redundant information and would thus be naturally connected in any Bayesian network, the algorithm was informed by inclusion of a structure file, indicating that there should be no interaction between the treatment node and the treatmentyn node. The Bayesian analysis was repeated again using the 24 cluster nodes and a condition node representing only treatment type. Finally, network analysis was carried out using the 24 cluster nodes and a treatment representing the presence or absence of treatment (treatmentyn) only. For each network analysis, 10 searches were carried out, each producing a consensus graph from the top 100 scoring networks. The consensus graph is created in a post processing step, carried out by Banjo, that weights links based on the ranking of the constituent individual graphs (Sladeczek, Hartemink and Robinson, 2008). These 10 consensus graphs were then combined using a custom perl script (Smith, 2016).

**Classifier and Kaplan Meier survival analysis**

Prognostic capability of clusters directly connected to presence or absence of treatment in the Bayesian networks was assessed using an independent clinical ovarian cancer gene expression data set accessed from GEO, GSE9891 (Tothill, *et al,* 2008). Both expression data and clinical data, including overall survival, were available for this dataset. Genes within the three clusters connected to the treatmentyn

node were identified as of interest from the Bayesian network analysis. The equivalents for the genes within these clusters were selected in the clinical data set by converting both Illumina IDs and Affymetrix IDs to Entrez IDs and finding the intersection, using a custom script (change_ids.R, Appendix B). The expression values of these intersecting genes were used to hierarchically cluster the patients in the clinical data set, using a Euclidian metric (clinical_clustering, Appendix C). This produced hierachial dendrograms, one for each cluster directly connected to the treatment presence or absence (treatmentyn) condition. These dendrograms were cut at the highest point to produce two groups of patients for each gene cluster.

Kaplan-Meier graphs were then plotted to assess the effect of these genes on patient progression free survival (time until relapse) and overall survival (time until death) using the Tothill, *et al.* (2008) dataset (Kaplan_meier.R, Appendix D). The prognostic power was assessed by calculating the appropriate statistics: both p value, identifying statistically significant differences between the clusters survival, and chi-square. Due to multiple testing, the p value considered as statistically significant was adjusted to $\leq 0.0167$ ($p \leq 0.05/3$) according to the Bonferroni correction.

## RESULTS

**Differentially expressed genes in each cell line and treatment condition cluster in groups of varying sizes, with distinct temporal expression patterns.** The differentially expressed genes in the OV1002 treated with carboplatin condition were clustered into 8 groups by the COUNTERPOINT visualisation tool (Figure 1A). Cluster 1 (Figure 1B) was made up of a large number of genes, with genes showing slightly low gene expression across the entire 14 days, with little temporal variation. Cluster 2 (Figure 1C) similarly contained many genes, the largest of the OV1002 carboplatin clusters. The vast majority of genes within the cluster showed slightly high expression with a minor incline towards day 14. A few genes in this cluster showed an expression peak (~+1.0 log fold change (LFC)) at day 4. Both Clusters 1 and 2 were large in size, and showed the most divergence - or largest range of expression values - across individual genes. Cluster 3 (Figure 1D) contained only 3 genes, which varied significantly across the 14 days, with normal expression on days 1, 4 and 14, but moderately high expression on days 2 and 7 (~+0.5 LFC). Cluster 4 (Figure 1E) contained 8 genes, which were expressed highly on day 1 before their expression dropped on day 2 and then gradually rose again until day 14 (~+0.75 LFC). Cluster 5 (Figure 1F) contained only 2 genes, which were both expressed at low levels on day 1, then gradually increased, peaking at day 7 (~+0.5 LFC), before declining slightly on day 14 back to a normal level. Cluster 6 (Figure 1G) showed normal expression on day 1 before rising to a peak of high expression on day 4 (~+1.6 LFC) , and then dropping back to normal expression on day 14. Cluster 7 (Figure 1H) has normal expression until day 2, and then continually declines until a trough on day 14 (~-1 LFC). Cluster 8 (Figure 1I) shows rather similar

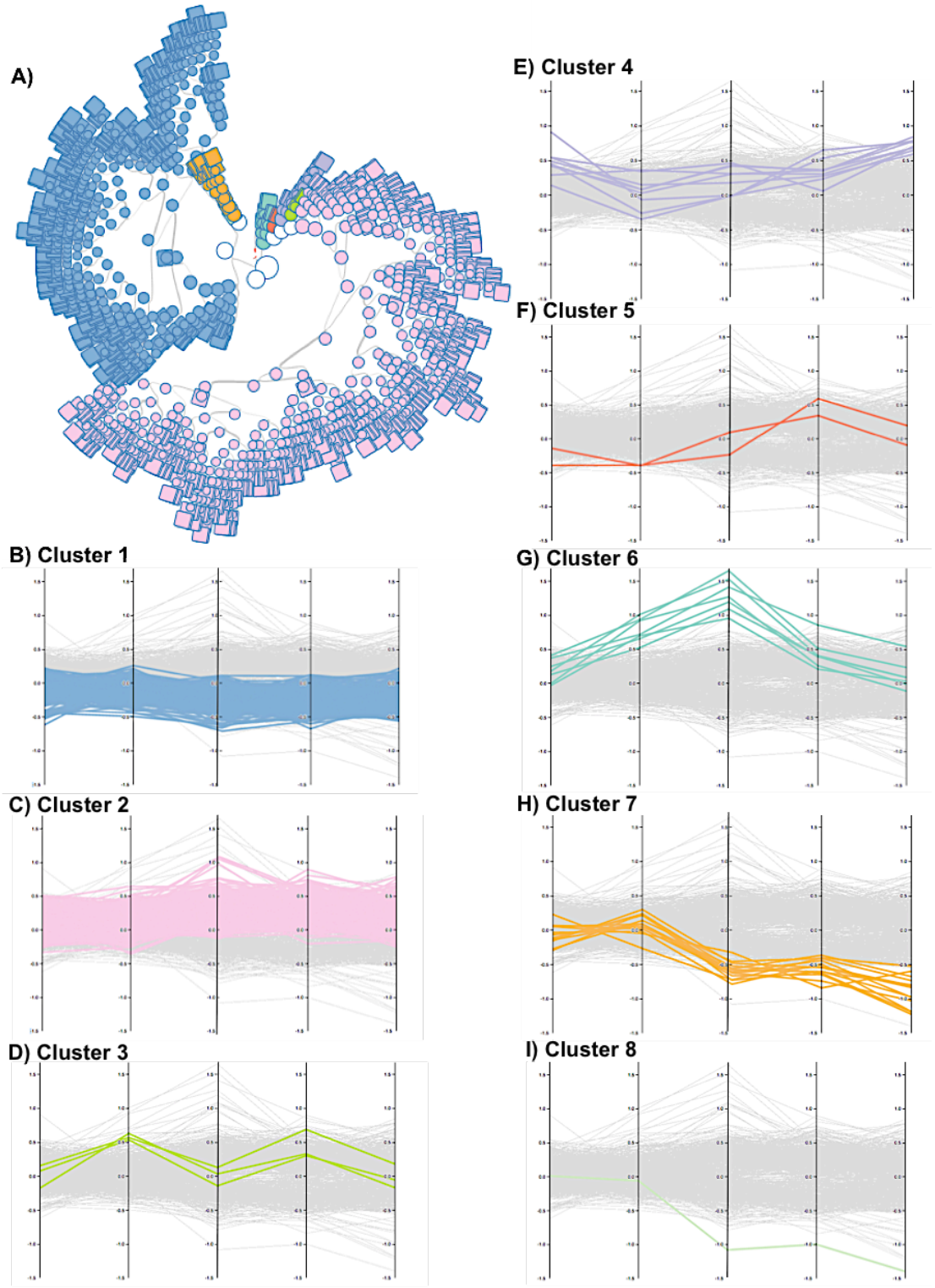



**Figure 1: Differentially expressed genes in the OV1002 carboplatin condition, clustered based on Euclidian metric.** COUNTERPOINT visualisation of (A) Dendrogram of all 8 clusters, with each square representing a gene and each colour representing a different cluster. (B-I) Temporal gene expression of log fold change for individual clusters. The y axis represent Day 1, Day 2, Day 4, Day 7 and Day 14, from left to right and each range from +1.5 to -1.5 log fold change. (B) Cluster 1. (C) Cluster 2. (D) Cluster 3. (E) Cluster 4. (F) Cluster 5. (G) Cluster 6. (H) Cluster 7. (I) Cluster 8.

temporal expression to cluster 7, however its expression declines at a faster rate, with its trough also occurring on day 14, but reaching a lower value (~-1.4 LFC).

The differentially expressed genes in the condition OV1002 treated with carboplatin and paclitaxel were divided into 6 clusters (Figure 2A). Cluster 1 (Figure 2B) has normal expression on day 1, with a decline on day 2 to a slightly low level (~ -1 LFC) which is maintained until day 14. Cluster 2 (Figure 2C) exhibits normal expression on day 1 that rises to a high peak on day 4 (~+2.25 LFC), which then drops back to normal expression on d

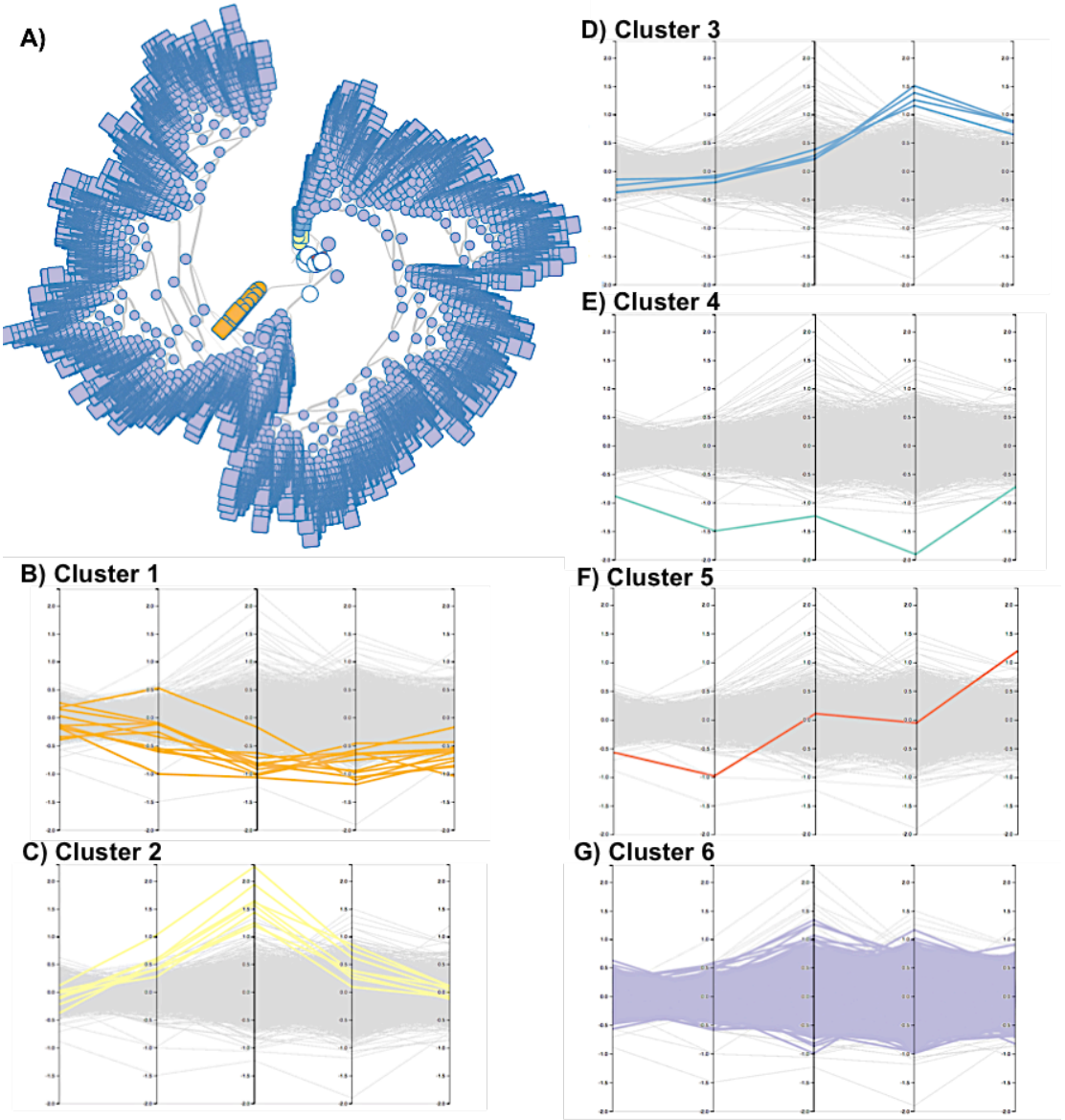



**Figure 2: Differentially expressed genes in the OV1002 carboplatin and paclitaxel condition, clustered based on Euclidian metric.** COUNTERPOINT visualistaion of (A) Dendrogram of all 6 clusters, with each square representing a gene and each colour representing a different cluster. (B-I) Temporal gene expression of log fold change for individual clusters. The y axis represent Day 1, Day 2, Day 4, Day 7 and Day 14, from left to right and each range from +2.0 to -2.0 log fold change. (B) Cluster 1. (C) Cluster 2. (D) Cluster 3. (E) Cluster 4. (F) Cluster 5. (G) Cluster 6.

ay 14. Cluster 3 (Figure 2D) has normal expression on day 1 that then slowly increases until a peak at day 7 (~+1.5 LFC), which is followed by a slight decline to day 14. Cluster 4 (Figure 2E) consists of a single gene that has consistently low levels across the 14 days, with a trough (~-1.7 LFC) at day 7. Cluster 5 (Figure 2F) is also made up of a single gene whose expression is low on day 1 (~-0.5 LFC) and decreases slightly to day 2, but then steadily increases until a peak high expression level on day 14 (~+1.25 LFC). Cluster 6 (Figure 2E) is the largest cluster of the OV1002 treated with carboplatin and paclitaxel clusters whose genes display consistently normal expression across the 14 day. This cluster exhibits the most variation of the clusters across genes with some peaks and troughs notable around days 4 and 7.

