6. Statistics of Sparsely Sampled Curves

Marit Holden and Lars Holden

Abstract We develop new statistical methods for analyzing sparsely sampled curves that vary in time. The typical dataset is differences in log gene expressions from case-control pairs for a large number of genes sampled relative to time of diagnosis. We focus on weak signals in the gene expression in many genes instead of strong signals in a few genes. The methods are based on moving windows in time, hypothesis testing, dimension reductions and randomization of the time to observation.

Keywords Sparsely samples curves | case control differences | time to diagnosis | hypothesis testing | randomization

INTRODUCTION

In this chapter, we describe statistical methods for analyzing the data described in the previous paragraphs. The methods may, however, be applied more generally. We only need observations on irregular points in time from one or several strata from many different response variables, and we test whether the response variables are stationary and whether there are differences between the strata. The approaches are based on using a moving window in time and hypothesis testing.

The lack of repeated samples, in addition to irregularly sampled data and the quite small number of case-control pairs (around 400–500 or less), is a challenge, especially in analyses stratified on important clinical parameters. The aim of some of the analyses is therefore to show that there are changes over time or between strata, but without identifying single differentially expressed genes or gene sets. Normally, when identifying such genes or gene sets, we test multiple hypotheses and consequently need to adjust for this, for example by controlling the false discovery rate (FDR). For some of our datasets, we would then find no or very few genes. We therefore need approaches with fewer tests that include many genes in
order to avoid the problem of low power, noisy data, and multiple testing. Such approaches can be viewed as an effective method for dimension reduction in studies of functional genomics.

In such approaches, we need to decide how many genes to include when testing different kinds of hypotheses. Assume the genes have been ranked so that those that are varying most significantly have the highest rank. In (Holden 2015a, Holden 2015b), we showed that if there is a difference in average value of $X_{g,c}$ between the strata for some of the genes, but we do not know which genes, and the difference is normally distributed, then the statistical tests are strongest for a small rank. We concluded that if the distribution has heavier tails than the normal distribution, we should focus on the few genes with the strongest signal. On the other hand, if the distribution has a less heavy tail, for example a constant difference in the average value, then the statistical test is strongest for a larger rank, often larger than (closer to) the number of genes with a difference in average value between the strata.

This chapter discusses the analysis of functions $f_{a,g}(t)$ for different strata $a, b, ...$ and many genes $g$ where the function varies with time $t$. These functions may be denoted as trajectories. The challenge in our case is that the functions are sparsely sampled, the different strata are sampled at different points in time, and there is considerable noise in the sampling. We wish to identify the time dependency in the data and the differences between the strata. We expect to find small differences in many genes instead of larger differences in one or a few genes. This chapter summarizes some of the applied work.

In order to study how gene expression profiles vary with time in the years before or after a cancer diagnosis, several datasets showing gene expression in blood have been collected. Each such dataset consists of cases diagnosed with cancer and healthy controls. Each case and control belong to one case-control pair. The case and control of a case-control pair each gave one blood sample, and they are matched by time of blood sampling and year of birth. In the statistical analyses, we use the differences of the log2 gene expression levels, $X_{g,c}$, for each case-control pair $c$ and gene $g$. Here, $c = 1, ..., M$, where $M$ is the number of case-control pairs, and $g = 1, ..., N_g$, where $N_g$ is the number of genes. As the time intervals between blood sampling and cancer diagnosis vary from case to case, the case-control pairs provide information on the sparsely sampled curves describing gene expression over time some years before or after diagnosis.

Let $t_c$ be the time interval between blood sampling and cancer diagnosis for case-control pair $c$, where $t_1 \leq t_2 \leq ... \leq t_M$. We assume the log2 gene expressions $X_{g,c}$ follow a smooth function in time $f_{s(c),g}(t_c) = E[X_{g,c}]$, where $s(c)$ is the stratum
of case $c$. We estimate the function $f_{s(c),g}(t)$ by taking an average of the observations $X_{g,c}$ from stratum $s(c)$ in an interval in time. The variance of $X_{g,c}$ is estimated from the variance of the observations in the same interval.

