Reproducibility Assessment of Independent Component Analysis of Expression Ratios from DNA microarrays.

David Philip Kreil* David J. C. MacKay†

Technical Report

Revision 1.0, compiled 16th October 2002

*Department of Genetics, University of Cambridge, Downing Site, Cambridge CB2 3EH
†Inference Group, Cavendish Laboratory, Madingley Road, Cambridge CB3 0HE
Abstract

Independent Component Analysis (ICA) relies on a non-Gaussian distribution of the underlying latent variables that are to be uncovered. As a consequence, characteristics of ICA algorithms can depend on properties of the examined data, making them specific to the given application domain.

DNA microarrays allow the measurement of transcript activity for thousands of genes in parallel. Using multiple measurement 'channels' per gene, moreover, different samples can directly be compared. Most commonly, a particular sample of interest is studied next to a neutral control. It is then typical to focus on the channel ratios.

In contrast to the very high number of measurement variables (~10^4, scaling with the number of examined genes), an everyday data set will be limited to only a few samples (~10-100). Little is known about the expected distributions of the underlying latent variables. We therefore test results from applications of variational Bayesian ICA in this domain for robustness.

Since it is common in the field to transform or preprocess the data, we first examined alternative transforms and data selections for the smallest modelling reconstruction errors. Logratio data are reconstructed better than non-transformed ratio data by our linear model with a Gaussian error term.

Any comparison of ICA results must allow for its invariance under rescaling and permutation of the extracted signatures, which hold the loadings of the original variables (gene transcript ratios) on particular latent variables. We introduce a method to optimally match corresponding signatures between sets of results.

The stability of signatures is then examined after (1) repetition of the same analysis run with different random number generator seeds, and (2) repetition of the analysis with partial data sets. Both the effect of dropping a proportion of the gene transcript ratios, and of dropping measurements for several samples has been studied. Signatures with a high relative data power were very likely to be retained, resulting in an overall stability of the analyses.
Acknowledgments

We wish to thank Johan Rung (EMBL EBI) for discussion and advice regarding preparation of the datasets used in this study.

D. Kroll acknowledges support by a Medical Research Council Research Training Fellowship (GS/555).
Contents

Abstract ................................................... i
Acknowledgments ....................................... iii
Contents .................................................. v
List of Tables .......................................... vii
List of Figures .......................................... ix

1 Introduction ........................................... 1

2 Methods ............................................... 3
2.1 Data sets ............................................. 3
2.2 Preprocessing ........................................ 3
2.3 Algorithms ........................................... 4
2.4 Study details ........................................ 9

3 Results ................................................ 11
3.1 Alternative data selection and transforms ............. 11
3.2 Reproducibility ..................................... 18
3.3 Effects of different random number generator seeds .. 18
List of Tables

2.1 Outline of experiments testing reproducibility ............... 9

3.1 Standard deviations about zero of the reconstruction error for data transform alternatives ......................... 15

3.2 Set #1 signatures identified in all pairwise matches to signature sets obtained with different random number generator seeds, for different thresholds ........................................... 25

3.3 Reference set signatures identified in all pairwise matches to signature sets obtained after deleting 20% of the variables, for different thresholds .................................................. 29

3.4 Set #1 signatures identified in all pairwise matches to signature sets obtained after randomly excluding entire measurements, for different thresholds ................................................. 31

3.5 Set #1 signatures identified in all pairwise matches to signature sets obtained after randomly excluding entire measurements, for different thresholds (II) .............................. 35
List of Figures

2.1 Signature set comparison, algorithm overview .................. 5
2.2 Iterative Proportional Fitting, algorithm overview .......... 6
2.3 Permutation matrix construction, algorithm overview ....... 7
2.4 Similarity matrix analysis, algorithm overview ............. 8

3.1 Histogram of experimental error estimates for (non-transformed) ratio data ........................................ 12
3.2 Histogram of experimental error estimates for log-ratio data . 13
3.3 Scatterplots of error measures for (non-transformed) ratios, showing the worst 5% of the data ..................... 14
3.4 Scatterplots of error measures for (non-transformed) ratios, excluding the worst 5% of the data .................... 15
3.5 Scatterplots of error measures for log-ratios, showing the worst 5% of the data ..................................... 16
3.6 Scatterplots of error measures for log-ratios, excluding the worst 5% of the data .................................... 17
3.7 Similarity matrix for different random number generator seeds 19
3.8 Pairwise identification percentages for different random number generator seeds ........................................ 20
Chapter 1

Introduction

Independent Component Analysis (ICA) relies on a non-Gaussian distribution of the underlying latent variables that are to be uncovered. As a consequence, certain properties of algorithms for ICA can depend on characteristics of the data examined, which will be different according to the application domain of the analysis.

DNA microarrays allow the simultaneous measurement of transcript levels for thousands of genes (8). These give an indication of which genes have been turned on in a given sample. Using multiple 'channels' per measurement, several transcript levels are usually determined for each gene. Most commonly, one channel is used for a neutral control, and one other channel measures the transcript level of a particular sample under investigation. It is then typical to focus on the channel ratios. Samples can be from different tissue types, developmental stages, disease classes, or specifically genetically engineered organisms.

Only recent studies have applied ICA to such data (5, 6, 7). In contrast to the very high number of measurement variables (∼10⁴, scaling with the number of examined genes), an everyday data set will consist of only a few samples (∼10 – 100). Little is known about the expected distributions of the underlying latent variables. Testing results from applications of ICA in
this domain for robustness is therefore of interest.

