Probabilistic estimation of microarray data reliability and underlying gene expression

Bilke, Sven; Breslin, Thomas; Sigvardsson, Mikael

Published in:
BMC Bioinformatics

DOI:
10.1186/1471-2105-4-40

2003

Link to publication

Citation for published version (APA):

General rights
Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

• Users may download and print one copy of any publication from the public portal for the purpose of private study or research.
• You may not further distribute the material or use it for any profit-making activity or commercial gain
• You may freely distribute the URL identifying the publication in the public portal

Take down policy
If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.
Methodology article

Probabilistic estimation of microarray data reliability and underlying gene expression
Sven Bilke*1, Thomas Breslin1 and Mikael Sigvardsson2

Address: 1Complex Systems Division, Department of Theoretical Physics, University of Lund, Sölvegatan 14A, SE-223 62 Lund, Sweden and 2The Laboratory for Cell Differentiation Studies, Department for Stem Cell Biology, BMC B12, SE-22185 Lund, Sweden

Email: Sven Bilke* - sven@thep.lu.se; Thomas Breslin - thomas@thep.lu.se; Mikael Sigvardsson - Mikael.Sigvardsson@stemcell.lu.se

* Corresponding author

Abstract

Background: The availability of high throughput methods for measurement of mRNA concentrations makes the reliability of conclusions drawn from the data and global quality control of samples and hybridization important issues. We address these issues by an information theoretic approach, applied to discretized expression values in replicated gene expression data.

Results: Our approach yields a quantitative measure of two important parameter classes: First, the probability \( P(\sigma|S) \) that a gene is in the biological state \( \sigma \) in a certain variety, given its observed expression \( S \) in the samples of that variety. Second, sample specific error probabilities which serve as consistency indicators of the measured samples of each variety. The method and its limitations are tested on gene expression data for developing murine B-cells and a t-test is used as reference. On a set of known genes it performs better than the t-test despite the crude discretization into only two expression levels. The consistency indicators, i.e. the error probabilities, correlate well with variations in the biological material and thus prove efficient.

Conclusions: The proposed method is effective in determining differential gene expression and sample reliability in replicated microarray data. Already at two discrete expression levels in each sample, it gives a good explanation of the data and is comparable to standard techniques.

Background

A broad variety of algorithms has been developed and used to extract biologically relevant information from gene expression data. Among others commonly used are visual inspection [1], hierarchical and k-means clustering [2], self organizing maps [3,4] and singular value decomposition [5,6]. These methods aim mainly at identifying predominant patterns and thus groups of "cooperating" genes based on the assumption that related genes have similar expression patterns.

Compared to the amount of work devoted to efficient methods to extract information from the data, somewhat less attention has been paid to the question of the reliability of the generated results. The ANOVA analysis [7] allows estimation, and thus elimination, of some systematic error sources. Bootstrapping cluster analysis estimates the stability of cluster assignments [8] based on artificial data-sets generated with ANOVA coefficients. Some authors also considered the question of how well a certain oligo [10] is suited to measure the mRNA expression level of the related gene.

Some work has gone towards the ambitious task of learning topological properties or qualitative features of the genetic regulatory network from expression profiles, see...
e.g. [11]. A major limiting factor in these attempts is the comparative sparseness of available data. It is therefore reasonable to consider reduced models, for example a Boolean representation of the gene activity. It is known that many biological properties, for instance stability and hysteresis, can be modeled by the dynamics of such reduced models [12–14].

In this work we investigate the possibility of reducing complexity of gene expression data by discretizing the expression levels. The approach we present enables a new way of extracting biologically relevant information from the data in the following way: A biological variety, i.e. a biological system defined by the investigator, is represented by several samples which are subjected to gene expression analysis. If gene expression levels are discretized into \( n \) values, and the variety is represented by \( m \) samples, the number of observable expression states for a gene are limited to \( n^m \). These observed states \( S \) are modeled as being derived from a smaller number of underlying, biological states \( \sigma \), through a measurement process. Rather than making static assignments \( S \rightarrow \sigma \) we calculate conditional probabilities \( P(\sigma|S) \). The number of possible expression profiles for a gene over a set of varieties is limited and the probability of each expression profile is easily calculated. Since the model we use considers both the underlying biology and the measurement process it also generates a measure of sample coherence in each biological variety.

