Gene expression

A statistical model providing comprehensive predictions for the mRNA differential display

David Meintrup\(^1\) and Ellen Reisinger\(^2\)*

\(^1\)Institut EIT 1, Universitaet der Bundeswehr Muenchen, 85577 Neubiberg, Germany and \(^2\)Department of Physiology II, Universitaet Freiburg, 79104 Freiburg, Germany

Received on June 30, 2005; revised on August 9, 2005; accepted on August 16, 2005

Advance Access publication August 18, 2005

ABSTRACT

Motivation: Differential display (DD) or arbitrarily primed fingerprinting serves to identify differentially expressed genes, but these techniques cannot determine how many of the theoretically available genes have been uncovered. Previous mathematical models are unsatisfying as they are not suitable to analyze experimental data.

Results: In the present study, we provide a statistical model based on the redundancy of cDNA fragments amplified during DD experiments. This model is applicable to any DD and predicts (1) the total number of genes expressed in a sample cell type or tissue, (2) the number of differentially expressed genes, (3) the coverage obtained with any given number of primer combinations. In a DD experiment comparing two developmental stages of the post natal rat inner ear, we estimated the total number of differentially expressed genes accessible by DD to be 445, and the number of primer combinations required to uncover 90% of these to be 127.

Availability: The algorithms were implemented in Matlab\textsuperscript{TM} (The Mathworks, Inc., Natick, MA) environment and are available at www.physiologie.uni-freiburg.de/download.html

Contact: ellen.reisinger@physiologie.uni-freiburg.de

1 INTRODUCTION

The mRNA differential display (DD) is a reliable and highly sensitive technique to compare gene expression in different cell types or between distinct developmental stages of a given cell type (Liang and Pardee, 1992; Liang et al., 1993; McClelland et al., 1995; Wan et al., 1996). cDNA fragments are amplified in a PCR using arbitrary primers combined with one- or two-base anchored oligo(dT)-primers each. These primers bind multiply in the transcriptome in a partly degenerated way (Liang and Pardee, 1992; Bauer et al., 1993). Mixtures of DNA fragments resulting from DD-PCRs are displayed after electrophoresis and autoradiography. Selected cDNA fragments are recovered from the gel and identified by cycle sequencing. As the intensity of a band in the display reflects the mRNA amount of the respective gene, the DD allows selective identification of genes that differ in their expression level.

To date, proven arbitrary primers as well as reliable protocols are available (see Systems and Methods section), and since genomic sequences of common laboratory animals are on hand, most sequences of the identified bands match with public database entries, making this technique easier to handle than some years ago. However, at present little is known about the coverage, i.e. the portion of differentially expressed genes detected by the DD.

Several efforts were undertaken to predict the number of arbitrary primers required to cover most expressed genes in a cell. First estimations by Liang and Pardee (1992) predicted 240 primer combinations required to statistically cover all mRNA sequences expressed in one type of cell. These predictions, however, suffered from inaccurate estimations on the total number of genes in higher organisms, and did not account for redundancy, i.e. the possibility that the same cDNA might be ‘hit’ by several primer combinations. Recently, a highly elaborated mathematical model (Yang and Liang, 2004) predicted the probability for a single cDNA to be amplified at least once after a certain number of DD-PCRs, therefore being independent of redundancy or the total number of genes. This model is based on predictions on the hybridization of primers and cDNA under certain experimental conditions and considers the degeneracy of primer binding in detail. However, it exhibits major disadvantages: first, no experimental data on the degenerate binding of the 5’ end of arbitrary primers other than 13mers are provided, making the model barely applicable to a DD using longer or shorter arbitrary primers. Second, changes in experimental conditions will alter the binding properties of the primers and affect the basic assumptions of this model. Third, at least some of the parameters required for predictions in this model may not be verified experimentally. Fourth, as small changes in the parameters result in major deviations in the predicted coverage, this model gives inaccurate predictions. Moreover, none of the models proposed so far is applicable to so-called ‘RAP-PCR’ (‘RNA fingerprinting by arbitrarily primed PCR’), where two arbitrary primers are combined with each other (Welsh et al., 1992).