Since it is common in the field to transform or preprocess the data, we first examined alternative transforms and data selections for the smallest modelling reconstruction errors.

A test for robustness requires comparing ICA results, e.g., the signatures, which hold the loadings of the original variables (gene transcript ratios) on particular latent variables. ICA is invariant under rescaling and permutation of these signatures, and any comparison of two sets of signatures must allow for that. To this end we introduced a method to optimally match corresponding signatures from two sets to one another.

We then examined stability of signatures after (1) repetition of the same analysis run with different random number generator seeds, and (2) repetition of the analysis with partial data sets. We studied both the effect of dropping a proportion of the gene transcript ratios, and of dropping measurements for several samples.

Chapter 2

Methods

2.1 Data sets

We used data published by Hughes et al. (4), who provide one of the most extensive data sets publicly available, and moreover performed 63 'neutral' vs 'neutral' experiments, examining the natural variation seen in unaltered wild type yeast. It is this subset of 63 samples that this study uses.

2.2 Preprocessing

The data were first preprocessed to remove all the ratios of genes for which there occurred values that were not a number or infinite. Data for 1464 such genes were dropped, reducing the size of the final data matrix to 63 × 4870.

The data as provided by Rosetta Inc. (4) contains log_{10} transformed channel ratios and estimates of their experimental errors.

For an examination of non-transformed ratios, both the data and the experimental error estimates have been appropriately reverse transformed.
2.3 Algorithms

Independent Component Analysis (ICA) was performed using Ensemble Learning as implemented in MatLab by Mislin (7). With $s$ enumerating samples (e.g., different tissues or experimental conditions) and $i$ enumerating the original input variables (gene transcript ratios), we get a decomposition

$$D_{si} = A_{si} B_{li} + 
u_{si}$$

(2.1)

where $l$ enumerates the latent variables, and $\nu$ allows for Gaussian noise. The matrix $B$ ($\sigma_l$) holds the signatures; a signature $\sigma_l$ ($\sigma_{li}$) shows how much each of the original input variables contributes to a particular latent variable. The amounts required to reconstruct the data in latent variable space are given by matrix $A$.

Signatures from two ICA runs were compared to one another using an Iterative Proportional Fitting (IPF) procedure to obtain a best guess permutation matrix giving the required reordering for an optimal match. For two signatures $\sigma$ and $\tau$, a similarity measure

$$s(\sigma, \tau) = \frac{(\sum \sigma_l \tau_l)^2}{\sum \sigma_l^2 \sum \tau_l^2}$$

(2.2)

was defined, which is invariant under rescaling of the signatures. Here, the sums run over all original variables (gene transcript ratio).

To rank latent variables according to their contribution to the reconstruction, we calculate the relative data power for a latent variable;

$$p_l = \frac{\sum_{s,i} (a_{si} b_{li})^2}{\sum_{s,i} a_{si}^2} \frac{\sum_{i} b_{li}^2}{\sum_{i} b_{li}^2}$$

(2.3)

function compare_signature_sets(Sigma, Tau);
Sigma = sort_signatures_by_relative_data_power(Sigma);
Tau = sort_signatures_by_relative_data_power(Tau);
P_IPF = permutation_matrix_by_IPF(Sigma, Tau);
P_constr = construct_best_guess_permutation_matrix(P_IPF);
Tau_new = P_constr * Tau;
S = similarity_matrix(S, Tau_new);
stats = analyse_similarity_matrix(S);
return(stats);

Figure 2.1: Signature set comparison, algorithm overview.
function permutation_matrix_by_IPF(Sigma, Tau);
Sigma = normalize_gene_ratio_variance_across_samples(Sigma);
Tau = normalize_gene_ratio_variance_across_samples(Tau);
S = similarity_matrix(Sigma, Tau);
mu = 1;
alphas = ones;
betas = ones;
repeat
    M = S ** mu;
    M = rescale_rows(M, alphas);
    M = rescale_columns(M, betas);
    mu = log_increase(mu);
repeat
    alphas = update_row_scaling(M, alphas);
    M = rescale_rows(M, alphas);
    betas = update_column_scaling(M, betas);
    M = rescale_columns(M, betas);
until average_squared_element_change_in_M < maxInnerDelta
until average_squared_element_change_in_M < maxDelta
return(M);

Figure 2.2: Iterative Proportional Fitting algorithm overview. The procedure aborts after a specified maximum number of iterations. A faster converging yet possibly less robust alternative is obtained by initializing M = diag and updating M = M ** mu.

2.3 Algorithms

function construct_best_guess_permutation_matrix(P_IPF);
map = zeros;
P = P_IPF;
repeat
    for row_index = 1 to number_of_rows(P)
        alias row = row_of(P, row_index);
        column_index = index(max(row), row);
        alias column = column_of(P, column_index);
        if max(row) > max(column) and max(row) > threshold
            column = zeros;
            row = zeros;
            map(row_index) = column_index;
        end_if
end_for
until no_changes_to_P;
for row_index = 1 to number_of_rows(P)
    column_index = index(row_index, map);
    if not column_index
        free_column_index = index(0, map);
        map(free_column_index) = row_index;
    end_if
end_for
for row_index = 1 to number_of_rows(P)
    alias row = row_of(P, row_index);
    column_index = index(row_index, map);
    row = zeros;
    row(column_index) = 1;
end_for
return(P);

