Report Number: WUCSE-2006-42

2006-01-01

A comprehensive analysis of the effect of microarray data

Monika Ray, Johannes Freudenberg, and Weixiong Zhang

Background: Microarray data preprocessing, such as differentially expressed (DE) genes selection, is performed prior to higher level statistical analysis in order to account for technical variability. Preprocessing for the Affymetrix GeneChip includes background correction, normalisation and summarisation. Numerous preprocessing methods have been proposed with little consensus as to which is the most suitable. Furthermore, due to poor concordance among results from cross-platform analyses, protocols are being developed to enable cross-platform reproducibility. However, the effect of data analysis on a single platform is still unknown. The objective of our study is two-fold: first to determine whether there is consistency in... **Read complete abstract on page 2.**

A comprehensive analysis of the effect of microarray data

Monika Ray, Johannes Freudenberg, and Weixiong Zhang

Complete Abstract:

Background: Microarray data preprocessing, such as differentially expressed (DE) genes selection, is performed prior to higher level statistical analysis in order to account for technical variability. Preprocessing for the Affymetrix GeneChip includes background correction, normalisation and summarisation. Numerous preprocessing methods have been proposed with little consensus as to which is the most suitable. Furthermore, due to poor concordance among results from cross-platform analyses, protocols are being developed to enable cross-platform reproducibility. However, the effect of data analysis on a single platform is still unknown. The objective of our study is two-fold: first to determine whether there is consistency in the results obtained from a single platform; and second to investigate the effect of preprocessing on DE genes selection, analysed on four datasets. Results: Results indicate that microarray analysis is subjective. The lists of DE genes are variable and dependent on the preprocessing method used. Furthermore, the characteristics of the dataset, and the type of DE genes identification method used, greatly affect the outcome. Despite using a single platform, there is a lot of variability in the results. Conclusions: This is the first comprehensive analysis using multiple datasets generated from a single platform and involving many DE genes selection methods to assess the effect of data preprocessing on downstream analysis. Results indicate that preprocessing methods affect downstream analysis. Results are also affected by the kind of data and statistical analysis tools used. Our study reveals that there are inconsistencies in results obtained from a single platform. These issues have been overlooked in past reports.

2006-42

A comprehensive analysis of the effect of microarray data

Authors: Monika Ray, Johannes Freudenberg, Weixiong Zhang

Corresponding Author: mray@cse.wustl.edu

Abstract: Background: Microarray data preprocessing, such as differentially expressed (DE) genes selection, is performed
prior to higher level statistical analysis in order to account for technical variability. Preprocessing for the Affymetrix GeneChip includes background correction, normalisation and summarisation. Numerous preprocessing
methods have been proposed with little consensus as to which is the most suitable. Furthermore, due to poor concordance among results from cross-platform analyses, protocols are being developed to enable cross-platform
reproducibility. However, the effect of data analysis on a single platform is still unknown. The objective of our study is two-fold: first to determine whether there is consistency in the results obtained from a single platform; and second to investigate the effect of preprocessing on DE genes selection, analysed on four datasets.
Results: Results indicate that microarray analysis is subjective. The lists of DE genes are variable and dependent
on the preprocessing method used. Furthermore, the characteristics of the dataset, and the type of DE genes identification method used, greatly affect the outcome. Despite using a single platform, there is a lot of variability in the results.
Conclusions: This is the first comprehensive analysis using multiple datasets generated from a single platform and involving many DE genes selection methods to assess the effect of data preprocessing on downstream analysis. Results indicate that preprocessing methods affect downstream analysis. Results are also affected by the kind of data and statistical analysis tools used. Our study reveals that there are inconsistencies in results obtained from a single platform. These issues have been overlooked in past reports.

Type of Report: Other

A comprehensive analysis of the effect of microarray data preprocessing methods on differentially expressed transcript selection

Monika Ray[1],Johannes Freudenberg[3],Weixiong Zhang[1,2*]

[1]Department of Computer Science and Engineering, Washington University in Saint Louis, Saint Louis, MO 63130-4899, USA

[2]Department of Genetics, Washington University in Saint Louis, Saint Louis, MO 63130-4899, USA

[3]Biomedical Informatics, Cincinnati Children's Hospital Medical Center, 3333 Burnet Avenue, Cincinnati OH, 45229, USA

Email: Monika Ray - mray@cse.wustl.edu; Johannes Freudenberg - Johannes.Freudenberg@cchmc.org; Weixiong Zhang*- zhang@cse.wustl.edu;

*Corresponding author

Abstract

Background: Microarray data preprocessing, such as differentially expressed (DE) genes selection, is performed prior to higher level statistical analysis in order to account for technical variability. Preprocessing for the Affymetrix GeneChip includes background correction, normalisation and summarisation. Numerous preprocessing methods have been proposed with little consensus as to which is the most suitable. Furthermore, due to poor concordance among results from cross-platform analyses, protocols are being developed to enable cross-platform reproducibility. However, the effect of data analysis on a single platform is still unknown. The objective of our

study is two-fold: first to determine whether there is consistency in the results obtained from a single platform; and second to investigate the effect of preprocessing on DE genes selection, analysed on four datasets.

Results: Results indicate that microarray analysis is subjective. The lists of DE genes are variable and dependent on the preprocessing method used. Furthermore, the characteristics of the dataset, and the type of DE genes identification method used, greatly affect the outcome. Despite using a single platform, there is a lot of variability in the results.

Conclusions: This is the first comprehensive analysis using multiple datasets generated from a single platform and involving many DE genes selection methods to assess the effect of data preprocessing on downstream analysis. Results indicate that preprocessing methods affect downstream analysis. Results are also affected by the kind of data and statistical analysis tools used. Our study reveals that there are inconsistencies in results obtained from a single platform. These issues have been overlooked in past reports.

Background

Since its invention in 1995 [1], microarray technology has become a principal tool for high-throughput gene expression detection and analysis in basic science research and clinical studies. A simple search on PubMed will return over 10,000 papers on microarray analysis published within the last two years. A plethora of gene expression data on model systems and human diseases has been collected and organised at NCBI and many other microarray data repositories. Large microarray data have been produced by many different microarray platforms, including the widely used Affymetrix GeneChip [2].

Concomitant with the production of massive amounts of data, there has been a proliferation of computational methods for quantifying the level of expression on a DNA chip for various microarray platforms. Even with the same microarray platform, the quality of microarray gene expression can also be significantly impacted by the data preprocessing methods adopted. Several studies have called into question the validity of microarray results mainly due to the disparities between results obtained by different groups analysing similar samples [3–5]. Therefore, efforts have been directed to the understanding of the causes of discrepancies across platforms and the development of protocols that would allow for cross-platform

comparisons and reproducibility [6–8]. Unfortunately, the effect of data preprocessing methods, even for the same microarray platform, has not been well understood nor adequately addressed so far.

An important aspect of our investigation is that, although we used multiple datasets, we focused on data generated from a single platform. Hence, one of the main objectives of this paper is to provide the first comprehensive analysis of data preprocessing methods on the *same microarray platform*. By focusing on one microarray platform, we are able to avoid complications from cross-platform gene expression detection and minimise the variability in the results. Furthermore, if there is a huge discrepancy despite using a single platform, then there will be even greater disparities in a cross-platform study.

Along with the increased use of microarray technology, preprocessing methodology for Affymetrix GeneChip – the market leader – is a quickly expanding research subject. In order to compare preprocessing methods, the affycomp webtool [9] was established in 2003, where over 40 microarray preprocessing methods have been submitted by users [10]. In our study, we focused on the Affymetrix GeneChip platform because it is widely used and a large number of data preprocessing methods has been developed for it. Gene expression measurements are confounded by different sources of systematic variation. Sources of noise include differences in dye efficiencies, scanner malfunction, uneven hybridisation, array design, experimenter and other extraneous factors. In order to compare the mRNA expression levels of each probe on a microarray chip as well as compare microarrays from different experiments for the purposes of performing higher-order analyses such as clustering, prediction, and building regulatory networks, it is necessary to remove unwanted variations while retaining intrinsic biological variations. Thus, data preprocessing is a critical step that affects accuracy and validity of downstream analyses. The four main preprocessing steps for Affymetrix GeneChip are - background correction, normalisation, perfect match (PM) probe correction and summarisation. New processing algorithms that claim to handle the characteristics of microarrays well, are continuously being developed. However, there still is no consensus on which processing method is the most efficient [11, 12] or how a preprocessing method be chosen.

Differentially expressed (DE) gene ranking or selection is the next course of action after data generation. The use of microarrays to determine bona fide changes in gene expression between experimental paradigms is confounded by noise due to variability in measurement. Since scientists are eager to use microarrays to identify marker genes and understand the pathogenesis of diseases, there should be an acute awareness of what procedures are being used to identify the molecular signature of a phenotype. Therefore, the second

objective of our analysis is to assess the variability in measurement by investigating the effect of different background correction, normalisation, PM correction and summarisation techniques on downstream analysis, in particular, DE gene selection. Most literature on comparison of microarray processing methods usually analyse either background correction or normalisation. Furthermore, not only do these studies focus on just one dataset, they use tightly controlled calibration data derived from spike-in or dilution studies [10, 13, 14] or data from very simple organisms [15]. However, to date, we have not found any literature that has thoroughly examined the influence of each of these preprocessing stages on DE gene selection on multiple, diverse, real-life datasets. If the DE gene selection stage is highly affected by these preprocessing mechanisms, all subsequent stages will also be affected.

Based on biological assumptions and microarray chip design, many types of preprocessing algorithms have been developed to date [16, 17]. Background correction is the process of correcting probe intensities on an array using information only from that array. Normalisation is the process of removing non-biological variability across arrays. Summarisation is the process of combining the preprocessed PM probes to compute an expression measure for each probeset on the array. In our paper, we focus on the two most widely used preprocessing pipelines, Microarray Suite 5.0 (MAS5.0, now GeneChip Operating System - GCOS) developed by Affymetrix [18], and Robust Multiarray Analysis (RMA) [14]. We used RMA and MAS5.0 for background correction, and constant normalisation and quantile normalisation [14] methods, for normalisation. MAS5.0 and RMA were applied for PM correction, while MAS5.0 (Tukey Biweight) and RMA (medianpolish) were employed for summarisation. Since GC - Robust Multichip Average (GCRMA) [19] is a popular method for background correction and many consider it as one of the best current preprocessing methods [10, 20], we conclude our study with the application of this technique.

The gene selection tools investigated in this study were extraction of differential gene expression (EDGE) [21] and RankGene [22]. These two packages were chosen mainly due to their widespread usage. At a false discovery rate level of 3%, EDGE provides a 198% increase over the highly popular SAM software [21] and therefore, was preferred, in our study, over SAM. Finally, we want to verify whether the choice of a preprocessing method has a greater effect on gene selection than the choice of gene selection tool, which was the conclusion reached by [23].

Following is a brief overview of the study. Four diverse, non-time series, human datasets generated from Affymetrix DNA chip were used in this study - Alzheimer's disease (AD) [24], breast cancer [25], Duchenne

muscular dystrophy (DMD) [26] and chronic myelogenous leukemia (CML) [27]. These particular datasets were chosen since they capture intrinsic gene expression variations underlying these diverse human diseases, and they have as large a number of samples as one can expect from clinical experiments. On each of the four different datasets we applied sixteen different preprocessing strategies resulting from the combining all possible combinations - two for each of the four stages of data preprocessing. In addition, four strategies with GCRMA at the background correction stage were also applied (see Table1). A set of processing methods for each stage, for example, RMA (R) for background correction, constant (C) for normalisation, MAS 5.0 (M) for PM correction and RMA (R) for summarisation refers to a 'combination'. The order of data preprocessing is first background correction, then normalisation, followed by PM correction, and finally summarisation. The four letters in each combination refer to the methods used in the four preprocessing stages in order. R refers to RMA, C refers to constant, Q refers to quantile, M refers to MAS5.0, and G refers to GCRMA. DE gene selection using EDGE as well as RankGene was performed on each combination. In order to determine which method had the greatest effect on DE gene selection, comparisons were made between combinations which differed in a method for a single stage, for example, between RCMR and RCRR (method for PM correction differs). Number of common transcripts was used as the measure of similarity between methods. If the methods are highly dissimilar, then the number of overlapping transcripts would be small. The results have been presented as graphs representing the number of overlapping transcripts as well as via hierarchical clustering.

Results

Combinations were divided into seven groups/sets of four comparisons - one comparison for each of the four processing stages. The results are presented as bar graphs representing the number of overlapping transcripts. Each bar represents one comparison between two combinations. All the graphs have seven sets of four bars (excluding the last group on the far right of each graph). In any particular set, in each of the comparisons, only one of the four preprocessing methods is changed. In any graph, the heights of the bars should be compared within the set of four comparisons (i.e. one group) and not across the whole graph. Greater the effect of the method, smaller will be the number of common transcripts between the two preprocessing combinations, implying greater dissimilarity between the combinations. Figures 1-20 show the amount of overlap between any two combinations in each of the four datasets. In RankGene, for each combination, 500 transcripts were ranked according to the ranking criteria of choice. In EDGE, the

number of significant transcripts were chosen based on the p value. In the case of EDGE selected transcripts, the total number of transcripts was different in each combination as can be seen in Table 1. Therefore, in order to compare the degree of overlap of two combinations, the arithmetic mean between the two combinations is computed. Therefore, the graphs on EDGE have the mean on the y-axis. Again, the bar with the smallest mean indicates the greatest dissimilarity between the two combinations. Heatmaps of the different comparisons were generated on each dataset with the EDGE selected transcripts (figures 21-24) to determine if there was any discernable pattern. All the probe lists obtained from different combinations and different gene selection/ranking methods are included as additional data files.

Table 2 shows the number of significant transcripts using EDGE only. Results indicate that preprocessing technique has a greater effect on gene selection than the kind of gene selection methods used. For example, if the number of significant transcripts in RCRR and RQRR is compared, the huge difference in the number of transcripts can easily be observed, despite using the same gene selection method.

The relative difference in the heights of the bars in the graphs is phenomenal across datasets for the same gene selection method. From figures 16-20, one can observe that the preprocessing methods have a greater similarity when applied to the DMD dataset. However, there is a lot of disparity among the preprocessing combinations when applied to the AD dataset. This clearly indicates that the characteristics of datasets, i.e. the amount of variability in gene expression, have an effect on the preprocessing method. If any one comparison is analysed across all datasets for any one gene selection method, for example, if the height of the MCMM-RCMM bar is observed across the different datasets, the bar height varies across datasets. This indicates that the effect of a preprocessing technique also depends on the kind of dataset. However, the same comparison across different gene selection methods on the same dataset does not produce a drastic difference in the result. These observations along with the results in Table 2 indicate that the preprocessing techniques play a greater role on DE gene selection than the gene selection method applied, whether it be RankGene or EDGE. This confirms the conclusion attained by [23].

On any dataset, the method that has the maximum effect varies across different gene selection techniques. Only in the breast cancer dataset, it seems that normalisation has the greatest impact. However, since the stage depends on the kind of technique used in that stage, we cannot reach any definitive conclusion about whether it is the stage or the technique that has the greatest effect. Analysis of this is left for future

research.

Bar graphs were not created for all possible comparisons between combinations due to lack of space. Therefore, hierarchical clustering was performed primarily to enable visualisation of comparisons between all pairs of combinations. Pattern detection is easier via clustering. However, as can be seen from figures 21-24, the hierarchical clustering does not show any definite pattern with respect to the preprocessing techniques.

Initially we had only applied RMA and MAS5.0 for background correction. When no conclusive results surfaced, we applied the GCRMA method. As GCRMA has grown in popularity, we presumed that it would eliminate the discrepancies observed in the results thus far in our study. However, as can be observed from figures 1-24, GCRMA did not help in resolving matters. GCRMA and RMA belong to the same family of processing methods. However, if GCRM and RCRM or GQRM and RQRM is compared, the number of overlapping transcripts is sometimes high, as in the case of the breast cancer dataset, sometimes low, as in the case of the AD dataset, and sometimes there is no overlap at all, as in the case of the CML dataset. This is also true for all the other comparisons involving GCRMA. In all the bar graphs, the far right end consists of comparisons between GCRMA, RMA and MAS5. Again, there does not seem to be any consistency in the results obtained after applying GCRMA. Results tend to be variable depending on the data analysed and the DE gene selection method employed.

Discussion

Our study is the first systematic and comprehensive analysis of the effect of *all* the data preprocessing stages on DE gene selection analysed on multiple human datasets. There are three key conclusions from this study. The first is that different preprocessing techniques have varying degrees of effect on downstream analysis based on the kind of data analysed. Preprocessing techniques are not robust enough to handle the amount of variability contained in microarray data. Second, the choice of gene selection methods used to evaluate the effect of preprocessing, also affects the outcome. Third, even with using a single platform, the discrepancies between results are substantial.

It is a cause for great concern if the results from the same dataset, using the same gene selection tool on the same platform are not more in concordance. This implies that preprocessing methods have a great

impact on statistical analyses and this in turn affects the reproducibility of results. If there is little or no consensus among results from the same platform, cross-platform discrepancies cannot be effectively surmounted. Therefore, efforts should be directed towards methods that would increase consistency among results generated from the same platform before extending them to cross-platform analysis. Only statistical analysis techniques and inter-laboratory experiment protocols are not the causes of cross-platform disparities as mentioned by [6, 7]. Along with standardising research protocols and statistical analysis tools, attention should be paid to data preprocessing methods.

Some reports state that background correction or normalisation have the greatest effect [10, 23, 28] on DE gene selection. As there are no standard evaluation methods, different reports can provide conflicting results. Investigators evaluating processing methods should also address data quality and assessment tools used. Recently, there has been a deluge of research articles, editorials and commentaries on the reliability of microarray results [29–32]. Ioannidis [29] goes so far as to say that given information on a gene and 200 patients, he can show that this gene affects survival ($p < 0.05$) even if it does not! Due to poor accuracy, sensitivity, specificity and reproducibility, microarrays have not yet passed the Food and Drug Administration (FDA) regulations for routine use in clinical settings.

Miklos and Maleszka [3] state that different analyses of the same microarray data lead to different gene prioritisations. Another group performed a multiple random validation strategy to re-analyse data from 7 of the largest microarray studies to identify a molecular signature of cancer, and discovered that molecular signatures are unstable [33]. Disparities between genes selected from microarray analysis and biologically relevant genes (i.e. those that are truly relevant) are great even on simple organisms such as yeast [34]. Hence, the incongruity cannot be a characteristic of complex multi-cellular organisms. On the other hand many studies claim that their list of genes is significant based on p value, which is typically less than 0.05. However, statistical significance does not imply biological significance. Due to this emotional dependence on p values, the majority of the published results may be irrelevant if not false [35, 36].

This study, along with results in other reports, indicates that microarray analysis is subjective and highly unstable. So how can these problems be overcome? Just developing new algorithms for data preprocessing, without addressing the reasons why the currents methods fail, will not solve the problems related to microarray analysis. No matter how sophisticated higher analyses become, one cannot ignore the fact that microarray chip design also needs to be improved. The correct design of probes on a DNA chip is essential.

Additionally, due to cross-platform and inter-laboratory discrepancies, one should be wary of generating or accepting results derived from meta-analysis data. Furthermore, evaluation of bioinformatics protocols should be done on real-life datasets in addition to the well-controlled, calibration datasets. Most of the previous work on the analysis of processing algorithms was performed on the dilution or spike-in data from Affymetrix [10, 13, 14]. However, the variance of the average chip intensity in such data is much lower than those measured in real-life, and not well-controlled, datasets. Furthermore, the use of a limited number of spike-in genes is not sufficient for a comprehensive analysis of both microarray technology and data analysis methods. Such issues with spike-in data will cast doubts on the applicability of such data for the development of analytical tools for use in diverse gene expression profiles. Large sample size is crucial. Although it is expensive and time consuming to gather a large number of samples, it is the price one has to pay for accuracy and reproducibility. It is also essential to understand the limitations of the technology being used. One thing the current microarray technology is still unable to deal with is the reliable detection of transcripts with low expression. Therefore, attempting to do this without improved technology is inadvisable. Characteristics of high-throughput data need to be well understood and addressed accordingly in the design and execution of experiments. Miron and Nadon [31] term this as 'inferential literacy'. Previous studies have indicated that the high rates of irreproducible results are due to research findings being defended only by statistical significance. Hence, the use and development of other methods of analysis, such as the one suggested by [37], should be encouraged. Avoidance of bias in the manipulation of the analyses as well as having clear study designs will help scientists who plan on embarking onto the field of microarray analysis and prevent them from performing experiments that would most likely erroneous.

In conclusion, efforts should be directed to (i) improved chip design, (ii) 'inferential literacy', (iii) large sample sizes, (iv) random sampling of data for validation, (v) evaluation of tools on real-life as well as calibration datasets, (vi) multiple and completely independent validation of results, (vii) standardisation of evaluation techniques and laboratory protocols, and finally (viii) application of better methods of significance analysis. The good news is that a fraction of the scientific community is realising the problems surrounding microarray analysis and appropriately addressing these issues. However, it will be a while before everyone understands and utilises the fruits of this labour.