**METHODS FOR ANALYZING COMPLEX CURVES**

**Moving window in time**

A central part of the statistical methodology is to examine how gene expression varies with time. By dividing the entire time period into shorter time periods and computing different kind of statistics for each such time period, we simplify the problem to examining how these statistics vary with time. We use overlapping time periods by using a moving window in time. The statistics computed for a time period are independent of time. As the distribution of the gene expressions may vary with time, the lengths of the time periods should be chosen such that we obtain as short time periods as possible. However, to obtain as good estimates as possible for each time period, there should be as many case-control pairs as possible within each time period. There is a trade-off between these two wishes. To conclude, we define $M - L + 1$ time periods $[t_1, t_L], [t_2, t_{L+1}], \ldots, [t_{M-L+1}, t_M]$ where $L$ is chosen such that we obtain short time periods with many case-control pairs. Typically, we let $L = M/4$.

Compared to an approach where the time periods are not overlapping, an advantage with the moving window approach is that we are better able to identify the points in time relative to diagnosis where changes in gene expression occur.

**Randomization for estimating null distributions and p-values**

In most hypothesis tests, we compute p-values by estimating the null distribution for the statistic of the hypothesis test by randomizing the data, i.e. interchanging covariates (time to diagnosis, case/control, etc.) between the patients. In the randomization we preserve critical properties of the genes (level of expression, complex correlation between genes, etc.) and randomize only what is connected to the changes in time, stratum or case/control status. This randomization defines the null distribution for the test statistic that is used when finding the p-value.

We can randomize the data either by randomizing the case and control in each case-control pair, by randomizing the case-control pairs between the periods, or by randomizing between the two strata within the time period. Note that all three randomization strategies maintain the correlation structure between the genes for
each case-control pair. Also note that each randomization of the data leads to a different ordering of the genes if the genes are ordered according to the statistic of the hypothesis test.

The p-value of the test is set to \( \frac{K + 1}{N_S + 1} \), where \( N_S \) is the total number of randomizations and \( K \) is the number of randomizations out of \( N_S \) with a more extreme statistic than the statistic for the real data. When we test one hypothesis for each gene, we take multiple testing into account by using the Benjamin-Hochberg procedure for controlling the false discovery rate, FDR (Reiner et al. 2003).

Extracting information from a time period

When extracting information for a time period, we use the gene expression data in that time period and ignore information about time when we compute different kinds of statistics, identify differentially expressed genes etc. We have used the following statistics for a time period:

**Ordered standard deviation, mean and weight for a time period**

Let \( m_{p,g} \) be the sample mean and \( s_{p,g} \) be the sample standard deviations for the differences in \( \log_2 \) gene expressions for gene \( g \) in time period \( p \). The variable \( m_{p,g} \) is an approximation to \( f_{s_{p,g}}(t_c) = E\{X_{g,c}\} \) and \( s_{p,g} \) an approximation to the standard deviation of \( X_{g,c} \) in the interval. Let \( m_{p,g,a} \) be the sample mean and \( s_{p,g,a} \) (\( s_{p,g,b} \)) be the sample standard deviations for the differences in \( \log_2 \) gene expression for gene \( g \) in time period \( p \) for stratum \( a \) (\( b \)). We define the statistics \( s_{p,(g)}, m_{p,(g)}, \) and \( w_{p,(g)} \) from these sample means and standard deviations as follows:

