Temporal Analysis of Oncogenesis Using MicroRNA Expression Data

Thomas Zichner,[1,2] Zelmina Lubovac,[1] Björn Olsson[1]

[1]Bioinformatics Research Group, Systems Biology Centre, Department of Life Sciences, University of Skövde, Box 408, S-54128 Skövde, Sweden
[2]Institut für Informatik, Friedrich-Schiller-Universität Jena, Ernst-Abbe-Platz 2, D-07743 Jena, Germany
tzi@minet.uni-jena.de, zelmina.lubovac@his.se, bjorn.olsson@his.se

Abstract: MicroRNAs (miRNAs) have rapidly become the focus of many cancer research studies. These small non-coding RNAs have been shown to play important roles in the regulation of oncogenes and tumor suppressors. It has also been demonstrated that miRNA expression profiles differ significantly between normal and cancerous cells, which indicates the possibility of using miRNAs as markers for cancer diagnosis and prognosis. However, not much is known about the regulation of miRNA expression. One of the issues worth investigating is whether deregulations of miRNA expression in cancer cells occur according to some pattern or in a random order. We therefore selected two approaches, previously used to derive graph models of oncogenesis using chromosomal imbalance data, and adapted them to miRNA expression data. Applying the adapted algorithms to a breast cancer data set, we obtained results indicating the temporal order of miRNA deregulations during tumor development. When analyzing the specific deregulations appearing at different time points in the derived model, we found that several of the deregulations identified as early events could be supported through literature studies.

1 Introduction

One of the important issues that have dominated cancer research during the last decade has been to identify molecular biomarkers [Lu05], i.e., indicators of cancer staging and tumor subtypes. MicroRNAs (miRNAs) have been seen as potential biomarkers that not only may serve for diagnostic and prognostic purposes, but are also assumed to have a great therapeutic potential in cancer [Sa08].

MicroRNA expression profiles have been used in previous work to classify tumors and differentiate between normal and cancerous tissue [Io05][Lu05][Vo06]. To the best of our knowledge, however, there are no existing approaches that consider changes in miRNA expression patterns during cancer progression, in order to reveal the possible temporal ordering of aberrant miRNA expression. Therefore, we selected two existing methods for deriving graph models of oncogenesis, previously applied to comparative genomic hybridization data, and applied them to miRNA expression data. The purpose with the adapted methods is to derive models illustrating the temporal order of events

during cancer progression. A deeper understanding of miRNAs during tumor progression, including the temporal order of their deregulations, may lead to novel methods regarding the prediction of survival of cancer patients and the choice of treatment, as well as for cancer subtype prediction.

2 Method

2.1 Data set

To perform temporal analyses of miRNAs a data set generated in [Io05] was used. It contains the expression levels of 157 human miRNAs and 69 human precursor miRNAs in 109 breast cancer samples (primary tumor samples as well as human breast cancer cell lines) and in normal breast tissue. The normal samples consisted of six pools of five normal breast tissues each and four additional single breast tissues, of which we used only two because of an observed unreasonable deviation of the other two. Further details about the used data set can be obtained from [Io05]. The raw data can be obtained from ArrayExpress [Br03], which can be accessed via http://www.ebi.ac.uk/microarray-as/aer/. The ID of the used data set is E-TABM-23. For our analyses, we used the normalized data set kindly provided by Marilena V. Iorio.

2.2 Determining aberrant expression

The first step of the analysis was to determine the subset of miRNAs that are most likely to have aberrant expression in breast cancer compared to normal tissue. In all further analyses we focused only on this subset of miRNAs. To derive this set, the two-sample Kolmogorov-Smirnov test as well as the Wilcoxon rank sum test was applied to each miRNA's expression profile. For both tests, the null hypothesis is that the expression values of a certain miRNA are derived from the same distribution for the normal as well as the cancer samples. All microRNAs with $p < 0.05$ in both tests were considered as deregulated and included into the subset. The reason for choosing the Kolmogorov-Smirnov and Wilcoxon rank sum tests is that they do not make any assumptions concerning the value distribution.

The next step was to determine in which breast cancer tumor samples each miRNA was aberrantly expressed. The assumption is that expression values in tumor samples which differ by more than two standard deviations σ from the mean expression value μ in normal tissue can be considered as deregulated. This method is commonly used in microarray analysis to identify differentially expressed genes [CQB03][Ka06].