Microarray technology is indeed a valuable tool. However, it is important to minimise and separate noise from meaningful information in the data by employing proper experiment and data analysis protocols.

Processing of noise masquerading as knowledge or eliminating valuable information while removing noise, will lead to faulty or overoptimistic conclusions. In the absence of proper evaluation tools and standardised experiment procedures, microarray analysis has become more of an art than science; more like an act of faith than scientific inference. There is an elephant in the room.

Materials and Methods

Data

For the purpose of generalising results, we make use of four datasets, all of which were generated from Affymetrix DNA chip. The first is Alzheimer's Disease (AD) data, generated on Affymetrix GeneChip (HG-U133A) by [24]. The dataset consists of 9 controls and 22 affected individuals. The affected individuals had varying degrees of AD severity ranging from incipient to severe. In our analysis, we compared the nine controls versus the seven incipient cases. The second dataset is on breast cancer, generated from Affymetrix GeneChip (HG-U95A)and initially analysed in [25] and [38]. Breast cancer core biopsies were taken from patients found to be resistant or sensitive to docetaxel treatment. The data consists of 24 tumour samples, 10 of which are sensitive to docetaxel and 14 of which are resistant to docetaxel treatment. The third dataset is on Duchenne muscular dystrophy (DMD), generated from Affymetrix GeneChip (HG-U95A) and analysed in [26]. In this dataset, quadricep skeletal muscle biopsies were taken from 12 DMD patients and 12 unaffected control patients. The fourth dataset is on the response of chronic myelogenous leukemia (CML) patients to imatinib (Gleevec) treatment, produced from Affymetrix GeneChip (HG-U95A) and analysed in [27]. All four datasets are available to the public on PubMed. AD data can be directly downloaded from [39], breast cancer data from [40], CML data from [41] and DMD data from [42].

Data Preprocessing

Computation of gene expression measure is a four step procedure - background correction, normalisation, perfect match (PM) probe correction and finally summarisation. In most cases, PM and summarisation are combined into one step. Let X be the raw probe intensities across all arrays and E be the final probeset expression measures. Then if B is the background correction operation on probes on each array, N is the operation which normalises across arrays and S is the operation that combines probes to compute an

expression measure, the gene expression measure can be formulated as [28]-

$$E = S(N(B(X))) \tag{1}$$

Twenty different combinations of background correction methods, normalisation methods, PM correction methods and summarisation methods were considered to determine which method has the greatest effect on gene selection.

Background Correction

For the purposes of our analysis, we used the RMA, GCRMA and MAS5.0 background correction methods.

MAS 5.0 Background Correction: This method was developed by Affymetrix [18]. Briefly, each chip is split up into 16 zones of equal size. For each zone k, the background b_k and the noise n_k are defined as the mean and standard deviation of the lowest 2% of zone k's probe intensities, respectively. The probe specific background value $b(x, y)$ and noise value $n(x, y)$ are defined as a weighted sum over all b_k and n_k, respectively, where x and y denote the position of a probe on the chip. The corrected signal is the raw signal reduced by $b(x, y)$ where physically possible, otherwise it is $n(x, y)$. Please refer to [18] for more details.

RMA Background Correction: Robust Multi-array Average (RMA) is a widely used alternative preprocessing strategy for Affymetrix GeneChips [14]. The procedure is based on the assumption that the observed probe signal O consists of a normally distributed background component N and an exponentially distributed signal component S such that

$$O = N + S,\ N \ \sim\ \mathrm{N}(\mu, \sigma^2),\ S \ \sim\ \mathrm{Exp}(\alpha) \tag{2}$$

The parameters α, μ, and σ^2 are estimated from the data, and the raw intensities are replaced with the estimated expected value $\hat{E}(S|O = o)$ given the observed value o.

GCRMA Background Correction: GCRMA is a further development of RMA that takes the probe specific hybridisation affinities into account [19]. It also acknowledges the use of mismatch (MM) probes (see below). The following model is fitted (where PM refers to perfect match probe) -

$$PM \quad = \quad O_{PM} + N_{PM} + S$$

$$MM \quad = \quad O_{MM} + N_{MM} + \phi S \tag{3}$$

where O_{PM} and O_{MM} represent optical noise, N_{PM} and N_{MM} represent non-specific binding noise, and S is the actual signal of interest. ϕ is a value between zero and one and accounts for the fact that MM probes tend be less sensitive than their corresponding PM probes but often still measure a specific signal. The parameters N_{PM} and N_{MM} are assumed to be a function of the probe specific hybridisation affinity which is pre-computed based on either a calibration data set or the experimental data – in our study we choose the latter. Equation (3) is then fitted using an empirical Bayes approach. Please refer to [19] for more details.

Normalisation

For the purposes of our analysis, we used constant or global normalisation and quantiles normalisation. The former is similar to the one used by Affymetrix, the latter is part of the RMA preprocessing strategy.

Global Normalisation: The term "global normalisation" refers to a family of methods where each probe intensity is scaled by a chip-specific factor such that a given summary statistic m'_j is equal for all chips after scaling. That is $m'_j = m$, for $j = 1, \ldots, J$, where J is the number of samples. Commonly used summary statistics include sum, median, and mean. In global normalisation, the mean of the first sample is used (without loss of generality), $m = m_1$. Thus, the chip specific scaling factor is computed as follows-

$$f_i = \frac{m_1}{m_j}, j = 1, \ldots, J, \tag{4}$$

where m_j is the mean of all probe intensities in sample j.

Quantiles Normalisation: Quantiles normalisation assumes that the intensities of each chip originate from the same underlying distribution. That is, the quantiles for each chip should be the same. However, biases in the signal generating process result in chip-specific distributions. The goal of quantiles normalisation is to remove these biases by transforming the data such that each quantile is the same across all chips. Based on this rationale, the following algorithm was proposed [14].

- Sort probe intensities X_j for each sample j.
- For each quantile (or rank) i, compute the mean $m_i = \frac{1}{J} \sum_j x_{(i)j}$.
- Replace $x_{(i)j}$ by m_i for each sample j.

- Restore the original order of X_j for each sample j.

Perfect Match Correction

Each perfect match (PM) probe on an Affymetrix GeneChip is complemented by a mismatch (MM) probe which has a different base as its 13^{th} nucleotide. Hence, a mismatch probe is identical to the perfect match sequence except for a single incorrect base in the middle of the oligomer. The rationale for including MM probes in the chip design is to provide a tool for measuring the unspecific binding contribution of the signal. In our study, we applied the MAS5.0 and RMA PM correction methods. The MAS5.0 PM correction is performed as follows. First, compute the ideal mismatch IM_{ij} for each probe pair i in each sample j and then subtract IM_{ij} from the PM probe intensity PM_{ij}, where the ideal mismatch is equal to the mismatch probe intensity MM_{ij}, if $MM_{ij} < PM_{ij}$, otherwise IM_{ij} is computed by downscaling the corresponding perfect match signal. Please refer to [18] for more details. In RMA preprocessing, mismatch probes are simply ignored due to the fact that these probe signals also have probe-specific signal contributions, which would dilute the PM corrected signal.

Summarisation

Affymetrix DNA microarrays contain multiple probes for each target transcript. In order to establish a single expression value for each probeset (i.e. a set of probes corresponding to the same transcript), individual probe values must be summarised. Affymetrix provides chip definition files (CDFs) which can be used to determine probeset assignments. However, recent studies suggest that the manufacturer's probe design and annotation may be outdated and erroneous [43]. Therefore, redefined CDFs (version 6, reference database: RefSeq) were obtained from [44] and used instead of the manufacturer's CDFs. For the purposes of our analysis, we used the MAS5 ("Tukey Biweight") and RMA ("medianpolish") summarisation methods.

Tukey Biweight: This algorithm is described in [18] in detail. Briefly, probe values are log-transformed after PM correction. Next, for each probeset,

1. For each probe j, the deviation u_j from the median weighted by the median deviation from the median is computed.

2. For each probe j, the weight $w_j = (1 - u^2)^2$ is computed.

3. The weighted mean TBI is computed where

$$TBI = \frac{\sum_j w_j x_j}{\sum_j w_j} \tag{5}$$

TBI is the final summary value.

Medianpolish: This algorithm is used as part of the RMA framework [14]. Briefly, a two-way ANOVA-like model is fitted:

$$\log_2(y_{ij}) = \alpha_i + \mu_j + \epsilon_{ij}, \tag{6}$$

where α_i is the probe affinity effect of probe i, $\sum_i \alpha_i = 0$, μ_j represents the expression level for array j, and ϵ_{ij} is an independent identically distributed error term with zero mean. The estimate $\hat{\mu}_j$ is the wanted expression value for the respective probe set on array j. The model parameters are estimated using a robust procedure that iteratively estimates the error matrix $\hat{\epsilon}_{ij}$ by repeatedly subtracting row medians and column medians in an alternating fashion until reaching convergence.

Clustering

For a given gene selection method twenty gene lists were obtained, one for each preprocessing combination. Next, a 20×20 similarity matrix was computed by counting the number of common genes for each pair of strategies. In case of EDGE, this number was normalized by $\frac{1}{2}\left(\frac{1}{l_1} + \frac{1}{l_2}\right)$ where l_1 and l_2 are the lengths of the first and second gene list, respectively, in order to account for the varying lengths of significantly changed gene expression generated by each preprocessing strategy. The similarity matrix S was then converted to a distance matrix D by subtracting S from the maximum value m of S, $\mathrm{D} = m - \mathrm{S}$. Next, hierarchical clustering based on D was computed using function hclust() of the statistical package R [45]. Briefly, this function initially assigns each preprocessing method its own cluster. Next, the algorithm iteratively combines the two most similar clusters until the all-cluster is formed comprising all preprocessing methods. After each joining of two clusters, cluster distances are recomputed using the Lance-Williams dissimilarity update formula. In our implementation, we use complete linkage, i.e. the distance between the most dissimilar pair of preprocessing methods, one from each of the joining clusters. The clustering was then combined with a heatmap using the R function heatmap() on S such that red indicates high similarity and white indicates no similarity.

Differentially Expressed Gene Identification

Significant gene identification or selection of differentially expressed (DE) genes has been a subject of active research for many years, resulting in a multitude of algorithms. Significant gene extraction falls into two broad categories - *wrapper* methods and *filter* methods. In wrapper gene selection methods, the DE gene identification phase is integrated with the classification phase. The main objective of such an algorithm is to find a gene subset that will result in the highest classification accuracy. In filter methods, the DE gene extraction phase is independent of the classification phase. After significant gene selection, one can use any classifier on that subset of genes and tune the model to achieve high accuracy.

Wrapper methods attempt to remove redundancy and irrelevancy in the list of significant features/genes. However, if the objective is to identify new genes responsible for a phenotype, then redundancy should not be eliminated. In the process of eliminating redundant genes, one can lose a potentially new marker gene. Since our study focused on concept discovery, where the objective is to discover new genes, we only considered filter methods. Most of the previous studies on preprocessing techniques used filter methods, such as ANOVA and the ratio of expression levels, to obtain a list of DE genes. In this study, we used two packages for the identification of DE genes - RankGene [22] and EDGE [21].

RankGene is a C++ programme for analysing gene expression data, feature selection and ranking genes based on the predictive power of each gene to classify samples into functional or disease categories [22]. It supports eight measures for quantifying a gene's ability to distinguish between classes: information gain, twoing rule, sum minority, max minority, Gini index, sum of variances, t-statistics, and one-dimensional support vector machines. The first six of the eight methods are quite commonly used statistical learning techniques. One-dimensional SVM measures the effectiveness of a gene by calculating the accuracy of single feature SVM classifiers. The t-statistics measure was first used in [46] and is a commonly used technique. Since we are not evaluating different gene selection methods but rather investigating the effect of preprocessing on DE gene selection strategies, which gene selection technique we use is of little concern. In our analysis, we used information gain, twoing rule, t-statistic and sum minority. Information gain, twoing rule, and sum minority are statistical impurity measures, which quantify the best possible class predictability that can be obtained by dividing the full range of expression of a given gene into two disjoint intervals. All samples in one interval belong to one class (e.g, normal) and all samples in the other interval belong to the other class (e.g, affected). The difference among the various measures is used to quantify the

error in prediction. For a given choice of measure, RankGene minimises the error over all possible thresholds that partition the gene into two intervals.

EDGE is an open-source software programme which identifies DE genes based on the optimal discovery procedure (ODP) [21]. It maximises the expected number of true positives for each fixed level of expected false positives. While most gene selection methods consider only one feature at a time, the ODP method uses information from the entire data set when testing each feature, which is its greatest asset. It uses a modified t-statistic to select significant genes. EDGE ranks all the genes in the dataset but also selects a subset of genes considered significant at a particular p value. EDGE is a newer tool compared to RankGene and its main advantage is the ability to analyse time series data. The t-statistic of RankGene is different from that of EDGE. The modified t-statistic of EDGE does not analyse each gene individually for its significance, but rather takes into account the correlation among different genes.

While EDGE is a gene selection as well as gene ranking tool, RankGene is only a ranking tool. However, we chose to analyse the top 500 transcripts from the ranked list. By choosing these different gene selection packages for analysis in this paper, majority of the spectrum of possible filter methods available for gene selection is covered. By utilising these two DE gene identification softwares, we have applied five different gene selection algorithms.

List of abbreviations

- DE - differentially expressed
- PM - perfect match
- MAS5.0 - Microarray Suite 5.0
- RMA - Robust Multiarray Analysis
- GCRMA - GC - Robust Multichip Average
- EDGE - extraction of differential gene expression
- SAM - significance analysis of microarrays
- AD - Alzheimer's disease

- DMD - Duchenne muscular dystrophy
- CML - chronic myelogenous leukemia
- FDA - Food and Drug Administration
- MM - mismatch
- IM - ideal mismatch
- CDFs - chip definition files
- D - distance matrix
- S - similarity matrix
- ANOVA - analysis of variance between groups
- SVM - support vector machines
- ODP - optimal discovery procedure
- RCRR - RMA background correction, constant normalisation, RMA PM correction, RMA summarisation
- RQRR - RMA background correction, quantile normalisation, RMA PM correction, RMA summarisation
- MQRR - MAS5.0 background correction, quantile normalisation, RMA PM correction, RMA summarisation
- RQMM - RMA background correction, quantile normalisation, MAS5.0 PM correction, MAS5.0 summarisation
- MCMM - MAS5.0 background correction, constant normalisation, MAS5.0 PM correction, MAS5.0 summarisation
- RCMM - RMA background correction, constant normalisation, MAS5.0 PM correction, MAS5.0 summarisation
- MCRR - MAS5.0 background correction, constant normalisation, RMA PM correction, RMA summarisation

- MQMM - MAS5.0 background correction, quantile normalisation, MAS5.0 PM correction, MAS5.0 summarisation
- RCRM - RMA background correction, constant normalisation, RMA PM correction, MAS5.0 summarisation
- RQRM - RMA background correction, quantile normalisation, RMA PM correction, MAS5.0 summarisation
- MQRM - MAS5.0 background correction, quantile normalisation, RMA PM correction, MAS5.0 summarisation
- RQMR - RMA background correction, quantile normalisation, MAS5.0 PM correction, RMA summarisation
- MCMR - MAS5.0 background correction, constant normalisation, MAS5.0 PM correction, RMA summarisation
- RCMR - RMA background correction, constant normalisation, MAS5.0 PM correction, RMA summarisation
- MCRM - MAS5.0 background correction, constant normalisation, RMA PM correction, MAS5.0 summarisation
- MQMR - MAS5.0 background correction, quantile normalisation, MAS5.0 PM correction, RMA summarisation
- GCRM - GCRMA background correction, constant normalisation, RMA PM correction, MAS5.0 summarisation
- GCRR - GCRMA background correction, constant normalisation, RMA PM correction, RMA summarisation
- GQRM - GCRMA background correction, quantile normalisation, RMA PM correction, MAS5.0 summarisation
- GQRR - GCRMA background correction, quantile normalisation, RMA PM correction, RMA summarisation

Additional Files

Additional file 1 — adedge.zip

This file contains the list of transcripts obtained by using EDGE on the AD dataset. Although all the transcripts are ranked, the number of significant transcripts at a particular cut-off value of p is mentioned at the top of the file. Each file is named with the preprocessing combination employed on the data. The nomenclature is explained in background section, last paragraph. Individual files (.txt extension) can be viewed with Microsoft Excel or Windows Notepad.

Additional file 2 — adrankcriteria1.zip

This file contains the list of transcripts, from the AD data, ranked according to RankGene criteria 1 - information gain. Each file is named with the preprocessing combination employed on the data. The nomenclature is explained in background section, last paragraph. Individual files (.txt extension) can be viewed with Microsoft Excel or Windows Notepad.

Additional file 3 — adrankcriteria2.zip

This file contains the list of transcripts, from the AD data, ranked according to RankGene criteria 2 - twoing rule. Each file is named with the preprocessing combination employed on the data. The nomenclature is explained in background section, last paragraph. Individual files (.txt extension) can be viewed with Microsoft Excel or Windows Notepad.

Additional file 4 — adrankcriteria3.zip

This file contains the list of transcripts, from the AD data, ranked according to RankGene criteria 3 - sum minority. Each file is named with the preprocessing combination employed on the data. The nomenclature is explained in background section, last paragraph. Individual files (.txt extension) can be viewed with Microsoft Excel or Windows Notepad.

Additional file 5 — adrankcriteria7.zip

This file contains the list of transcripts, from the AD data, ranked according to RankGene criteria 7 - t-statistic. Each file is named with the preprocessing combination employed on the data. The

nomenclature is explained in background section, last paragraph. Individual files (.txt extension) can be viewed with Microsoft Excel or Windows Notepad.

Additional file 6 — bcedge.zip

This file contains the list of transcripts obtained by using EDGE on the breast cancer dataset. Although all the transcripts are ranked, the number of significant transcripts at a particular cut-off value of p is mentioned at the top of the file. Each file is named with the preprocessing combination employed on the data. The nomenclature is explained in background section, last paragraph. Individual files (.txt extension) can be viewed with Microsoft Excel or Windows Notepad.

Additional file 7 — bcrankcriteria1.zip

This file contains the list of transcripts, from the breast cancer data, ranked according to RankGene criteria 1 - information gain. Each file is named with the preprocessing combination employed on the data. The nomenclature is explained in background section, last paragraph. Individual files (.txt extension) can be viewed with Microsoft Excel or Windows Notepad.

Additional file 8 — bcrankcriteria2.zip

This file contains the list of transcripts, from the breast cancer data, ranked according to RankGene criteria 2 - twoing rule. Each file is named with the preprocessing combination employed on the data. The nomenclature is explained in background section, last paragraph. Individual files (.txt extension) can be viewed with Microsoft Excel or Windows Notepad.

Additional file 9 — bcrankcriteria3.zip

This file contains the list of transcripts, from the breast cancer data, ranked according to RankGene criteria 3 - sum minority. Each file is named with the preprocessing combination employed on the data. The nomenclature is explained in background section, last paragraph. Individual files (.txt extension) can be viewed with Microsoft Excel or Windows Notepad.

Additional file 10 — bcrankcriteria7.zip

This file contains the list of transcripts, from the breast cancer data, ranked according to RankGene criteria 7 - t-statistic. Each file is named with the preprocessing combination employed on the data. The nomenclature is explained in background section, last paragraph. Individual files (.txt extension) can be viewed with Microsoft Excel or Windows Notepad.

Additional file 11 — cmledge.zip

This file contains the list of transcripts obtained by using EDGE on the CML dataset. Although all the transcripts are ranked, the number of significant transcripts at a particular cut-off value of p is mentioned at the top of the file. Each file is named with the preprocessing combination employed on the data. The nomenclature is explained in background section, last paragraph. Individual files (.txt extension) can be viewed with Microsoft Excel or Windows Notepad.

Additional file 12 — cmlrankcriteria1.zip

This file contains the list of transcripts, from the CML data, ranked according to RankGene criteria 1 - information gain. Each file is named with the preprocessing combination employed on the data. The nomenclature is explained in background section, last paragraph. Individual files (.txt extension) can be viewed with Microsoft Excel or Windows Notepad.

Additional file 13 — cmlrankcriteria2.zip

This file contains the list of transcripts, from the CML data, ranked according to RankGene criteria 2 - twoing rule. Each file is named with the preprocessing combination employed on the data. The nomenclature is explained in background section, last paragraph. Individual files (.txt extension) can be viewed with Microsoft Excel or Windows Notepad.

Additional file 14 — cmlrankcriteria3.zip

This file contains the list of transcripts, from the CML data, ranked according to RankGene criteria 3 - sum minority. Each file is named with the preprocessing combination employed on the data. The

nomenclature is explained in background section, last paragraph. Individual files (.txt extension) can be viewed with Microsoft Excel or Windows Notepad.