- \( s_{p,(g)} = s_{p,g} \), where \( s_{p,g} \) has rank \( g \) when the \( s_{p,g} \)'s for period \( p \) are sorted in increasing order. Rank 1 corresponds to the smallest of the \( s_{p,g} \)'s for period \( p \).
- \( m_{p,(g)} = |m_{p,g}| \), where \( |m_{p,g}| \) has rank \( g \) when the \( |m_{p,g}| \)'s for period \( p \) are sorted in decreasing order. Rank 1 corresponds to the largest of the \( |m_{p,g}| \)'s for period \( p \).
- Let \( w_p = \frac{m_{p,g,a} - m_{p,g,b}}{\sqrt{s_{p,g,a}^2 + s_{p,g,b}^2}} \) be the weight for gene \( g \) in time period \( p \), i.e. a measure of the difference between the two strata. \( w_{p,(g)} = |w_p| \) where \( |w_p| \) has rank \( g \) when the \( |w_p| \)'s for period \( p \) are sorted in decreasing order. Rank 1 corresponds to the largest of the \( |w_p| \)'s for period \( p \).

In some approaches, we use the sample mean \( m_{p,g} \) directly, without ranking.
The number of differentially expressed genes for a time period

We identify the significantly differentially expressed genes in the time period using the Bioconductor R-package limma, linear models for microarrays (Ritchie et al. 2015), where the response is $X_{g,c}$ i.e. the difference in log2 gene expression between the case and the control of a case-control pair.

In the next section, we describe how these statistics can be used for examining how gene expression varies over time, between strata, and between cases and controls.

Using information extracted for time periods

Finding signal in the data

The objective is to be able to identify small changes that vary slowly in time and/or between strata, by using a large number of genes in each hypothesis test and predictor.

For examining whether there are differences between cases and controls, between strata or in time, we test following hypotheses using the three statistics’ standard deviation, mean and weight defined in the previous section:

**H0-case-ctrl:** The expectation of $X_{g,c}$ is zero. This means that there is no difference between the expectations of the log2 gene expression values for the cases and controls. It implies that $f_{S(c),g}(t_c) = 0$. If the null hypothesis is false, the expectation will be different from zero for some periods and genes. We test the hypothesis by using the statistic $m_{p,(g)}$.

**H0-time:** The distribution of $X_{g,c}$ is not associated with the time to diagnosis. This means that the expectation and standard deviation of $X_{g,c}$ are the same in all time periods. It implies that $f_{S(t),g}(t_c)$ is constant in time. If the null hypothesis is false, the standard deviation for some periods will be lower than the standard deviations for the entire time period for some genes. Also, the absolute value of the expectation for some periods will be higher than the absolute value of the expectation for the entire time period for some genes. We test the hypothesis first by using the statistic $s_{p,(g)}$ and then by using the statistic $m_{p,(g)}$.

**H0-node:** The expectation of $X_{g,c}$ is not associated with stratum. This means that the expectations for the two strata are equal for all genes $g$ and time to diagnosis $t$. It implies that $f_{S(c),g}(t_c)$ is the same for all strata. If the null hypothesis is false, the difference in expectation will be different from zero for some periods and genes. We test the hypothesis by using the statistic $w_{p,(g)}$. 
The null distribution of each statistic will be estimated by randomizing the data, and we compute p-values by comparing the statistic for the data to the estimated null distribution.

We will reject the H0-case-ctrl hypothesis if the hypothesis $m_{p,g}>0$ is rejected for at least one time period $p$ and rank $g$, where $g$ belongs to a subset of the $N_g$ ranks. In practice, we have chosen to let the subset of ranks consist of ranks between approximately 1% and 25% of the number of genes, so that the subset contains both relatively low and high ranks. This means that H0-case-ctrl is rejected based on a very large number of hypotheses, that are also highly positively correlated, and we therefore needed to adjust for multiple testing. The approaches for rejecting the H0-time and H0-node hypotheses are similar. Besides rejecting the three null hypotheses, the hypothesis tests for the statistics for each time period and rank can be used for illustrating how the p-values are associated with the time to diagnosis.

Changes in the number of differentially expressed genes over time

In this approach, we examine changes in gene expression over time by examining how the number of differentially expressed genes between cases and controls varies with time. The time curve in this case consists of the number of differentially expressed genes in each time period. Such time curves give an indication of when there is a large difference between cases and controls before or after diagnosis. For comparing different strata, we can compare the time curves of the strata.