For the further work, we classified each miRNA in each tumor sample as under-, normal, or over-expressed, depending on whether the expression value was lower than μ-2σ, between $\mu - 2\delta$ and $\mu + 2\delta$, or greater than $\mu + 2\delta$.

2.3 Creating a set of events

The approach to determine aberrant expression described in section 2.2 results in a matrix D showing single deregulations, i.e., single cases of under- and over-expression. Assuming k as the number of deregulated miRNAs (i.e., the size of the miRNA subset) and m as the number of tumor samples, matrix D is defined by:

$$D = (d_{ij})_{\substack{1 \le i \le k \\ 1 \le j \le m}}$$

where d_{ij} = -1 if miRNA i is under-expressed in sample j, d_{ij} = 0 if it is normally expressed, and d_{ij} = 1 if it is over-expressed.

For further analyses, we distinguished between deregulations (i.e., the combination of a particular miRNA and the kind of deregulation) instead of just miRNAs, because a miRNA may be over-expressed in some tumor samples and under-expressed in some others. This distinction is necessary since there might be different causes for the different kinds of deregulations. Each pair of a miRNA and a type of deregulation is also referred to as an *event*. Thus, instead of k miRNAs, $2k$ events are considered.

Since it is quite difficult to make reliable assumptions about events which occur very rarely we restricted the further analyses to events that were present in at least 15% of the tumor samples. This resulted in an event matrix E:

$$E = (e_{ij})_{\substack{1 \le i \le n \\ 1 \le j \le m}}$$

where e_{ij} = 1 if event i is present in sample j, and e_{ij} = 0 otherwise. Furthermore, n denotes the number of events and $\sum_{j=1}^{m} e_{ij} \ge 0.15 \cdot m$ for all $1 \le i \le n$.

2.4 Temporal analysis and generating graph models of oncogenesis

Following the approaches by Höglund et al. [Hö01][Hö05] and Beerenwinkel et al. [Be05], the miRNA data set was analyzed and graph models representing the events during cancer progression generated. The following text explains the steps of the methods.

The approach for analyzing temporal relations described here is adapted from [Hö01]. Considering all tumor samples that show a certain deregulation, the average number of simultaneously observed events is calculated. If there are only a few such events, the considered deregulation is assumed to be an early event in the oncogenesis. Otherwise, i.e., if there are many simultaneously occurring events, it is assumed to be a late event. This reasoning is based on the well-established knowledge that tumors in late stages have usually accumulated a large number of mutations, leading to genomic instability [To02].

Instead of considering gains and losses of chromosomal parts, as in [Hö01], we considered the over- and under-expression of miRNA genes as events. First, for each

tumor sample, the number of miRNAs which are considered to have aberrant expression in the sample is calculated. In the following, this number is called NDPT - number of deregulations per tumor sample. Secondly, for each event, the NDPT is recorded for every tumor sample in which the event is present. By determining the median NDPT, the time of occurrence (TOC) of an event can be estimated.

To identify possible patterns in the data set, we performed a principal component analysis (PCA). PCA is a multivariate method frequently used to search for underlying structures in the data. It is a technique to reduce multidimensional data sets to lower dimensions. Briefly, principal components are linear combinations of the original variables, orthogonal, and ordered with respect to their variance so that the first principal component has the largest variance. The idea of applying PCA is to retain just those characteristics of a data set that contribute most to its variance. We performed the PCA with the deregulations as variables and the tumor samples as observations. To show the results we plotted all deregulations in relation to the first two principal components, which explain more than 40% of the total variance in the miRNA data set.

As the third analysis, we built oncogenetic tree mixture models, a graph-based representation of oncogenetic pathways, using the approach proposed by Beerenwinkel et al. [Be05a]. In addition to the temporal ordering, tree models also indicate possible alternative pathways of tumor development, characterized by different combinations (and/or orderings) of events. They thus have the potential to provide insights about tumor subtypes. Details about the approach by Beerenwinkel et al. can be found in [Be05a][Ra05]. It is a further development of an approach originally proposed by Desper et al. [De99], which is based on an algorithm for finding minimum weight branching trees. We used the software package mtreemix [Be05b] (http://mtreemix.bioinf.mpi-sb.mpg.de/) to learn the tree models with the event matrix E as input.