Differentially expressed genes in HOX424 carboplatin and paclitaxel clustered into the most groups, 10 in total (Figure 3A). Two of these groups had large numbers of genes but the rest were small, containing 4 or less genes. Cluster 1 (Figure 3B) was one of the larger groups and showed slightly above normal gene expression, with a small peak (<+1 LFC) on day 7. At day 14 there appears to be some divergence, with some genes showing a second peak but others dropping to below normal expression (~-0.8 LFC). Cluster 2 (Figure 3C) contains only 4 genes, which on day 1 have normal gene expression, but which rises to a peak at day 2. They then drop to a large trough (~-1 LFC) on day 7, and returning to normal on day 14. Cluster 3 (Figure 3D) contains 2 genes and shows an inverted pattern of cluster 2, with normal expression on day 1, a small trough at day 2 (~-0.5 LFC) followed by a high peak at day 7 (up to ~+1.5 LFC) and a return to normal on day14. Cluster 4 (Figure 3E) shows a similar temporal pattern but lacks the trough at day 2, instead showing normal to just below average expression until day 4, after which there is a sharp increase in expression (~+1.15 LFC) on day 7 and then a drop to just below average on day 14. Cluster 5

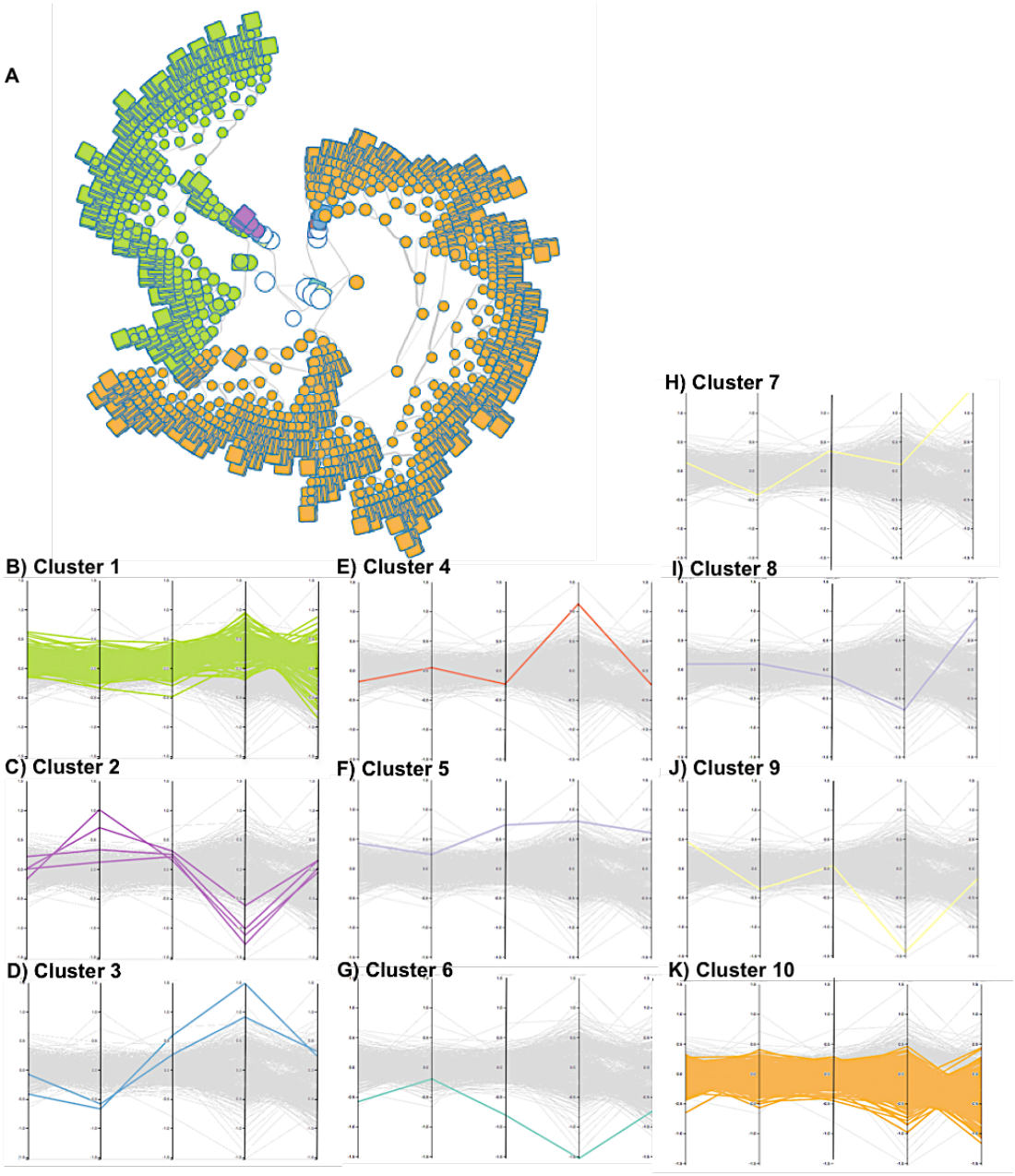



**Figure 3: Differentially expressed genes in the HOX424 carboplatin and paclitaxel condition, clustered based on Euclidian metric.** COUNTERPOINT visualisation of (A) Dendrogram of all 10 clusters, with each square representing a gene and each colour representing a different cluster. (B-I) Temporal gene expression of log fold change for individual clusters. The y axis represent Day 1, Day 2, Day 4, Day 7 and Day 14, from left to right and each range from +1.5 to -1.5 log fold change. (B) Cluster 1. (C) Cluster 2. (D) Cluster 3. (E) Cluster 4, up late. (F) Cluster 5. (G) Cluster 6. (H) Cluster 7. (I) Cluster 8. (J) Cluster 9. (K) Cluster 10.

(Figure 3F) is also made of a single gene and has a much smoother temporal gene expression pattern. It has a slightly above average gene expression on day 1, which drops slightly at day 2 and then rises very gradually to a low peak (~+0.8 LFC) at day 7 and then drops slightly (~+0.6 LFC) on day 14. Cluster 6 (Figure 3G) is another single gene cluster which starts on day 1 with expression below normal (~-0.6 LFC). This rises slightly on day 2 and then drops at a steady pace to a deep trough on day 7

(<-1.5 LFC) and then rises slightly to day 14 (~-0.75 LFC). Cluster 7 (Figure 3H) has a much more dynamic temporal gene expression profile, made up of a single gene. On day 1 it has a normal gene expression which then drops on day 2 (to ~-0.4 LFC), but rises on day 4 (to ~+0.3 LFC). It drops back to a normal level on day 7 and then rises to a large peak (~+1.5 LFC) on day 14. Cluster 8 (Figure 3I) contains a single gene that is expressed at normal level until day 4, before dropping to a trough (~-0.7 LFC) at day 7 and then rising (to ~+0.8 LFC,) well above normal expression, on day 14. Cluster 9 (Figure 3J) begins with high expression on day 1 (~+0.5 LFC), which drops to slightly low expression on day 2 (~-0.35 LFC) and rises to normal expression on day 4. There is then a large drop in expression to a trough at day 7 (~-1.4 LFC) and then a rise back to normal expression by day 14. The final cluster (Figure 3K) is the largest of the 10, containing 527 genes. This cluster shows similar expression to cluster 1 but at a lower level. It begins at normal expression on day 1 and remains there until day 4. On day 7 there is partial divergence, with some genes increasing their expression and others decreasing. The same things happen at day 14, but with more of a trend to decreasing expression. From hence forth, cluster names shall be abbreviated such that 'ov' represents OV1002 platinum sensitive cells, 'hx' represents HOX424 platinum resistant cells, 'c' represents cells treated with carboplatin only and 'ct' represents cells treated with carboplatin and paclitaxel in combination, with the following number representing the cluster number.

**Cell line and the presence or absence of treatment has a strong statistical influence on gene expression pattern.** When both cell line and treatment type were included as variable nodes within Bayesian network analysis, gene expression was directly dependent on cell line for a large number of the gene clusters - 19 of the 24

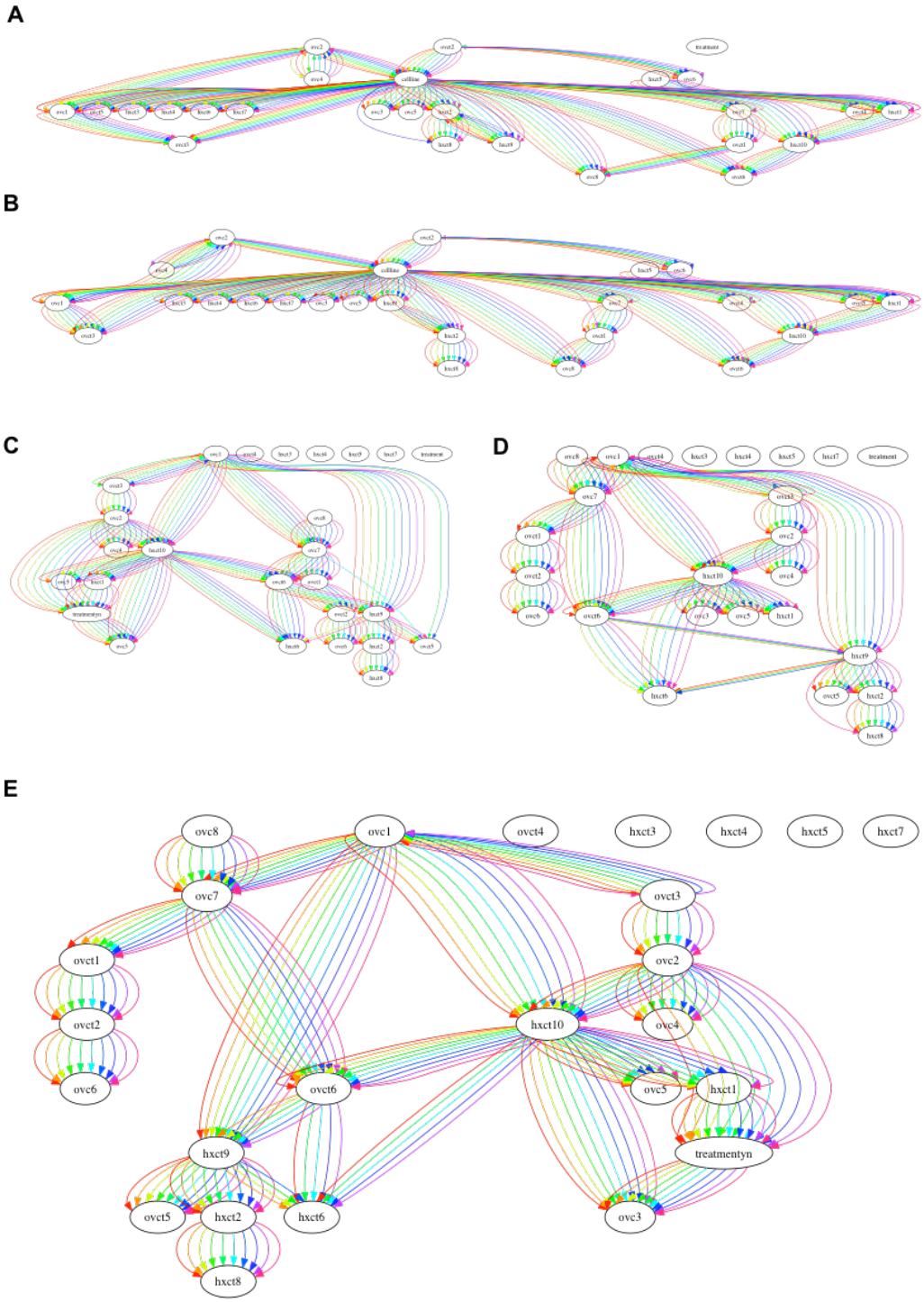



**Figure 4: Static Bayesian Networks modelling statistical dependency relationships between 24 gene clusters and different condition variables**. Each network is a representation of 10 searches combined with each colour representing links found by an individual search. Each combined graph models the relationship between different sets of variables. The nodes are coded so that 'ov' represents OV1002 platinum sensitive cells, 'hx' represents HOX424 platinum resistant cells, 'c' represents those treated with carboplatin only and 'ct' represents those treated with carboplatin and paclitaxel.(A) 24 nodes representing each of the gene clusters, plus one node representing cell line (cellline) and one node representing treatment type (treatment). (B) 24 nodes representing all gene clusters plus a node representing cell line (cellline). (C) 24 nodes representing gene clusters in addition to a node representing treatment type (treatment), and one node representing the presence or absence of treatment (treatmentyn). (D) 24 nodes representing each of the clusters plus one node representing treatment type (treatment). (E) 24 nodes representing the gene clusters plus a node representing the presence or absence of treatment (treatmentyn).