We demonstrate the feasibility of this approach for a binary discretization of gene expression. For the discretization step we use the absent/present classification provided by the Affymetrix software [9]. The outcome of our method on a data set covering gene expression in developing murine B-cells is compared to the results of a standard analysis. We show that even with the crude discretization into only two expression levels the method is competitive to statistical methods based on continuous expression levels.

**Methods**

**The Model**

A major step in the analysis of gene expression data is to separate the biological content of the data from measurement and sample specific errors. In other words given an observation, i.e. the expression values of a gene in several samples representing the same biological variety, one wants to conclude on the biological state \( \sigma \), which generated the observation. This can be expressed as a conditional probability,

\[
P(\sigma|S) \quad (1)
\]

that a gene is in a certain biological state \( \sigma \) given the corresponding observed state \( S \). In the application on which we demonstrate the method we consider three different biological varieties: pro, pre, and mature B-cells. The samples in each variety are different cell lines arrested at the corresponding stage of development.

In this work we take an information theoretic point of view to estimate this probability: The information of interest, the state \( \sigma \), is "transmitted" in a noisy measurement process and potentially distorted (Figure 1). Using Bayes’ theorem, the desired conditional probability Eq. (1) can be expressed as:

\[
P(\sigma|S) = \frac{P(S|\sigma)P_\sigma}{P_S}. \quad (2)
\]

On the right hand side of this equation, \( P(S|\sigma) \) is the probability to observe state \( S \) if the underlying biological state is \( \sigma \). In a sense, \( P(S|\sigma) \) describes the noise characteristic of the measurement process. In the following we will show how this conditional probability, and the other probabilities on the r.h.s. of Eq. (2) can be estimated.

Given a set of \( m \) samples representing the same biological variety, differences in the expression level of a gene between the samples can arise from two independent sources:

1. **Random variation** within the variety. This may be caused by temporal differences in response to the stimuli, slightly different environmental conditions, genotypic differences between samples, etc.

2. **Sample specific errors.** These are mainly caused by the measurement process, e.g. differences in the treatment of the mRNA, scratched arrays, and so on. However, outlier samples, cultured under considerably different conditions, also contribute to sample specific errors.

A separation of these two contributions is possible only with an appropriate model for the variation of gene expression between the samples. In the choice of model, one has considerable freedom within the bounds set by biological plausibility. A limiting factor on the biological model comes from the type and amount of available data. The data used in this work contains only four samples for each variety. For the model we propose this is the **minimum** number of samples required to estimate the model parameters.

In the discretization of gene expression levels, we use only two discrete values, 0 and 1, for the expression of a gene in a sample. This means that the number of observable states, \( S \), in a variety consisting of \( m \) samples is \( 2^m \). With
The underlying distribution on the left hand side can be described by the probabilities for each underlying state, \(P(\sigma_i)\), \(P(\sigma_1)\), and \(P(\sigma_2)\) (see text). This distribution is then distorted by sample specific errors, \(P_{0 \rightarrow 1}^i\) and \(P_{1 \rightarrow 0}^i\), resulting in an experimentally observed distribution, depicted on the right hand side.

### Figure 1

Schematic diagram illustrating the transition from underlying to observed distributions of states, in the case of \(m = 4\) samples. The underlying distribution on the left hand side can be described by the probabilities for each underlying state, \(P(\sigma_i)\), \(P(\sigma_1)\), and \(P(\sigma_2)\), and \(P(\sigma_3)\) (see text). This distribution is then distorted by sample specific errors, \(P_{0 \rightarrow 1}^i\) and \(P_{1 \rightarrow 0}^i\), resulting in an experimentally observed distribution, depicted on the right hand side.

The biological rationale for this model is given by the following example: If one considers a biological variety such as cells in the retina of the eye, then a certain number of crucial genes ought to be expressed in all samples. Such genes might include rhodopsin, a molecule that responds to light. In contrast, genes such as the hemoglobin family, which are typical of erythrocytes, ought not to be expressed in the retina. A third class of genes could be considered as independent of the system in the sense that their expression is not directly related to the biological system. Such genes may vary in expression both due to environmental and genetic differences between the samples.