Additional file 15 — cmlrankcriteria7.zip

This file contains the list of transcripts, from the CML data, ranked according to RankGene criteria 7 - t-statistic. Each file is named with the preprocessing combination employed on the data. The nomenclature is explained in background section, last paragraph. Individual files (.txt extension) can be viewed with Microsoft Excel or Windows Notepad.

Additional file 16 — dmdedge.zip

This file contains the list of transcripts obtained by using EDGE on the DMD dataset. Although all the transcripts are ranked, the number of significant transcripts at a particular cut-off value of p is mentioned at the top of the file. Each file is named with the preprocessing combination employed on the data. The nomenclature is explained in background section, last paragraph. Individual files (.txt extension) can be viewed with Microsoft Excel or Windows Notepad.

Additional file 17 — dmdrankcriteria1.zip

This file contains the list of transcripts, from the DMD data, ranked according to RankGene criteria 1 - information gain. Each file is named with the preprocessing combination employed on the data. The nomenclature is explained in background section, last paragraph. Individual files (.txt extension) can be viewed with Microsoft Excel or Windows Notepad.

Additional file 18 — dmdrankcriteria2.zip

This file contains the list of transcripts, from the DMD data, ranked according to RankGene criteria 2 - twoing rule. Each file is named with the preprocessing combination employed on the data. The nomenclature is explained in background section, last paragraph. Individual files (.txt extension) can be viewed with Microsoft Excel or Windows Notepad.

Additional file 19 — dmdrankcriteria3.zip

This file contains the list of transcripts, from the DMD data, ranked according to RankGene criteria 3 - sum minority. Each file is named with the preprocessing combination employed on the data. The nomenclature is explained in background section, last paragraph. Individual files (.txt extension) can be viewed with Microsoft Excel or Windows Notepad.

Additional file 20 — dmdrankcriteria7.zip

This file contains the list of transcripts, from the DMD data, ranked according to RankGene criteria 7 - t-statistic. Each file is named with the preprocessing combination employed on the data. The nomenclature is explained in background section, last paragraph. Individual files (.txt extension) can be viewed with Microsoft Excel or Windows Notepad.

Acknowledgements

This research was funded in part by NSF grants EIA-0113618 and IIS-0535257, and Monsanto Co.

References

1. Schena M, Shalon D, Davis RW, Brown PO: **Quantitative monitoring of gene expression patterns with a complementary DNA microarray**. *Science* 1995, **270**(5235):467–470.
2. Affymetrix: [http://www.affymetrix.com/index.affx].
3. Miklos GLG, Maleszka R: **Microarray reality checks in the context of a complex disease**. *Nature biotechnology* 2004, **22**(5):615–621.
4. Yauk CL, Berndt ML, Williams A, Douglas GR: **Comprehensive comparison of six microarray technologies**. *Nucleic Acids Research* 2004, **32**(15):e124.
5. Tan PK, Downey TJ, Spitznagel ELJ, Xu P, Fu D, Dimitrov DS, Lempicki RA, Raaka BM, Cam MC: **Evaluation of gene expression measurements from commercial microarray platforms**. *Nucleic Acids Research* 2003, **31**(19):5676–5684.

6. Larkin JE, Frank BC, Gavras H, Sultana R, Quackenbush J: **Independence and reproducibility across microarray platforms.** *Nature Methods* 2005, **2**(5):337–344.

7. Irizarry RA, Warren D, Spencer F, Kim IF, Biswal S, Frank BC, Gabrielson E, Garcia JG, Geoghegan J, Germino G, Griffin C, Hilmer SC, Hoffman E, Jedlicka AE, Kawasaki E, Martinez-Murillo F, Morsberger L, Lee H, Petersen D, Quackenbush J, Scott A, Wilson M, Yang Y, Ye SQ, Yu W: **Multiple-laboratory comparison of microarray platforms.** *Nature Methods* 2005, **2**(5):345–350.

8. Wang Y, Barbacioru C, Hyland F, Xiao W, Hunkapiller KL, Blake J, Chan F, Gonzalez C, Zhang L, Samaha RR: **Large scale real-time PCR validation on gene expression measurements from two commercial long-oligonucleotide microarrays**. *BMC Genomics* 2006, **7**(59).

9. Irizarry RA: **Affycomp II, A Benchmark for Affymetrix GeneChip Expression Measures** [http://affycomp.biostat.jhsph.edu/].

10. Irizarry RA, Wu Z, Jaffee HA: **Comparison of Affymetrix GeneChip expression measures**. *Bioinformatics* 2006, **22**(7):789–794.

11. Qin L, Beyer RP, Hudson FN, Linford NJ, Morris DE, Kerr KF: **Evaluation of methods for oligonucleotide array data via quantitative real-time PCR**. *BMC Bioinformatics* 2006, **7**(23).

12. Harr B, Schlotterer C: **Comparison of algorithms for the analysis of Affymetrix microarray data as evaluated by co-expression of genes in known operons**. *Nucleic Acids Research* 2006, **34**(2).

13. Bolstad BM, Irizarry RA, Astrand M, Speed TP: **A comparison of normalization methods for high density oligonucleotide array data based on variance and bias**. *Bioinformatics* 2003, **19**(2):185–193.

14. Irizarry RA, Hobbs B, Collin F, Beazer-Barclay YD, Antonellis KJ, Scherf U, Speed TP: **Exploration, normalization, and Summaries of High-density Oligonucleotide Array Probe Level Data**. *Biostatistics* 2003, **4**(2).

15. Ding Y, Wilkins D: **The effect of normalisation on microarray data analysis**. *DNA and Cell Biology* 2004, **23**:635–642.

16. Quackenbush J: **Microarray data normalisation and transformation**. *Nature Genetics* 2002, **32**:496–501.

17. Smyth GK, Speed T: **Normalisation of cDNA microarray data**. *Methods* 2003, **31**:265–273.

18. Affymetrix: **Statistical algorithms description document** [http://www.affymetrix.com/support/technical/whitepapers/sadd_whitepaper.pdf].

19. Wu Z, Irizarry RA, Gentleman R, Murillo FM, Spencer F: **A model-based background adjustment for oligonucleotide expression arrays**. *J. Am. Stat. Assoc.* 2004, **99**:909–917.

20. University of Rochester Medical centre Fgc: **Microarray Gene Expression Analysis Tools** [http://fgc.urmc.rochester.edu/data_analysis.html].

21. Storey JD, Dai JY, Leek JT: **The optimal discovery procedure for large-scale significance testing, with applications to comparative microarray experiments**. *University of Washington Biostatistics Working Paper Series, Working paper 260* 2005.

22. Su Y, M MT, Pavlovic V, Schaffer M, Kasif S: **RankGene: Identification of diagnostic genes based on expression data**. *Bioinformatics* 2003, **19**(12):1578–1579.

23. Hoffmann R, Seidl T, Dugas M: **Profound effect of normalisation on detection of differentially expressed genes in oligonucleotide microarray data analysis**. *Bioinformatics* 2006, **22**(7):789–794.

24. Blalock EM, Geddes JW, Chen NM K Cand Porter, Markesbery WR, Landfield PW: **Incipient Alzheimer's disease: Microarray correlation analyses reveal major transcriptional and tumor suppressor responses**. *Proc Natl Acad Sci USA* 2004, **101**:2174–2178.

25. Chang JC, Wooten EC, Tsimelzon A, Hilsenbeck SG, Gutierrez YL Tham, Kalidas M, C M, Elledge R, Mohsin S, Osborne CK, Chamness GC, Allred DC, Lewis MT, Wong H, O Connell P: **Patterns of Resistance and Incomplete Response to Docetaxel by Gene Expression Profiling in Breast Cancer Patients**. *Journal of Clinical Oncology* 2005, **23**(6):1169–1177.

26. Haslett JN, Sanoudou D, Kho AT, Bennett RR, Greenberg SA, Kohane IS, Beggs AH, Kunkel LM: **Gene expression comparison of biopsies from Duchenne muscular dystrophy (DMD) and normal skeletal muscle**. *Proc Natl Acad Sci USA* 2002, **99**(23):15000–15005.

27. Crossman LC, Mori M, Hsieh Y, Park BS, Lange T, Paschka P, Harrington CA, Krohn K, Niederwieser DW, Hochhaus A, Druker BJ, Deininger MW: **In chronic myeloid leukemia white cells from cytogenetic responders and non-responders to imatinib have very similar gene expression signatures**. *Haematologica* 2005, **90**:459–464.

28. Bolstad BM: **Comparing the effects of background, normalization and summarization on gene expression estimates** [http://www.stat.berkeley.edu/users/bolstad/stuff/components.pdf].

29. Ioannidis JPA: **Microarrays and molecular research: noise discovery?** *Lancet* 2005, **365**:454–455.

30. Shields R: **MIAME, we have a problem**. *Trends in genetics* 2006, **22**(2):65–66.

31. Miron M, Nadon R: **Inferential literacy for experimental high-throughput biology**. *Trends in genetics* 2006, **22**(2):84–89.

32. Draghici S, Khatri P, Eklund AC, Szallasi Z: **Reliability and reproducibility issues in DNA microarray measurements**. *Trends in genetics* 2006, **22**(2):101–109.

33. Michiels S, Koscielny S, Hill C: **Prediction of cancer outcome with microarrays: A multiple random validation strategy**. *Lancet* 2005, **365**:488–492.

34. Birrell GW, Brown JA, Wu HI, Giaeverdagger G, Chudagger AM, Davisdagger RW, Brown JM: **Transcriptional response of Saccharomyces cerevisiae to DNA-damaging agents does not identify the genes that protect against these agents**. *Proc Natl Acad Sci USA* 2002, **99**(13):8778–8783.

35. Halloran PF, Reeve J, Kaplan B: **Lies, Damn Lies, and Statistics: The Perils of the P Value**. *American Journal of transplantation* 2006, **6**:10–11.

36. Ioannidis JPA: **Why most published research findings are false**. *PLOS Medicine* 2005, **2**(8):e124.

37. Wacholder S, Chanock S, Garcia-Closas M, Elghormli L, Rothman N: **Assessing the probability that a positive report is false: an approach for molecular epidemiology studies**. *J Natl Cancer Inst.* 2004, **96**(6):434–442.

38. Chang JC, Wooten EC, Tsimelzon A, Hilsenbeck SG, Gutierrez MC, Elledge R, Mohsin S, Osborne CK, Chamness GC, Allred DC, O Connell P: **Gene expression profiling for the prediction of therapeutic response to docetaxel in patients with breast cancer**. *Lancet* 2003, **362**:362–369.

39. PubMed: [ftp://ftp.ncbi.nih.gov/pub/geo/DATA/supplementary/series/GSE1297/].

40. PubMed: [ftp://ftp.ncbi.nih.gov/pub/geo/DATA/supplementary/series/GSE349/].

41. PubMed: [http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE2535].

42. PubMed: [http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE1004].

43. Dai M, Wang P, Boyd AD, Kostov G, Athey B, Jones EG, Bunney WE, Myers RM, Speed TP, Akil H, Watson SJ, Meng F: **Evolving gene/transcript definitions significantly alter the interpretation of GeneChip data**. *Nucleic Acids Research* 2005, **33**(20).

44. Molecular, Behavioral Neuroscience Institute UoM: **Brainarray** [http://brainarray.mhri.med.umich.edu/Brainarray/Database/CustomCDF/CDF_download_v6.asp].

45. R Development Core Team: *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria 2006, [http://www.R-project.org]. [ISBN 3-900051-07-0].

46. Golub TR, Slonim DK, Tamayo P, Huard C, Gaasenbeek M, Mesirov JP, Coller H, Loh ML, Downing JR, Caligiuri MA, Bloomfield CD, Lander ES: **Molecular classification of cancer: class discovery and class prediction by gene expression monitoring**. *Science* 1999, **286**:531–537.

Figures

Figure 1 - RankGene - Amount of overlap between preprocessing combinations using information gain ranking criteria on Alzheimer's disease data

This figure shows the number of common transcripts between two preprocessing combinations. In each of the eight sets, one stage of the 4-stage preprocessing is changed and the amount of similarity is determined by the degree of overlap.

Figure 2 - RankGene - Amount of overlap between preprocessing combinations using twoing rule ranking criteria on Alzheimer's disease data

This figure shows the number of common transcripts between two preprocessing combinations. In each of the eight sets, one stage of the 4-stage preprocessing is changed and the amount of similarity is determined by the degree of overlap.

Figure 3 - RankGene - Amount of overlap between preprocessing combinations using sum minority ranking criteria on Alzheimer's disease data

This figure shows the number of common transcripts between two preprocessing combinations. In each of the eight sets, one stage of the 4-stage preprocessing is changed and the amount of similarity is determined by the degree of overlap.

Figure 4 - RankGene - Amount of overlap between preprocessing combinations using t-statistic ranking criteria on Alzheimer's disease data

This figure shows the number of common transcripts between two preprocessing combinations. In each of the eight sets, one stage of the 4-stage preprocessing is changed and the amount of similarity is determined by the degree of overlap.

Figure 5 - Amount of overlap between preprocessing combinations using EDGE on Alzheimer's disease data

This figure shows the number of common transcripts between two preprocessing combinations. In each of the eight sets, one stage of the 4-stage preprocessing is changed and the amount of similarity is determined by the degree of overlap. Significant genes are selected on $p < 0.01$.

Figure 6 - RankGene - Amount of overlap between preprocessing combinations using information gain ranking criteria on Breast cancer data

This figure shows the number of common transcripts between two preprocessing combinations. In each of the eight sets, one stage of the 4-stage preprocessing is changed and the amount of similarity is determined by the degree of overlap.

Figure 7 - RankGene - Amount of overlap between preprocessing combinations using twoing rule ranking criteria on Breast cancer data

This figure shows the number of common transcripts between two preprocessing combinations. In each of the eight sets, one stage of the 4-stage preprocessing is changed and the amount of similarity is determined

by the degree of overlap.

Figure 8 - RankGene - Amount of overlap between preprocessing combinations using sum minority ranking criteria on Breast cancer data

This figure shows the number of common transcripts between two preprocessing combinations. In each of the eight sets, one stage of the 4-stage preprocessing is changed and the amount of similarity is determined by the degree of overlap.

Figure 9 - RankGene - Amount of overlap between preprocessing combinations using t-statistic ranking criteria on Breast cancer data

This figure shows the number of common transcripts between two preprocessing combinations. In each of the eight sets, one stage of the 4-stage preprocessing is changed and the amount of similarity is determined by the degree of overlap.

Figure 10 - Amount of overlap between preprocessing combinations using EDGE on Breast cancer data

This figure shows the number of common transcripts between two preprocessing combinations. In each of the eight sets, one stage of the 4-stage preprocessing is changed and the amount of similarity is determined by the degree of overlap. Significant genes are selected on a $p < 0.001$.

Figure 11 - RankGene - Amount of overlap between preprocessing combinations using information gain ranking criteria on chronic myelogenous leukemia data

This figure shows the number of common transcripts between two preprocessing combinations. In each of the eight sets, one stage of the 4-stage preprocessing is changed and the amount of similarity is determined by the degree of overlap.

Figure 12 - RankGene - Amount of overlap between preprocessing combinations using twoing rule ranking criteria on chronic myelogenous leukemia data

This figure shows the number of common transcripts between two preprocessing combinations. In each of the eight sets, one stage of the 4-stage preprocessing is changed and the amount of similarity is determined by the degree of overlap.

Figure 13 - RankGene - Amount of overlap between preprocessing combinations using sum minority ranking criteria on chronic myelogenous leukemia data

This figure shows the number of common transcripts between two preprocessing combinations. In each of the eight sets, one stage of the 4-stage preprocessing is changed and the amount of similarity is determined by the degree of overlap.

Figure 14 - RankGene - Amount of overlap between preprocessing combinations using t-statistic ranking criteria on chronic myelogenous leukemia data

This figure shows the number of common transcripts between two preprocessing combinations. In each of the eight sets, one stage of the 4-stage preprocessing is changed and the amount of similarity is determined by the degree of overlap.

Figure 15 - Amount of overlap between preprocessing combinations using EDGE on chronic myelogenous leukemia data

This figure shows the number of common transcripts between two preprocessing combinations. In each of the eight sets, one stage of the 4-stage preprocessing is changed and the amount of similarity is determined by the degree of overlap. Significant genes are selected on a $p < 0.01$.

Figure 16 - RankGene - Amount of overlap between preprocessing combinations using information gain ranking criteria on Duchenne muscular dystrophy data

This figure shows the number of common transcripts between two preprocessing combinations. In each of the eight sets, one stage of the 4-stage preprocessing is changed and the amount of similarity is determined

by the degree of overlap.

Figure 17 - RankGene - Amount of overlap between preprocessing combinations using twoing rule ranking criteria on Duchenne muscular dystrophy data

This figure shows the number of common transcripts between two preprocessing combinations. In each of the eight sets, one stage of the 4-stage preprocessing is changed and the amount of similarity is determined by the degree of overlap.

Figure 18 - RankGene - Amount of overlap between preprocessing combinations using sum minority ranking criteria on Duchenne muscular dystrophy data

This figure shows the number of common transcripts between two preprocessing combinations. In each of the eight sets, one stage of the 4-stage preprocessing is changed and the amount of similarity is determined by the degree of overlap.

Figure 19 - RankGene - Amount of overlap between preprocessing combinations using t-statistic ranking criteria on Duchenne muscular dystrophy data

This figure shows the number of common transcripts between two preprocessing combinations. In each of the eight sets, one stage of the 4-stage preprocessing is changed and the amount of similarity is determined by the degree of overlap.

Figure 20 - Amount of overlap between preprocessing combinations using EDGE on Duchenne muscular dystrophy data

This figure shows the number of common transcripts between two preprocessing combinations. In each of the eight sets, one stage of the 4-stage preprocessing is changed and the amount of similarity is determined by the degree of overlap. Significant genes are selected on a $p < 0.01$.

Figure 21 - Hierarchical clustering of preprocessing combinations on EDGE selected transcripts on AD

Strength of the similarity is illustrated via colour intensity. Red indicates high similarity and white indicates no similarity.

Figure 22 - Hierarchical clustering of preprocessing combinations on EDGE selected transcripts on breast cancer

Strength of the similarity is illustrated via colour intensity. Red indicates high similarity and white indicates no similarity.

Figure 23 - Hierarchical clustering of preprocessing combinations on EDGE selected transcripts on CML

Strength of the similarity is illustrated via colour intensity. Red indicates high similarity and white indicates no similarity.

Figure 24 - Hierarchical clustering of preprocessing combinations on EDGE selected transcripts on DMD

Strength of the similarity is illustrated via colour intensity. Red indicates high similarity and white indicates no similarity.

Tables

Table 1 - Processing methods and the stages they were applied in

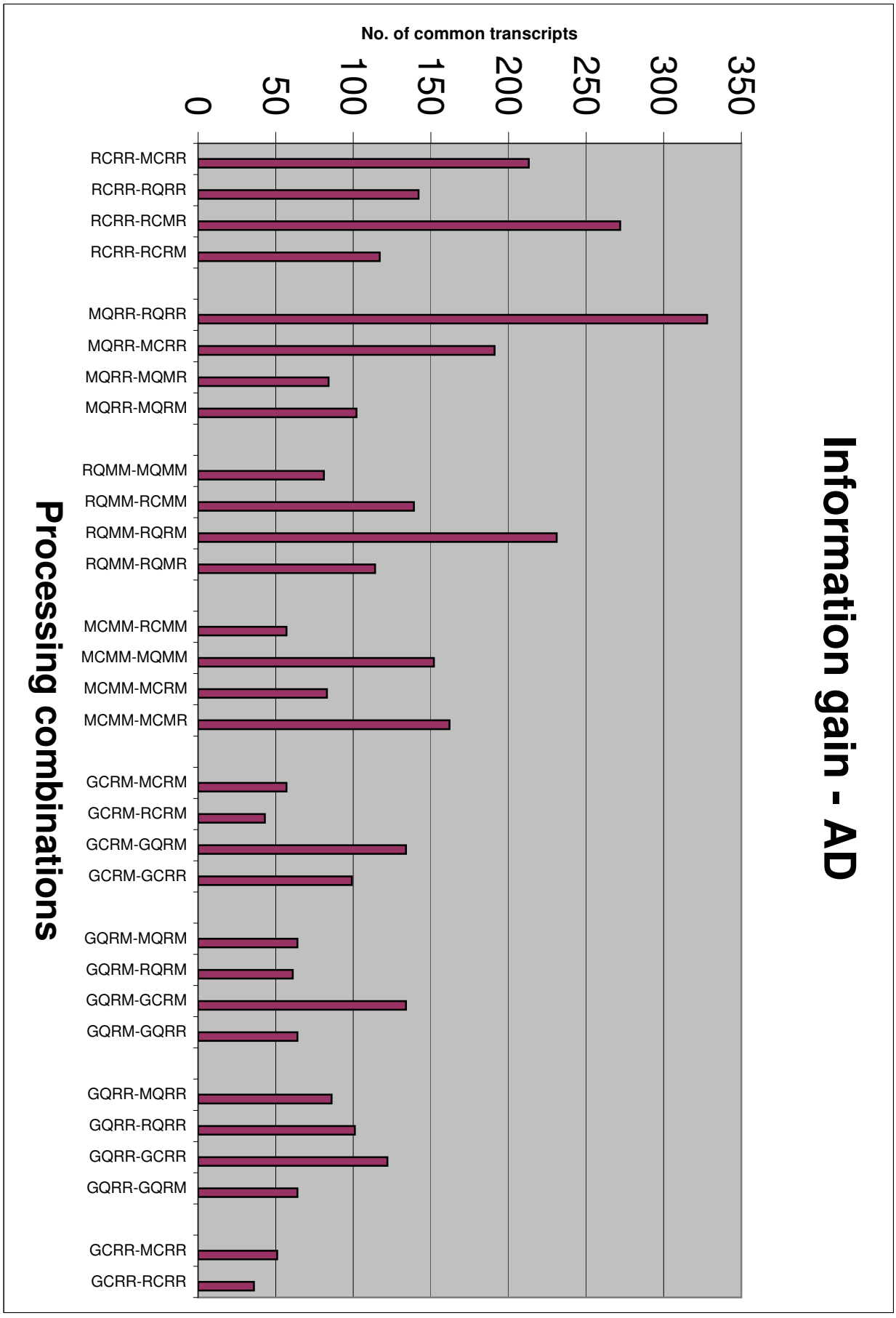
x refers to the stage the method was applied in. Letters in brackets refer to the abbreviation of that method.