Figure 2.3: Construction of a best guess permutation matrix algorithm overview. This procedure assumes that the input signatures have been sorted by a measure of relative importance (e.g., relative data power). The first loop is also aborted after a specified maximum number of iterations.
function analyse_similarity_matrix(S);
    ok = zeros;
good = zeros;
M = S;
row_index = 0;
repeat
    i = highest_diagonal_element_index_next(row_index,M);
    if M(i,i) >= max(row_of(M,i)) and
       M(i,i) >= max(column_of(M,i)) and
       M(i,i) >= 0.1 then
        ok(i) = 1;
    if M(i,i) >= 0.6 then
        good(i) = 1;
    end_if
    row_of(M,i) = zeros;
    column_of(M,i) = zeros;
end_if
until no_changes_to_M;
return(ok,good,S);

Figure 2.4: Similarity matrix analysis. Algorithm overview. If a diagonal element similarity score is less than 0.1, it is considered too bad for a match.

highest_diagonal_element_index_next will in successive calls return all row indices of its matrix argument so that the corresponding diagonal element values decrease. Once the set of all row indices has been exhausted, the function will restart with the index for the highest such element. By allowing multiple iterations, clear cases can be dealt with first, removing ambiguities for matches with weaker similarity score. The loop is aborted after a specified maximum number of iterations.

2.4 Study details

The full analysis was first repeated with ten different random number generator seeds, yielding $\binom{10}{2} = 45$ pairwise comparisons.

In the following other tests, to separate the effect caused by a changed random number generator seed, analyses were all run after using the same seed to initialize the random number generator (RNG).

Six independent subsets of the complete data were then each generated by dropping data for a different random selection of 974 genes (20%). The obtained signatures were compared to the signatures of the full analysis.

Four independent subsets of the complete data were generated by dropping a different random selection of thirteen samples (20%). The four signatures obtained were matched to one another, again yielding six comparisons.

<table>
<thead>
<tr>
<th>N</th>
<th>pw</th>
<th>type</th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
<td>45</td>
<td>RNG seed variation</td>
</tr>
<tr>
<td>6</td>
<td>6</td>
<td>dropping a random 5% of the original variables</td>
</tr>
<tr>
<td>6</td>
<td>6</td>
<td>dropping a random 20% of the original variables</td>
</tr>
<tr>
<td>6</td>
<td>6</td>
<td>dropping a random 35% of the original variables</td>
</tr>
<tr>
<td>6</td>
<td>6</td>
<td>dropping a random 50% of the original variables</td>
</tr>
<tr>
<td>4</td>
<td>6</td>
<td>dropping a random 5% of measurement samples</td>
</tr>
<tr>
<td>4</td>
<td>6</td>
<td>dropping a random 20% of measurement samples</td>
</tr>
<tr>
<td>4</td>
<td>6</td>
<td>dropping a random 35% of measurement samples</td>
</tr>
<tr>
<td>4</td>
<td>6</td>
<td>dropping a random 50% of measurement samples</td>
</tr>
</tbody>
</table>

Table 2.1: Outline of experiments testing reproducibility. The table shows the type of experiment, how many independent data sets have been generated (N), and the number of pairwise comparisons made (pw).
The experiments dropping certain original variables or measurements were also done dropping 5%, 35%, and 50% of the data, respectively (see Table 2.1 on the preceding page).

After optimal reordering according to the permutation matrix constructed using IPF, the similarity measure (2.2) was calculated for all signature pairs, and analysed as to whether individual signatures could be matched up. We visualized the resulting similarity matrices using Hinton plots.

This algorithm was implemented in Matlab.

Chapter 3

Results

3.1 Alternative data selection and transforms

Often, log-ratios are analysed rather than non-transformed ratios. We have examined reconstruction errors for both alternatives.

It is further customary in the field to normalize data with respect to the experimental error estimate for each data point. We prefer to explicitly model data and errors, using the non-normalized data directly. It has been observed before that the distribution of reconstruction errors has heavy tails (7) compared to a Gaussian distribution. This is also true for the experimental error estimates in our data set (Figs 3.1 and 3.2).

A Gaussian distribution would, however, be a useful approximation for reasonably compact error distributions. Plotting error measures as data values for non-transformed ratios (Figs 3.3 and 3.4) and for log-transformed ratios (Figs 3.5 and 3.6) shows that by exclusion of the data points with the worst 5% of experimental error estimates, one can indeed get a compact range of errors, avoiding more complex error models in a first approximation.

The plots also suggest that, in the Rosetta data set, several data points with particularly high errors have actually been generated by range truncation; extreme values appear to have been set to the finite values ±2 on the log
Figure 3.1: Histogram of experimental error estimates for (non-transformed) ratio data. The horizontal axis corresponds to the magnitudes of the experimental error estimates, the heights of the bars to their frequencies. Note that the presentation is semi-logarithmic. The distribution of errors shows a heavy tail with a wide range of observed values in the data set. In the above histogram, moreover, all error values larger than 3000 have been combined into the last bin. The full range of the error values is $\sim 0.1, \ldots, 38543$ (not shown).

Figure 3.2: Histogram of experimental error estimates for log-ratio data. The horizontal axis corresponds to the magnitudes of the experimental error estimates, the heights of the bars to their frequencies. Note that the presentation is semi-logarithmic. The distribution of errors shows a heavy tail with a wide range of observed values in the data set. In the above histogram, moreover, all error values larger than 15 have been combined into the last bin. The full range of the error values is $\sim 0.8, \ldots, 167$ (not shown).

scale. Clearly, we want to exclude these from a quantitative model. All further results have been obtained after pruning the data points with the worst 5\% of experimental error estimates.