The model discussed so far is depicted graphically in the left part of Figure 1, where a possible distribution of the relative frequencies of the three biological states is depicted, for the case of \(m = 4\) samples. The distribution can be described by three numbers: the probabilities \(P(\sigma_1)\) and \(P(\sigma_2)\), which contribute to the frequencies of the states \(S = (1,1,1,1)\) and \(S = (0,0,0,0)\), and \(P(\sigma_3)\) which contributes to both the frequency of mixed states and the two states above. Describing the mixed states with only one parameter \(P(\sigma_3)\) implies that the biological variation is modeled evenly and identically distributed independently for each sample. In a second step, the measurement process with possible sample specific errors is modeled as statistically independent between samples. For each sample \(i\), we define two parameters, \(P_{0 \rightarrow 1}^i\) and \(P_{1 \rightarrow 0}^i\), denoted sample specific error probabilities.

To introduce the full formalism of our current model we start by considering a simple example, again for \(m = 4\) samples. An observed state \(S, S = (S_1, S_2, S_3, S_4) = (1,0,1,0),\)
may be generated by the gene being in state $\sigma_i$ with the probability:

$$P(\sigma_i) \left(1 - P_{1 \rightarrow 0}^i\right) \left(1 - P_{2 \rightarrow 0}^i\right) \left(1 - P_{3 \rightarrow 0}^i\right) \left(1 - P_{4 \rightarrow 0}^i\right),$$

or it may be generated by the gene being in state $\sigma_0$ with the probability:

$$P(\sigma_0) \left(P_{0 \rightarrow 1}^i\right) \left(1 - P_{1 \rightarrow 0}^i\right) \left(P_{3 \rightarrow 1}^i\right) \left(1 - P_{4 \rightarrow 0}^i\right),$$

or it may be generated by the gene being in state $\sigma_i$ with the probability:

$$P(\sigma_i) \left(1 - P_{1 \rightarrow 0}^i\right) \left(1 - P_{2 \rightarrow 0}^i\right) \left(P_{3 \rightarrow 1}^i\right) \left(1 - P_{4 \rightarrow 0}^i\right),$$

or it may be generated by the gene being in state $\sigma_0$ with the probability:

$$P(\sigma_0) \left(P_{0 \rightarrow 1}^i\right) \left(P_{1 \rightarrow 0}^i\right) \left(1 - P_{2 \rightarrow 0}^i\right) \left(1 - P_{4 \rightarrow 0}^i\right).$$

With the briefer notation,

$$P_i^{1 \rightarrow 0} \equiv P_{1 \rightarrow 0}^i \delta_{s_i,0} + \left(1 - P_{1 \rightarrow 0}^i\right) \delta_{s_i,1},$$

$$P_i^{0 \rightarrow 1} \equiv P_{0 \rightarrow 1}^i \delta_{s_i,1} + \left(1 - P_{0 \rightarrow 1}^i\right) \delta_{s_i,0},$$

where $\delta$ refers to the Kronecker delta (i.e. $\delta_{j,k} = 1$ if $j = k$ and 0 otherwise), we may express the distribution of observed states, in the general case of binary discretization and 0 otherwise), we may express the distribution of observed states, in the general case of binary discretization and 0 otherwise), we may express the distribution of observed states, in the general case of binary discretization and 0 otherwise), we may express the distribution of observed states, in the general case of binary discretization.

We estimate the parameters $P(\sigma_1)$, $P(\sigma_0)$, $P(\sigma_i)$ and $\left\{P_i^{1 \rightarrow 0}, P_i^{0 \rightarrow 1}\right\}$ are estimated from the observed distribution of states (right side of Figure 1) by Levenberg-Marquardt [15] chi-square minimization of the unweighted error to the theoretical distribution Eq. (3). Using Eq. (2), and the parameters estimated as above, we believe that a gene belongs to the underlying states $\sigma_0$, $\sigma_i$, $\sigma_r$, given the $2^M$ = 16 observable states S, can now be expressed as:

$$P(\sigma_i | S) = \frac{P(S | \sigma_i) P(\sigma_i)}{P(S)}$$

$$P(\sigma_0 | S) = \frac{P(S | \sigma_0) P(\sigma_0)}{P(S)}$$

$$P(\sigma_r | S) = \frac{P(S | \sigma_r) P(\sigma_r)}{P(S)}$$

Once the probability that a gene is in a certain biological state $\Sigma^i \in \sigma_i$, $\sigma_0$, $\sigma_r$ has been calculated for all variants $i = 1, \ldots, v$, one can calculate the probability that a gene exhibits a certain expression profile over a set of different varieties by taking the product

$$P\left(\sum_{i=1}^{v} \sum_{j=1}^{u} S_i \right) = \prod_{i=1}^{v} P\left(\sum_{j=1}^{u} S_i \right).$$