nodes share a link with cell line (Figure 4A). In comparison, the treatment type node is the only one to have no links. Accordingly, the removal of the treatment node has no effect on the rest of the network and the large level of links with cell line are retained (Figure 4B). The removal of cell line as a node and introduction of one node representing type of treatment and another node representing presence or absence of treatment causes a dramatic change of shape to the Bayesian network (Figure 4C). The presence or absence of treatment shares links with 3 other nodes, ovc2 (OV1002 treated with carboplatin, cluster 2), ovc3 (OV1002 treated with carboplatin, cluster 3) and hxct1 (HOX424 treated with carboplatin and paclitaxel, cluster 1). In contrast, the treatment type node shares no links with other nodes. There are several other nodes that also have no links, all of which are cluster nodes of genes differentially expressed when treated with carboplatin and paclitaxel; ovct4, hxct3, hxct4, hxct5 and hxct7. When the presence or absence of treatment node is removed, few differences are noted, and the treatment type node continues to share no links with any other nodes (Figure 4D). The final network calculated – which included all 24 gene cluster nodes and an additional node representing the presence or absence of treatment (Figure 4E) - showed few differences to the network produced when the treatment type node is included. Treatmentyn retained its links with ovc2, ovc3 and hxct1 whilst the nodes ovct4, hxct3, hxct4, hxct5 and hxct7 still lacked any links with other nodes. This final network was the one used to choose gene clusters whose prognostic ability would be tested.

**Genes in the clusters ovc2, ovc3 and hxct1 can be used to cluster patients based on expression, splitting them into groups which have differential survival patterns.** Kaplan Meier plots show the difference in survival patterns between patient

groups split based on expression of genes in ovc2, ovc3 or hxct1. In each case, the cluster with the better survival was coloured blue and called cluster 1, and the other cluster was coloured green and called cluster 2, by convention. Patients clustered and split based on expression of ovc2 show differential progression free survival, $p=0.0361$, $\chi^2=4.4$, df=1 (Figure 5A). As $p\geq0.0167$, the difference is deemed insignificant. However the two clusters do appear visually different. Cluster 1 and cluster 2 have similar relapse rates for the first ~10 months, after which relapse rate of cluster 1 plateaus significantly whereas cluster 2 continues to have a steady relapse rate. At the end of the study, ~60% of cluster 1 have yet to relapse, where as only 20% of patients in cluster 2 have yet to relapse. The same genes in the ovc2 cluster, differentiate patients to produce significantly different overall survival patterns (Figure 5B). The differential overall survival has $p=0.00264$ ($p\leq0.0167$), $\chi^2=9$, df=1, making the different survival profiles significantly different. ~50% of the patients in Cluster 1 survived until the end of the study, whereas only 20% of the patients in Cluster 2 remained alive at the end of the study.

The ovc3 cluster splits patients into two groups that have insignificant differential progression free survival, $p=0.0515$ ($p\geq0.0167$), $\chi^2= 3.8$, df=1 (Figure 5C). They both have a similar relapse rate until ~20 months at which point cluster 2 continues at a similar relapse rate, before leveling out at ~25 months, whereas cluster 1 begins to slow after ~20 months and levels out at ~55 days. In terms of overall survival, ovc3 is able to split the patients into two groups that have differential survival, $p=0.0083$ ($p\leq0.0167$), $\chi^2=7$, df=1 (Figure 5D). Cluster 2 has a consistently high death rate with all patients having died or left the study by ~60 months. Comparatively, although cluster 2 has an initially high death rate, this begins to plateau at ~60 days and 30% of patients are alive at the end of the study.

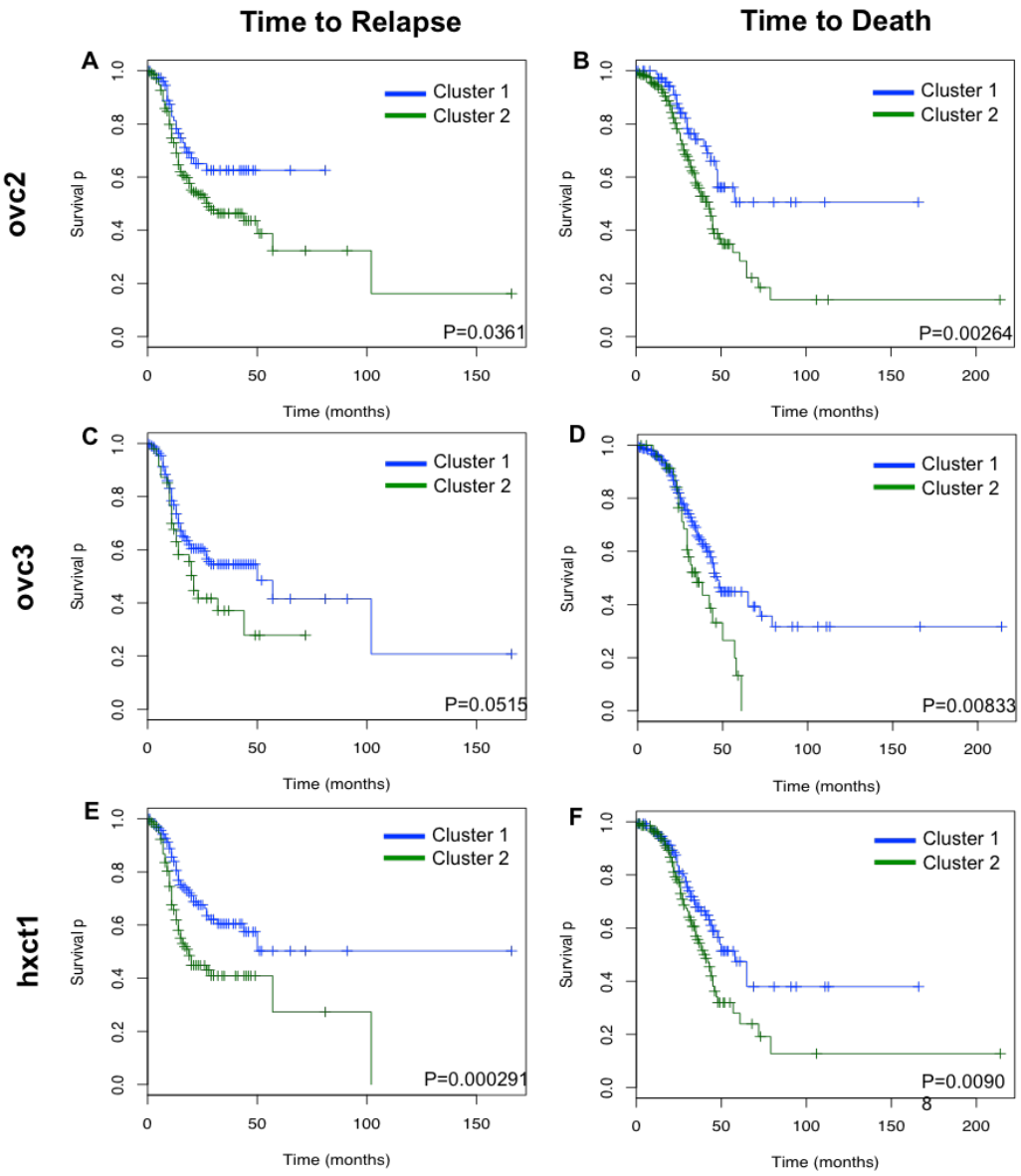



**Figure 5: Kaplan Meier Plots modeling survival of Tothill *et al.* (2005) clinical data set based on clustering using 3 sets of genes.** (A) Time to relapse of patients, split based on expression of genes in ovc2 cluster, p=0.0361, $\chi^2$= 4.4 on df=1. (B) Time to death of patients, split based on expression of genes in ovc2, p= 0.00264, $\chi^2$= 9 on df=1. (C) Time until relapse of patients, split based on expression of genes in ovc3, p=0.0515, $\chi^2$= 3.8 on df=1. (D) Time to death of patients, split based on expression of genes in ovc3, p=0.00833, $\chi^2$= 7 on df=1. (E) Time to relapse of patients, split based on expression of genes in hxct1, p=0.000291, $\chi^2$= 13.1 on df=1. (F) Time to death of patients, split based on expression of genes in hxct1, p=0.00908, $\chi^2$= 6.8 on df=1.

The hxct1 group clusters the patients into two groups that have both a significantly different progression free survival (p=0.000291 ($p \leq 0.0167$), $\chi^2$= 13.1, df=1) and overall survival (p=0.00908 ($p \leq 0.0167$), $\chi^2$=6.8,df=1). For progression free survival (Figure 5E), cluster 2 has a much higher relapse rate until ~25 months, after which it

plateaus, but all patients have relapsed by ~105months. In contrast, cluster 1 has a slower relapse rate and ~50% of patients have yet to relapse at the end of study (>150 months). The hxct1 cluster is also prognostic of overall survival (Figure 5F). Cluster 1 has a slightly slower death rate than cluster 2, with 40% of patients still surviving at last contact in cluster 1, compared to 10% in cluster 2.

## DISCUSSION

This study aimed to enquire whether Bayesian network analysis could identify gene clusters *in vitro* that were able to prognose a clinical dataset *in vivo*. We used two cell lines with different platinum sensitivity to look at gene expression following administration of drugs. Unlike other studies, we used a temporal gene expression series, with expression values measured at several time points following treatment with carboplatin or carboplatin and paclitaxel. Differentially expressed genes in each condition were clustered using a visualisation tool. This method allowed decision of clustering parameters based on visual feedback. For each condition group, a different amount of clusters - each with distinct temporal gene expression profiles - were defined using the tool. Across all three condition groups, genes tended to cluster into one or two clusters comprising the majority of the genes which had a relatively mild temporal expression pattern. The rest of the clusters contained much fewer genes, sometimes only one, but these genes were much more dynamic in their temporal gene expression. The larger groups did however contain genes with more divergent temporal expression patterns. For example clusters ovct6 (Figure 2G), hxct1 (Figure 3B) and hxct10 (Figure 3K) all appear to show conflict in their temporal pattern at one or more time points. This could be because the group contains too many genes and would benefit from further clustering. However during the clustering processes, splitting of these groups was impossible without splitting appropriately clustered groups. This could be seen as one flaw of this clustering method.

As previously stated, it was the smaller clusters of genes that tended to show the more dynamically interesting expression patterns. These may have more potential as biomarkers, as the level of difference between control and condition expression is greater, leading to more obvious changes *in vivo* when measuring in patients.

Examples of more distinctive gene expression patterns with more variable LFC values can be found in all of the conditions. For example, cluster ovc8 (Figure 1I) contains a single gene encoding Adrenomedullin (ADM) ,whose expression drops dramatically from normal on day 1 following treatment, to -1.5 LFC 14 days after treatment. *ADM* has many functions associated with cancer and tumour progression. These include stimulating the up-regulation of vasodilation and angiogenesis, acting as an apoptosis survival factor and repressing the immune response (Zudaire, Martínez and Cuttitta, 2003). Thus, the down regulation seen in the xenograft - which assumabley helps in reducing blood supply to tumours, aiding the immune response's reaction to the harmful cells and increasing cell death of the proliferating carcinoma - would be a positive reaction that would hinder tumour growth. There is no documented direct evidence or reasoning as to why the administration of carboplatin may induce this reaction automatically. ADM has previously been identified as a target of cancer treatment (Nikitenko, *et al.,* 2006), and its inhibition has been shown to increase sensitivity to chemotherapeutic agents including carboplatin (Chen, *et al*., 2012). However, in this case, it appears the cells are naturally responding in an advantageous manner, and research into the mechanism behind this may be insightful. Enhancing this effect may be a potential method of therapy.