Comparing the distributions of the reconstruction error for models of log-ratio and non-transformed ratios, one could see that the log-ratio model had lower reconstruction errors both in log-ratio space and in non-transformed ratio space (Table 3.1 on page 15). A linear model in log-ratio space corresponds to multiplicative effects for the non-transformed ratios. Such effects have been proposed for gene expression that is regulated by multiple tran-
Figure 3.3: Scatterplots of error measures for (non-transformed) ratios, showing the worst 5% of the data. The graphs plot the relationships between various error measures and the (non-transformed) ratio data. The error measures shown are: experimental error estimates (empirical data errors), standard deviations for the reconstructed data from Ensemble Learning, and reconstruction errors (original minus data reconstructed from the model).

All further results have been obtained from log-ratio data.

Figure 3.4: Scatterplots of error measures for (non-transformed) ratios, excluding the worst 5% of the data. A random 5% subset of the remaining data has been plotted. The scales of the panels show the ranges of the entire remaining data. Also see legend of Fig. 3.3 on the facing page.

<table>
<thead>
<tr>
<th>(a)</th>
<th>Model</th>
<th>ratios</th>
<th>log-ratios</th>
<th>(b)</th>
<th>Model</th>
<th>ratios</th>
<th>log-ratios</th>
</tr>
</thead>
<tbody>
<tr>
<td>ratio scale</td>
<td>0.039</td>
<td>0.028</td>
<td>0.038</td>
<td>ratio scale</td>
<td>0.38</td>
<td>0.03</td>
<td></td>
</tr>
<tr>
<td>log-ratio scale</td>
<td>0.025</td>
<td>0.012</td>
<td>2.42</td>
<td>log-ratio scale</td>
<td>1.72</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 3.1: Standard deviations about zero of the reconstruction error for data transform alternatives. After application of the ICA model to the (non-transformed) ratio data, or to log-ratios, the standard deviations about zero of the reconstruction error have been determined in both scales (log and non-transformed). For the relative errors, only non-zero data points have been considered.

(a) absolute reconstruction error; (b) relative reconstruction error
Figure 3.5: Scatterplots of error measures for log-ratios, showing the worst 5% of the data. The graphs plot the relationships between various error measures and the log-ratio data. The error measures shown are: experimental error estimates (empirical data errors), standard deviations for the reconstructed data from Ensemble Learning, and reconstruction errors (original minus data reconstructed from the model).

Figure 3.6: Scatterplots of error measures for log-ratios, excluding the worst 5% of the data. A random 5% subset of the remaining data has been plotted. The scales of the plots show the ranges of the entire remaining data. Also see legend of Fig. 3.5 on the facing page.
3.2 Reproducibility

In each study, the repeated analyses yielded pairwise similarity matrices which were fairly alike. Examples and summaries may therefore be shown, only.

3.3 Effects of different random number generator seeds

Figure 3.7 on the facing page shows a similarity matrix obtained from the analysis of the full data set with different random number generator seeds. Out of the ten runs, a pair of signature sets that yielded one of the worst matches has been chosen for this example. Only latent variables with a relative data power of at least $2 \times 10^4$ have been included, leaving 54 of the original 63. The high similarity scores along the diagonal show good pairwise matches, while occasional large off-diagonal elements indicate similarity to other signatures than the assigned optimal match.

Examining all 45 pairwise matches, we found that for $(88 \pm 4)\%$ of the signatures a best pairwise match can be identified (range $78\% - 96\%$), most of which have a 'good' score of 0.6 or higher [$75 \pm 5\%$, range $67\% - 85\%$].

One wonders whether exclusion of latent variables with low relative data power could improve the picture. Indeed, using a cutoff value of $\sim 1\%$, even $(96 \pm 7)\%$ of the remaining signatures can be matched pairwise (range $90\% - 100\%$), most of which have a 'good' score [(90$\pm 8$)$\%$, range $82\% - 100\%$]. Increasing the cutoff above that does not further help these statistics. Only a handful of signatures remain. If only a few of these are reversed in their order by relative data power in a pair of compared sets, so that a signature is amongst the $N$ signatures of highest data power to be considered in one set, but not its matching partner in the other, no match can be found, and a bad score results (e.g., at a cutoff value of 2%, one mismatch in a set

Figure 3.7: Hinton plot of the similarity matrix for sets of signatures obtained from analysis of the full data sets with different random number generator seeds. The area of the white blocks represents the values of the corresponding matrix entries, which can range from zero (no similarity) to one (perfect match). For a perfect match, two signatures must be collinear. For reference, the relative data power (POW) is also displayed. Diagonal elements are coloured black if the match can be considered optimal by heuristic criteria (see Methods).
3.3 Effects of different random number generator seeds

of 9 already gives a penalty of more than 10%). A threshold of 6.5% for relative data power leaves, e.g., only three signatures in set #1, and four from set #4. To perform IPF, the three signatures with the highest relative data power have to be chosen from each set. This, however, discards signature #4 from set #4, which is the best match to signature #3 of set #1 (data not shown). Figure 3.8 on the preceding page displays how well signatures could be matched pairwise between sets for various thresholds.

Examination which signatures could be identified in a pairwise match also showed that, for small or heterogeneous sets, this varies non-trivially with the choice of the threshold parameter; Whether signatures with extreme loadings on particular original variables (gene transcript ratios) are included, affects the subsequent normalization of this variable.