In this way, the probabilistic state analysis also generates a clustering: For a given expression profile over the varieties, e.g. $\sigma_0, \sigma_1, \ldots, \sigma_i$, we may extract those genes for which this expression profile is the most probable. In fact this is a “soft” clustering, in that an expression profile can belong to several clusters simultaneously with different probabilities. Moreover the genes clustered to a biologically interesting expression profile can be ranked by the probability of Eq. (4).

**Experimental data preparation**

All cells were grown in RPMI medium supplemented with 7.5% fetal calf serum, 10 mM HEPES, 2 mM pyruvate, 50 mM 2-mercaptoethanol and 50 mg gentamicin per ml (complete RPMI media) (all purchased from Life Technologies AB, T = E4by, Sweden) at 37 = B0C and 5% CO2. RNA was prepared using Trizol (GIBCO) and 7.5 = B5g of total RNA was annealed to a T7-oligo T primer by denaturation at 70 = B0C for 10 minutes followed by 10 minutes of incubation of the samples on ice. First strand synthesis was performed for 2 hours at 42 = B0C using 20 U of Superscript Reverse Transcriptase (GIBCO) in buffers and nucleotide mixes according to the manufacturers instructions. This was followed by a second strand synthesis for 2 hours at 16 = B0C, using RNAaseH, E coli DNA polymerase I and E coli DNA ligase (all from GIBCO), according to the manufacturers instructions. The obtained double stranded cDNA was then blunted by the addition of 20 U of T4 DNA polymerase and incubation for 5 minutes at 16 = B0C. The material was then purified by Phenol:Chloroform:isoamyl alcohol extraction followed by precipitation with NH4Ac and Ethanol. The cDNA was then used in an in vitro transcription reaction for 6 h at 37 = B0C using a T7 IVT kit and biotin labeled ribonucleotides. The obtained cRNA was purified from unincorporated nucleotides on a RNeasy column (Qiagen). The eluted cRNA was then fragmented by incubation of the products for two hours in fragmentation buffer (40 mM Tris-acetate, pH 8.1, 100 mM KOAc, 150 mM MgOAc). 20 = B5g of the final fragmented cRNA was then hybridized to affymetrix chip U74Av2 (Affymetrix) in 200 = B5l hybridization buffer (100 mM MES-buffer, pH 6.6, 1 M NaCl, 20 mM EDTA, 0.01 Herring sperm DNA (100 = B5 g/ml) and Acetylated BSA (500 = B5 g/ml) in an Affymetrix Gene Chip Hybridization oven 320. The chip was then developed by the addition of FITC-streptavidin followed by washing using an Affymetrix Gene Chip Fluidics Station 400. Scanning was performed using a Hewlett Packard Gene Array Scanner.
Results
To evaluate the method we used both real and synthetic data. The experimental data was generated with Affymetrix microarrays for the study of differentiating murine B-cells at different stages in the differentiation process. In this publication the data is only used to demonstrate the feasibility of the proposed method. The biological implications of this study are published elsewhere [20].

Synthetic data and the effect of correlation
For synthetic data, generated with the model parameters $P(\sigma_0) = 0.45, P(\sigma_1) = 0.35, P(\sigma_2) = 0.2$ and $P_{1\rightarrow0} = P_{0\rightarrow1} = 0.02$ for all samples $i$, parameter estimates are, as expected, given with low errors. The parameter values were chosen as typical values from the estimates on real data, see next section, and the result was verified for sample sizes $m = 4, m = 5$, and $m = 6$ (data not shown).

An assumption of simple model used to derive Eq. (3) is that randomly varying genes vary independently in the samples of a variety. Hence we investigated how severely this assumption influences the estimation of the model parameters.