Another example of a small cluster with dynamic temporal gene expression would be ovct5 (Figure 2F), containing only Interferon alpha inducible protein 6 (IFI6). IFI6 has normal expression on day 1 following treatment with carboplatin and paclitaxel, which rises to a peak of +1.5 LFC at day 14. IFI6 produces a peptide that has been said to possess antiapoptitc effects (Schaar, *et al.,* 2005). However, there is little literature evidence of previous links to cancer, ovarian or otherwise. This gene may be promoting the growth and spread of the tumour, and its high differential

expression may warrant further research into the action and effects of IFI6 in ovarian cancer, providing support for its consideration as a biomarker or drug target.

An example of a hxct cluster with dynamic expression would be cluster hxct7 (Figure 3H), containing FBJ murine osteorsarcoma viral oncogene homolog B (*FOSB*). *FOSB* encodes leucine zippers that are able to couple to JUN proteins, forming a transcription factor, AP-1 (Mahner, *et al.,* 2008). This has been shown to have an effect on the growth of tumours in breast cancer. It appears that low levels of *FOSB* are found in poorly differentiated mammary cells (Milde-Langosch, *et al.,* 2003), but higher levels are found in more established breast tumours (Bamberger, *et al.,* 1999). The rise in the expression of *FOSB* over time visible in our ovarian expression data could indicate that a similar effect is occurring in this instance. Further research into the gene expression pattern of *FOSB* in ovarian cancer may reveal the pattern and potential use of *FOSB* in controlling progression of the tumour. *FOSB* has previously been found to be prognostic in ovarian cancer (Kataoka, *et al.,* 2012), so although not used as a prognostic gene in our experiment, could also warrant further research into its prognostic power.

In general, through visual clustering, it is easier to identify genes of interest that are significantly different in their temporal expression from others and which may be of interest in terms of their cause and effect and whether they may be of use as a therapy target.

The networks built out of these gene clusters helped identify the causal dependencies of these temporal expression patterns on different variables such as cell line and treatment type. Our first Bayesian network, which models the dependency of the gene clusters on both cell line and treatment type infers that the expression

patterns of the majority of clusters was directly dependent on cell line (Figure 4A). This suggests that, as the major difference between cell lines was the presence or absence of platinum sensitivity, expression patterns are highly dependent on resistance to carboplatin - the platinum based chemotherapeutic agent. However, this explanation may not be conclusive, as the cell lines had other differences such as carcinoma type: OV1002 was derived from high-grade serous adenocarcinoma, whereas HOX424 is derived from clear cell/endometroid carcinoma. Previous studies have found that histology of the tumour can cause differential expression between tumours (Liowska, *et al.,* 2014) and thus some of the dependence of gene expression pattern on cell line may be caused by different cell histologies rather than chemotherapeutic sensitivities. Ideally, to be able to definitively avoid this intervening factor, xenografts would be derived from identical sources with induced chemotherapeutic sensitivities to avoid confounding variables. If it was found that cancer histology has a major effect, this may support opinions that the heterogeneity of cancer means treatment can not be general and must instead be personalised to the individuals cancer type.

One unexpected result of the initial Bayesian network is that no clusters show a dependency on treatment type, and on its removal from the network analysis no change is seen to the network structure (Figure 4B). This indicates that either treatment type is having no effect on the expression pattern of the genes, or its effect is markedly weaker than cell line. Thus when they are modelled together, the cell line dependencies override any effect of treatment type. This could be due to treatment effecting very few genes when compared to resistance - which may be caused by a number of genes simultaneously - and perhaps none of the genes effected by the treatment were assay in the xenograft microarray.

When cell line was removed as a variable and treatment was split into nodes representing the presence or absence of treatment and the type of treatment, we saw that the presence or absence of treatment node had several direct dependencies with gene cluster nodes, but treatment type still shared no links with other nodes (Figure 4C). This is surprising as it may have been hypothesised that given paclitaxel is used with the purpose of treating platinum resistant ovarian cancer (Yusuf, *et al.,* 2003), that treatment type, and the presence or absence of paclitaxel may have had a larger effect on gene expression, especially across the two cell lines. The lack of connection could be attributed to a low response rate to paclitaxel. Some studies have shown platinum resistant patients to have a response rate of only 30% (Mantia-Smaldone, Edwards and Vlad, 2011), and thus, it may not have as dramatic effect as originally expected. Another possible explanation is that, cells can become resistant to paclitaxel, at times simultaneously to platinum resistance occurrence (Stordal, *et al.,* 2012). We can not rule out the possibility of paclitaxel resistance in the HOX424 cell line which could have led to a low response to treatment with the medication.

In addition to the node representing the presence or absence of treatment, several gene cluster nodes also showed no direct links to any other nodes, gene clusters or conditions. These included ovct4, hxct3, hxct4, hxct5 and hxct7. These gene clusters have no dependency, direct or indirect, on treatment type, or even the presence of medication. These represent a set of genes that are currently not being affected by medication and therefore could indicate previously unused therapy targets. As previously stated, hxct7 comprises of *FOSB*, a gene known to have cancer promoting properties, but our Bayesian network suggests it is currently being unaffected by medication, reinforcing the idea of it as a potential drug target.

In contrast, the presence or absence of treatment node had several dependent gene expression clusters: ovc2, ovc3 and hxct1. The ovc2 and hxct1 clusters were both larger clusters of genes, with milder expression patterns. This is markedly different from the smaller ovc3 cluster, which contained only 3 genes and had a much more distinct temporal pattern.

The three clusters with direct dependency on the presence or absence of treatment were used to cluster an independent data set and assess the prognostic ability of the gene cluster. All three gene clusters are able to split clinical data into groups with significantly divergent overall survival clusters (Figure 5). However only hxct1 was able to split the clinical data into groups with significantly different progression free survival (Figure 5C). Several studies that have used supervised and semi-supervised machine learning methods to identify gene profiles in an attempt to prognose patients (Spentzos, *et al.,* 2004; Denkert*, et al.,* 2009), were able to successfully prognose overall survival, but not progression free survival, similarly to our findings. It may be that overall survival and prediction of death is reliant on more dramatic gene expression changes that are easily detectable when predicting using gene expression analysis, and that changes associated with relapse are more subtle and thus more difficult to identify. Such subtle changes may be lost, especially in the larger clusters, when cluster means are calculated.

The success of the prognosis testing helps highlight Bayesian network analysis as a powerful method of choosing significant genes of interest in prognosis testing. One benefit of Bayesian network analysis is that it deals well with noisy data (Friedman, *et al.,* 2000), which may provide an advantage over other methods. Koussounadis, *et al.,* (2014) used the same gene expression data set to select genes for

prognosis of the Tothill, *et al.,* (2008) data set, but selected genes based on those with enriched KEGG pathways. They were not able to prognose cells with any groups of genes as small as ovc3, but did manage to identify gene clusters that could prognose progression free survival. This could be as the groups that were able to do this were of a smaller size, but still comprised of multiple genes, and thus the subtle expression changes previously hypothesised to be necessary for progression free survival prediction were more visible.

The ovc3 cluster contains *TUBA4A* which encodes tubulin alpha 4a, a major component of microtubules. *TUBA4A* has previously been shown to interact with the tumour suppressor Adenomatous polyposis coli (APC) (Zumbrunn, *et al.,* 2001). *TUBA4A* has also been shown to have its expression supressed in breast cancer cells in response to certain treatments (Dezsö, *et al.,* 2014). The prognostic ability of this gene highlights it as significant in the progression of the disease and it may be a good candidate for further study as to its involvement within cancer, and potential as a drug target.

It was unexpected that the single gene in cluster ovc3 was able to predict prognosis with a better significance than hxct1, which contains 254 genes. This is in some ways surprising as one might expect a larger number of genes to have a higher combined predictive power. However this may not be the case because, as previously mentioned, the larger groups (including hxct1, Figure 3B) had shown some divergence in gene expression pattern amongst their individual constituent genes. Therefore, when a mean was made of the expression values, some of the more dynamic expression values may have been silenced, leading to loss of important prognostic gene expression patterns, and hence, lower prognostic power. This may be an example of a disadvantage of the visual clustering method, as further splitting of

the group may have led to more distinct mean gene expression patterns, and more powerful prognostic ability.

The identified gene clusters are examples of promising research avenues that could help produce a prognostic gene cluster. However, despite interesting and statistically significant results from the Kaplan Meier analysis, the prognostic ability was only shown on one data set, allowing little generalisation to a wider patient population. Repetition of the clustering and Kaplan Meier analysis on other clinical datasets would confirm the gene clusters prognostic capability and further validate further study into use of these genes in a clinical setting.

## CONCLUDING REMARKS

In support of our hypothesis, this study has demonstrated the ability and utility of Bayesian networks as a method of selecting prognostically powerful genes. We were able to find several gene clusters that could predict prognosis *in vivo*, using analysis of *in vitro* data. However, to further justify the *in vivo* prognostic ability - and hence the potential use of such genes for prognosis in a more clinical setting - validation in multiple independent clinical data sets would be necessary.

# REFERENCES


Agarwal, R. and Kaye, s. B., 2003. Ovarian cancer: strategies for overcoming resistance to chemotherapy. *Nature Reviews Cancer,* [online] 3(7), pp.502-516. Available at: <http://www.nature.com/nrc/journal/v3/n7/full/nrc1123.html> [Accessed 16 March 2016].

Bamberger, A. M., Methner, C., Lisboa, B. W., Städtler, C., Schulte, H. M., Löning, T. and Milde-Langosch, K., 1999. Expression pattern of the AP-1 family in breast cancer: association with fosB expression with a well-differentiated, receptor-positive tumor phenotype. *International Journal of Cancer,* [online] 84(5), pp.533-538. Available at: <http://www.ncbi.nlm.nih.gov/pubmed/10502734> [Accessed 16 April 2016].

Beer, L. A., Wang, H., Tang, H. Y., Cao, Z., Chang-Wong, T., Tanyi, J. L., Zhang, R., Liu, Q. and Speicher, D. W., 2013. Identification of multiple novel protein biomarkers shed by human serous ovarian tumors into the blood of immunocomprimised mice and verified in patient sera. *PLoS One,* [online] 8(3). Available at: <http://www.ncbi.nlm.nih.gov/pubmed/23544127> [Accessed 15 March 2016]

Cai, S. Y., Yang, T., Chen, Y., Wang, J. W., Li, L. and Xu, M. J., 2015. Gene expression profiling of ovarian carcinomas and prognostic analysis of outcome. *Journal of Ovarian Cancer Research,* [online] 8(15). Available at: <http://ovarianresearch.biomedcentral.com/articles/10.1186/s13048-015-0176-9> [Accessed 15 April 2016]

Chen, P., Pang, X., Zhang, Y. and He, Y., 2012. Effect of inhibition of the adrenomedullin gene on the growth and chemosensitivity of ovarian cancer cells. *Oncology Reports* [online] 27(5), pp.1461-1466. Available from: <http://www.spandidos-publications.com/or/27/5/1461> [Accessed 14 April 2016]

Coveney, C., Boocock, D. J., Rees, R. C., Deen, S. and Ball, G. R., 2015. Data Mining of Gene Arrays for Biomarkers of Survival in Ovarian Cancer. *Microarrays,* [online] 4(3), pp.324-338. Available at: <http://www.mdpi.com/2076-3905/4/3/324/htm> [Accessed 16 April 2016].

Dezsö, Z., Oestreicher, J., Weaver, A., Santiago, S., Agoulnik, S., Chow, J., Oda, Y., and Funahashi, Y., 2014. Gene Expression Profiling Reveals Epithelial Mesenchymal Transition (EMT) Genes Can Selectively Differentiate Eribulin Sensitive Breast Cancer Cells. *PLoS One,* [online] 9(8). Available at: <http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0106131> [Accessed 16 Aril 2016]

du Bois, A., Lück, H., Meier, W., Adams, H., Möbus, V., Costa, S., Bauknecht, T., Richter, B., Warm, M., Schröder, W., Olbricht, S., Nitz, U., Jackisch C., Emons, G., Wagner, U., Kuhn, W. and Pfisterer, J., 2003. A Randomized Clinical Trial of Cisplatin/Paclitaxel Versus Carboplatin/Paclitaxel as First-Line Treatment of Ovarian Cancer. *Journal of the National Cancer Institute,* [online] 95(17), pp.1320-1329. Available at: <http://jnci.oxfordjournals.org/content/95/17/1320.full.pdf+html> [Accessed 16 March 2016].