3 Results

3.1 Determining aberrant expression

The subset of miRNAs that are most likely to be deregulated in breast cancer was first selected, as explained in section 2.2. The two statistical tests, applied to find significant deregulations, identified 48 miRNAs that are significant according to both tests. This set of miRNAs was used in further analysis. The next step was to determine in which of the breast cancer tumor samples a certain miRNA is aberrantly expressed, as explained in section 2.2. The distribution of the number of deregulations per miRNA (NDPM) as well as the distribution of the number of deregulations per tumor sample (NDPT) is shown in Figure 1. A table showing the p-values and the number of tumor samples in which each miRNA is considered as over- or under-expressed is available from the authors.

Figure 1 and the table resulted in some observations regarding specific miRNAs. The miRNA mir-210 has the lowest p-value in both tests. The most down-regulated miRNA is mir-19a, while the most up-regulated miRNA is mir-21. It is also apparent that the variance in the number of deregulations per miRNA is large.

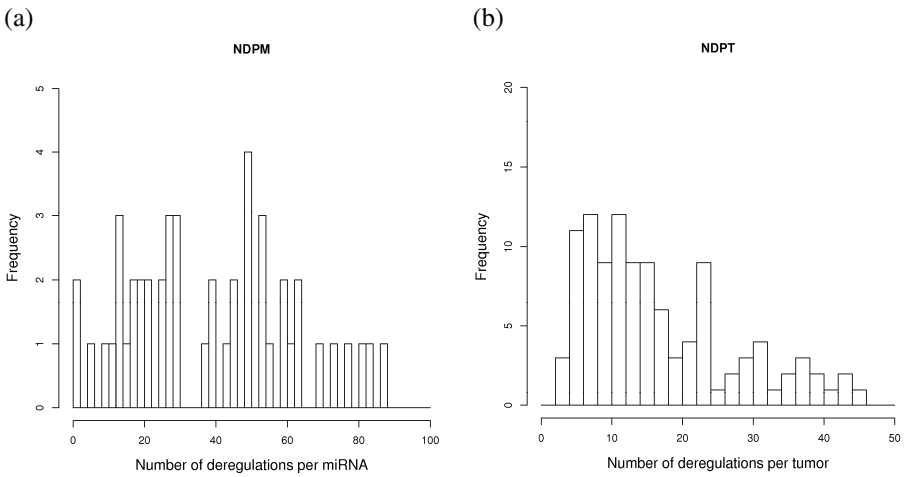

Figure 1: Distribution of the number of deregulations per miRNA (a) and the number of deregulations per tumor sample (b). No distinction was made between over- and under-expression.

In the following, we examined events, instead of considering only the miRNAs. As already described, an event is defined as a pair consisting of a specific miRNA and a specific kind of deregulation, i.e., either over- or under-expression. Events are here denoted by the miRNA name followed by a plus or minus sign. Thus, mir-125b-1− signifies the event that mir-125b-1 is under-expressed. Note that an event is not the same as an observation, since the same event (e.g., mir-125b-1−) may be observed in a large number of tumor samples. As we only considered events that were observed in at least 15% of the samples, the resulting set used in further analysis consisted of 36 events.

3.2 Temporal analysis

We here adapted the approach from [Hö01], with the aim to reveal the temporal ordering of miRNA deregulation events. As described in 2.4, the events are ordered according to the time of occurrence, estimated by the number of co-occurring events. The results, shown in Figure 2, indicate an evident temporal ordering of the events. On average, events like mir-125b-1− co-occur with significantly fewer events than, for instance, event mir-208−. Thus, it can be assumed that mir-125b-1− is an early event compared to mir-208−, which indicates that mir-125b-1 may play an important role in the onset of tumor development.