To assess the influence of correlations between randomly varying genes we generated a data set consisting of four bits, i.e. samples, with the same parameters as above. In the random patterns a correlation was introduced between the third and fourth bit by changing the value of of the fourth to that of the third with a certain probability. We define this probability as the correlation factor. The correlation was introduced before distorting the patterns with error probabilities. We then plotted the mean error in the estimation of parameters over 500 runs of synthetically generated data for correlation factors in the range $\{0, 0.02, ..., 0.98\}$.

Figure 2 shows the error in the estimation of the parameters describing the underlying distribution. We notice that even for fully correlated patterns the estimation error is less than 20% of the correct values. The estimation of the probability for biologically varying genes is somewhat worse, for fully correlated patterns the error is almost 50%. For real data one can, however, expect a much smaller correlation. The average error in the estimates of the error probabilities, as seen in Figure 3, shows the expected behavior: The average error grows with the correlation for the uncorrelated samples, while the estimate for the correlated observations is almost unaffected. Intuitively, the model compensates for the correlation by increasing $P(\sigma_1)$ and $P(\sigma_0)$ as well as the error probabilities and lowering $P(\sigma_2)$. For correlation factors above 0.50, due to the compensation effect, the model deteriorates in explaining the data. This can be seen in the sum $P(\sigma_0) + P(\sigma_1) + P(\sigma_2)$ which initially drops from almost 1 to 0.99 as the correlation factor rises from 0 to 0.50 and then from 0.99 to 0.96 for correlation factors in the range 0.50 to 0.98 (data not shown). We hence conclude that it is reasonable not to impose the condition $P(\sigma_0) + P(\sigma_1) + P(\sigma_2) = 1$ in the model, as this sum indicates if samples are strongly correlated in genes whose expression vary around the threshold of discretization.

In summary, for not too large correlations in the biological variance the algorithm gives a good quantitative estimate of the model parameters. In the case of large correlations the qualitative picture given by the estimated parameters is still reliable.

Real data
Differentiating B-cells are characterized by phenotypic markers into different stages of development. Here we chose to study the expressional differences between three such stages; pro, pre and mature B-cells. For each of these three varieties we used four different cell lines arrested at the corresponding stage of development. Measurements we performed with Affymetrix array containing probesets for 12488 genes and ESTs on each sample. The discretization of expression levels was given by the Affymetrix GeneChip absent present calls [9].

Our algorithm was used to estimate the parameters $P(\sigma_1)$, $P(\sigma_0)$ and $P(\sigma_2)$, describing the biological distribution and the error probabilities (see Table 1). Theoretically, one expects the three biological probabilities to sum up to one. In our model, Eq. (3), we do not explicitly impose this condition. Nevertheless, the sum of the independently estimated parameters is close to one. This indicates that our model is a reasonable approximation of the biological system and the measurement process.

The error probabilities from Eq. (3) can be used as a consistency index for the samples in a given variety. In the last variety (mature B-cells) the maximum error probability is notably higher. This effect is likely to be explained by the different anatomical origins of the cell lines representing this group. No such differences exist in the other groups since they all originate in the bone marrow which is the only anatomical site for B cell development in the adult animal [16]. In contrast, the mature B cell can reside in several other sites such as spleen, lymph-nodes and intestine which may affect the gene expression profile in these cells [17,18]. With only four samples, it is not unlikely that these effects show up in the error probabilities and not only in the random variation parameter $P(\sigma_2)$.

Comparison to conventional t-test on known genes
To determine how well biologically relevant information can be extracted from the discretized data, we compare it
with another statistical method based on continuous expression values. We use our method to identify differences in gene expression between two varieties in the following way. A gene that goes up between variety \(i\) and variety \(j\) is characterized by the states \(\sigma_0^i, \sigma_1^i\) or \(\sigma_0^j, \sigma_1^j\). Hence the belief that a gene goes up is given by the probability:

\[
P(\text{up between variety } i \text{ and } j) = P(\sigma_1^i | \sigma_0^i) P(\sigma_0^j) + P(\sigma_0^i | \sigma_1^i) P(\sigma_1^j).
\]

suppressing the conditional probabilities, \(P(\cdot | S)\), for brevity. Similarly, the belief that a gene goes down is given by the probability:

\[
P(\text{down between variety } i \text{ and } j) = P(\sigma_1^i | \sigma_0^i) P(\sigma_0^j) + P(\sigma_0^i | \sigma_1^i) P(\sigma_1^j).
\]