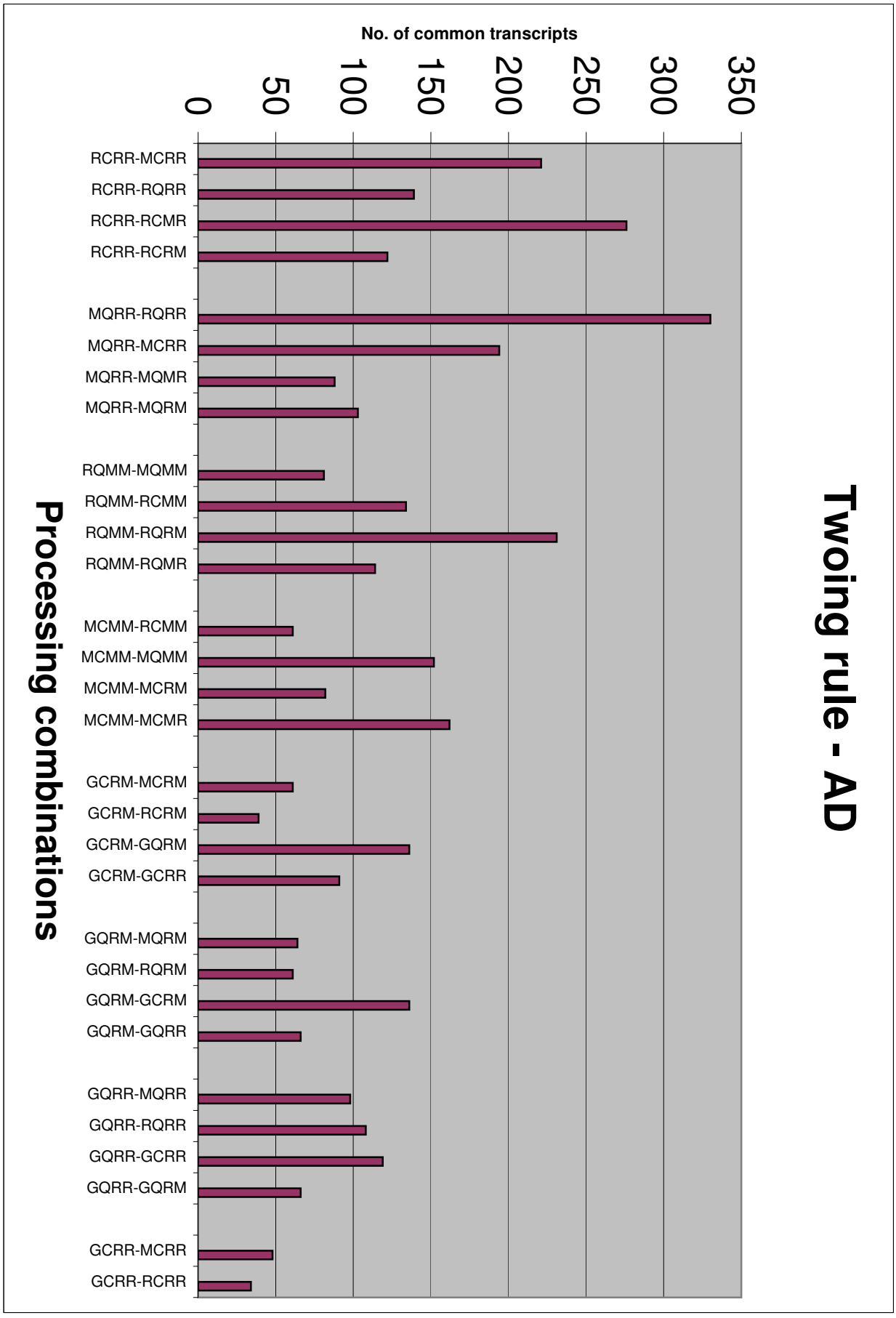
Processing Method	Background correction(stage 1)	Normalisation(stage 2)	PM correction(stage 3)	Summarisation(stage 4)
RMA (R)	x		x	x
MAS5.0 (M)	x		x	x
GCRMA (G)	x			
Constant (C)		x		
Quantiles (Q)		x		

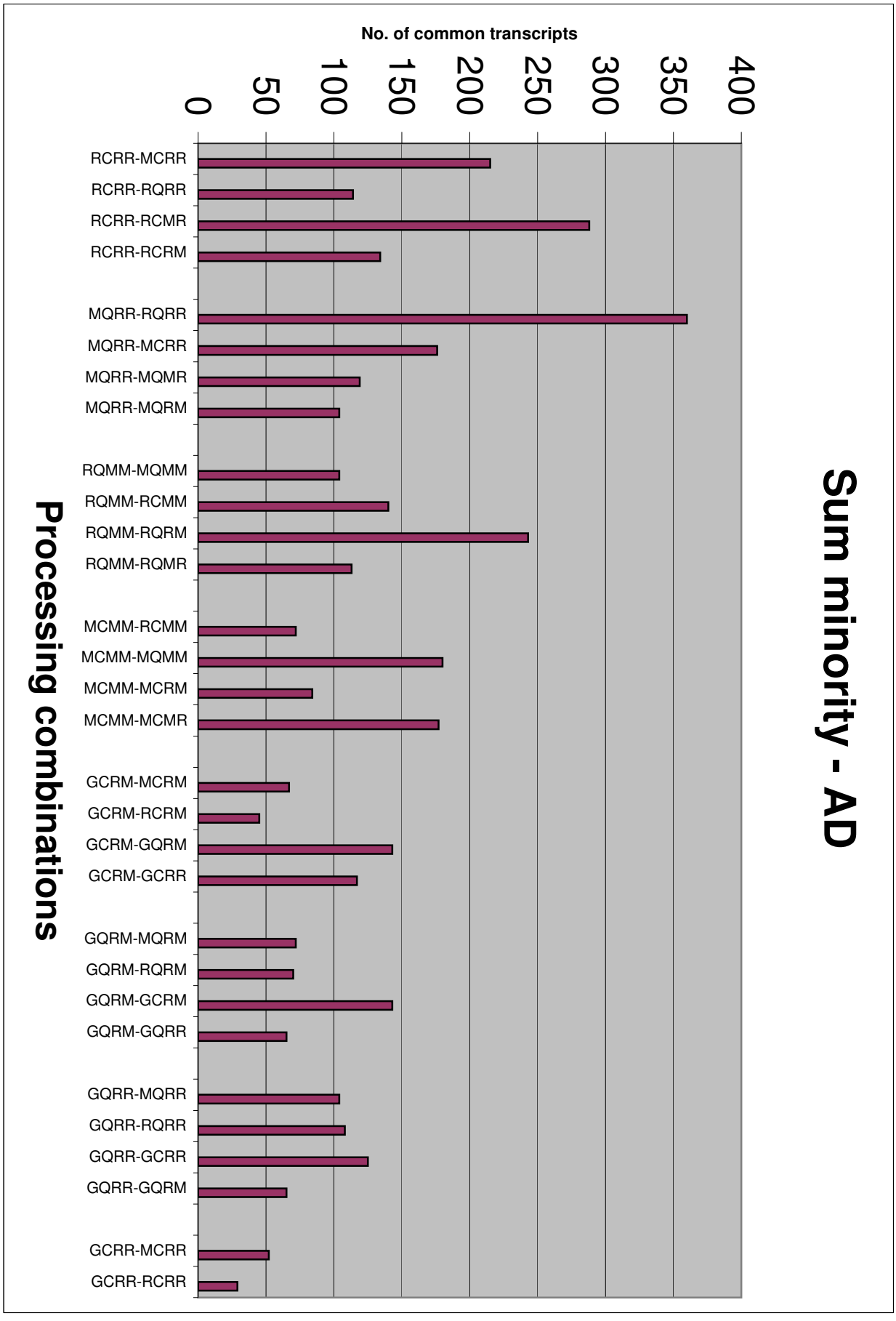
Table 2 - Number of significant transcripts selected by EDGE

Significant transcripts at $p < 0.001$ for breast cancer and $p < 0.01$ for all others. The order of preprocessing is first background correction, then normalisation, followed by PM correction and finally summarisation. These four stages refer to the four letters in each preprocessing combination in order. R refers to RMA, C refers to constant, Q refers to quantile, M refers to MAS5.0, and G refers to GCRMA.

Preprocessing Combination	AD	Breast Cancer	CML	DMD
RCRR	115	567	110	1370
RQRR	61	966	35	2224
MQRR	41	680	28	2364
RQMM	65	517	37	1631
MCMM	76	1352	52	1821
RCMM	104	573	215	878
MCRR	67	644	17	1682
MQMM	86	476	54	2004
RCRM	99	644	182	1483
RQRM	61	746	18	1885
MQRM	67	573	29	1848
RQMR	44	1194	16	2230
MCMR	66	752	80	1683
RCMR	135	899	136	1057
MCRM	59	672	22	1419
MQMR	56	543	75	2134
GCRM	73	1593	493	1514
GCRR	142	1506	178	1559
GQRM	90	3181	45	2142
GQRR	44	5156	17	2084


Information gain - AD
No. of common transcripts
0
50
100
150
200
250
300
350
Processing combinations
RCRR-MCRR
RCRR-RQRR
RCRR-RCMR
RCRR-RCRM
MQRR-RQRR
MQRR-MCRR
MQRR-MQMR
MQRR-MQRM
RQMM-MQMM
RQMM-RCMM
RQMM-RQRM
RQMM-RQMR
MCMM-RCMM
MCMM-MQMM
MCMM-MCRM
MCMM-MCMR
GCRM-MCRM
GCRM-RCRM
GCRM-GQRM
GCRM-GCRR
GQRM-MQRM
GQRM-RQRM
GQRM-GCRM
GQRM-GQRR
GQRR-MQRR
GQRR-RQRR
GQRR-GCRR
GQRR-GQRM
GCRR-MCRR
GCRR-RCRR


Twoing rule - AD
No. of common transcripts
0
50
100
150
200
250
300
350
Processing combinations
RCRR-MCRR
RCRR-RQRR
RCRR-RCMR
RCRR-RCRM
MQRR-RQRR
MQRR-MCRR
MQRR-MQMR
MQRR-MQRM
RQMM-MQMM
RQMM-RCMM
RQMM-RQRM
RQMM-RQMR
MCMM-RCMM
MCMM-MQMM
MCMM-MCRM
MCMM-MCMR
GCRM-MCRM
GCRM-RCRM
GCRM-GQRM
GCRM-GCRR
GQRM-MQRM
GQRM-RQRM
GQRM-GCRM
GQRM-GQRR
GQRR-MQRR
GQRR-RQRR
GQRR-GCRR
GQRR-GQRM
GCRR-MCRR
GCRR-RCRR


Sum minority - AD
No. of common transcripts
0
50
100
150
200
250
300
350
400
Processing combinations
RCRR-MCRR
RCRR-RQRR
RCRR-RCMR
RCRR-RCRM
MQRR-RQRR
MQRR-MCRR
MQRR-MQMR
MQRR-MQRM
RQMM-MQMM
RQMM-RCMM
RQMM-RQRM
RQMM-RQMR
MCMM-RCMM
MCMM-MQMM
MCMM-MCRM
MCMM-MCMR
GCRM-MCRM
GCRM-RCRM
GCRM-GQRM
GCRM-GCRR
GQRM-MQRM
GQRM-RQRM
GQRM-GCRM
GQRM-GQRR
GQRR-MQRR
GQRR-RQRR
GQRR-GCRR
GQRR-GQRM
GCRR-MCRR
GCRR-RCRR

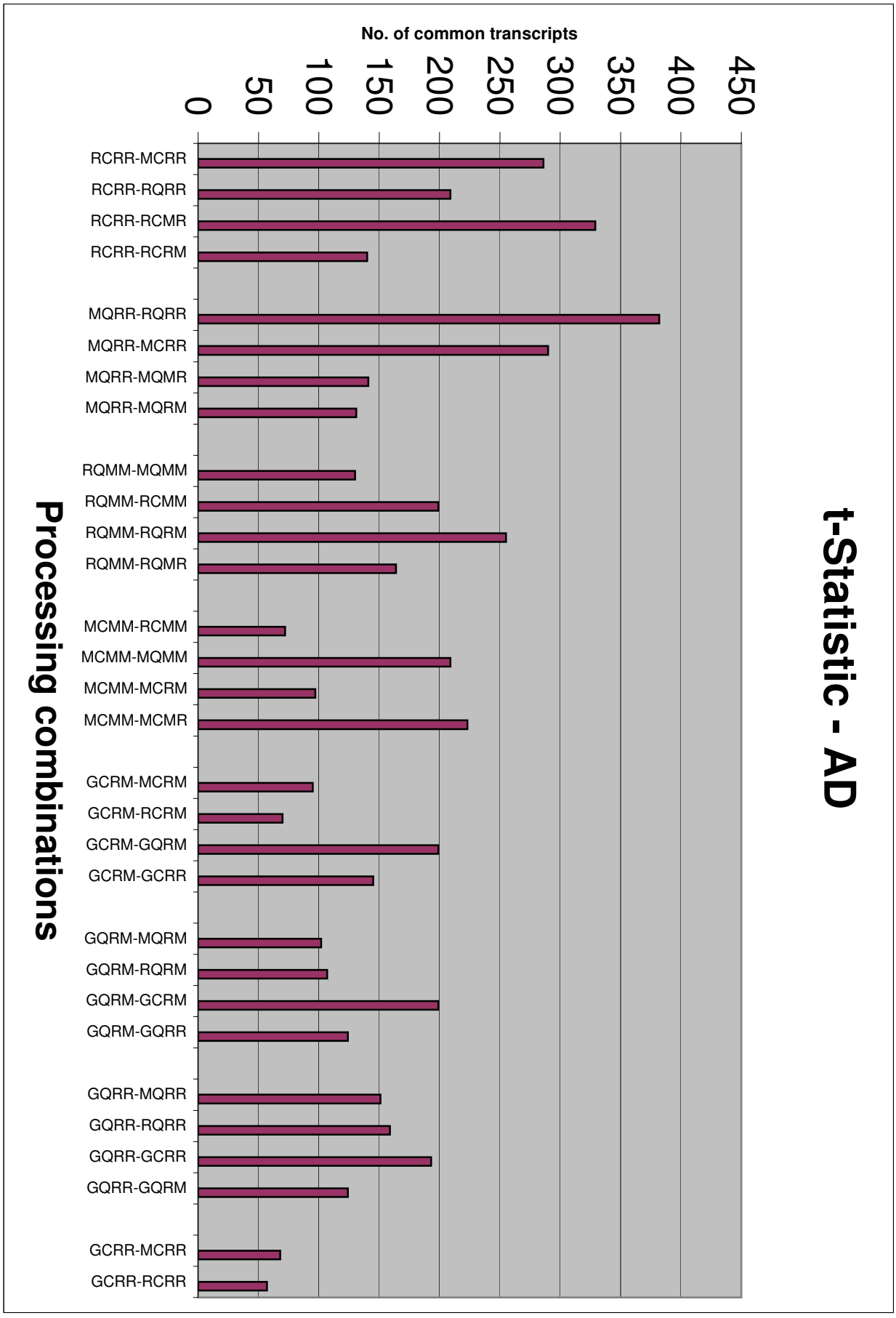
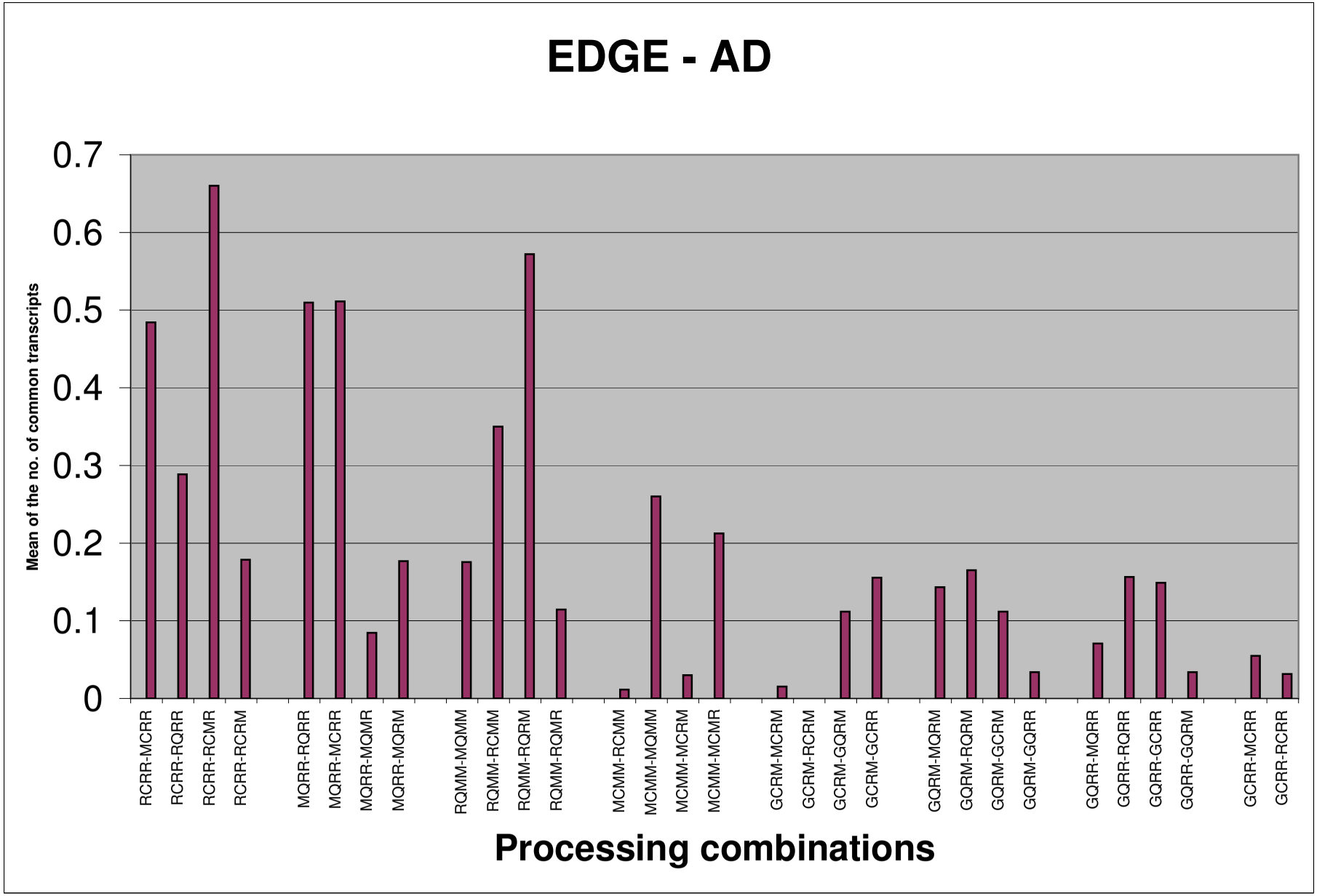
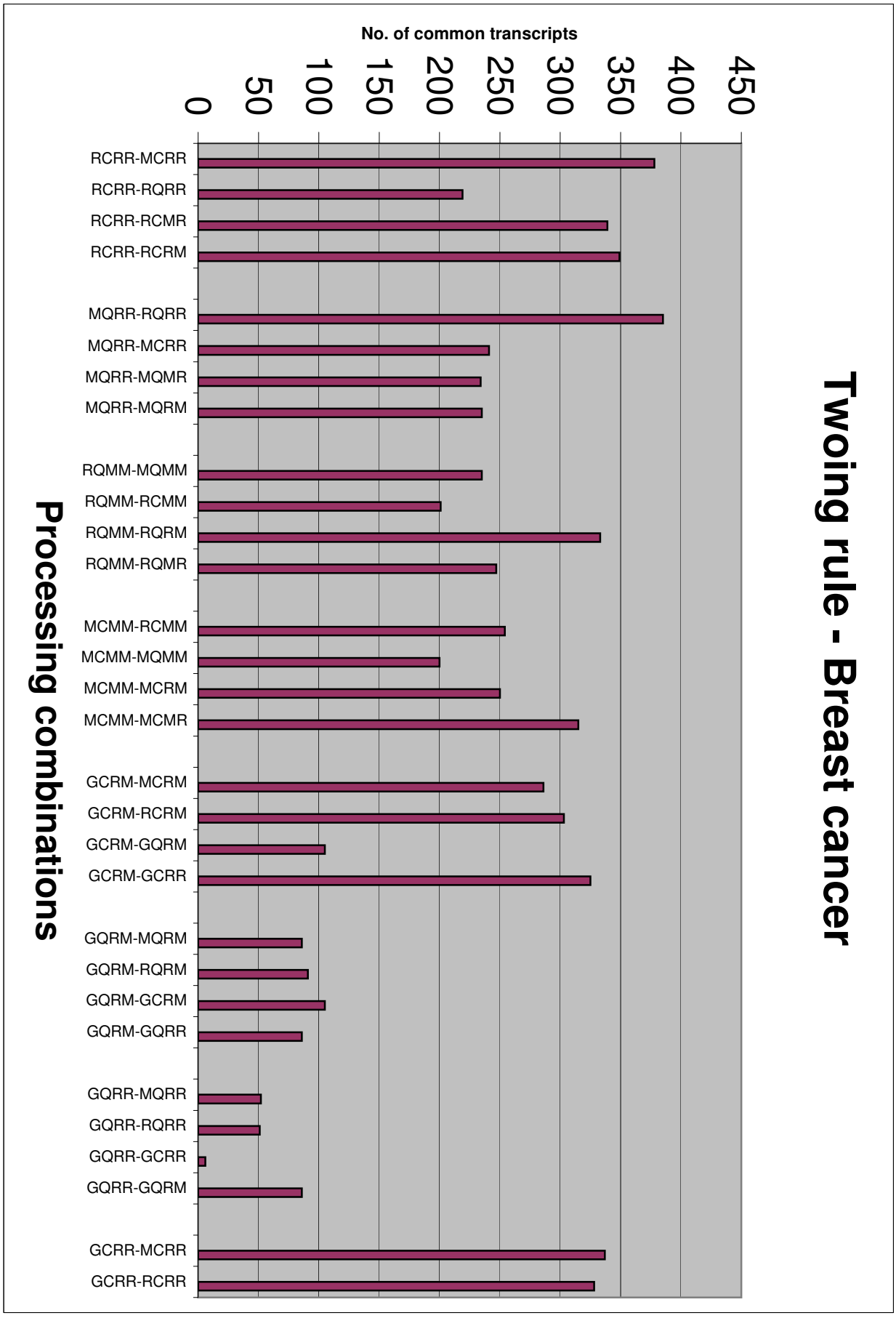
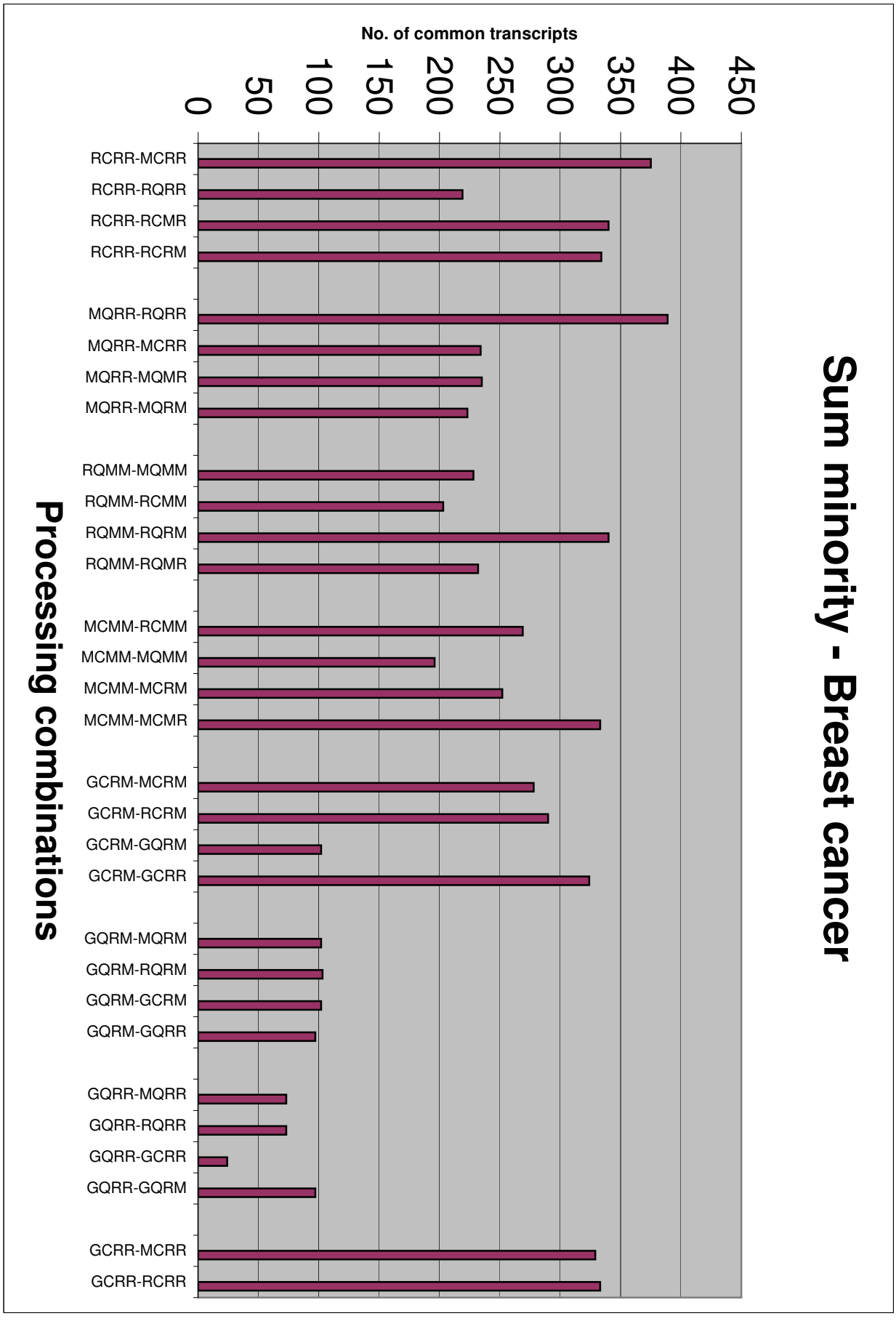