PCA was performed to determine the underlying structure behind the data. The aim is to calculate the largest principal components which can describe most of the overall variance shown in the data. The number of principal components equals the number of variables, i.e., in our case the number of events. The PCA resulted in 36 components, of which the first two explain about 45% of the total variance, and the first three components explain more than the half of the total variance. All considered events are plotted in relation to the first two components in Figure 3. It can be seen that most events

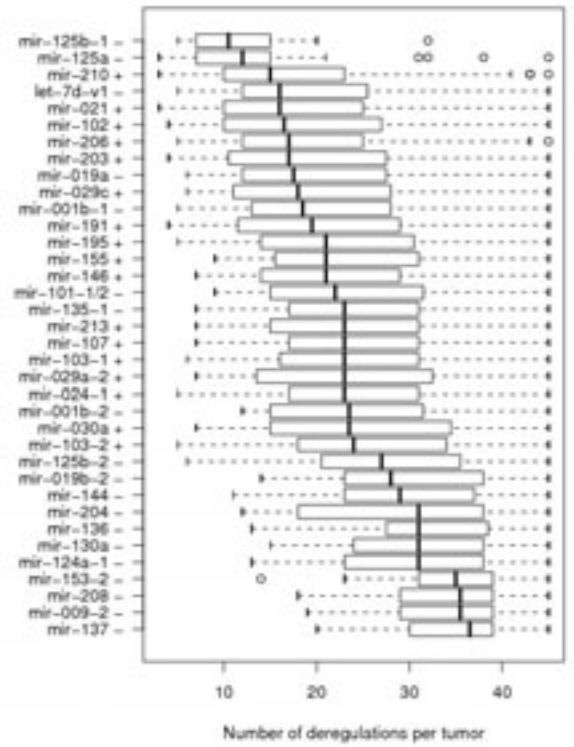

Figure 2: Estimated time of occurrence of the events. The sign '+' after a miRNA name indicates over-expression, while '−' indicates under-expression. Vertical bars indicate the median of the number of deregulations per tumor (NDPT) of the corresponding samples, boxes show the boundaries of the 25th and 75th percentiles, and the outer lines indicate the non-outlier minimum and maximum values.

differ only in relation to the second component, except for the events we assumed to occur as a cluster (see Figure 3). Further, there are also sets of events with almost identical values regarding the first two components.

3.4 Building oncogenetic tree mixture models

As a final step in the analysis, we built oncogenetic tree mixture models [Be05a] using the software package mtreemix developed by Beerenwinkel et al. [Be05b]. The first tree model that we derived shows all considered events. The second model shows a subset of events generated by a method proposed in [Br82]. These two models are not shown due to space limitations, but are available from the authors on request. The third model shows a subset of 15 events (Figure 4), which were selected because they were more separated in the PCA-plot (Figure 3) and because they were also identified in [Io05]. The number of trees K in the mixture model was set to two (i.e., one non-noise component) for the first two models. In the third model K was set to three according to the estimation of mtreemix select.

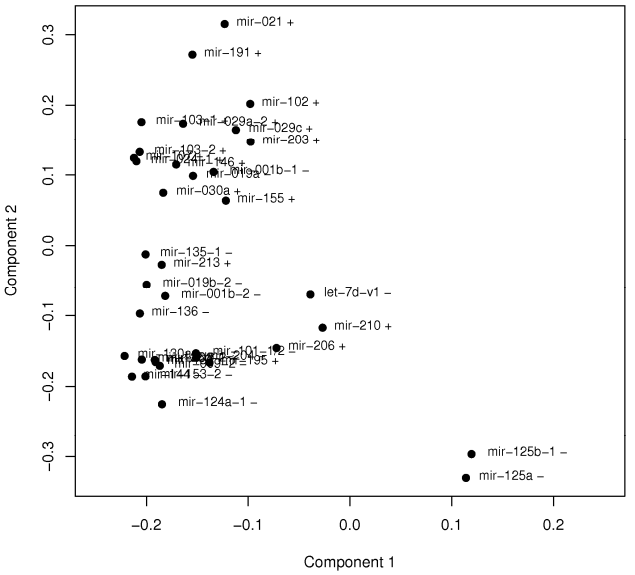

Figure 3: Considered events plotted in relation to the first two components of the PCA. An interesting observation is that in the co-occurrence statistics (data not shown) mir-206+, mir-210+, let-7d-v1−, mir-125a−, and mir-125b-1− are separated and form a cluster, i.e., they occur very often together, but rather seldom in the presence of other deregulations.

From the figures it can be seen that the mixture model that includes all events is not very stable (several edges were not present in any of the 1000 bootstrap iterations). Also, it may be observed that there are several edges which are present in all three or at least two models. For instance:

wild type → mir-021+ wild type → mir-102+ mir-125a− → mir-125b-1−

The edge "wild type → mir-021+" is present in all models. This applies also to the edge between mir-125a− and mir-125b-1−; however, in this case both directions are present.