Having multiple pairwise comparison available, the question arises whether the same latent variables could stably be identified in each of the pairwise matches. Examining how many signatures of a given set could be identified in pairwise matches to at least two other sets, yields in the worst case a reduction by 8% (see Fig. 3.9 on the following page). For an arbitrarily picked set #1, we also determined which signatures could be identified in all the pairwise comparisons to the remaining three sets. In the worst case (for a cutoff value of $2 \times 10^{-4}$), this gave a reduction of more than 40%. For higher cutoff values (> 0.5%), however, no further reduction was seen (data not shown). For a minimum relative data power threshold of $\sim 1\%$, in all sets more than 90% of the signatures could be identified in at least two other sets. Also, more than 80% of the signatures from set #1 were identified in all pairwise matches to the other nine sets.

Similarly for identified matches with a 'good' score (see Fig. 3.10 on page 23): A combination of two pairwise matches yielded a reduction of 9% compared to plain pairwise scores. The combination of all pairwise matches for set #1 gave a reduction of $\sim 15\%$. For a minimum relative data power threshold of $\sim 1\%$, in all sets 85% of the signatures could be identified in at least two other sets with a good score ($\geq 0.9$). Moreover, two-thirds of the
WT63: Typical reproducibility (RNG seed variation)

Figure 3.9: Identification percentages for different random number generator seeds. For 45 attempts of pairwise matching of signatures from otherwise identical ICA runs with different random number generator seeds, always two pairs of matched sets have been combined. This means that this plot displays how many signatures could be identified in at least two comparisons. Another curve shows how many signatures from an arbitrarily picked set could be identified in all the pairwise matches to the remaining nine sets. Signature sets were restricted by different thresholds of relative data power before the comparison. For each such threshold, the graph shows the means and error bars (standard deviations) of the percentages of the signatures that could be identified. The number of signatures above the threshold is also displayed (the total number, not the percentage is plotted).

WT63: Typical reproducibility (RNG seed variation)

Figure 3.10: Identification percentages for different random number generator seeds, good matches. For 45 attempts of pairwise matching of signatures from otherwise identical ICA runs with different random number generator seeds, always two pairs of matched sets have been combined. This means that this plot shows how many signatures could be identified in at least two comparisons and had a good similarity score of 0.6 or higher. Another curve shows how many signatures from an arbitrarily picked set could be identified in all the pairwise matches to the remaining nine sets with a good score. Signature sets were restricted by different thresholds of relative data power before the comparison. For each such threshold, the graph shows the means and error bars (standard deviations) of the percentages of the signatures that could be identified. The number of signatures above the threshold is also displayed (the total number, not the percentage is plotted).
signatures from set #1 were identified in all pairwise matches to the other nine sets with a good score.

The apparent irregularity of the curves showing the number of signatures found in all pairwise comparisons reflects that this quantity is derived for a single set, set #1, while the other plots average the results for multiple sets.

It was also interesting to check which particular signatures of set #1 could be identified as the relative data power threshold was varied. From Table 3.2 one can see that not only are the lists of signatures at higher cutoffs almost subsets of those for lower cutoffs, but that, moreover, for a high enough cutoff (~1%) the list of identified signatures is contiguous, starting with the signature of highest relative data power.

For practical purposes of assessing reliability of a single set of signatures from this experiment, one could therefore pick a threshold of ±1.5%. The higher the relative data power of a signature, the more likely it is to be reproducible. Only occasionally, the closest match to a signature of high relative data power in terms of the analysis occurs with a much lower relative data power value. A second, verification run is likely to expose such cases and may hence be advisable.

3.4 Effects of excluding random subsets of the original variables

Comparison of signatures obtained after exclusion of a random 20% of the original input variables to those signatures from ICA of the full (reference) data set gives a similarly robust picture (Fig. 3.11 on page 26, and Figs. 3.12 and 3.13). Although, again, there is some variation by cutoff value, a threshold of ~1% will give signatures that remain identifiable after exclusion of a random 20% of the original variables. At the 1% threshold, over 90% of signatures could be identified in all six pairwise matches, most of which were good. In general, the higher the relative data power of a signature, the

<table>
<thead>
<tr>
<th>cutoff</th>
<th>N</th>
<th>identified signatures with good (ok) similarity scores</th>
</tr>
</thead>
<tbody>
<tr>
<td>$1.2 \times 10^{-3}$</td>
<td>60</td>
<td>1 2 3 4 5 6 7 8 9 10 11 12 (14) 17 18 19 20 21 22 24 25</td>
</tr>
<tr>
<td></td>
<td></td>
<td>26 27 29 30 32 33 (34) (36) (46) (41) 42 43 47 (51) 57 (58)</td>
</tr>
<tr>
<td>$2.0 \times 10^{-4}$</td>
<td>54</td>
<td>1 (3) 4 5 6 7 8 10 11 12 (14) (15) (16) 17 19 20 22 23 24 26</td>
</tr>
<tr>
<td></td>
<td></td>
<td>27 28 29 30 32 33 (34) (36) (41) 42 43 44 47</td>
</tr>
<tr>
<td>$1.2 \times 10^{-3}$</td>
<td>48</td>
<td>1 2 (3) 4 5 6 7 8 9 10 11 12 (15) (16) 17 18 19 20 22 23</td>
</tr>
<tr>
<td></td>
<td></td>
<td>24 25 26 27 28 29 30 32 (36) (41) 42 43 (44)</td>
</tr>
<tr>
<td>$2.0 \times 10^{-3}$</td>
<td>42</td>
<td>1 2 (3) 4 5 6 7 8 9 10 11 12 17 18 19 20 21 (21) 23 24 25</td>
</tr>
<tr>
<td></td>
<td></td>
<td>26 27 28 29 30 32 (36) (41)</td>
</tr>
<tr>
<td>$2.5 \times 10^{-3}$</td>
<td>36</td>
<td>1 2 (3) 4 5 7 8 9 10 11 12 17 18 19 20 23 24 25 26 27</td>
</tr>
<tr>
<td></td>
<td></td>
<td>28 29</td>
</tr>
<tr>
<td>$3.0 \times 10^{-3}$</td>
<td>30</td>
<td>1 2 3 4 5 (6) 7 9 10 11 12 17 18 19 20 23 24 25 27 28</td>
</tr>
<tr>
<td>$4.0 \times 10^{-3}$</td>
<td>24</td>
<td>2 3 4 (6) 7 8 9 10 11 12 (17) 18 19 20 (20) 23</td>
</tr>
<tr>
<td>$7.0 \times 10^{-3}$</td>
<td>18</td>
<td>1 3 4 (6) 7 8 9 10 12 17</td>
</tr>
<tr>
<td>$1.1 \times 10^{-2}$</td>
<td>12</td>
<td>1 2 3 4 5 (6) 7 8 9 (11)</td>
</tr>
</tbody>
</table>