t-Statistic - AD
No. of common transcripts
0
50
100
150
200
250
300
350
400
450
Processing combinations
RCRR-MCRR
RCRR-RQRR
RCRR-RCMR
RCRR-RCRM
MQRR-RQRR
MQRR-MCRR
MQRR-MQMR
MQRR-MQRM
RQMM-MQMM
RQMM-RCMM
RQMM-RQRM
RQMM-RQMR
MCMM-RCMM
MCMM-MQMM
MCMM-MCRM
MCMM-MCMR
GCRM-MCRM
GCRM-RCRM
GCRM-GQRM
GCRM-GCRR
GQRM-MQRM
GQRM-RQRM
GQRM-GCRM
GQRM-GQRR
GQRR-MQRR
GQRR-RQRR
GQRR-GCRR
GQRR-GQRM
GCRR-MCRR
GCRR-RCRR

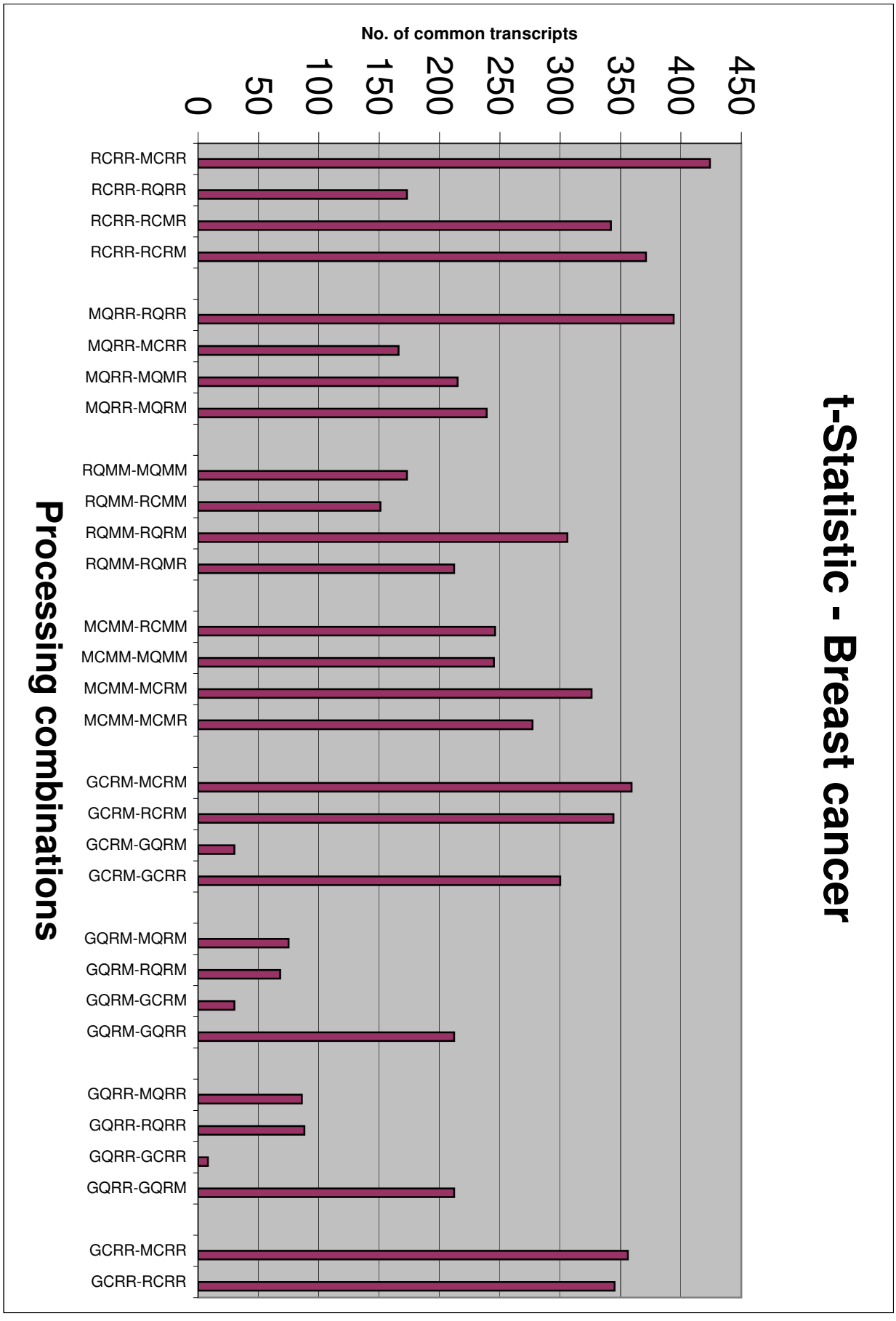
Figure 5

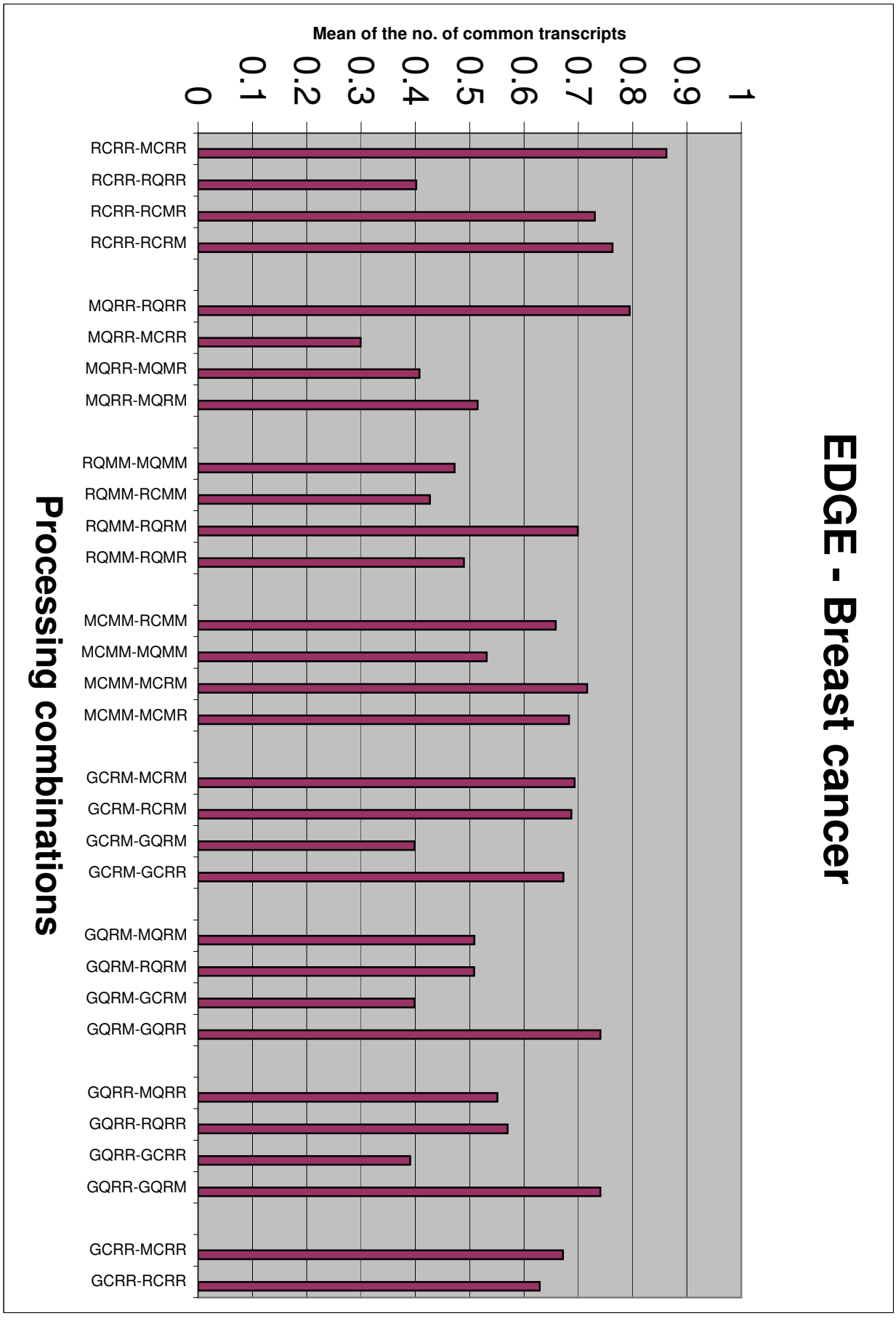

EDGE - AD
Mean of the no. of common transcripts
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
RCRR-MCRR
RCRR-RQRR
RCRR-RCMR
RCRR-RCRM
MQRR-RQRR
MQRR-MCRR
MQRR-MQMR
MQRR-MQRM
RQMM-MQMM
RQMM-RCMM
RQMM-RQRM
RQMM-RQMR
MCMM-RCMM
MCMM-MQMM
MCMM-MCRM
MCMM-MCMR
GCRM-MCRM
GCRM-RCRM
GCRM-GQRM
GCRM-GCRR
GQRM-MQRM
GQRM-RQRM
GQRM-GCRM
GQRM-GQRR
GQRR-MQRR
GQRR-RQRR
GQRR-GCRR
GQRR-GQRM
GCRR-MCRR
GCRR-RCRR
Processing combinations

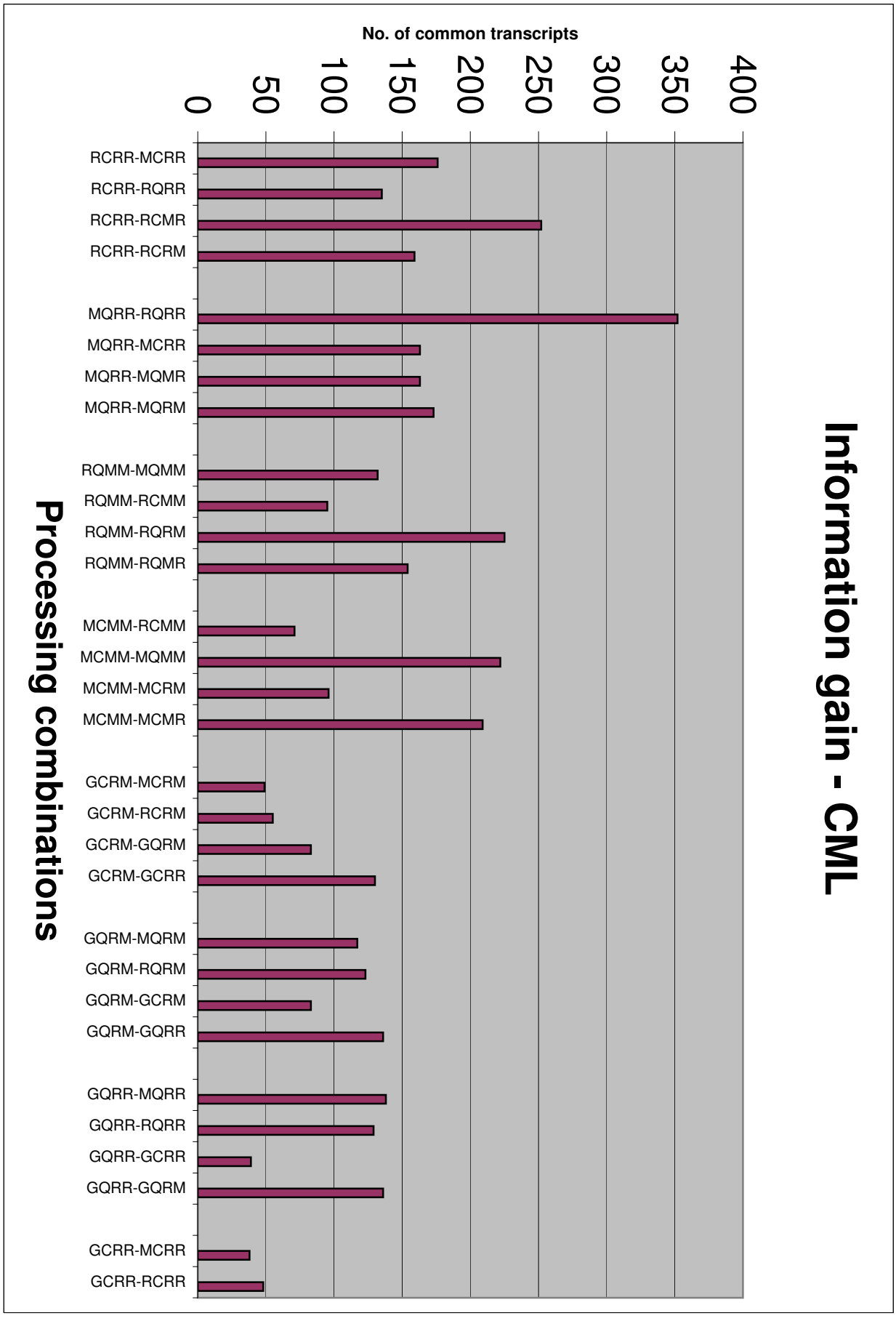
Information gain - Breast cancer
No. of common transcripts
0
50
100
150
200
250
300
350
400
Processing combinations
RCRR-MCRR
RCRR-RQRR
RCRR-RCMR
RCRR-RCRM
MQRR-RQRR
MQRR-MCRR
MQRR-MQMR
MQRR-MQRM
RQMM-MQMM
RQMM-RCMM
RQMM-RQRM
RQMM-RQMR
MCMM-RCMM
MCMM-MQMM
MCMM-MCRM
MCMM-MCMR
GCRM-MCRM
GCRM-RCRM
GCRM-GQRM
GCRM-GCRR
GQRM-MQRM
GQRM-RQRM
GQRM-GCRM
GQRM-GQRR
GQRR-MQRR
GQRR-RQRR
GQRR-GCRR
GQRR-GQRM
GCRR-MCRR
GCRR-RCRR