3.5 Evaluation

We evaluated our results by applying the methods to randomly altered data sets. For each miRNA the order of expression values (tumor samples) was randomly altered. Figure 5 shows the results of temporal order analysis and PCA for the randomized data. In contrast to the original data set, it can be easily seen that there is no detectable temporal order of events (Figure 5a). There is only a difference of about 2 between the median numbers of deregulations per tumor (NDPT) of the "earliest" and the "latest" event, compared to a difference of about 27 when considering the original data set. Additionally, there are two large sets of events (14 and 15 items, respectively) which have the same median NDPT, and thus considered as simultaneously occurring, which also indicates the absence of a significant temporal order.

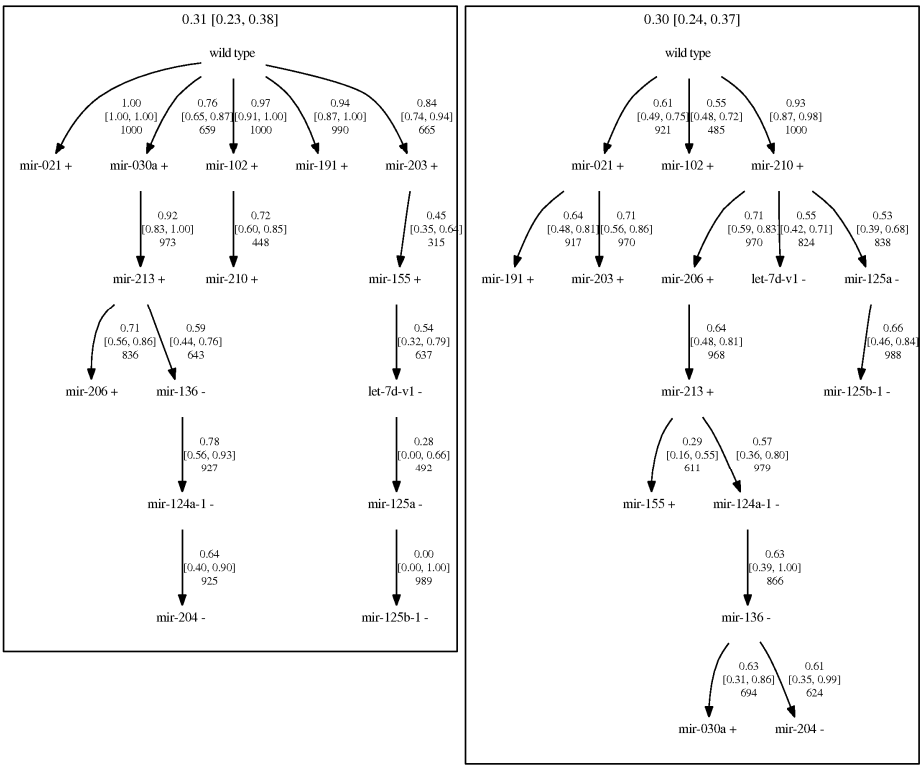

Figure 4: Oncogenetic tree mixture model of 15 events selected based on the PCA results. The star-shaped noise component is not shown. Edges are annotated with the transition probability (with confidence interval, CI) and the bootstrap samples count as a measure of stability. The weight of each mixture component (with CI) is given at the top of the box.

Also the PCA of the altered data shows much less underlying structure (Figure 5b). The events plotted in relation to the first two components are more uniformly distributed. The first three components explain only ~19% of the total variance, which is much less than in the original data set, where the first three components explained >50% of the variance.

4 Discussion

We performed several analyses to derive information about temporal and occurrence relations between miRNA deregulations. All analyses, especially the comparison to randomized data set, show that there is an underlying structure behind the used data set. This means that the observed events do not occur randomly.