When testing whether the number of differentially expressed genes are different for two strata, we use $n_s$, the number of genes that are differentially expressed between cases and controls in at least one time period, as test statistic for stratum $s$.

When comparing stratum $a$ and stratum $b$, we cannot directly compare $n_a$ and $n_b$ if $M_a$, the number of case-control pairs in stratum $a$, is much larger than $M_b$, the number of case-control pairs in stratum $b$. To use the same number of case-controls pairs when comparing the strata, we test the following hypothesis.

$H_{0-strata1}$: The number of differentially expressed genes between cases and controls is different for stratum $a$ and $b$. Assume $M_b < M_a$. We want to estimate the null distribution for $n_a$ when the sample size of stratum $a$ is $M_b$, and then compare this distribution to $n_b$. The null distribution is found using simulation by repeatedly sample $M_b$ case-control pairs from stratum $a$, and then compute the number of differentially expressed genes for each sampled dataset. The p-value of the hypothesis test is computed by comparing the samples of the null distribution to $n_b$. 
Identifying significant genes based on area between curves

In this section we also estimate the function \( f_{s(c)g}(t_c) = E[X_{g,c}] \) by taking the mean of observations of \( X_{g,c} \) in an interval of time. However, here we analyze the properties of the function in the entire time period at the same time instead of for each interval in time separately. Then we are able to analyze all the data from a gene instead of focusing on properties in an interval. This may be used to identify genes that are significantly different between strata and predict strata from the gene observations.

We estimate the area between two curves \( f_{a,g}(t) \) and \( f_{b,g}(t) \) by the test statistics \( V_g = |f_{a,g}(t) - f_{b,g}(t)| = \int |f_{a,g}(t) - f_{b,g}(t)| \, dt \) where the two strata are denoted \( a \) and \( b \), respectively. This is equal to the weighted sum of the absolute value of the differences in average gene expression between the two strata in each time interval, where the weight depends on the length of the time interval. The area is large if there is a large difference between the curves and we neglect whether this is due to different average value or one is increasing and the other is decreasing in time.

We test the following hypothesis:

**H0-strata2**: The functions \( f_{a,g}(t) \) and \( f_{b,g}(t) \) are equal. For each gene \( g \), we compare the observed \( V_g \) with the same variable from a simulated distribution where we resample the variables \( X_{g,c} \) for all the genes simultaneously by randomizing the stratum \( s(c) \) for the case-control pairs. We maintain the observations for each gene and the number of observations from each stratum. We have made \( N_g \) simultaneous tests and need to use the methods for adjusting for multiple testing.

Prediction of stratum based on local statistics

The weights \( w_{p,(g)} \) can be estimated for each rank \( g \) from data in period \( p \) for a training dataset. The stratum of the case of a new case-control pair, i.e. a case-control pair that does not belong to the training set, can then be predicted based on the score

\[
z = \sum_{s=1}^{n} \delta_{p,(s)} |w_{p,(s)}| x_{(g)},
\]

where \( x_{(g)} \) is the difference in \( \log_2 \) gene expressions of the new case-control pair and \( \delta_{p,(g)} \) is 1 if the weight \( w_{p,g} \) is positive, and –1 otherwise, where \( |w_{p,g}| = w_{p,(g)} \). The \( n \) genes with highest absolute value of the weights are used for computing the score, where \( n \) is a number less than or equal to the number of genes, \( N_g \). Large values of \( z \) indicate that the new case belongs to stratum \( a \). If \( z > c \), for some arbi-
trary threshold $c$ we conclude that the new case belongs to stratum $a$, otherwise we conclude that the new case belongs to stratum $b$. We may set $c = 0$ if it is not more important to avoid false classification in one stratum relative to the other and if

$$\sum_{g=1}^{\delta} w_{p(g)} m_{p(g),a} + m_{p(g),b} \approx 0,$$

where $m_{p(g),a}$ and $m_{p(g),b}$ are the sample means that are used when computing $w_{p(g)}$. Increasing (decreasing) $c$ results in fewer false positives (negatives) at the cost of more false negatives (positives).