Eckstein, N., 2011.Platinum resistance in breast and ovarian cancer cell lines. *Journal of Experimental and Clinical Cancer Research,* [online] 20. Available at: <http://jeccr.biomedcentral.com/articles/10.1186/1756-9966-30-91> [Accessed 16 April 2016]

Faratian, D., Zweemer, A. J., Nagumo, Y., Sims A. H., Muir, M., Dodds, M., Mullen, P., Um, I., Kay, C., Hasmann, M., Harrison, D. J.,and Langdon, S. P., 2011. Trastuzumab and pertuzumab produce changes in morphology and estrogen receptor signalling in ovarian cancer xenografts revealing new treatment strategies. *Clinical Cancer Research,* [online] 17(13), pp.4451-4461. Available at: <http://clincancerres.aacrjournals.org/content/17/13/4451.long> [Accessed 2 April 2016].

Friedman, N., Linial, M., Nachman, I. and Pe'er, D., 2000. Using Bayesian Networks to Analyze Expression Data. *Journal of Computational Biology,* [online] 7(3/4), pp.601-620. Available at: <http://online.liebertpub.com/doi/pdf/10.1089/106652700750050961> [Accessed 2 April 2016].

Gevaert, O., De Smet, F., Timmerman, D., Moreau, Y. and De Moor, B., 2006. Predicting the prognosis of breast cancer by integrating clinical and microarray data with Bayesian networks.

*Bioinformatics,* [online] 22(14), pp.e184-190. Available at: <http://bioinformatics.oxfordjournals.org/content/22/14/e184.short> [Accessed 18 March 2016].

Hedenfalk, I., Duggan, D., Chen, Y., Radmacher, M., Bittner, M., Simon, R., Meltzer, P., Gusterson, B., Esteller, M., Kallioniemi, O. P., Wilfond, B., Borg, A., Trent, J., Raffeld, M., Yakhini, Z., Ben-Dor, A., Dougherty, E., Kononen, J., Bubendorf, L., Fehrie, W., Pittaluga, S., Gruvberger, S., Loman, N., Johannsson, O., Olsson, H. and Sauter, G., 2001. Gene-expression profiles in hereditary breast cancer. *New England Journal of Medicine,* [online] 344(8), pp.539-548. Available at: <http://www.ncbi.nlm.nih.gov/pubmed/11207349> [Accessed 18 March 2016].

Jensen, F. V. and Nielsen, T. D., 2007. *Bayesian Networks and Decision Graphs*. 2nd ed. New York: Springer.

Kataoka, F., Tsuda, H., Arao, T., Nishimura, S., Tanaka, H., Nomura, H., Chiyoda, T., Hirasawa, A., Akahane, T., Nishio, H., Nishio, K. and Aoki, D., 2012. EGRI and FOSB gene expression in cancer stroma are independent prognostic indicators for epithelial ovarian cancer receiving standard therapy. *Genes Chromosomes and Cancer,* [online] 51(3), pp.300-312. Available at: http://www.ncbi.nlm.nih.gov/pubmed/22095904 [Accessed 16 April 2016].

Knox, R. J., Friedlos, F., Lydall, D. A. and Roberts, J. J., 1986. Mechanisms of Cytotoxicity of Anticancer Platinum Drugs: Evidence That *cis*-Diamminedichloroplatinum(II) and *cis*-Diammine-(1,1-cyclobutanedicarboxylato)platinum(II) Differ Only in the Kinetics of Their Interaction with DNA. *Cancer Research,* [online] 46(4 Pt2), pp.1972-1979. Available from: <http://cancerres.aacrjournals.org/content/46/4_Part_2/1972.full.pdf> [Accessed 16 April 2016]

Kobayashi, E., Ueda, Y., Matsuzaki, S., Yokoyama, T., Kimura, T., Yoshino, K., Fujita, M., Kimura, T. and Enomoto, T., 2012. Biomarkers for Screening, Diagnosis, and Monitoring of Ovarian Cancer. *Cancer Epidemiology, Biomarkers & Prevention,* [online] 21(11), pp.1902-1912. Available at: <http://cebp.aacrjournals.org/content/21/11/1902.full> [Accessed 16 March 2016].

Konstantinopoulos, P. A., Spentzos, D. and Cannistra, S. A., 2008. Gene expression profiling in epithelial ovarian cancer. *Nature Clinical Practice Oncology,* [online] 5(10), pp.577-587. Available at: < http://www.nature.com/nrclinonc/journal/v5/n10/full/ncponc1178.html> [Accessed 18 March 2016].

Koussandis, A., Langdon, S. P., Harisson, D. J. and Smith, V. A., 2014. Chemotherapy-induced dynamic gene expression changes *in vivo* are prognostic in ovarian cancer. *British Journal of Cancer,* [online] 110(12), pp.2975-2984. Available at: <http://www.nature.com/bjc/journal/v110/n12/full/bjc2014258a.html> [Accessed 2 April 2016].

Kumar, S., Mahdi, H., Bryant, C., Shah, J. P., Garg, G. and Munkarah, A., 2010. Clinical trials and progress with paclitaxel in ovarian cancer. *International Journal of Women's Health,* [online] 2, pp.411-427. Available at: <http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3024893/ > [Accessed 14 April 2016]

Lisowska K. M., Olbryt, M., Dudaladava, V., Pamula-Pilat, J., Kujawa, K., Grzybowska, E., Jarzab, M., Student, S., Rzepecka, I. K., Jarxab, B. and Kupryjańczyk, J., 2014. Gene expression analysis in ovarian cancer – faults and hints from DNA microarray study. *Frontiers in Oncology,* [online] 4(6). Available at: < http://journal.frontiersin.org/article/10.3389/fonc.2014.00006/full> [Accessed 18 March 2016].

Luvero, D., Milani, A. and Ledermann, J. A., 2014. Treatment options in recurrent ovarian cancer: latest evidence and clinical potential. *Therapeutic Advances in Medical Oncology,* [online] 6(5), pp.229-239. Available at: <http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4206613/> [Accessed 16 March 2016].

Mahner, S., Baasch, C., Schwarz, J., Hein, S., Wölber, L., Jänicke, F. and Milde-Langosch, K., 2008. C-Fos expression is a molecular predictor of progression and survival in epithelial ovarian carcinoma. *British Journal of Cancer,* [online] 99(8), pp. 1269-1275. Available at: <http://www.nature.com/bjc/journal/v99/n8/full/6604650a.html> [Accessed 16 April 2016]

Mantia-Smaldone, G. M., Edwards, R. P. and lad, A. M., 2011. Targeted treatment of recurrent platinum resistant ovarian cancer: current and emerging therapies. *Cancer Management and Research,* [online] 1, pp. 25-38. Available at: <http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3130354/> [Accessed 16 April 2016].

Milde-Langosch, K., KAppes, H., Riethdorf, S., Löning, T. and Bamberger, A. M., 2003. FosB is highly expressed in normal mammary epithelia, but down-regulated in poorly differentiated breast carcinomas. *Breast Cancer Research and Treatment,* [online] 77(3), pp.265-275. Available at: <http://www.ncbi.nlm.nih.gov/pubmed/12602926> [Accessed 16 April 2016].

Nguyen, L., Cardenas-Goicoechea, S. J., Gordon, P., Curtin, C., Momeni, M., Chuang, L. and Fishman D., 2013. Biomarkers for early detection of ovarian cancer. *Womens Health,* [online] 9(2), pp.171-185. Available at: <http://www.ncbi.nlm.nih.gov/pubmed/23477323> [Accessed 16 March 2016].

Nikitenko, L. L., Fox, S. B., Kehoe, S., Rees, M.C.P. and Bicknell, R., 2006. Adrenomedullin and tumour angiogenesis. *British Journal of Cancer,* [online] 94(1), pp.1-7. Available at: <http://www.nature.com/bjc/journal/v94/n1/full/6602832a.html> [Accessed 16 April 2016].

Nolen, B. M. and Lokshin, A. E., Multianylate assay systems in the differential diagnosis of ovarian cancer. *Expert Opinion on Medical Diagnostics,* [online] 6(2), pp.131-138. Available at: <http://www.ncbi.nlm.nih.gov/pubmed/22468148> [Accessed 15 April 2016]

Polanski, A., Polanska, J., Jarzab, M., Wiench, M. and Jarzab, B., 2007. Application of Bayesian networks for inferring cause-effect relations from gene expression profiles of cancer versus normal cells. *Mathematical Biosciences,* [online] 209(2), pp.528-546. Available at: <http://www.ncbi.nlm.nih.gov/pubmed/17467015> [Accessed 18 March 2016].

Raja, F. A., Chopra, N. and Ledermann, J. A., 2012. Optimal first-line treatment in ovarian cancer. *Annals of Oncology,* [online] Suppl 10, pp.x118-127. Available at: <http://annonc.oxfordjournals.org/content/23/suppl_10/x118.full> [Accessed 16 March 2016].

Schaar, C. G., Kluin-Nelemans, H. C., te Marvelde, C., le Cessie, S., reed, W. P., Fibbe, W. E., 2005. Interferon- alpha as maintainance therapy in patients with multiple myeloma. *Annals of Oncology,* [online] 16(4), pp. 634-639. Available at: <http://annonc.oxfordjournals.org/content/16/4/634.full> [Accessed 14 April 2016].

Segal, E., Shapira, M., Regev, A., Pe'er D., Botstein, D., Koller, D. and Friedman, N., 2003. Module networks: identifying regulatory modules and their condition-specific regulators from gene expression data. *Nature Genetics,* [online] 34(2), pp.166-176. Available at: <http://www.nature.com/ng/journal/v34/n2/full/ng1165.html> [Accessed 2 April 2016].

Shapira, I., Oswald, M., Lovecchio, J., Khalili, H., Menzin, A., Whyte, J., Dos Santos, L., Liang, S., Bhuiya, T., Keogh, M., Mason C., Sultan. K., Budman, D., Gregersen, P. K. and Lee, A. T., 2014. Circulating biomarkers for detection of ovarian cancer and predicting cancer outcomes. *British Journal of Cancer,* [online] 110(4), pp.976-983. Available at: <http://www.nature.com/bjc/journal/v110/n4/full/bjc2013795a.html> [Accessed 16 March 2016].

Siegel, R. L., Miller, K. D. and Jemal, A., 2015. Cancer Statistics, 2015. *CA: A Cancer Journal for Clinicians,* [online] 65(1), pp.5-29. Available at: <http://www.ncbi.nlm.nih.gov/pubmed/25559415> [Accessed 16 March 2016].

Sims, A. H., Zweemer, A. J., Nagumo, Y., Faratian, D., Muir, M., Dodds, M., Um, I., Kay, C., Hasmann, M., Harrison, D. J. and Langdon, S. P., 2012. Defining the molecular response to trastuzumab, pertuzumab and combination therapy in ovarian cancer. *British Journal of Cancer,* [online] 106(11), pp.1779-1789. Available at: <http://www.nature.com/bjc/journal/v106/n11/full/bjc2012176a.html> [Accessed 2 April 2016].

Sladeczek, J., Hartemink, A. J., and Robinson, J., 2008. *Banjo: User Guide.* [online] Duke University. Available at: <http://users.cs.duke.edu/~amink/software/banjo/documentation/banjo.user.pdf> [Accessed 16 March 2016]

Smith, V. A., Yu, J., Smulders, T. V., Hartemink, A. J. and Jarvis, E. D., 2006. Computational Inference of Neural Information Flow Networks. *PLoS Computational Biology,* [online] 2(11), pp.1436-1449. Available from: http://synergy.st-andrews.ac.uk/vannesmithlab/files/2015/08/Smith_et_al_PLoSCB06.pdf [Accessed 10 April 2016].

Smith, V. A., 2016. combine_dot.pl. [email] (Personal communication, 11 March 2016)

Spentzos, D., Levine, D. A., Ramoni, M. F., Joseph, M., Gu, X., Boyd, J., Libermann, T. A. and Cannistra, S. A., 2004. Gene expression signature with independent prognostic significance in epithelial ovarian cancer. *Journal of Clinical Oncology,* [online] 22(23), pp.4700-4710. Available at: <http://www.ncbi.nlm.nih.gov/pubmed/15505275> [Accessed 15 April 2016]

Stordal, B., Hamon, M., McEneaney, V., Roche, S., Gillet, J. P., O'Leary, J. J., Gottesman, M. and Clynes, M., 2012. Resistance to paclitaxel in a cisplatin-resistant ovarian cancer cell line is mediated by P-glycoprotein. *PLoS One,* [online] 7(7). Available at: <http://www.ncbi.nlm.nih.gov/pubmed/22792399> [Accessed 15 April 2016]

Tothill, R. W., Tinker, A. V., George, J., Brown, R., Fox, S. B., Lade, S., Johnson, D. S., Trivett, M. K., Etemadmoghadam, D., Locandro, B., Traficante, N., Fereday, S., Hung, J. A., Chiew, Y. E., Haviv, I., Australia Ovarian Cancer Study Group, Gertig, D., DeFazio, A. and Botwell, D. D., 2008. Novel molecular subtypes of serous and endometretroid ovarian cancer linked to clinical outcome. *Clinical Cancer research,* [online] 14(16), pp.5198-5208. Available at: <http://www.ncbi.nlm.nih.gov/pubmed/18698038> [Accessed 16 April 2016]

van de Vijver, M. J., He, Y. D., van't Veer, L. J., Dai, H., Hart, A. A., Voskuil, D. W., Schreiber, G. J., Peterse, J. L., Roberts, C., Marton, M. J., Parrish, M., Atsma, D., Witteveen, A., Glas, A., Delahaye, L., van der Velde, T., Baterlink, H., Rodenhuis, S., Rutgers, E. T., Friend, S. H. and Bernards, R., 2002. A gene-expression signature as a predictor of survival in breast cancer. *New England Journal of Medicine,* [online] 347(25), pp.1999-2009. Available at: <http://www.ncbi.nlm.nih.gov/pubmed/12490681> [Accessed 18 March 2016].