There is an agreement between the derived temporal order (i.e., the results from the temporal analysis according to [Hö01][Hö05]) as well as the oncogenetic trees built according to [Be05a] and the principal component analysis. The order of the events in

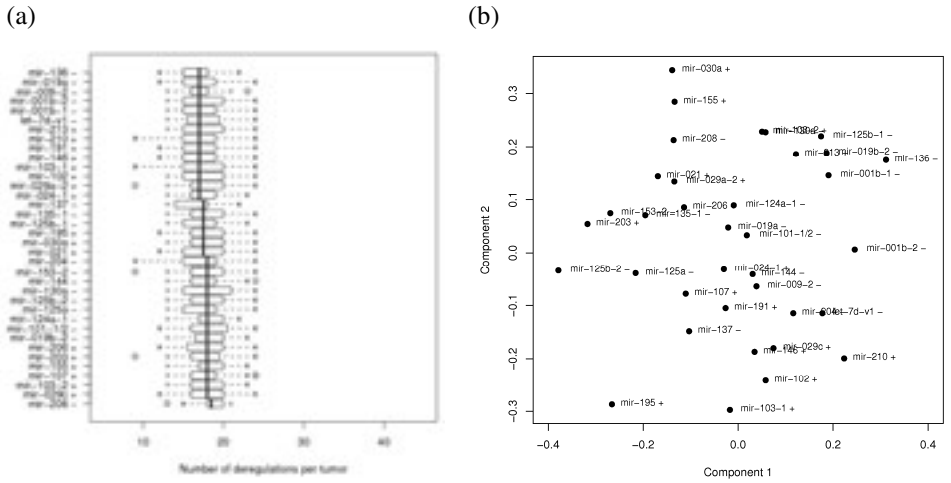

Figure 5: Results of the different analyses of the randomized data set.

relation to the second principal component correlates slightly with the observed time of occurrence. A high value of the second component indicates early occurrence, whereas a low value indicates a later occurring event. According to all three analyses, mir-21+ and mir-191+ are, among others, considered as the initial events in tumor progression. This assumption is also supported by literature [Io05][Vo06], because both miRNAs have been identified to be consistently over-expressed in cancer.

The events mir-206+, mir-210+, let-7d-v1−, mir-125a−, and mir-125b-1− are peculiar in some of the analyses (including the co-occurrence statistics, which are not shown). For instance, they are slightly separated in the PCA plots (Figure 3). These are among the very few events which show a variation in the first principal component compared to the other events. We assume that these events form a cluster, i.e., they occur very often together, but rather seldom in the presence of other deregulations. For all these miRNAs important roles in cancer have already been shown [Ad07][Ca08][Io05][Vo06].

There are also some disagreements in the results. For example, it is obvious that only small parts of the oncogenetic tree models agree with each other. Examples include edges like “wild type → mir-021+” or “wild type → mir-102+”. But many events occur in totally different places within the trees, i.e., in some as early and in some as late events, and always with different events as predecessor and successor. It is likely that the data set, although sufficiently large and informative to derive the temporal order, is not sufficiently large to derive accurate tree models. Since tree models branch out in different directions from the root node, the corresponding subsets of data become increasingly small, which rapidly leads to a lack of data for accurate modeling. Thus, it is to be expected that only the edges closest to the root will show consistency between different trees, unless a very large data set is used. The most important future work will therefore be to evaluate the tree models on larger data sets, as well as on different types of cancer.

References

[Ad07] Adams, B.D.; Furneaux, H.; White, B.A.: The micro-ribonucleic acid (miRNA) mir-206 targets the human estrogen receptor-alpha (eralpha) and represses eralpha messenger RNA and protein expression in breast cancer cell lines. Mol Endocrinol, 21(5):1132-1147, May 2007.

[Be05a] Beerenwinkel, N.; Rahnenführer, J.; Däumer, M.; Hoffmann, D.; Kaiser, R.; Selbig, J.; Lengauer, T.: Learning multiple evolutionary pathways from cross-sectional data. J Comput Biol 12(6):584-598, 2005.

[Be05b] Beerenwinkel, N.; Rahnenführer, J.; Kaiser, R.; Hoffmann, D.; Selbig, J.; Lengauer, T.: Mtreemix: a software package for learning and using mixture models of mutagenetic trees. Bioinformatics, 21(9):2106-2107, May 2005.

[Br82] Brodeur, G.M.; Tsiatis, A.A.; Williams, D.L.; Luthardt, F.W.; Green, A.A.: Statistical analysis of cytogenetic abnormalities in human cancer cells. Cancer Genet Cytogenet 7(2):137-152, 1982.