Table 3.2: Set #1 signatures identified in all pairwise matches to signature sets obtained with different random number generator seeds, for different thresholds. The signatures of an arbitrarily picked set (#1) of ICA results were determined which could be identified in all the pairwise comparisons to the signature sets obtained from ICA with different random number generator seeds. This table lists these signatures for various relative data power thresholds, and also shows which of them had a good similarity score (≥ 0.6). N gives the total number of signatures in the sets that were compared, i.e., this number of signatures were above the cutoff for all sets.
Figure 3.11: Hinton plot of the similarity matrix for ICA results on an 80% subset of variables, matched to the reference of the set of signatures obtained from analysis of the full data set. The area of the white blocks represents the values of the corresponding matrix entries, which can range from zero (no similarity) to one (perfect match). For a perfect match, two signatures must be collinear. For reference, the relative data power (POW) is also displayed. Diagonal elements are coloured black if the match can be considered optimal by heuristic criteria, N, B.; The worst match from the six comparisons was chosen for this figure, for reasons of demonstration.
WT63: Typical reproducibility on 80% subsets of original variables

![Graph showing reproducibility on 80% subsets of original variables.](image)

Figure 3.12: Identification percentages for results from different original variable subsets, good matches. For six attempts of pairwise matching of signatures from ICA of the entire data set (the reference) to results of ICA runs on different 80% subsets of the original input variables, always two pairs of matched sets have been combined. This means that this plot shows how many signatures could be identified in at least two comparisons and had a good similarity score of 0.6 or higher. Another curve shows how many signatures of the reference set could be identified in all the six pairwise matches with a good score. Signature sets were restricted by different thresholds of relative data power before the comparison. For each such threshold, the graph shows the means and error bars (standard deviations) of the percentages of the signatures that could be identified. The number of signatures above the threshold is also displayed (the total number, not the percentage is plotted).

<table>
<thead>
<tr>
<th>cutoff</th>
<th>$N$</th>
<th>Identified signatures with good (ok) similarity scores</th>
</tr>
</thead>
<tbody>
<tr>
<td>$1,2 \times 10^4$</td>
<td>60</td>
<td>1 2 3 4 5 6 7 8 9 10 11 12 17 18 19 (20) 21 23 24 26 27</td>
</tr>
<tr>
<td></td>
<td></td>
<td>28 29 30 32 33 (38) (39) (42) 43 46 47 49 (53) (54) 57 58</td>
</tr>
<tr>
<td>$2,0 \times 10^4$</td>
<td>54</td>
<td>1 2 3 4 5 6 7 8 9 10 11 12 17 18 19 (20) 21 23 24 26 27 28</td>
</tr>
<tr>
<td></td>
<td></td>
<td>29 30 32 33 36 (38) (39) (40) 43 46 47 49 (51)</td>
</tr>
<tr>
<td>$1,2 \times 10^3$</td>
<td>48</td>
<td>1 2 3 4 5 6 7 8 9 10 11 12 17 18 19 (20) 21 22 23 24 26</td>
</tr>
<tr>
<td></td>
<td></td>
<td>27 28 29 32 33 (39) (41) 42 43 (46)</td>
</tr>
<tr>
<td>$2,0 \times 10^3$</td>
<td>42</td>
<td>1 2 3 4 5 6 7 8 9 10 11 12 (15) 17 18 19 (20) 23 24 26</td>
</tr>
<tr>
<td></td>
<td></td>
<td>27 28 30 32 33 (39)</td>
</tr>
<tr>
<td>$2,5 \times 10^3$</td>
<td>36</td>
<td>1 2 3 4 5 6 7 9 10 11 12 (14) 17 18 (21) (22) 23 24 26 27</td>
</tr>
<tr>
<td></td>
<td></td>
<td>28 29 30 32 (35)</td>
</tr>
<tr>
<td>$3,6 \times 10^3$</td>
<td>30</td>
<td>1 2 3 4 7 10 11 12 (14) 17 18 19 (20) 23 24 26 27 28</td>
</tr>
<tr>
<td>$4,3 \times 10^3$</td>
<td>24</td>
<td>1 2 3 4 6 7 9 10 11 12 (15) (16) 17 18 19 (20) 23</td>
</tr>
<tr>
<td>$7,5 \times 10^3$</td>
<td>18</td>
<td>1 2 3 4 5 7 9 10 11 12</td>
</tr>
<tr>
<td>$1,1 \times 10^2$</td>
<td>12</td>
<td>1 2 3 4 5 6 7 8 9 10 (11)</td>
</tr>
</tbody>
</table>

Table 3.3: Reference set signatures identified in all pairwise matches to signature sets obtained after deleting 30% of the variables for different thresholds. Compared to a reference set calculated using the entire data, the signatures were determined which could be identified in all the pairwise comparisons to signatures obtained after deleting 30% of the original variables. This table lists these signatures for various relative data power thresholds, and also shows which of them had a good similarity score ($\geq 0.6$). $N$ gives the total number of signatures in the sets that were compared, i.e., this number of signatures were above the cutoff for all sets.
more likely it is to be robust. Since the lists in this table show signatures that could be identified in all six pairwise matches, only minor variation with cutoff threshold can be expected due to the effect of the different set choice for each experiment on the subsequent normalization. Signatures #1 through #11 seemed to be particularly stable.