The available datasets are too small to be divided into a training and validation set. When predicting the stratum of the cases in a dataset, we should therefore use a leave-one-out or $k$-fold cross validation approach. When using the leave-one-out approach, we predict the stratum of case $j$ using weights $w_{p(g)}$ that have been estimated using the dataset where case-control pair $j$ has been excluded. The $k$-fold cross validation approach is similar, except that we divide the dataset into $k$ folds and predict the stratum of the cases in fold $f$ using weights $w_{p(g)}$ that have been estimated using the dataset where the case-control pairs in fold $f$ have been excluded.

**EXAMPLES OF USE OF THE METHODS**

In this section, we give examples from papers where the methods described above have been used.

**Finding signal in the data**

In a previous methodological study, time was categorized in three non-overlapping periods (Lund et al. 2016); see Chapter 8. The aim in that study was to show that there is signal in the data, but without showing where in time the changes in gene expression occurred or which genes were involved. The main idea used in the paper is that genes can be grouped into curve groups, each curve group corresponding to genes with a similar development over time (Figure 6.1). Based on these curve groups, we tested a set of hypotheses that determined whether there is development in gene expression levels over time, and whether this development varies among different strata. For a breast cancer dataset in the Norwegian Women and Cancer (NOWAC) postgenome cohort, the curve group analysis revealed that development of gene expression levels varied in the last years before breast cancer diagnosis, and that this development differed by lymph node status and participa-
tion in the Norwegian Breast Cancer Screening Program. The effect of the participation may be due to different treatment for the participating women representing the majority of the population.

Figure 6.1. Examples of curve groups according to time to diagnosis. Example of two different curve groups: curve group ‘123’ (upper panel, gene expression values increasing with time) and curve group ‘132’ (lower panel, highest gene expression value in the middle time period). In the left panels curves with the gene expression differences for 20 genes from the given curve group are plotted. For illustrational purposes, the curves have been estimated from the data using splines. In the middle panels the data for one of the 20 genes are shown with the corresponding spline-estimated curve. The points represent the differences in gene expression for each case-control pair. The mean value in each or the three time periods is shown in red. The right panels are similar to the middle panels except that the data points that are plotted are the mean values computed over the 20 genes in the left panel.

Partly to be able to better identify the points in time relative to cancer diagnosis where changes in gene expression occur, we developed methods based on a moving window in time. This approach includes time in a more continuous manner than the approach based on curve groups. In Holden et al. 2017 we used moving windows and randomization for the same dataset as we used in Lund et al. 2016, Chapter 8. The null hypotheses of no differences between cases and controls, no time-dependent changes, and no differences between different strata were all rejected. The main conclusion of the analyses was that there are time-dependent changes of the blood transcriptome up to eight years before breast cancer diagnosis.
Changes in the number of differentially expressed genes over time

In Chapter 9 we examined the changes in gene expression after diagnosis for a breast cancer dataset in the Norwegian Women and Cancer (NOWAC) post-genome cohort. We stratified stage in invasive or metastatic breast cancer, and vital status in dead or alive at the end of follow-up, and observed a significant increase of differentially expressed genes among women with metastatic disease who later died both compared to invasive cases that survived ($p = 0.001$) and to metastatic cases that survived ($p = 0.024$). To illustrate the difference between strata, we made heatmaps for the most differentially expressed genes over time; see Figure 6.2 below.

We also observed a second transient increase in blood gene expression a few years after diagnosis in metastatic cases, hypothetically representing a capitulation of the immune system.

Identifying significant genes based on area between curves

The analysis of complex curves was extended in order to identify genes that are significantly different between two strata. The method tests differences in a non-parametric time development relative to time of diagnosis of the gene expressions from different strata using the area between the curves in a long time period.