[Br03] Brazma, A.; Parkinson, H.; Sarkans, U.; Shojatalab, M.; Vilo, J.; Abeygunawardena, N.; Holloway, E.; Kapushesky, M.; Kemmeren, P.; Lara, G.G.; Oezcimen, A.; Rocca-Serra, P.; Sansone, S-A.: Arrayexpress-a public repository for microarray gene expression data at the EBI. Nucleic Acids Res 31(1):68-71, 2003.

[Ca08] Camps, C.; Buffa, F.M.; Colella, S.; Moore, J.; Sotiriou, C.; Sheldon, H.; Harris, A.L.; Gleadle, J.M.; Ragoussis, J.: hsa-mir-210 is induced by hypoxia and is an independent prognostic factor in breast cancer. Clin Cancer Res, 14(5):1340-1348, Mar 2008.

[CQB03] Causton, H.C.; Quackenbush, J.; Brazma, A.: Microarray Gene Expression Data Analysis: A Beginner's Guide. Wiley-Blackwell, 2003.

[De99] Desper, R.; Jiang, F.; Kallioniemi, O.P.; Moch, H.; Papadimitriou, C.H.; Schäffer, A.A.: Inferring tree models for oncogenesis from comparative genome hybridization data. J Comput Biol 6(1):37-51, 1999.

[Hö01] Höglund, M.; Gisselsson, D., Mandahl, N.; Johansson, B.; Mertens, F.; Mitelman, F.; Säll, T.: Multivariate analyses of genomic imbalances in solid tumors reveal distinct and converging pathways of karyotypic evolution. Genes Chromosomes Cancer 31(2):156-171, 2001.

[Hö05] Höglund, M.; Frigyesi, A.; Säll, T.; Gisselsson, D.; Mitelman, F.: Statistical behavior of complex cancer karyotypes. Genes Chromosomes Cancer 42(4):327-341, 2005.

[Io05] Iorio, M.V; Ferracin, M.; Liu, C-G.; Veronese, A.; Spizzo, R.; Sabbioni, S.; Magri, E.; Pedriali, M.; Fabbri, M.; Campiglio, M.; Ménard, S.; Palazzo, J.P.; Rosengerg, A.; Musiani, P.; Volinia, S.; Nenci, I.; Calin, G.A.; Querzoli, P.; Negrini, M.; Croce, C.M.: MicroRNA gene expression deregulation in human breast cancer. Cancer Research 65(16):7065-7070, 2005.

[Ka06] Kawada, J.I.; Kimura, H.; Kamachi, Y.; Nishikawa, K.; Taniguchi, M.; Nagaoka, K.; Kurahashi, H.; Kojima, S.; Morishima, T.: Analysis of gene-expression profiles by oligonucleotide microarray in children with influenza. J Gen Virol 87(6):1677-1683, 2006.

[Lu05] Lu, J.; Getz, G.; Miska, E.A.; Alvarez-Saavedra, E.; Lamb, J.; Peck, D.; Sweet-Cordero, A.; Elbert, B.L.; Mak, R.H.; Fernando, A.A.; Downing, J.R.; Jacks, T.; Horvitz, H.R.; Golub, T.R.: MicroRNA expression profiles classify human cancers. Nature 435(7043):834-838, 2005.

[Ra05] Rahnenführer, J.; Beerenwinkel, N.; Schulz, W.A.; Hartmann, C.; von Deimling, A.; Wullich, B.; Lengauer, T.: Estimating cancer survival and clinical outcome based on genetic tumor progression scores. Bioinformatics 21(10):2438-2446, 2005.

[Sa08] Sassen, S.; Miska, E.A.; Caldas, C.: MicroRNA-implications for cancer. Virchows Arch, 452(1):1-10, 2008.

[To02] Tomlinson, I.; Sasieni, P.; Bodmer, W.: How many mutations in a cancer? Am J Pathol:160:755–8, 2002.

[Vo06] Volinia, S.; Calin, G.A.; Liu, C-G.; Amps, S.; Cimmino, A.; Petrocca, F.; Visone, R.; Iorio, M.; Roldo, C.; Ferracin, M.; Prueitt, R.L.; Yanaihara, N.; Lanza, G.; Scarpa, A.; Vecchione, A.; Negrini, M.; Harris, C.C.; Croce, C.M.: A microRNA expression signature of human solid tumors defines cancer gene targets. Proc Natl Acad Sci U S A 103(7):2257-2261, 2006.