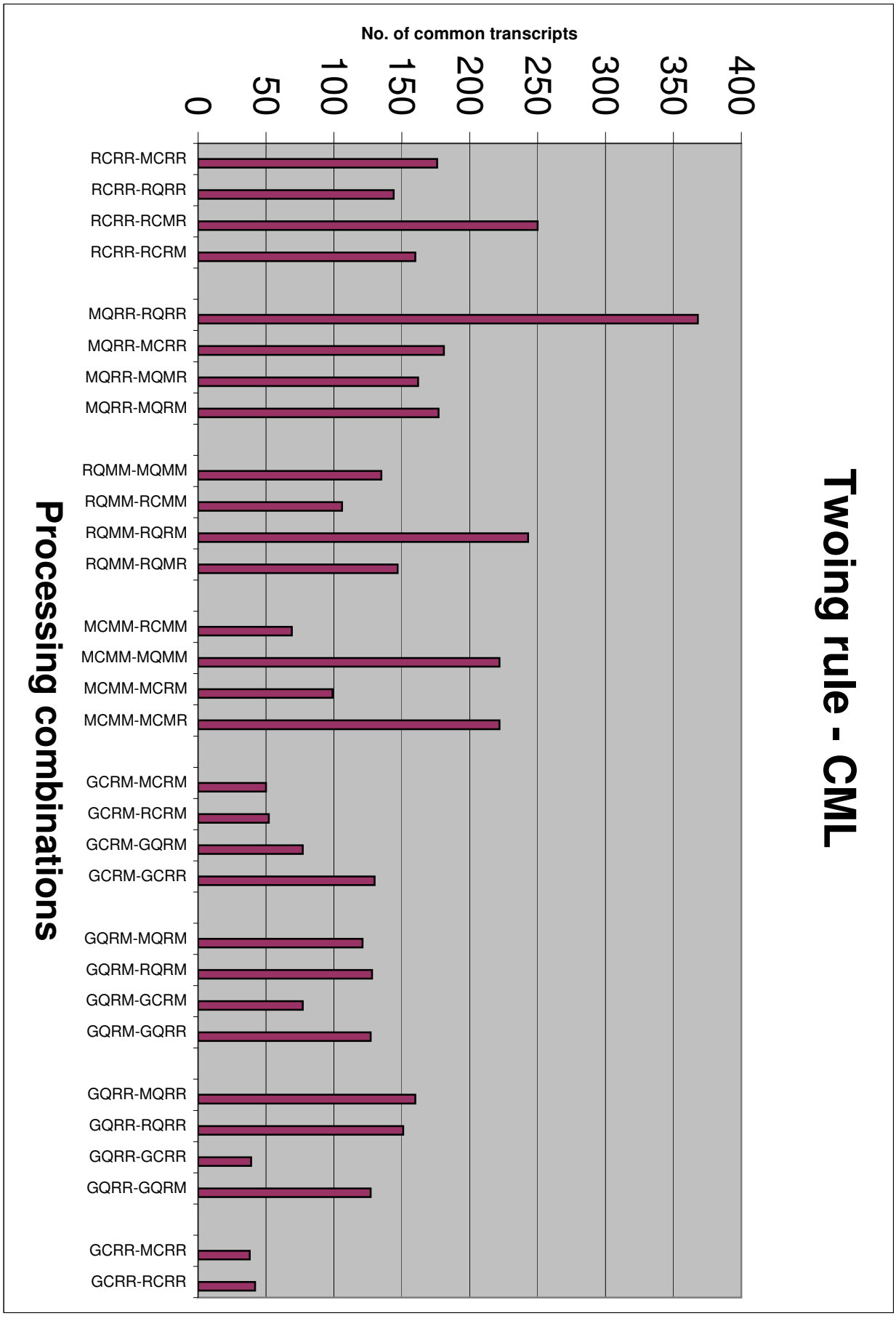

Twoing rule - Breast cancer
No. of common transcripts
0
50
100
150
200
250
300
350
400
450
Processing combinations
RCRR-MCRR
RCRR-RQRR
RCRR-RCMR
RCRR-RCRM
MQRR-RQRR
MQRR-MCRR
MQRR-MQMR
MQRR-MQRM
RQMM-MQMM
RQMM-RCMM
RQMM-RQRM
RQMM-RQMR
MCMM-RCMM
MCMM-MQMM
MCMM-MCRM
MCMM-MCMR
GCRM-MCRM
GCRM-RCRM
GCRM-GQRM
GCRM-GCRR
GQRM-MQRM
GQRM-RQRM
GQRM-GCRM
GQRM-GQRR
GQRR-MQRR
GQRR-RQRR
GQRR-GCRR
GQRR-GQRM
GCRR-MCRR
GCRR-RCRR

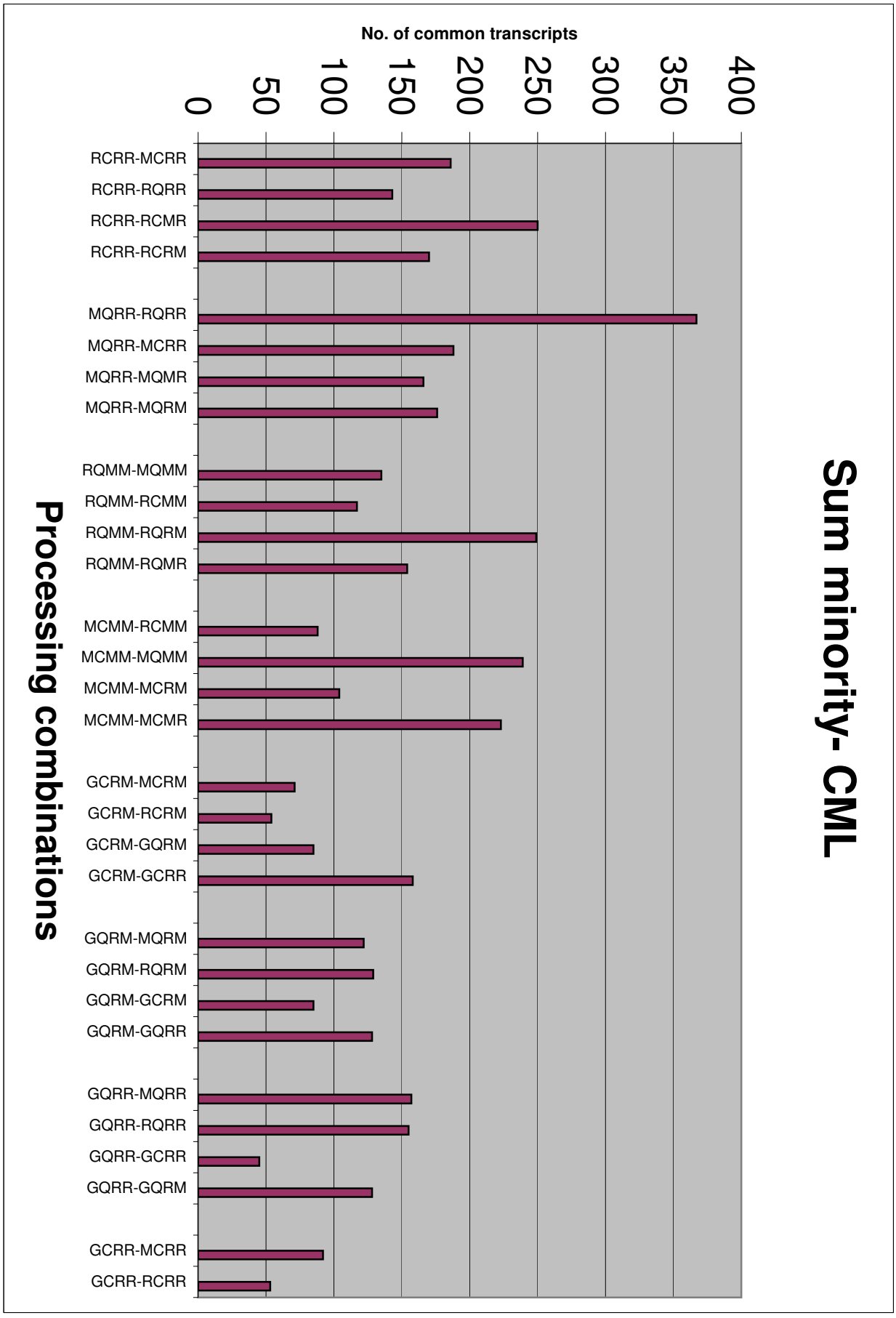

Sum minority - Breast cancer
No. of common transcripts
0
50
100
150
200
250
300
350
400
450
Processing combinations
RCRR-MCRR
RCRR-RQRR
RCRR-RCMR
RCRR-RCRM
MQRR-RQRR
MQRR-MCRR
MQRR-MQMR
MQRR-MQRM
RQMM-MQMM
RQMM-RCMM
RQMM-RQRM
RQMM-RQMR
MCMM-RCMM
MCMM-MQMM
MCMM-MCRM
MCMM-MCMR
GCRM-MCRM
GCRM-RCRM
GCRM-GQRM
GCRM-GCRR
GQRM-MQRM
GQRM-RQRM
GQRM-GCRM
GQRM-GQRR
GQRR-MQRR
GQRR-RQRR
GQRR-GCRR
GQRR-GQRM
GCRR-MCRR
GCRR-RCRR

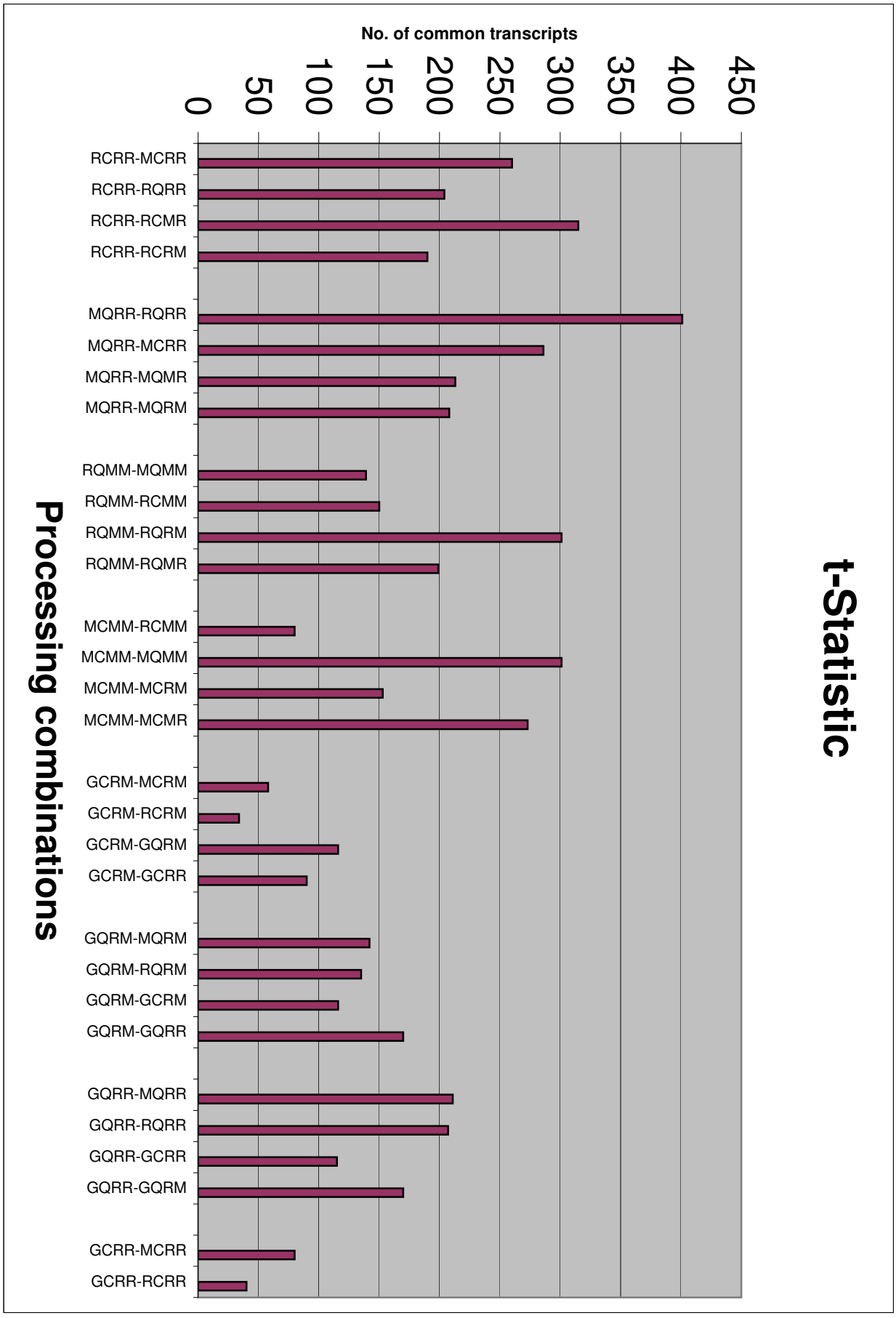

t-Statistic - Breast cancer
No. of common transcripts
0
50
100
150
200
250
300
350
400
450
Processing combinations
RCRR-MCRR
RCRR-RQRR
RCRR-RCMR
RCRR-RCRM
MQRR-RQRR
MQRR-MCRR
MQRR-MQMR
MQRR-MQRM
RQMM-MQMM
RQMM-RCMM
RQMM-RQRM
RQMM-RQMR
MCMM-RCMM
MCMM-MQMM
MCMM-MCRM
MCMM-MCMR
GCRM-MCRM
GCRM-RCRM
GCRM-GQRM
GCRM-GCRR
GQRM-MQRM
GQRM-RQRM
GQRM-GCRM
GQRM-GQRR
GQRR-MQRR
GQRR-RQRR
GQRR-GCRR
GQRR-GQRM
GCRR-MCRR
GCRR-RCRR


EDGE - Breast cancer
Mean of the no. of common transcripts
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
Processing combinations
RCRR-MCRR
RCRR-RQRR
RCRR-RCMR
RCRR-RCRM
MQRR-RQRR
MQRR-MCRR
MQRR-MQMR
MQRR-MQRM
RQMM-MQMM
RQMM-RCMM
RQMM-RQRM
RQMM-RQMR
MCMM-RCMM
MCMM-MQMM
MCMM-MCRM
MCMM-MCMR
GCRM-MCRM
GCRM-RCRM
GCRM-GQRM
GCRM-GCRR
GQRM-MQRM
GQRM-RQRM
GQRM-GCRM
GQRM-GQRR
GQRR-MQRR
GQRR-RQRR
GQRR-GCRR
GQRR-GQRM
GCRR-MCRR
GCRR-RCRR


Information gain - CML
No. of common transcripts
0
50
100
150
200
250
300
350
400
Processing combinations
RCRR-MCRR
RCRR-RQRR
RCRR-RCMR
RCRR-RCRM
MQRR-RQRR
MQRR-MCRR
MQRR-MQMR
MQRR-MQRM
RQMM-MQMM
RQMM-RCMM
RQMM-RQRM
RQMM-RQMR
MCMM-RCMM
MCMM-MQMM
MCMM-MCRM
MCMM-MCMR
GCRM-MCRM
GCRM-RCRM
GCRM-GQRM
GCRM-GCRR
GQRM-MQRM
GQRM-RQRM
GQRM-GCRM
GQRM-GQRR
GQRR-MQRR
GQRR-RQRR
GQRR-GCRR
GQRR-GQRM
GCRR-MCRR
GCRR-RCRR


Twoing rule - CML
No. of common transcripts
0
50
100
150
200
250
300
350
400
Processing combinations
RCRR-MCRR
RCRR-RQRR
RCRR-RCMR
RCRR-RCRM
MQRR-RQRR
MQRR-MCRR
MQRR-MQMR
MQRR-MQRM
RQMM-MQMM
RQMM-RCMM
RQMM-RQRM
RQMM-RQMR
MCMM-RCMM
MCMM-MQMM
MCMM-MCRM
MCMM-MCMR
GCRM-MCRM
GCRM-RCRM
GCRM-GQRM
GCRM-GCRR
GQRM-MQRM
GQRM-RQRM
GQRM-GCRM
GQRM-GQRR
GQRR-MQRR
GQRR-RQRR
GQRR-GCRR
GQRR-GQRM
GCRR-MCRR
GCRR-RCRR


Sum minority- CML
No. of common transcripts
0
50
100
150
200
250
300
350
400
Processing combinations
RCRR-MCRR
RCRR-RQRR
RCRR-RCMR
RCRR-RCRM
MQRR-RQRR
MQRR-MCRR
MQRR-MQMR
MQRR-MQRM
RQMM-MQMM
RQMM-RCMM
RQMM-RQRM
RQMM-RQMR
MCMM-RCMM
MCMM-MQMM
MCMM-MCRM
MCMM-MCMR
GCRM-MCRM
GCRM-RCRM
GCRM-GQRM
GCRM-GCRR
GQRM-MQRM
GQRM-RQRM
GQRM-GCRM
GQRM-GQRR
GQRR-MQRR
GQRR-RQRR
GQRR-GCRR
GQRR-GQRM
GCRR-MCRR
GCRR-RCRR


No. of common transcripts
0
50
100
150
200
250
300
350
400
450
RCRR-MCRR
RCRR-RQRR
RCRR-RCMR
RCRR-RCRM
MQRR-RQRR
MQRR-MCRR
MQRR-MQMR
MQRR-MQRM
RQMM-MQMM
RQMM-RCMM
RQMM-RQRM
RQMM-RQMR
MCMM-RCMM
MCMM-MQMM
MCMM-MCRM
MCMM-MCMR
GCRM-MCRM
GCRM-RCRM
GCRM-GQRM
GCRM-GCRR
GQRM-MQRM
GQRM-RQRM
GQRM-GCRM
GQRM-GQRR
GQRR-MQRR
GQRR-RQRR
GQRR-GCRR
GQRR-GQRM
GCRR-MCRR
GCRR-RCRR
Processing combinations
t-Statistic