The question arises whether even more data could have been removed without affecting the stability of the signatures of high relative data power. Removing 35% or even 50% of the original variables actually has similar results in that signatures with higher relative data power are more likely to be conserved. In each case, two-thirds of the signatures could be identified in all six pairwise comparisons, and they all had a good score (data not shown).

So, do these numbers improve if less data is removed? Interestingly, there is no such trend (data not shown). Rather, it seems that removing small or even large amounts of the original variables has a similar impact on the signatures as any other minute changes, such as picking a different seed for the random number generator, essentially leaving between two thirds and 90% of recoverable signatures, most of which with a good score. Apparently, the robust signatures found are characterized by many genes, so that dropping even a larger proportion of genes does not preclude picking up these signatures.

### 3.5 Effects of excluding random subsets of measurement samples

Already Fig. 3.14 on the next page shows that in a setting of relatively few measurement samples and many variables, removal of the latter has less serious consequences than removal of the former (cf. Fig. 3.11 on page 26). At high enough relative data power (~1%), however, many signatures can still be identified (Fig. 3.15 on page 32 and Fig. 3.16). At a threshold of 1%, in pairwise matches, about two thirds of the signatures were always identified (half of which with a good score) after 20% of measurement samples had been removed.

**Figure 3.14:** Hinton plot of the similarity matrix for ICA results on two different 80% subsets of measurement samples. The area of the white blocks represents the values of the corresponding matrix entries, which can range from zero (no similarity) to one (perfect match). For a perfect match, two signatures must be collinear. For reference, the relative data power (POW) is also displayed. Diagonal elements are colored black if the match can be considered optimal by heuristic criteria. N.B.: The worst match from the six comparisons was chosen for this figure.
WT63: Typical reproducibility on 80% subsets of original variables

![Graph](image)

**Figure 3.15:** Identification percentages for results from different measurement sample subsets. For six attempts of pairwise matching of signatures from ICA of different 80% measurement sample subsets, always two pairs of matched sets have been combined. This means that this plot shows how many signatures could be identified in at least two comparisons. Another curve shows how many signatures from an arbitrarily picked set could be identified in all the pairwise matches to the remaining three sets. Signature sets were restricted by different thresholds of relative data power before the comparison. For each such threshold, the graph shows the means and error bars (standard deviations) of the percentages of the signatures that could be identified. The number of signatures above the threshold is also displayed (the total number, not the percentage is plotted).

WT63: Typical reproducibility on 80% subsets of original variables

![Graph](image)

**Figure 3.16:** Identification percentages for results from different measurement sample subsets, good matches. For six attempts of pairwise matching of signatures from ICA of different 80% measurement sample subsets, always two pairs of matched sets have been combined. This means that this plot shows how many signatures could be identified in at least two comparisons and had a good similarity score of 0.6 or higher. Another curve shows how many signatures from an arbitrarily picked set could be identified in all the pairwise matches to the remaining three sets with a good score. Signature sets were restricted by different thresholds of relative data power before the comparison. For each such threshold, the graph shows the means and error bars (standard deviations) of the percentages of the signatures that could be identified. The number of signatures above the threshold is also displayed (the total number, not the percentage is plotted).
## 3.5 Effects of excluding random subsets of measurement samples

<table>
<thead>
<tr>
<th>p</th>
<th>cutoff</th>
<th>N</th>
<th>identified signatures with good (ok) similarity scores</th>
</tr>
</thead>
<tbody>
<tr>
<td>50%</td>
<td>2.5 × 10³</td>
<td>30</td>
<td>(1) (2) (3) (6) (15) (16) (21)</td>
</tr>
<tr>
<td>50%</td>
<td>3.6 × 10³</td>
<td>21</td>
<td>(1) (2) (4) (6) (15)</td>
</tr>
<tr>
<td>50%</td>
<td>4.3 × 10³</td>
<td>18</td>
<td>2 (3) (15)</td>
</tr>
<tr>
<td>50%</td>
<td>7.5 × 10³</td>
<td>12</td>
<td>2 (9) (11)</td>
</tr>
<tr>
<td>65%</td>
<td>2.8 × 10³</td>
<td>36</td>
<td>1 3 (4) (5) 6 (7) (8) (9) (11) (16) (20) (22) (23) (26) (31)</td>
</tr>
<tr>
<td>65%</td>
<td>2.5 × 10³</td>
<td>30</td>
<td>1 3 (5) 6 (9) (11) (14) (19) (20) (21) (22) (26) (28) (30)</td>
</tr>
<tr>
<td>65%</td>
<td>3.6 × 10³</td>
<td>21</td>
<td>3 (5) 6 (8) (9) (10) 11</td>
</tr>
<tr>
<td>65%</td>
<td>4.3 × 10³</td>
<td>18</td>
<td>1 3 (4) 10 (11)</td>
</tr>
<tr>
<td>65%</td>
<td>7.5 × 10³</td>
<td>12</td>
<td>1 3 (4) 6 (7) (8) (9) 11</td>
</tr>
</tbody>
</table>