The method was tested on case-control differences in log2 gene expressions in a post-diagnostic time period separating between the women who survived and the women who died of breast cancer. The method clearly showed non-linear changes, with rapid transient mostly increasing fold changes, in cases who later died. Survivors had no changes. For cases that died, this transient increase was followed by a regression towards the gene expression profiles of survivors. For 9786 genes, the integrated area from 18 months to 8 years was highly significant ($p < 0.00001$) among women who died. There were indications of a stronger relationship in metastatic cases alone.

Figure 6.3 below shows the $f_{s(c)}g(t_c)$ curves for the 100 most significant genes in the period after diagnosis for the women that survived and died of breast cancer.
Figure 6.2. Heat maps with 74 selected genes in each stratum; invasive versus metastatic and alive or dead at end of follow-up. The heat maps show log fold change for each gene (y-axis) for each quarter of the years after diagnosis (x-axis).
Figure 6.3. The 100 most significant genes in the period after diagnosis of breast cancer. The data is scaled such that variance in the data is 1 for all the genes. Hence, the vertical axis does not give information about the fold change.

This systems epidemiology approach provides a proof of concept for the use of gene expression as an individualized biomarker of prognosis related to death or not. Since we had no prior knowledge of the shape of differences in gene expressions as a function of time relative to diagnosis, we needed a non-parametric model that identified all possible changes in trajectories. The aim of the study was to explore single gene expression trajectories from immune cells in blood over the first years after diagnosis as predictors of later vital status, dead or alive.

Prediction of stratum based on local statistics
In Holden et al. 2017, we described Prediction of stratum based on local statistics, to illustrate how the predictive power of the test varies with time. In Figure 6.4 below, we observe that for screening detected cancers the probability of correct prediction of metastasis status was best in year 1 before diagnosis compared to year 3 and 4 before diagnosis for clinically detected cancers.
Figure 6.4. Prediction results. a) Correctly (green) or wrongly (red) classified cases plotted against time to diagnosis for the screening (upper panel) and the clinical group (lower panel). A circle is plotted above every fifth case. Long vertical lines are plotted to indicate the years. On the y-axis “with” means cases with metastases and “without” means cases without metastases. b) Fraction of correctly classified cases with (red) and without (black) metastases over time for the screening (upper panel) and the clinical group (lower panel). The fraction for each point in time is computed using a moving window of one year (clinical) or 100 days (screening). The resulting curve is then smoothed using a median filter with a window size of one year (clinical) or 100 days (screening).
Identifying significant genes with only one time period

The mortality of breast cancer is strongly associated with parity, i.e. the number of children to whom a woman has given birth (Lund 1990). In Lund 2018, we used one of the breast cancer datasets in the Norwegian Women and Cancer (NOWAC) postgenome cohort mentioned in the previous sections to examine whether there is an association between gene expression and parity. In that study we used only one time period, which means that we ignore information about time and as a consequence will mainly be able to identify genes with expressions that do not vary with time in the years before/after diagnosis. As there is a large body of evidence demonstrating the long-lasting protective effect of each full-term pregnancy (FTP) on the development of breast cancer (BC) later in life, this is reasonable.

We used the Bioconductor R-package Limma, linear models for microarrays (Ritchie et al. 2015), to identify the genes that were influenced by parity. In the linear model, the responses were $X_{g,c}$, the differences in the log2 gene expression for each case-control pair, while we included the parity of the control and the parity of the case as covariates. In the analyses, we merged parities 1–3 and 4–6 so that the parity data consisted of three different values: 0, 1–3, and 4–6. The merging was done in order to reduce the effect of the highest parities. We identified gene sets that were influenced by parity using Limma in the same way as we did for individual genes, by using enrichment scores for gene sets instead of differences in the log2 gene expressions as responses in the linear model. The enrichment scores for gene sets were obtained from the $X_{g,c}$s using the Bioconductor R-package gene set variation analysis, GSVA (Hänzelmann et al. 2013).