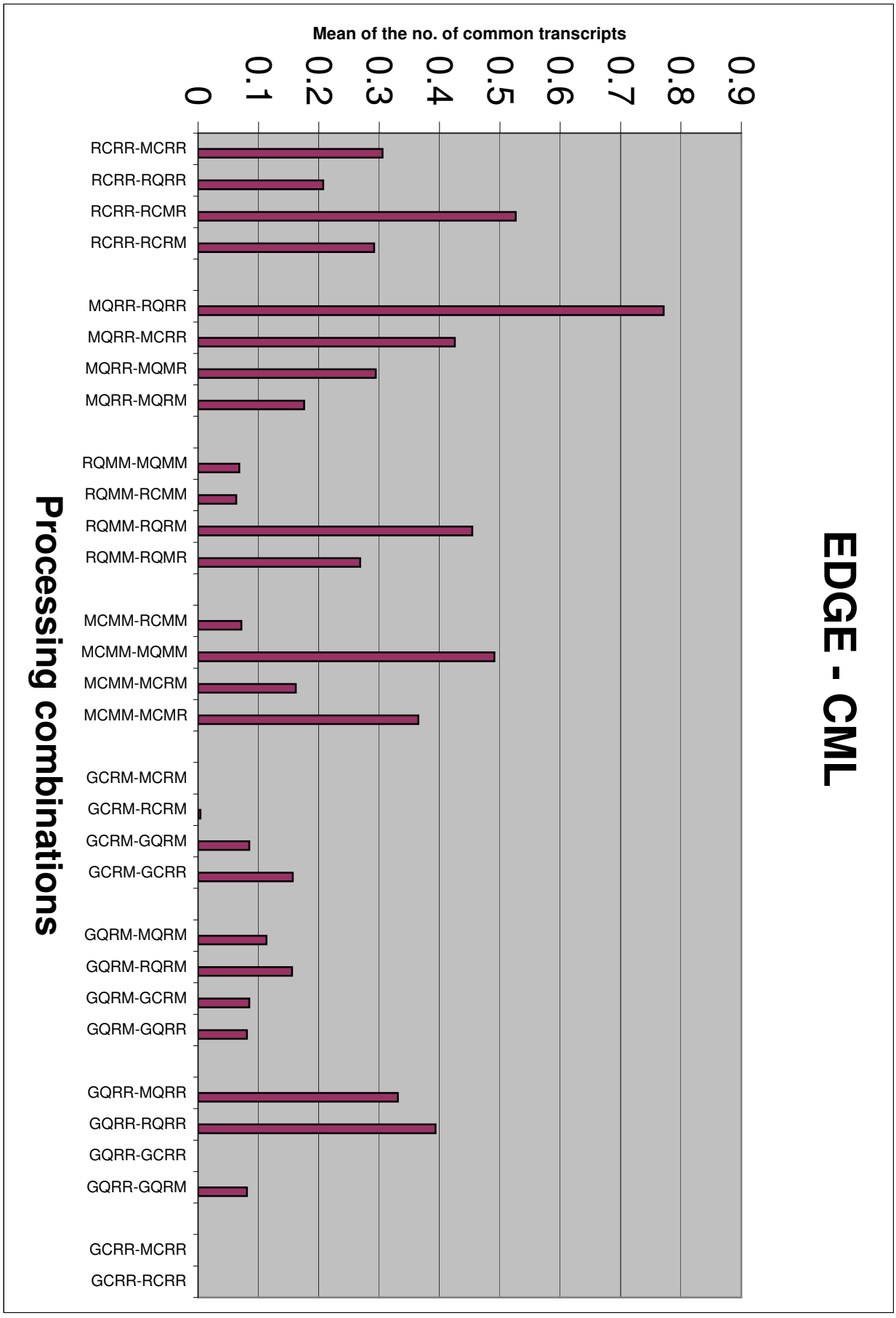
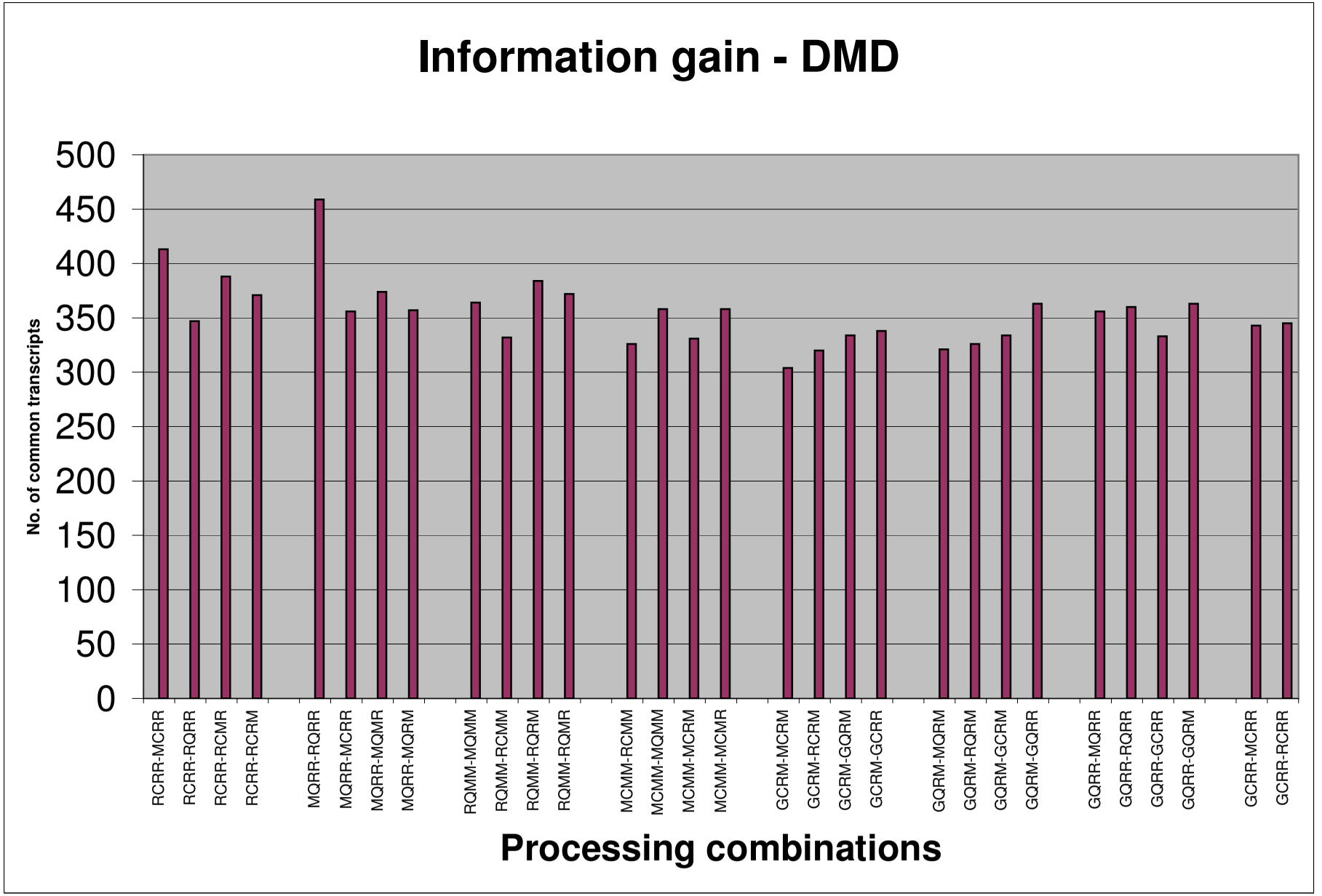
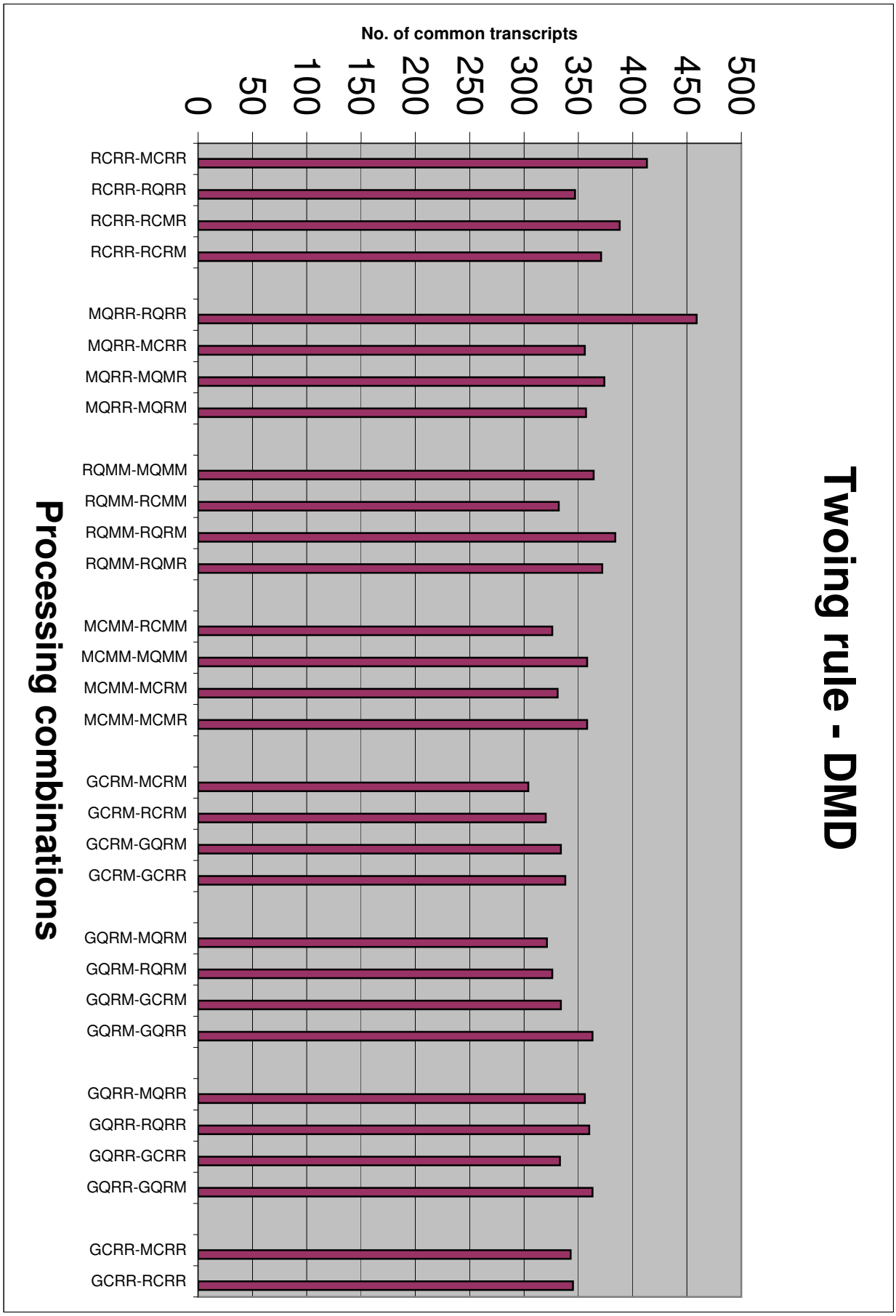
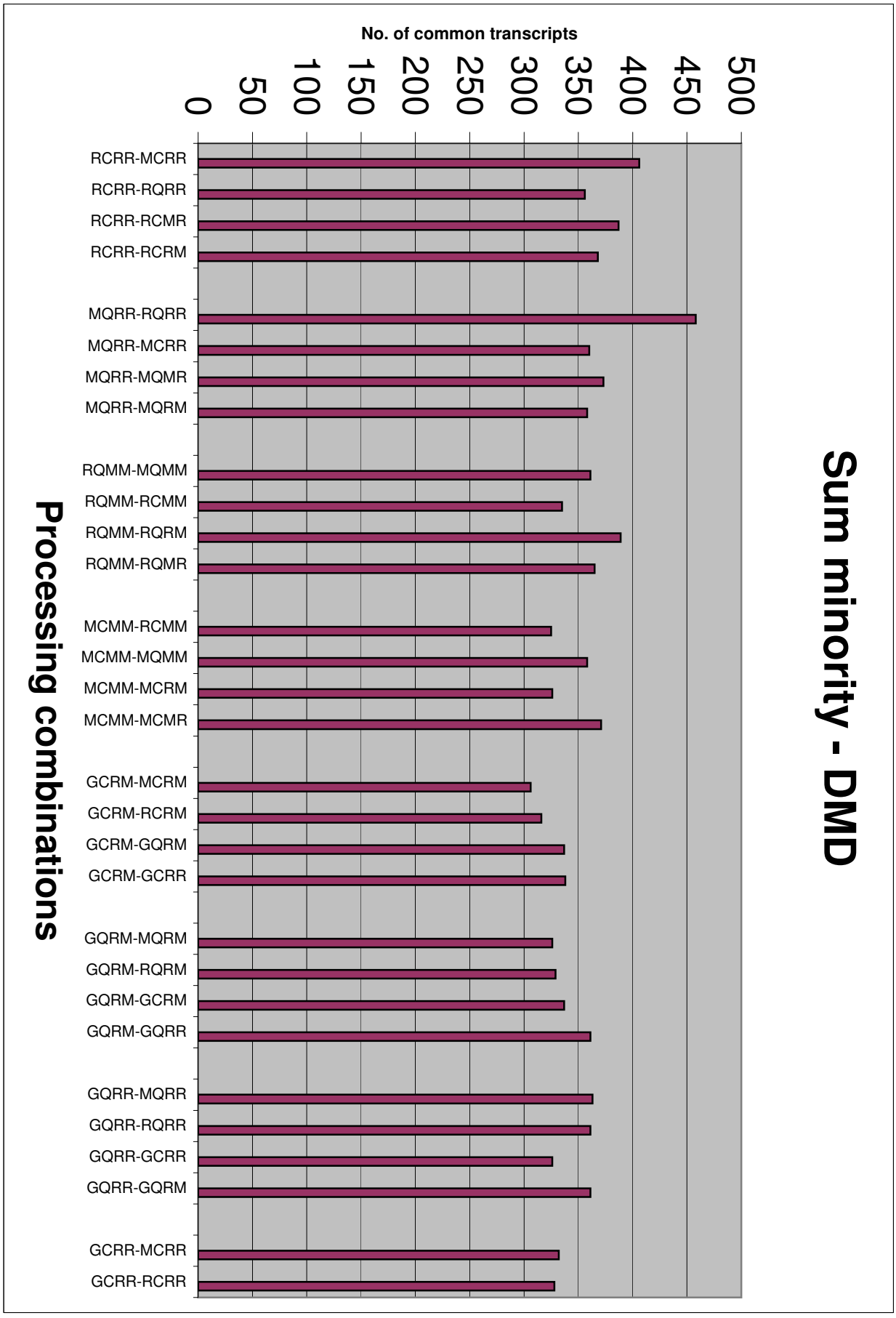
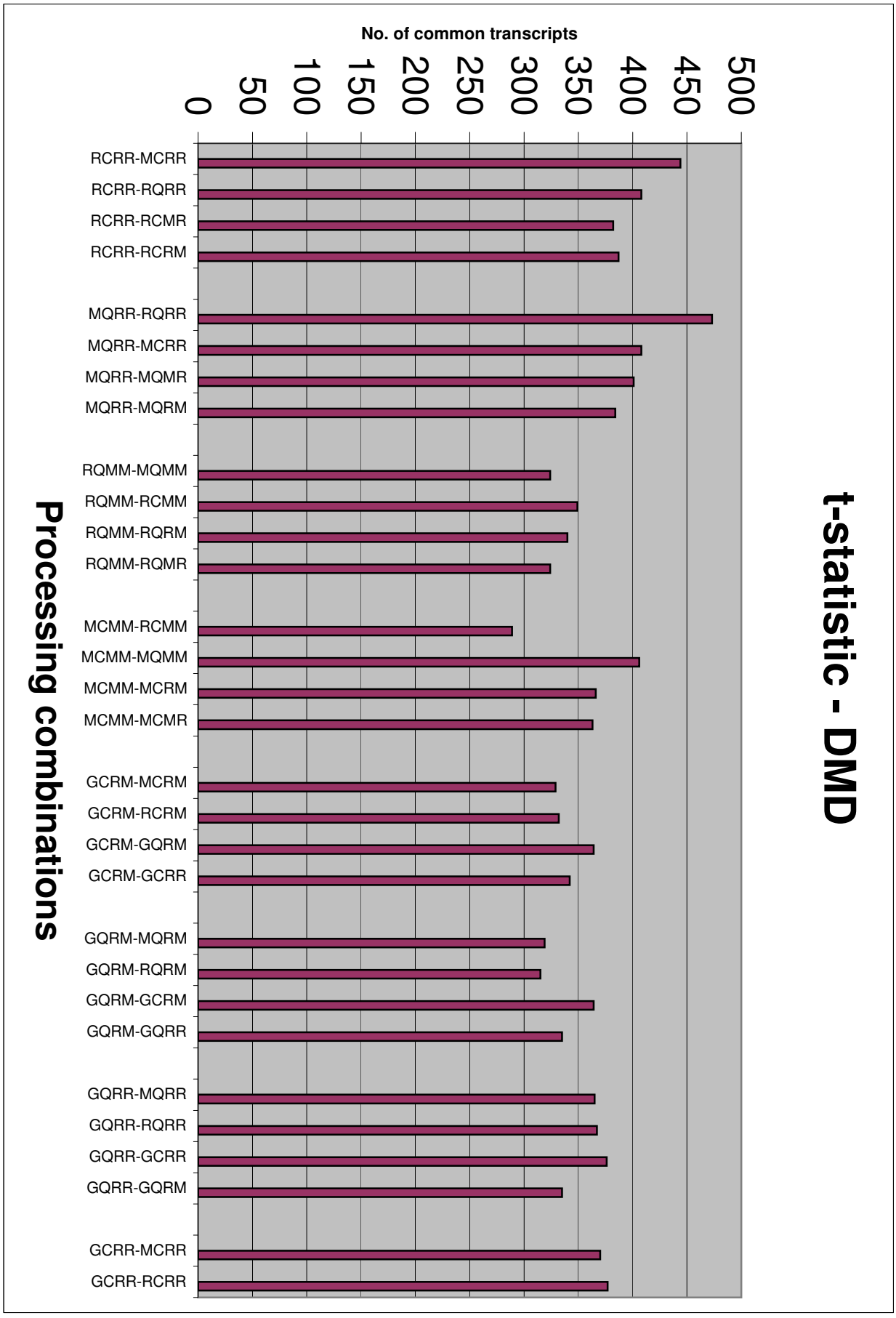

EDGE - CML
Mean of the no. of common transcripts
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
Processing combinations
RCRR-MCRR
RCRR-RQRR
RCRR-RCMR
RCRR-RCRM
MQRR-RQRR
MQRR-MCRR
MQRR-MQMR
MQRR-MQRM
RQMM-MQMM
RQMM-RCMM
RQMM-RQRM
RQMM-RQMR
MCMM-RCMM
MCMM-MQMM
MCMM-MCRM
MCMM-MCMR
GCRM-MCRM
GCRM-RCRM
GCRM-GQRM
GCRM-GCRR
GQRM-MQRM
GQRM-RQRM
GQRM-GCRM
GQRM-GQRR
GQRR-MQRR
GQRR-RQRR
GQRR-GCRR
GQRR-GQRM
GCRR-MCRR
GCRR-RCRR

Figure 16


Information gain - DMD
No. of common transcripts
500
450
400
350
300
250
200
150
100
50
0
RCRR-MCRR
RCRR-RQRR
RCRR-RCMR
RCRR-RCRM
MQRR-RQRR
MQRR-MCRR
MQRR-MQMR
MQRR-MQRM
RQMM-MQMM
RQMM-RCMM
RQMM-RQRM
RQMM-RQMR
MCMM-RCMM
MCMM-MQMM
MCMM-MCRM
MCMM-MCMR
GCRM-MCRM
GCRM-RCRM
GCRM-GQRM
GCRM-GCRR
GQRM-MQRM
GQRM-RQRM
GQRM-GCRM
GQRM-GQRR
GQRR-MQRR
GQRR-RQRR
GQRR-GCRR
GQRR-GQRM
GCRR-MCRR
GCRR-RCRR
Processing combinations


Twoing rule - DMD
No. of common transcripts
0
50
100
150
200
250
300
350
400
450
500
Processing combinations
RCRR-MCRR
RCRR-RQRR
RCRR-RCMR
RCRR-RCRM
MQRR-RQRR
MQRR-MCRR
MQRR-MQMR
MQRR-MQRM
RQMM-MQMM
RQMM-RCMM
RQMM-RQRM
RQMM-RQMR
MCMM-RCMM
MCMM-MQMM
MCMM-MCRM
MCMM-MCMR
GCRM-MCRM
GCRM-RCRM
GCRM-GQRM
GCRM-GCRR
GQRM-MQRM
GQRM-RQRM
GQRM-GCRM
GQRM-GQRR
GQRR-MQRR
GQRR-RQRR
GQRR-GCRR
GQRR-GQRM
GCRR-MCRR
GCRR-RCRR


Sum minority - DMD
No. of common transcripts
0
50
100
150
200
250
300
350
400
450
500
Processing combinations
RCRR-MCRR
RCRR-RQRR
RCRR-RCMR
RCRR-RCRM
MQRR-RQRR
MQRR-MCRR
MQRR-MQMR
MQRR-MQRM
RQMM-MQMM
RQMM-RCMM
RQMM-RQRM
RQMM-RQMR
MCMM-RCMM
MCMM-MQMM
MCMM-MCRM
MCMM-MCMR
GCRM-MCRM
GCRM-RCRM
GCRM-GQRM
GCRM-GCRR
GQRM-MQRM
GQRM-RQRM
GQRM-GCRM
GQRM-GQRR
GQRR-MQRR
GQRR-RQRR
GQRR-GCRR
GQRR-GQRM
GCRR-MCRR
GCRR-RCRR


t-statistic - DMD
No. of common transcripts
0
50
100
150
200
250
300
350
400
450
500
Processing combinations
RCRR-MCRR
RCRR-RQRR
RCRR-RCMR
RCRR-RCRM
MQRR-RQRR
MQRR-MCRR
MQRR-MQMR
MQRR-MQRM
RQMM-MQMM
RQMM-RCMM
RQMM-RQRM
RQMM-RQMR
MCMM-RCMM
MCMM-MQMM
MCMM-MCRM
MCMM-MCMR
GCRM-MCRM
GCRM-RCRM
GCRM-GQRM
GCRM-GCRR
GQRM-MQRM
GQRM-RQRM
GQRM-GCRM
GQRM-GQRR
GQRR-MQRR
GQRR-RQRR
GQRR-GCRR
GQRR-GQRM
GCRR-MCRR
GCRR-RCRR