Vogogias, T., 2016. COUNTERPOINT visualisation tool. [email] (Personal communication, 10 March 2016).

Yusuf, R. Z., Duan, Z., Lamendola, D. E., Penson, R. T. and Seiden, M. V., 2003. Paclitaxel resistance: molecular mechanisms and pharmacologic manipulation. *Current Cancer Drug Targets,* [online] 3(1), pp.1-19. Available at: <http://www.ncbi.nlm.nih.gov/pubmed/12570657> [Accessed 16 April 2016]

Zudaire, E., MArtínez, A. and Cuttitta, F., 2003. Adrenomedullin and cancer. *Regulatory Peptides,* [online] 112(1-3), pp.175-183. Available from: <http://www.sciencedirect.com/science/article/pii/S0167011503000375> [Accessed 14 April 2016]

Zumbrunn, J., Kinoshita, K., Hyman, A. A. and Näthke, I. S., 2001. Binding of the adenomatous polyposis coli protein to microtubules increases microtubule stability and is regulated by GSK3β phosphorylation. *Current Biology,* [online] 11(1), pp.44-49. Available at: <http://www.sciencedirect.com/science/article/pii/S0960982201000021> [Accessed 16 April 2016]

## APPENDIX A: DE_geneexpr_byday.R

```
# Import Biobase
library(Biobase)
# Import data as a matrix with the first column being the rownames and the data being
read ina s text, not factors
GSE49577 <- as.data.frame(read.table("/Users/hannah/Documents/basic_data.txt",
sep ="\t", header=T, row.names=1))
colnames(GSE49577)

#Subset by day and condition

OVCLCBD1 <- subset(GSE49577,
select=c(3,6,9,12,15,19,25,28,32,34,37,40,14,18,20,21))
OVCLCBD2 <- subset(GSE49577,
select=c(3,6,9,12,15,19,25,28,32,34,37,40,47,49,52,53))
OVCLCBD4 <- subset(GSE49577,
select=c(3,6,9,12,15,19,25,28,32,34,37,40,23,26,29,31))
OVCLCBD7 <- subset(GSE49577,
select=c(3,6,9,12,15,19,25,28,32,34,37,40,35,39,42,44))
OVCLCBD14 <- subset(GSE49577,
select=c(3,6,9,12,15,19,25,28,32,34,37,40,2,5,8,11))

OVCLCTD1 <- subset(GSE49577,
select=c(3,6,9,12,15,19,25,28,32,34,37,40,36,38,41,43))
OVCLCTD2 <- subset(GSE49577,
select=c(3,6,9,12,15,19,25,28,32,34,37,40,1,4,7,10))
OVCLCTD4 <- subset(GSE49577,
select=c(3,6,9,12,15,19,25,28,32,34,37,40,45,48,51,54))
OVCLCTD7 <- subset(GSE49577,
select=c(3,6,9,12,15,19,25,28,32,34,37,40,13,16,19,22))
OVCLCTD14 <- subset(GSE49577,
select=c(3,6,9,12,15,19,25,28,32,34,37,40,24,27,30,33))

HXCLCBD1 <- subset(GSE49577,
select=c(63,64,65,69,70,76,82,84,85,86,88,89,68,72,74,77))
HXCLCBD2 <- subset(GSE49577,
select=c(63,64,65,69,70,76,82,84,85,86,88,89,57,61,66,67))
HXCLCBD4 <- subset(GSE49577,
select=c(63,64,65,69,70,76,82,84,85,86,88,89,92,93,101))
HXCLCBD7 <- subset(GSE49577,
select=c(63,64,65,69,70,76,82,84,85,86,88,89,81,83,87,90))
HXCLCBD14 <- subset(GSE49577,
select=c(63,64,65,69,70,76,82,84,85,86,88,89,71,73,75,78))

HXCLCTD1 <- subset(GSE49577,
select=c(63,64,65,69,70,76,82,84,85,86,88,89,91,95,98,100))
HXCLCTD2 <- subset(GSE49577,
select=c(63,64,65,69,70,76,82,84,85,86,88,89,79,80))
```

```
HXCLCTD4 <- subset(GSE49577,
select=c(63,64,65,69,70,76,82,84,85,86,88,89,58,59,60,62))
HXCLCTD7 <- subset(GSE49577,
select=c(63,64,65,69,70,76,82,84,85,86,88,89,94,96,97,99))
HXCLCTD14 <- subset(GSE49577,
select=c(63,64,65,69,70,76,82,84,85,86,88,89,46,50,55,56))

# Import limma
library(limma)
# Look at new headers
head(GSE49577)
colnames(GSE49577)

# creating contrast matrices and top tables for each day and condition

colnames(OVCLCBD1)
design1 <- matrix(c(1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,2,2,2,2),
nrow=16, ncol=2)
colnames(design1) <- c("mean","OVCLCBD1")
design1
fit1 <- lmFit(OVCLCBD1, design1)
fit1b <- eBayes(fit1)
tableOVCLCBD1 <- topTable(fit1b, coef="OVCLCBD1", adjust.method="fdr",
number=100000)
tail(tableOVCLCBD1)
head(tableOVCLCBD1)
tableOVCLCBD1
write.table(tableOVCLCBD1,
"/Users/hannah/Documents/DEgenes_tables_updated_updated/tableOVCLCBD1.txt",
sep="\t")

colnames(OVCLCBD2)
design2 <- matrix(c(1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,2,2,2,2),
nrow=16, ncol=2)
colnames(design2) <- c("mean", "OVCLCBD2")
design2
fit2 <- lmFit(OVCLCBD2, design2)
fit2b <- eBayes(fit2)
tableOVCLCBD2 <- topTable(fit2b, coef="OVCLCBD2", adjust.method="fdr",
number=100000)
tail(tableOVCLCBD2)
head(tableOVCLCBD2)
tableOVCLCBD2
write.table(tableOVCLCBD2,
"/Users/hannah/Documents/DEgenes_tables_updated_updated/tableOVCLCBD2.txt",
sep="\t")

colnames(OVCLCBD4)
design3 <- matrix(c(1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,2,2,2,2),
nrow=16, ncol=2)
```

```
colnames(design3) <- c("mean", "OVCLCBD4")
design3
fit3 <- lmFit(OVCLCBD4, design3)
fit3b <- eBayes(fit3)
tableOVCLCBD4 <- topTable(fit3b, coef="OVCLCBD4", adjust.method="fdr",
number=100000)
tail(tableOVCLCBD4)
head(tableOVCLCBD4)
tableOVCLCBD4
write.table(tableOVCLCBD4,
"/Users/hannah/Documents/DEgenes_tables_updated_updated/tableOVCLCBD4.txt",
sep="\t")

colnames(OVCLCBD7)
design4 <- matrix(c(1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,2,2,2,2),
nrow=16, ncol=2)
colnames(design4) <-c("mean", "OVCLCBD7")
design4
fit4 <- lmFit(OVCLCBD7, design4)
fit4b <-eBayes(fit4)
tableOVCLCBD7 <- topTable(fit4b, coef="OVCLCBD7", adjust.method="fdr",
number=100000)
tail(tableOVCLCBD7)
head(tableOVCLCBD7)
tableOVCLCBD7
write.table(tableOVCLCBD7,
"/Users/hannah/Documents/DEgenes_tables_updated_updated/tableOVCLCBD7.txt",
sep="\t")

colnames(OVCLCBD14)
design5 <- matrix(c(1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,2,2,2,2),
nrow=16, ncol=2)
colnames(design5) <- c("mean", "OVCLCBD14")
design5
fit5 <- lmFit(OVCLCBD14, design5)
fit5b <-eBayes(fit5)
tableOVCLCBD14 <- topTable(fit5b, coef="OVCLCBD14", adjust.method="fdr",
number=100000)
tail(tableOVCLCBD14)
head(tableOVCLCBD14)
tableOVCLCBD14
write.table(tableOVCLCBD14,
"/Users/hannah/Documents/DEgenes_tables_updated_updated/tableOVCLCBD14.txt
", sep="\t")

colnames(OVCLCTD1)
design6 <- matrix(c(1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,2,2,2,2),
nrow=16, ncol=2)
colnames(design6) <- c("means", "OVCLCTD1")
design6
```

```
fit6 <- lmFit(OVCLCTD1, design6)
fit6b <-eBayes(fit6)
tableOVCLCTD1 <- topTable(fit6b, coef="OVCLCTD1", adjust.method="fdr",
number=100000)
tail(tableOVCLCTD1)
head(tableOVCLCTD1)
tableOVCLCTD1
write.table(tableOVCLCTD1,
"/Users/hannah/Documents/DEgenes_tables_updated_updated/tableOVCLCTD1.txt",
sep="\t")

colnames(OVCLCTD2)
design7 <- matrix(c(1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,2,2,2,2),
nrow=16, ncol=2)
colnames(design7) <- c("means", "OVCLCTD2")
design7
fit7 <- lmFit(OVCLCTD2, design7)
fit7b <- eBayes(fit7)
tableOVCLCTD2 <- topTable(fit7b, coef="OVCLCTD2", adjust.method="fdr",
number=100000)
tail(tableOVCLCTD2)
head(tableOVCLCTD2)
tableOVCLCTD2
write.table(tableOVCLCTD2,
"/Users/hannah/Documents/DEgenes_tables_updated_updated/tableOVCLCTD2.txt",
sep="\t")

colnames(OVCLCTD4)
design8 <-matrix(c(1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,2,2,2,2),
nrow=16, ncol=2)
colnames(design8) <- c("means", "OVCLCTD4")
design8
fit8 <- lmFit(OVCLCTD4, design8)
fit8b <-eBayes(fit8)
tableOVCLCTD4 <- topTable(fit8b, coef="OVCLCTD4", adjust.method="fdr",
number=100000)
tail(tableOVCLCTD4)
head(tableOVCLCTD4)
tableOVCLCTD4
write.table(tableOVCLCTD4,
"/Users/hannah/Documents/DEgenes_tables_updated_updated/tableOVCLCTD4.txt",
sep="\t")

colnames(OVCLCTD7)
design9 <- matrix(c(1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,2,2,2,2),
nrow=16, ncol=2)
colnames(design9) <- c("means", "OVCLCTD7")
design9
fit9 <-lmFit(OVCLCTD7, design9)
fit9b <- eBayes(fit9)
```

```
tableOVCLCTD7 <- topTable(fit9b, coef="OVCLCTD7", adjust.method="fdr",
number=100000)
tail(tableOVCLCTD7)
head(tableOVCLCTD7)
tableOVCLCTD7
write.table(tableOVCLCTD7,
"/Users/hannah/Documents/DEgenes_tables_updated_updated/tableOVCLCTD7.txt",
sep="\t")

colnames(OVCLCTD14)
design10 <- matrix(c(1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,2,2,2,2),
nrow=16, ncol=2)
colnames(design10) <- c("means", "OVCLCTD14")
design10
fit10 <- lmFit(OVCLCTD14, design10)
fit10b <- eBayes(fit10)
tableOVCLCTD14 <- topTable(fit10b, coef="OVCLCTD14", adjust.method="fdr",
number=100000)
tail(tableOVCLCTD14)
head(tableOVCLCTD14)
tableOVCLCTD14
write.table(OVCLCTD14,
"/Users/hannah/Documents/DEgenes_tables_updated_updated/tableOVCLCTD14.txt
", sep="\t")

colnames(HXCLCBD1)
design11 <- matrix(c(1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,2,2,2,2),
nrow=16, ncol=2)
colnames(design11) <- c("means", "HXCLCBD1")
design11
fit11 <- lmFit(HXCLCBD1, design11)
fit11b <- eBayes(fit11)
tableHXCLCBD1 <- topTable(fit11b, coef="HXCLCBD1", adjust.method="fdr",
number=100000)
tail(tableHXCLCBD1)
head(tableHXCLCBD1)
tableHXCLCBD1
write.table(HXCLCBD1,
"/Users/hannah/Documents/DEgenes_tables_updated_updated/tableHXCLCBD1",
sep="\t")

colnames(HXCLCBD2)
design12 <- matrix(c(1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,2,2,2,2),
nrow=16, ncol=2)
colnames(design12) <- c("means", "HXCLCBD2")
design12
fit12 <- lmFit(HXCLCBD2, design12)
fit12b <- eBayes(fit12)
tableHXCLCBD2 <- topTable(fit12b, coef="HXCLCBD2", adjust.method="fdr",
number=100000)
```

```
tail(tableHXCLCBD2)
head(tableHXCLCBD2)
tableHXCLCBD2
write.table(HXCLCBD2,
"/Users/hannah/Documents/DEgenes_tables_updated_updated/tableHXCLCBD2.txt",
sep="\t")

colnames(HXCLCBD4)
design13 <- matrix(c(1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,2,2,2),
nrow=15, ncol=2)
colnames(design13) <- c("means", "HXCLCBD4")
design13