We found that 756 genes showed linear trends in cancer-free controls, false discovery rate (FDR) 5%, but this was not the case for any of the genes in the breast cancer cases. Gene Set Enrichment Analysis, GSEA of immunologic gene sets, C7 collection in Molecular Signatures Database, MSigDB (GSEA MSigDB, Subramanian et al. 2005) revealed 215 significantly enriched human gene sets (FDR 5%). These marked differences in gene expression and enrichment profiles of immunologic gene sets between breast cancer cases and healthy controls suggest an important protective effect of the immune system on breast cancer risk.

REFERENCES

GSEA MSigDB, Gene Set Enrichment Analysis (GSEA). Molecular Signatures Database (MSigDB). Internet. UC San Diego and BROAD Institute; Available from: http://software.broadinstitute.org/gsea/index.jsp
Holden L. Classify strata. Oslo: Norwegian Computing Center; SAMBA/11/15; 2015b. pp 28. Available from: https://www.nr.no/directdownload/1426685952/classify_strata_holden2015.pdf

Holden L. Time development of gene expression; Oslo: Norwegian Computing Center; SAMBA/35/15; 2015a. pp. 13. Available from: https://www.nr.no/files/samba/smbi/note2015SAMBA3515TimeDevelopmentGenes.pdf

Holden M, Holden L, Olsen KS, Lund E. Local in Time Statistics for detecting weak gene expression signals in blood – illustrated for prediction of metastases in breast cancer in the NOWAC Post-genome Cohort. Advances in Genomics and Genetics. 2017; 7: 11–28. Available from: https://www.dovepress.com/local-in-time-statistics-for-detecting-weak-gene-expression-signals-in-peer-reviewed-article-AGG

Hänzelmann S, Castelo R, Guinney J. GSEA: gene set variation analysis for microarray and RNA-Seq data. BMC Bioinformatics. 2013 Jan 16; 14: 7. Available from: https://bmcbioinformatics.biomedcentral.com/articles/10.1186/1471-2105-14-7

Lund E. Childbearing in marriage and mortality from breast cancer in Norway. Int J Epidemiol. 1990; 19(3): 527–531. Available from: https://academic.oup.com/ije/article-abstract/19/3/527/760486?redirectedFrom=fulltext

Lund E, Holden L, Bøvelstad H, Plancade S, Mode N, Günther C-C et al. A new statistical method for curve group analysis of longitudinal gene expression data illustrated for breast cancer in the NOWAC Post-genome Cohort as a proof of principle. BMC Med Res Methodol. 2016 Mar 5; 16: 28. Available from: https://bmcmedresmethodol.biomedcentral.com/articles/10.1186/s12874-016-0129-z

Lund E, Nakamura A, Snapkov I, Thalabard JC, Olsen KS, Holden L, et al. Each pregnancy linearly changes immune gene expression in the blood of healthy women compared with breast cancer patients. Clin Epidemiol. 2018 Aug 6; 10: 931–940. Available from: https://www.dovepress.com/each-pregnancy-linearly-changes-immune-gene-expression-in-the-blood-of-peer-reviewed-article-CLEP

Reiner A, Yekutieli D, Benjamini Y. Identifying differentially expressed genes using false discovery rate controlling procedures. Bioinformatics. 2003 Feb 12; 19(3): 368–375. Available from: https://academic.oup.com/bioinformatics/article/19/3/368/258230

Ritchie ME, Phipson B, Wu D, Hu Y, Law CW, Shi W et al. limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Res. 2015 Apr 20; 43(7): e47. Available from: https://academic.oup.com/nar/article/43/7/e47/2414268

Subramanian A, Tamayo P, Mootha VK, et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc Natl Acad Sci USA. 2005 Oct 25; 102(43): 15545–15550. Available from: https://www.pnas.org/content/102/43/15545.long