Taking 1 - \(P(\text{up})\) thus yields the Bayesian p-value of a gene going up. To answer the same question when working on continuous expression data one possibility is to employ a one sided two sample t-test in the Welch approximation of unknown variances in the varieties. This enables testing, for each gene, whether the mean of expression is higher or lower in one variety than in another. For comparison of these two approaches we selected a set of genes based on their well documented expression pattern and biological functions in the developing B lymphocyte [16,19]. Several of these are functionally linked since they participate directly in somatic DNA rearrangement events occurring specifically at the pre-B cell stage or participate in the regulation of genes involved in this process and thus display restricted expression patterns (pre-B specific). A second set of genes were selected based on their expression in cells that are either committed to the B lineage (B-lineage specific genes, in pre-B and B-cells) or non committed to this developmental pathway (Not in B-lineage, expressed in pro-B cells) [21].

The result of this comparison is presented in Table 2. For 14 out of the 22 genes the two methods completely agree. Out of these 14 only one (Mb1) does not match the expected target profile. For the other 8 genes, where the two methods yield different results, the probabilistic state analysis gives the expected answer in 5 cases, which should be compared to the two cases, where the t-test gives the right answer. In one case (rag-1), neither of the two methods gives the expected result.
For the subset of genes considered here, our method has an advantage of 5 : 2 in giving the correct (i.e. expected) expression pattern. However, the number of samples is not big enough to draw firm conclusions from this result.

**Conclusions**

The method we have presented serves several purposes:

1. It gives a measure of the biological variation of the genes’ expression in different varieties.
2. It estimates each hybridizations’ global error probabilities. These parameters are very useful as they serve as quality/consistency indicators of the samples of each variety.

---

**Figure 3**

The average error in the estimation of the error probabilities \( \{ P_{i,j} \}^4_{i=1,j=1} \). For correlation factors above 0.2, \( P_{1,0}^1 \) and \( P_{1,0}^2 \) are notably raised. Patterns were these bits deviate from the other two are then not considered as random but rather caused by an error. This effect could only be avoided by introducing extra parameters for correlation between bits.

**Table 1: Typical parameter values. Summary of the estimated parameter values for the B-cell data.** Pe refers to the set of error probabilities, i.e., \( \left[ P_{i,j}^1 \right]_{i=1,j=1}^4 \).

<table>
<thead>
<tr>
<th></th>
<th>( P(\sigma_i) )</th>
<th>( P(\sigma_0) )</th>
<th>( P(\sigma_j) )</th>
<th>Max ( P_e )</th>
<th>Min ( P_e )</th>
<th>Median ( P_e )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pro B-cells</td>
<td>0.405</td>
<td>0.460</td>
<td>0.135</td>
<td>0.035</td>
<td>0.0002</td>
<td>0.020</td>
</tr>
<tr>
<td>Pre B-cells</td>
<td>0.395</td>
<td>0.450</td>
<td>0.155</td>
<td>0.047</td>
<td>0.003</td>
<td>0.028</td>
</tr>
<tr>
<td>Full B-cells</td>
<td>0.343</td>
<td>0.471</td>
<td>0.186</td>
<td>0.073</td>
<td>0.0007</td>
<td>0.022</td>
</tr>
</tbody>
</table>
3. Given the parameters above, it estimates the probability of a gene belonging to each of the three groups $\sigma_0$, $\sigma_1$, and $\sigma_2$. These probabilities in turn indicate whether the gene is likely to be below, fluctuating around or above the threshold of discretization.

4. Clustering of genes to expression profiles over a set of different varieties is achieved with Eq. (4). The probability, i.e. belief, that a gene belongs to a certain cluster is exactly quantified.

This novel approach is proven valuable for quantifying both data reliability and underlying gene expression in microarray experiments. Our method has been successfully applied in two different projects [22,20].

### Authors' Contributions

SB and TB have contributed to the development of the model and implemented the algorithms. MS has contributed the experimental data and biological expertise.

### Acknowledgments

SB and TB are supported by Knut and Alice Wallenberg Foundation through the SWEGENE consortium. MS is supported by the Swedish science council and the A. Osterlund foundation. The authors also extend their gratitude to Carsten Peterson, Patrik Edén, Jari Häkkinen and Markus Rignér, for fruitful discussions and careful proofreading of the manuscript.

### References