fit13 <- lmFit(HXCLCBD4, design13)
fit13b <- eBayes(fit13)
tableHXCLCBD4 <- topTable(fit13b, coef="HXCLCBD4", adjust.method="fdr",
number=100000)
tail(tableHXCLCBD4)
head(tableHXCLCBD4)
tableHXCLCBD4
write.table(HXCLCBD4,
"/Users/hannah/Documents/DEgenes_tables_updated_updated/tableHXCLCBD4.txt",
sep="\t")

colnames(HXCLCBD7)
design14 <- matrix(c(1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,2,2,2,2),
nrow=16, ncol=2)
colnames(design14) <- c("means", "HXCLCBD7")
design14
fit14 <- lmFit(HXCLCBD7, design14)
fit14b <- eBayes(fit14)
tableHXCLCBD7 <- topTable(fit14b, coef="HXCLCBD7", adjust.method="fdr",
number=100000)
tail(tableHXCLCBD7)
head(tableHXCLCBD7)
tableHXCLCBD7
write.table(HXCLCBD7,
"/Users/hannah/Documents/DEgenes_tables_updated_updated/tableHXCLCBD7.txt",
sep="\t")

colnames(HXCLCBD14)
design15 <- matrix(c(1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,2,2,2,2),
nrow=16, ncol=2)
colnames(design15) <- c("means", "HXCLCBD14")
design15
fit15 <- lmFit(HXCLCBD14, design15)
fit15b <- eBayes(fit15)
tableHXCLCBD14 <- topTable(fit15b, coef="HXCLCBD14", adjust.method="fdr",
number=100000)
tail(tableHXCLCBD14)
head(tableHXCLCBD14)
```

```
tableHXCLCBD14
write.table(HXCLCBD14,
"/Users/hannah/Documents/DEgenes_tables_updated_updated/tableHXCLCBD14.txt
", sep="\t")

colnames(HXCLCTD1)
design16 <- matrix(c(1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,2,2,2,2),
nrow=16, ncol=2)
colnames(design16) <- c("means", "HXCLCTD1")
design16
fit16 <- lmFit(HXCLCTD1, design16)
fit16b <- eBayes(fit16)
tableHXCLCTD1 <- topTable(fit16b, coef="HXCLCTD1", adjust.method="fdr",
number=100000)
tail(tableHXCLCTD1)
head(tableHXCLCTD1)
write.table(HXCLCTD1,
"/Users/hannah/Documents/DEgenes_tables_updated_updated/tableHXCLCTD1.txt",
sep="\t")

colnames(HXCLCTD2)
design17 <- matrix(c(1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,2,2), nrow=14,
ncol=2)
colnames(design17) <- c("means", "HXCLCTD2")
design17
fit17 <- lmFit(HXCLCTD2, design17)
fit17b <- eBayes(fit17)
tableHXCLCTD2 <- topTable(fit17b, coef="HXCLCTD2", adjust.method="fdr",
number=100000)
tail(tableHXCLCTD2)
head(tableHXCLCTD2)
write.table(tableHXCLCTD2,
"/Users/hannah/Documents/DEgenes_tables_updated_updated/tableHXCLCTD2.txt",
sep="\t")

colnames(HXCLCTD4)
design18 <- matrix(c(1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,2,2,2,2),
nrow=16, ncol=2)
colnames(design18) <- c("means", "HXCLCTD4")
design18
fit18 <- lmFit(HXCLCTD4, design18)
fit18b <- eBayes(fit18)
tableHXCLCTD4 <- topTable(fit18b, coef="HXCLCTD4", adjust.method="fdr",
number=100000)
tail(tableHXCLCTD4)
head(tableHXCLCTD4)
write.table(tableHXCLCTD4,
"/Users/hannah/Documents/DEgenes_tables_updated_updated/tableHXCLCTD4.txt",
sep="\t")
```

```
colnames(HXCLCTD7)
design19 <- matrix(c(1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,2,2,2,2),
nrow=16, ncol=2)
colnames(design19) <- c("means", "HXCLCTD7")
design19
fit19 <- lmFit(HXCLCTD7, design19)
fit19b <- eBayes(fit19)
tableHXCLCTD7 <- topTable(fit19b, coef="HXCLCTD7", adjust.method="fdr",
number=100000)
tail(tableHXCLCTD7)
head(tableHXCLCTD7)
write.table(tableHXCLCTD7,
"/Users/hannah/Documents/DEgenes_tables_updated_updated/tableHXCLCTD7.txt",
sep="\t")

colnames(HXCLCTD14)
design20 <- matrix(c(1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,2,2,2,2),
nrow=16, ncol=2)
colnames(design20) <- c("means", "HXCLCTD14")
design20
fit20 <- lmFit(HXCLCTD14, design20)
fit20b <- eBayes(fit20)
tableHXCLCTD14 <- topTable(fit20b, coef="HXCLCTD14", adjust.method="fdr",
number=100000)
tail(tableHXCLCTD14)
head(tableHXCLCTD14)
write.table(tableHXCLCTD14,
"/Users/hannah/Documents/DEgenes_tables_updated_updated/tableHXCLCTD14.txt
", sep="\t")

#Obtaining only significant genes

OVCLCBD1pval <- subset(tableOVCLCBD1, tableOVCLCBD1$adj.P.Val<0.05)
write.table(OVCLCBD1pval,"/Users/hannah/Documents/DEgenes_tables_updated_u
pdated/OVCLCBD1pval.txt", sep="\t")
OVCLCBD2pval <- subset(tableOVCLCBD2, tableOVCLCBD2$adj.P.Val<0.05)
write.table(OVCLCBD2pval,"/Users/hannah/Documents/DEgenes_tables_updated_u
pdated/OVCLCBD2pval.txt", sep="\t")
OVCLCBD4pval <- subset(tableOVCLCBD4, tableOVCLCBD4$adj.P.Val<0.05)
write.table(OVCLCBD4pval,"/Users/hannah/Documents/DEgenes_tables_updated_u
pdated/OVCLCBD4pval.txt", sep="\t")
OVCLCBD7pval <- subset(tableOVCLCBD7, tableOVCLCBD7$adj.P.Val<0.05)
write.table(OVCLCBD7pval,"/Users/hannah/Documents/DEgenes_tables_updated_u
pdated/OVCLCBD7pval.txt", sep="\t")
OVCLCBD14pval <- subset(tableOVCLCBD14, tableOVCLCBD14$adj.P.Val<0.05)
write.table(OVCLCBD14pval,"/Users/hannah/Documents/DEgenes_tables_updated_
updated/OVCLCBD14pval.txt", sep="\t")
OVCLCTD1pval <- subset(tableOVCLCTD1, tableOVCLCTD1$adj.P.Val<0.05)
write.table(OVCLCTD1pval,"/Users/hannah/Documents/DEgenes_tables_updated_u
pdated/OVCLCTD1pval.txt", sep="\t")
```

```
OVCLCTD2pval <- subset(tableOVCLCTD2, tableOVCLCTD2$adj.P.Val<0.05)
write.table(OVCLCTD2pval,"/Users/hannah/Documents/DEgenes_tables_updated_u
pdated/OVCLCTD2pval.txt", sep="\t")
OVCLCTD4pval <- subset(tableOVCLCTD4, tableOVCLCTD4$adj.P.Val<0.05)
write.table(OVCLCTD4pval,
"/Users/hannah/Documents/DEgenes_tables_updated_updated/OVCLCTD4pval.txt",
sep="\t")
OVCLCTD7pval <- subset(tableOVCLCTD7, tableOVCLCTD7$adj.P.Val<0.05)
write.table(OVCLCTD7pval,
"/Users/hannah/Documents/DEgenes_tables_updated_updated/OVCLCTD7pval.txt",
sep="\t")
OVCLCTD14pval <- subset(tableOVCLCTD14, tableOVCLCTD14$adj.P.Val<0.05)
write.table(OVCLCTD14pval,
"/Users/hannah/Documents/DEgenes_tables_updated_updated/OVCLCTD14pval.txt"
, sep="\t")
HXCLCBD1pval <- subset(tableHXCLCBD1, tableHXCLCBD1$adj.P.Val<0.05)
write.table(HXCLCBD1pval,
"/Users/hannah/Documents/DEgenes_tables_updated_updated/HXCLCBD1pval.txt",
sep="\t")
HXCLCBD2pval <- subset(tableHXCLCBD2, tableHXCLCBD2$adj.P.Val<0.05)
write.table(HXCLCBD2pval,
"/Users/hannah/Documents/DEgenes_tables_updated_updated/HXCLCBD2pval.txt",
sep="\t")
HXCLCBD4pval <- subset(tableHXCLCBD4, tableHXCLCBD4$adj.P.Val<0.05)
write.table(HXCLCBD4pval,
"/Users/hannah/Documents/DEgenes_tables_updated_updated/HXCLCBD4pval.txt",
sep="\t")
HXCLCBD7pval <- subset(tableHXCLCBD7, tableHXCLCBD7$adj.P.Val<0.05)
write.table(HXCLCBD7pval,
"/Users/hannah/Documents/DEgenes_tables_updated_updated/HXCLCBD7pval.txt",
sep="\t")
HXCLCBD14pval <- subset(tableHXCLCBD14, tableHXCLCTD14$adj.P.Val<0.05)
write.table(HXCLCBD14pval,
"/Users/hannah/Documents/DEgenes_tables_updated_updated/HXCLCBD14pval.txt"
, sep="\t")
HXCLCTD1pval <- subset(tableHXCLCTD1, tableHXCLCTD1$adj.P.Val<0.05)
write.table(HXCLCTD1pval,
"/Users/hannah/Documents/DEgenes_tables_updated_updated/HXCLCTD1pval.txt",
sep="\t")
HXCLCTD2pval <- subset(tableHXCLCTD2, tableHXCLCTD2$adj.P.Val<0.05)
write.table(HXCLCTD2pval,
"/Users/hannah/Documents/DEgenes_tables_updated_updated/HXCLCTD2pval.txt",
sep="\t")
HXCLCTD4pval <- subset(tableHXCLCTD4, tableHXCLCTD4$adj.P.Val<0.05)
write.table(HXCLCTD4pval,
"/Users/hannah/Documents/DEgenes_tables_updated_updated/HXCLCTD4pval.txt",
sep="\t")
HXCLCTD7pval <- subset(tableHXCLCTD7, tableHXCLCTD7$adj.P.Val<0.05)
```

```
write.table(HXCLCTD7pval,
"/Users/hannah/Documents/DEgenes_tables_updated_updated/HXCLCTD7pval.txt",
sep="\t")
HXCLCTD14pval <- subset(tableHXCLCTD14, tableHXCLCTD14$adj.P.Val<0.05)
write.table(HXCLCTD14pval,
"/Users/hannah/Documents/DEgenes_tables_updated_updated/HXCLCTD14pval.txt"
, sep="\t")

#making a column with desired data in it

OVCLCBD1pval$Day <- OVCLCBD1pval$LogFC
OVCLCBD2pval$Day <- OVCLCBD2pval$LogFC
OVCLCBD4pval$Day <- OVCLCBD4pval$LogFC
OVCLCBD7pval$Day <- OVCLCBD7pval$LogFC
OVCLCBD14pval$Day <- OVCLCBD14pval$LogFC

OVCLCTD1pval$Day <- OVCLCTD1pval$LogFC
OVCLCTD2pval$Day <- OVCLCTD2pval$aLogFC
OVCLCTD4pval$Day <- OVCLCTD4pval$LogFC
OVCLCTD7pval$Day <- OVCLCTD7pval$LogFC
OVCLCTD14pval$Day <- OVCLCTD14pval$LogFC

HXCLCBD1pval$Day <- HXCLCBD1pval$LogFC
HXCLCBD2pval$Day <- HXCLCBD2pval$LogFC
HXCLCBD4pval$Day <- HXCLCBD4pval$LogFC
HXCLCBD7pval$Day <- HXCLCBD7pval$LogFC
HXCLCBD14pval$Day <- HXCLCBD14pval$LogFC

HXCLCTD1pval$Day <- HXCLCTD1pval$LogFC
HXCLCTD2pval$Day <- HXCLCTD2pval$LogFC
HXCLCTD4pval$Day <- HXCLCTD4pval$LogFC
HXCLCTD7pval$Day <- HXCLCTD7pval$LogFC
HXCLCTD14pval$Day <- HXCLCTD14pval$LogFC

#Combining all days for each condition

OVCLCBall <-rbind(OVCLCBD1pval, OVCLCBD2pval, OVCLCBD4pval,
OVCLCBD7pval, OVCLCBD14pval)
OVCLCTall <- rbind(OVCLCTD1pval, OVCLCTD2pval, OVCLCTD4pval,
OVCLCTD7pval, OVCLCTD14pval)
HXCLCBall <- rbind(HXCLCBD1pval, HXCLCBD2pval, HXCLCBD4pval,
HXCLCBD7pval, HXCLCBD14pval)
HXCLCTall <- rbind(HXCLCTD1pval, HXCLCTD2pval, HXCLCTD4pval,
HXCLCTD7pval, HXCLCTD14pval)

#Making sure there are no duplicate genes

dim(OVCLCBall)
OVCLCBall <- unique(OVCLCBall)
dim(OVCLCBall)
```

```
dim(OVCLCTall)
OVCLCTall <- unique(OVCLCTall)
dim(OVCLCTall)

dim(HXCLCTall)
HXCLCTall <- unique(HXCLCTall)
dim(HXCLCTall)

#Outputting new tables

write.table(OVCLCBall,
"/Users/hannah/Documents/DEgenes_tables_updated_updated/OVCLCBall.txt",
sep="\t")
write.table(OVCLCTall,
"/Users/hannah/Documents/DEgenes_tables_updated_updated/OVCLCTall.txt",
sep="\t")
write.table(HXCLCBall,
"/Users/hannah/Documents/DEgenes_tables_updated_updated/HXCLCBall.txt",
sep="\t")
write.table(HXCLCTall,
"/Users/hannah/Documents/DEgenes_tables_updated_updated/HXCLCTall.txt",
sep="\t")
```