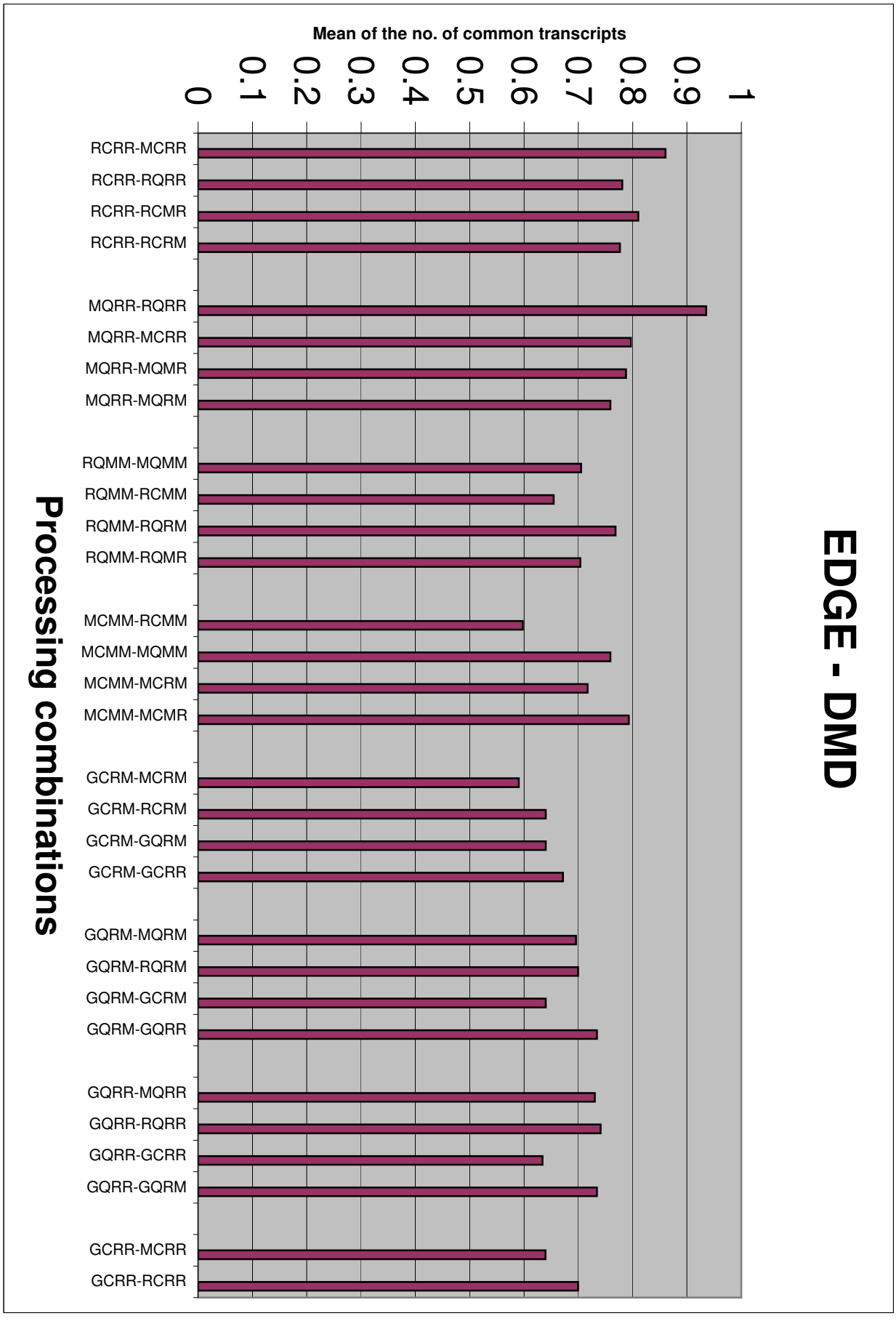

EDGE - DMD
Mean of the no. of common transcripts
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
Processing combinations
RCRR-MCRR
RCRR-RQRR
RCRR-RCMR
RCRR-RCRM
MQRR-RQRR
MQRR-MCRR
MQRR-MQMR
MQRR-MQRM
RQMM-MQMM
RQMM-RCMM
RQMM-RQRM
RQMM-RQMR
MCMM-RCMM
MCMM-MQMM
MCMM-MCRM
MCMM-MCMR
GCRM-MCRM
GCRM-RCRM
GCRM-GQRM
GCRM-GCRR
GQRM-MQRM
GQRM-RQRM
GQRM-GCRM
GQRM-GQRR
GQRR-MQRR
GQRR-RQRR
GQRR-GCRR
GQRR-GQRM
GCRR-MCRR
GCRR-RCRR

Figure 21

EDGE-AD

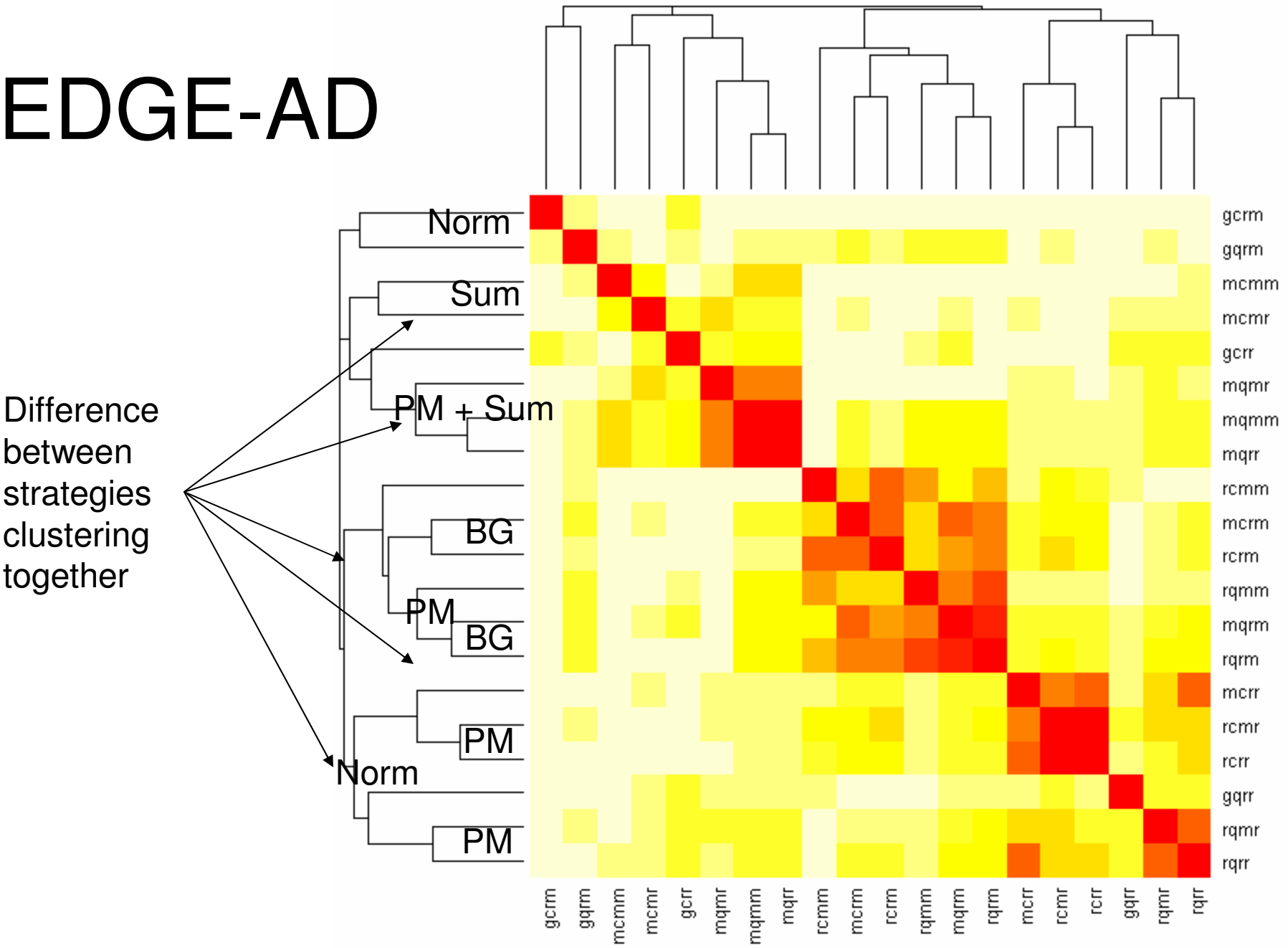


Figure 22

EDGE-BC

Difference between strategies clustering together

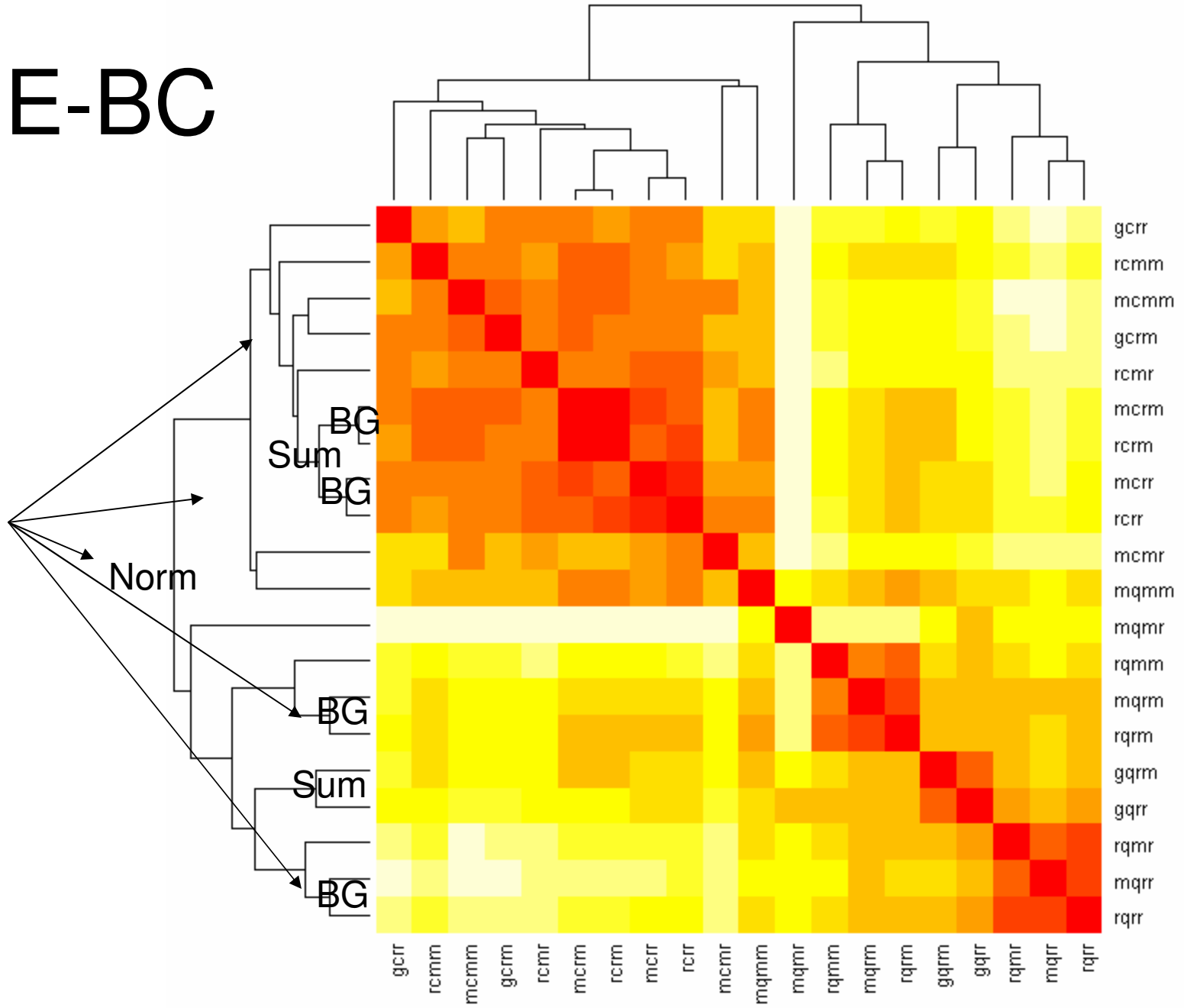


Figure 23

EDGE-CML

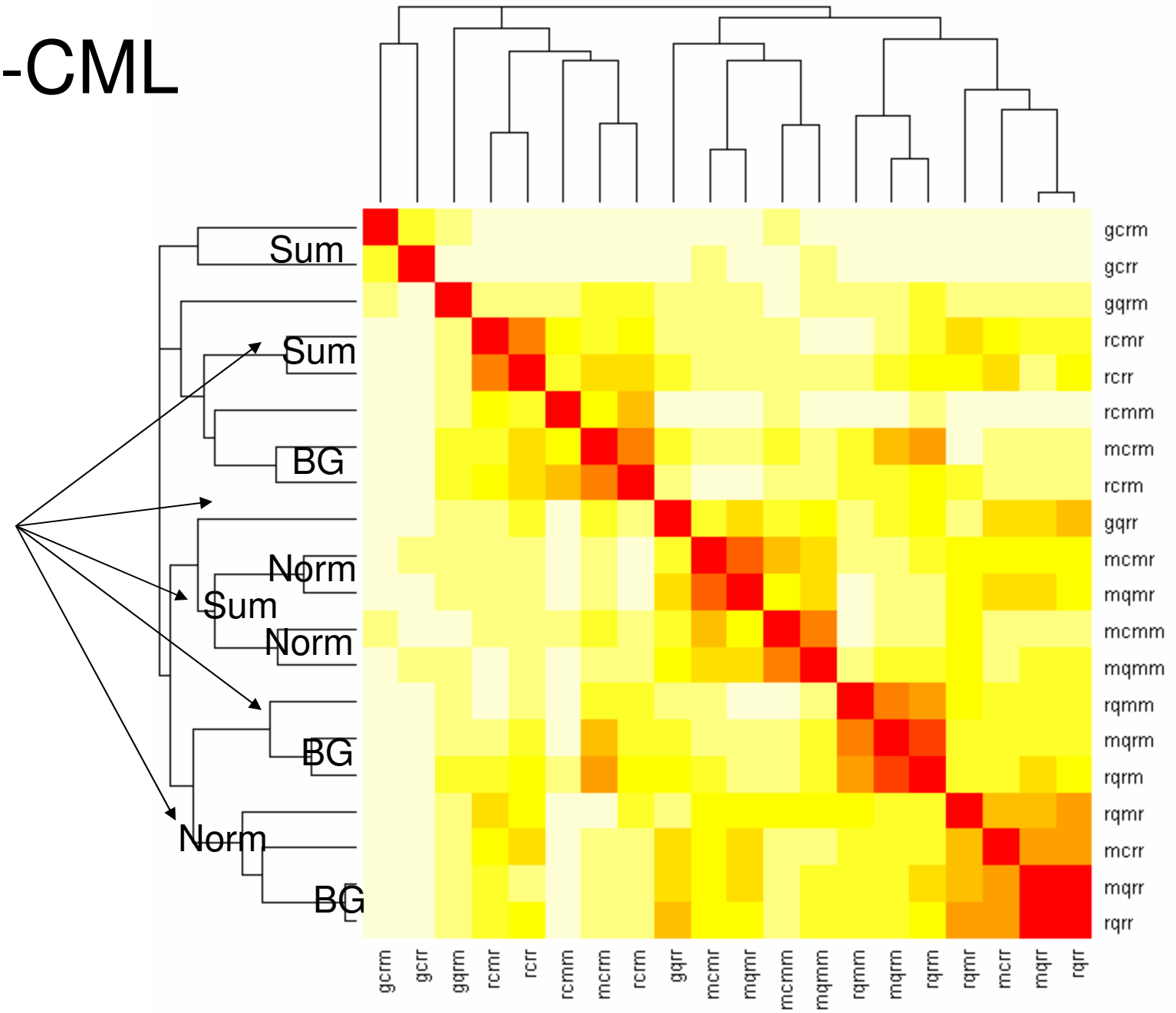
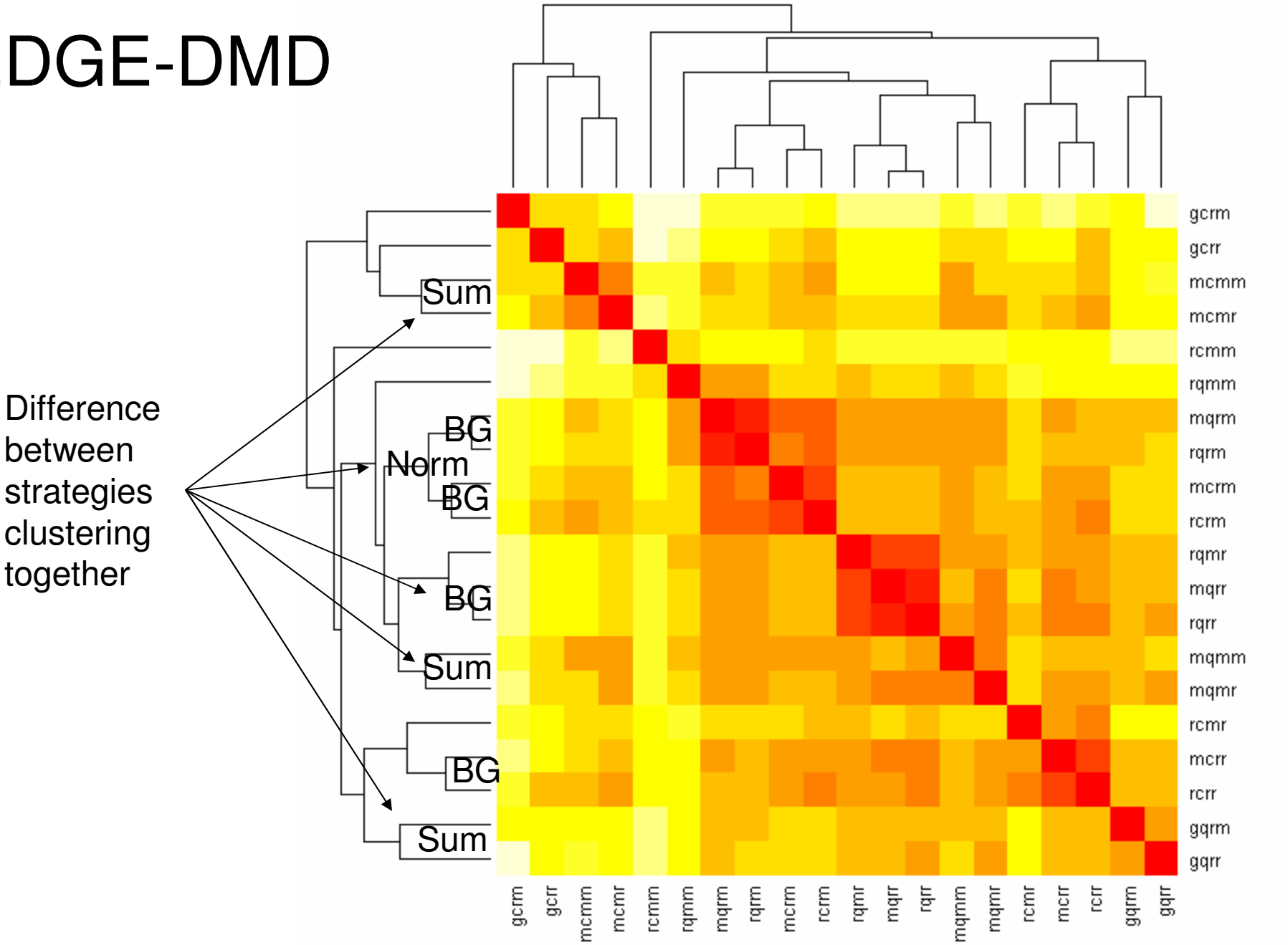


Figure 24

EDGE-DMD



Additional files provided with this submission:

Additional file 20 : dmdrankcriteria7.zip : 84Kb
http://genomebiology.com/imedia/9385023311079128/sup20.ZIP
Additional file 19 : dmdrankcriteria3.zip : 64Kb
http://genomebiology.com/imedia/1845735376107910/sup19.ZIP
Additional file 18 : dmdrankcriteria2.zip : 65Kb
http://genomebiology.com/imedia/3985417621079055/sup18.ZIP
Additional file 17 : dmdrankcriteria1.zip : 65Kb
http://genomebiology.com/imedia/1630983070107905/sup17.ZIP
Additional file 16 : dmdedge.zip : 2395Kb
http://genomebiology.com/imedia/2149892721079055/sup16.ZIP
Additional file 15 : cmlrankcriteria7.zip : 85Kb
http://genomebiology.com/imedia/2072270291107905/sup15.ZIP
Additional file 14 : cmlrankcriteria3.zip : 62Kb
http://genomebiology.com/imedia/6350425251079055/sup14.ZIP
Additional file 13 : cmlrankcriteria2.zip : 67Kb
http://genomebiology.com/imedia/1341764078107905/sup13.ZIP
Additional file 12 : cmlrankcriteria1.zip : 66Kb
http://genomebiology.com/imedia/3237730851079055/sup12.ZIP
Additional file 11 : cmledge.zip : 2218Kb
http://genomebiology.com/imedia/2819477051079061/sup11.ZIP
Additional file 10 : bcrankcriteria7.zip : 86Kb
http://genomebiology.com/imedia/1315001812107906/sup10.ZIP
Additional file 9 : bcrankcriteria3.zip : 61Kb
http://genomebiology.com/imedia/1036691770107905/sup9.ZIP
Additional file 8 : bcrankcriteria2.zip : 68Kb
http://genomebiology.com/imedia/4344430521079061/sup8.ZIP
Additional file 7 : bcrankcriteria1.zip : 68Kb
http://genomebiology.com/imedia/1583733981107898/sup7.ZIP
Additional file 6 : bcedge.zip : 2489Kb
http://genomebiology.com/imedia/1640226850107896/sup6.ZIP
Additional file 5 : adrankgenecriteria7.zip : 88Kb
http://genomebiology.com/imedia/1563496523107896/sup5.ZIP
Additional file 4 : adrankgenecriteria3.zip : 63Kb
http://genomebiology.com/imedia/2529471661078961/sup4.ZIP
Additional file 3 : adrankgenecriteria2.zip : 69Kb
http://genomebiology.com/imedia/7602571611078961/sup3.ZIP
Additional file 2 : adrankgenecriteria1.zip : 69Kb
http://genomebiology.com/imedia/1914150664107896/sup2.ZIP
Additional file 1 : adedge.zip : 3249Kb
http://genomebiology.com/imedia/1798347601078821/sup1.ZIP