Table 3.4: Set #1 signatures identified in all pairwise matches to signature sets obtained after randomly excluding entire measurements, for different thresholds. ICA was performed for four sets of data, in each of which a percentage p of the original 63 measurements had been retained randomly. The signatures of an arbitrary set (#1) of ICA results were then determined which could be identified in all the pairwise comparisons to the remaining three signature sets. This table lists these signatures for various relative data power thresholds, and also shows which of them had a good similarity score (≥ 0.8). N gives the total number of signatures in the sets that were compared, i.e., this number of signatures were above the cutoff for all sets.

<table>
<thead>
<tr>
<th>p</th>
<th>cutoff</th>
<th>N</th>
<th>identified signatures with good (ok) similarity scores</th>
</tr>
</thead>
<tbody>
<tr>
<td>95%</td>
<td>1.2 × 10³</td>
<td>54</td>
<td>1 2 3 4 5 6 7 8 9 (10) (11) (12) 13 16 18 (19) 20 21 22 (24) (25) 26 27 28 (30) 31 33 34 35 36 (39) 40 (41) (42) (43) 44 (45) 47 49 53 (54)</td>
</tr>
<tr>
<td>95%</td>
<td>2.0 × 10³</td>
<td>48</td>
<td>1 2 3 4 5 6 7 8 9 (10) (11) (12) (13) 16 18 (19) 20 21 22 (23) (24) (25) 26 27 28 (29) 31 34 35 36 38 (39) (40) (41) (42) 44 (45) 47</td>
</tr>
<tr>
<td>95%</td>
<td>1.2 × 10³</td>
<td>42</td>
<td>1 2 3 4 5 6 7 8 9 (10) (11) (12) (13) 14 (15) 16 18 (19) 20 21 22 (23) (24) (25) 26 27 28 (29) 31 33 34 35 (36) (38) (39) (40) (41)</td>
</tr>
<tr>
<td>95%</td>
<td>2.0 × 10³</td>
<td>36</td>
<td>1 2 3 4 5 6 7 8 9 (10) (11) (12) 13 16 18 19 20 21 22 (23) (24) (25) 26 27 28 (29) 31 (32) 34</td>
</tr>
<tr>
<td>95%</td>
<td>2.5 × 10³</td>
<td>30</td>
<td>1 2 3 4 5 6 7 8 9 10 (11) (12) (13) (14) (15) 16 (17) 18 19 20 21 22 (23) (24) 26 27 28 (29)</td>
</tr>
<tr>
<td>95%</td>
<td>3.6 × 10³</td>
<td>24</td>
<td>1 2 3 4 5 6 7 8 9 (10) (12) (13) 14 16 17 (18) 19 20 21 (23)</td>
</tr>
<tr>
<td>95%</td>
<td>4.3 × 10³</td>
<td>18</td>
<td>1 2 3 4 5 6 7 8 9 (11) (12) (13) (15) (16)</td>
</tr>
<tr>
<td>95%</td>
<td>7.5 × 10³</td>
<td>12</td>
<td>1 3 4 5 6 8</td>
</tr>
</tbody>
</table>

Table 3.5: Set #1 signatures identified in all pairwise matches to signature sets obtained after randomly excluding entire measurements, for different thresholds (II), Table 3.4 on the facing page, continued,
removed from each set. When fewer samples are removed, this proportion increases significantly (> 80% ok, two thirds with good score), and it is much smaller when even more data is dropped. Only 1/4 of signatures could always be identified when half of the measurement samples had been removed, and less than 10% of signatures could be matched with a good score; cf. Table 3.4 on page 34 and its continuation (Table 3.5). The fact that, by using a data set of only twice the size we can reliably extract three times as many signatures, and match six times as many signatures with a good score, highlights the importance of independent measurement samples in a situation where few such samples are available.

Chapter 4

Conclusion

ICA on yeast gene expression ratio data has proven to be fairly robust. The data set studied consisted of 63 measurements of expression ratios for 4870 genes. A relative data power threshold of ~ 1% gave a set of 12 signatures, all of which could be clearly identified in analysis reruns with different random number generator seeds, with the occasional exception of one or two signatures, depending on the exact choice of the relative data power threshold. Even after random exclusion of substantial amounts of data, most of the original 12 signatures could clearly be identified. This was demonstrated by removal of, e.g., up to 50% of the original input variables (particular gene expression ratios), or 20% of entire measurements (independent samples), after which most signatures could still be clearly identified, missing less than a third of signatures, depending on the choice of relative data power threshold. When only 20% of the original input variables (particular gene expression ratios) had been removed, almost all signatures were matched reliably (i.e., with only one or two exceptions, depending on the choice of data power threshold).

In summary, the observed robust signatures were characterized by many genes, which made them fairly immune to even crude omissions from the data. Eventually, the 63 measurements yielded ~ 10 reliably reproducible signatures. This is highly surprising and noteworthy, considering that the
wild type 'control' data has traditionally only been treated as noise. The obtained signatures are presently being studied for biological relevance.

This work should be repeated on independent data sets to see whether the recommended cutoff value of $\approx 1\%$ relative data power also transfers to other analyses, e.g., gene classification studies by expression profiles over time (2, 3). It can be expected, however, that at least another such threshold can be found.

Bibliography