## APPENDIX B: change_ids.R

```
source("https://bioconductor.org/biocLite.R")
biocLite()
biocLite("hgu133plus2.db")
library(hgu133plus2.db)

#import clinical dataset

GSE9891 <-
as.matrix(read.table("/Users/hannah/Downloads/originaltagsmatrixclinical.txt", sep
="\t", header=T, row.names=1))
dim(GSE9891)

#Convert to expression set and pull down entrez ids into new column description

GSE9891 <- ExpressionSet(assayData = GSE9891)
gnamesthill <- mget(rownames(exprs(GSE9891)), hgu133plus2ENTREZID,
ifnotfound=NA)
w.thill <- as.data.frame(exprs(GSE9891))
w.thill$DESCRIPTION <- unlist(gnamesthill)

#get cluster gene names

hxct1_gnames <-
read.table(file="/Users/hannah/Documents/HXCT_cluster_527/hxct_clust1.csv",
header=T, sep=",")
hxct1_gnames_df <- as.data.frame(hxct1_gnames)
head(hxct1_gnames)

#Look for matching ids and extract data into new table

ilmentrexids <- read.table(file="/Users/hannah/Documents/ilumentrezids.txt",
header=T, sep="\t")
ilmentrexids_1 <-
ilmentrexids_1 <- ilmentrexids[ilmentrexids$Probe_Id %in%
hxct1_gnames$Probe_Id,]

w.thill_1 <- w.thill[as.numeric(w.thill$DESCRIPTION) %in%
ilmentrexids_1$Entrez_Gene_ID,]
no_nas <- w.thill_1[complete.cases(w.thill_1),]

write.table(no_nas, file="/Users/hannah/Documents/thilldataset/hxct1newids.txt",
sep="\t", quote=F)


# Same for OVC2
ovc2_gnames <-
read.table(file="/Users/hannah/Documents/OVC_clusters_527/ovc_clust2.csv",
header=T, sep=",")
```

```
ovc2_gnames_df <- as.data.frame(ovc2_gnames)
head(ovc2)


ilmentrexids <- read.table(file="/Users/hannah/Documents/ilumentrezids.txt",
header=T, sep="\t")
ilmentrexidsov2_1 <- ilmentrexids[ilmentrexids$Probe_Id %in%
ovc2_gnames$Probe_Id,]

w.thill_1ov2 <- w.thill[as.numeric(w.thill$DESCRIPTION) %in%
ilmentrexidsov2_1$Entrez_Gene_ID,]
no_nasov2 <- w.thill_1ov2[complete.cases(w.thill_1ov2),]

write.table(no_nasov2, file="/Users/hannah/Documents/thilldataset/ovc2newids.txt",
sep="\t", quote=F)


#Same for OVC3
ovc3_gnames <-
read.table(file="/Users/hannah/Documents/OVC_clusters_527/ovc_clust3.csv",
header=T, sep=",")
ovc3_gnames_df <- as.data.frame(ovc3_gnames)
head(ovc3_gnames)


ilmentrexids <- read.table(file="/Users/hannah/Documents/ilumentrezids.txt",
header=T, sep="\t")
ilmentrexidsov3_1 <- ilmentrexids[ilmentrexids$Probe_Id %in%
ovc3_gnames$Probe_Id,]

w.thill_1ov3 <- w.thill[as.numeric(w.thill$DESCRIPTION) %in%
ilmentrexidsov3_1$Entrez_Gene_ID,]
no_nasov3 <- w.thill_1ov3[complete.cases(w.thill_1ov3),]
head(w.thill_1ov3)
write.table(no_nasov3, file="/Users/hannah/Documents/thilldataset/ovc3newids.txt",
sep="\t", quote=F)
```

## APPENDIX C: clinical_clustering.R

```
# Import Biobase
library(Biobase)
# Import data as a matrix with the first column being the rownames and the data being
read ina s text, not factors
hxct1 <-
as.matrix(read.table("/Users/hannah/Documents/thilldataset/hxct1newids.txt", sep
="\t", header=T, row.names=1))

class(hxct1)
dim(hxct1)
colnames(hxct1)
head(hxct1[,1:5])

hxct1 <- hxct1[,1:285]
hxct1exprSet <- ExpressionSet(assayData = hxct1)
class(hxct1exprSet)

#create distance

clineuclidhxct1 <- dist(t(exprs(hxct1exprSet)))

pdf('euclliddisthxct1.pdf')
image(as.matrix(clineuclidhxct1))
dev.off()

# hierarchically cluster

clusthxct1 <- hclust(clineuclidhxct1)

# Create dendrogram

pdf('euclidclusthxct1.pdf')
plot(clusthxct1)
dev.off()

#Cut tree into two at top level

cut2hx1 <- cutree(clusthxct1, (k=2))
cut2hx1

#Same ovc2
ovc2 <- as.matrix(read.table("/Users/hannah/Documents/thilldataset/ovc2newids.txt",
sep ="\t", header=T))

class(ovc2)
dim(ovc2)
colnames(ovc2)
head(ovc2[,1:5])
```

```
ovc2 <- ovc2[,1:285]
ovc2exprSet <- ExpressionSet(assayData = ovc2)
class(ovc2exprSet)


clineuclidovc2 <- dist(t(exprs(ovc2exprSet)))

pdf('eucliddistovc2.pdf')
image(as.matrix(clineuclidovc2))
dev.off()

clustovc2 <- hclust(clineuclidovc2)

pdf('euclidclustovc2.pdf')
plot(clustovc2)
dev.off()

cut2ovc2 <- cutree(clustovc2, (h=2))
cut2ovc2


#Same ovc3

ovc3 <- as.matrix(read.table("/Users/hannah/Documents/thilldataset/ovc3newids.txt",
sep ="\t", header=T))

class(ovc3)
dim(ovc3)
colnames(ovc3)
head(ovc3[,1:5])

ovc3 <- ovc3[,1:285]


clineuclidovc3 <- dist(ovc3)

pdf('eucliddistovc3.pdf')
image(as.matrix(clineuclidovc3))
dev.off()

clustovc3 <- hclust(clineuclidovc3)

pdf('euclidclustovc3.pdf')
plot(clustovc3)
dev.off()

cut2ovc3 <- cutree(clustovc3, (h=2))
cut2ovc3
```

## APPENDIX D: Kaplan_meier.R

```
source("https://bioconductor.org/biocLite.R")
biocLite()
biocLite("survival")
library(survival)
survival <- as.data.frame(read.table("/Users/hannah/Downloads/phenodata_T285.txt",
sep="\t", header=T, row.names=1))
survival$status <- survival$PATIENT_STATUS

survival$status <- gsub("D", as.numeric(1), survival$status, ignore.case = FALSE,
perl = FALSE, fixed = FALSE, useBytes = FALSE)
survival$status <- gsub("PF", as.numeric(0), survival$status, ignore.case=FALSE,
perl=FALSE, fixed=FALSE, useBytes=FALSE)
survival$status <- gsub("R", as.numeric(0), survival$status, ignore.case=FALSE,
perl=FALSE, fixed=FALSE, useBytes=FALSE)
head(survival)
survival$status
survival$cuthx1 <- cut2hx1
survival$cutovc2 <-cut2ovc2
survival$cutovc3 <-cut2ovc3
head(survival)
survival$cuthx1
survival$cutovc2

#T2R
mfit.cuthx1rel <- survfit(Surv(survival$T2R, status == 1)~survival$cuthx1, data =
survival)
pdf("/Users/hannah/Documents/kaplanmeier/km_cuthx1T2R.pdf", width=5,
height=5)
plot(mfit.cuthx1rel, col = c("blue","dark green"), main=list("Tothill285 Time to
Relapse - clustered on hxct1", cex=0.8), xlab="Time (months)", ylab="Survival p")
dev.off()

survdiff(Surv(survival$T2R, status == 1) ~ survival$cuthx1, data = survival)

# T2D /OVERALL SURVIVAL
mfit.cuthx1death <- survfit(Surv(survival$T2D, status == 1)~survival$cuthx1, data =
survival)
pdf("/Users/hannah/Documents/kaplanmeier/km_cuthx1T2D.pdf", width=5,
height=5)
plot(mfit.cuthx1death, col = c("blue","dark green"), main=list("Tothill285 Time to
Death - clustered on hxct1", cex=0.8), xlab="Time (months)", ylab="Survival p")
dev.off()

survdiff(Surv(survival$T2D, status == 1) ~ survival$cuthx1, data = survival)


#T2R
```

```
mfit.cutovc2rel <- survfit(Surv(survival$T2R, status == 1)~survival$cutovc2, data =
survival)
pdf("/Users/hannah/Documents/kaplanmeier/km_cutov1T2R.pdf", width=5,
height=5)
plot(mfit.cutovc2rel, col = c("blue","dark green"), main=list("Tothill285 Time to
Relapse - clustered on ovc2", cex=0.8), xlab="Time (months)", ylab="Survival p")
dev.off()

survdiff(Surv(survival$T2R, status == 1) ~ survival$cutovc2, data = survival)

# T2D /OVERALL SURVIVAL
mfit.cutovc2death <- survfit(Surv(survival$T2D, status == 1)~survival$cutovc2, data
= survival)
pdf("/Users/hannah/Documents/kaplanmeier/km_cutov1T2D.pdf", width=5,
height=5)
plot(mfit.cutovc2death, col = c("blue","dark green"), main=list("Tothill285 Time to
Death - clustered on ovc2", cex=0.8), xlab="Time (months)", ylab="Survival p")
dev.off()

survdiff(Surv(survival$T2D, status == 1) ~ survival$cutovc2, data = survival)

#T2R
mfit.cutovc3rel <- survfit(Surv(survival$T2R, status == 1)~survival$cutovc3, data =
survival)
pdf("/Users/hannah/Documents/kaplanmeier/km_cutov31T2R.pdf", width=5,
height=5)
plot(mfit.cutovc3rel, col = c("blue","dark green"), main=list("Tothill285 Time to
Relapse - clustered on ovc3", cex=0.8), xlab="Time (months)", ylab="Survival p")
dev.off()

survdiff(Surv(survival$T2R, status == 1) ~ survival$cutovc3, data = survival)

# T2D /OVERALL SURVIVAL
mfit.cutovc3death <- survfit(Surv(survival$T2D, status == 1)~survival$cutovc3, data
= survival)
pdf("/Users/hannah/Documents/kaplanmeier/km_cutov31T2D.pdf", width=5,
height=5)
plot(mfit.cutovc3death, col = c("blue","dark green"), main=list("Tothill285 Time to
Death - clustered on ovc3", cex=0.8), xlab="Time (months)", ylab="Survival p")
dev.off()

survdiff(Surv(survival$T2D, status == 1) ~ survival$cutovc3, data = survival)
```