In the present study, we provide a statistical model for mRNA DDs as well as RAP-PCRs that was developed on experimental data of a combined DD/RAP experiment consisting of 28 arbitrarily primed PCRs. Different from previous models, the model presented here is based on experimental observations of redundantly amplified genes, provided that a finite number of expressed genes in any examined tissue will result in repeated amplification of the same gene after a sufficient number of DD cycles. It is shown that this approach provides valid predictions for mRNA DD or RAP experiments, e.g. the total number of expressed genes in the examined tissue, the number of differentially expressed genes, and the course of coverage and redundancy.
supplemented with the Delta Differential Display Kit (Clontech), that is based on a long-distance DD-PCR described in Diachenko et al. (1996). Briefly, DD-PCRs were carried out in duplicate with 5 pmol of each primer and cDNA corresponding to 2.4 ng mRNA (higher concentration) or 0.6 ng mRNA (lower concentration) in a total volume of 20 μl consisting of 1× Klen Taq PCR–Buffer (Clontech), 50 μM dNTPs each, 50 nM [α-33P]dATP (2500 Ci/mmol, Amersham, Freiburg, Germany) and 0.4 μl Advantage Klen Taq Polymerase Mix (Clontech). The temperature cycle protocol comprised 5 min 94°C, 5 min 40°C, 5 min 68°C followed by two cycles with 1 min 94°C, 2 min 40°C and 5 min 68°C and 25 cycles with 45 s at 94°C, 45 s 49°C and 2 min 68°C and a final elongation step at 68°C for 7 min. The resulting DNA fragments were separated electrophoretically on a denaturing acrylamide gel and visualized through autoradiography.

2.3 Analysis of the DD
The number of bands in each fingerprint was counted visually at least three times using the mean value as the result. To reduce the number of false positives, only bands occurring in both duplicates were considered (Diachenko et al., 1996), and bands were regarded as differential only if both bands in the duplicate fingerprint were stronger than both bands in the other. Owing to our special interest in upregulated genes, only DNA bands with higher intensity at the later stage of development are indicated as differential; bands with higher intensity at the earlier stage are included in the non-differential bands. Differential bands were excised from the gel, reamplified and sequenced by cycle sequencing. If possible, reamplified PCR products were sequenced directly, otherwise the cDNA fragments were

### Table 1. Numbers of bands in the DD experiment

<table>
<thead>
<tr>
<th>Subexperiment no.</th>
<th>Combined primers</th>
<th>Sample size</th>
<th>Sample size (sum)</th>
<th>Differential bands</th>
<th>Differential bands (sum)</th>
<th>Non-redundant differential bands</th>
<th>Non-redundant differential bands (sum)</th>
<th>ML-estimator for p</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>P4 &amp; T5</td>
<td>41</td>
<td>41</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>1.022</td>
</tr>
<tr>
<td>2</td>
<td>P4 &amp; T7</td>
<td>48</td>
<td>89</td>
<td>5</td>
<td>10</td>
<td>5</td>
<td>10</td>
<td>1.012</td>
</tr>
<tr>
<td>3</td>
<td>P4 &amp; T3</td>
<td>41</td>
<td>130</td>
<td>5</td>
<td>15</td>
<td>4</td>
<td>14</td>
<td>1.154</td>
</tr>
<tr>
<td>4</td>
<td>P5 &amp; T5</td>
<td>47</td>
<td>177</td>
<td>10</td>
<td>25</td>
<td>10</td>
<td>24</td>
<td>1.1412</td>
</tr>
<tr>
<td>5</td>
<td>P5 &amp; T7</td>
<td>40</td>
<td>217</td>
<td>8</td>
<td>34</td>
<td>8</td>
<td>32</td>
<td>1.1567</td>
</tr>
<tr>
<td>6</td>
<td>P5 &amp; T3</td>
<td>34</td>
<td>251</td>
<td>8</td>
<td>42</td>
<td>6</td>
<td>38</td>
<td>1.1673</td>
</tr>
<tr>
<td>7</td>
<td>P6 &amp; T5</td>
<td>55</td>
<td>306</td>
<td>5</td>
<td>47</td>
<td>5</td>
<td>43</td>
<td>1.1536</td>
</tr>
<tr>
<td>8</td>
<td>P6 &amp; T7</td>
<td>48</td>
<td>354</td>
<td>7</td>
<td>54</td>
<td>6</td>
<td>49</td>
<td>1.1525</td>
</tr>
<tr>
<td>9</td>
<td>P6 &amp; T3</td>
<td>38</td>
<td>392</td>
<td>3</td>
<td>57</td>
<td>2</td>
<td>51</td>
<td>1.1454</td>
</tr>
<tr>
<td>10</td>
<td>P7 &amp; T5</td>
<td>47</td>
<td>440</td>
<td>13</td>
<td>70</td>
<td>10</td>
<td>61</td>
<td>1.1591</td>
</tr>
<tr>
<td>11</td>
<td>P7 &amp; T7</td>
<td>47</td>
<td>487</td>
<td>2</td>
<td>72</td>
<td>2</td>
<td>63</td>
<td>1.1478</td>
</tr>
<tr>
<td>12</td>
<td>P7 &amp; T3</td>
<td>45</td>
<td>532</td>
<td>6</td>
<td>78</td>
<td>6</td>
<td>69</td>
<td>1.1466</td>
</tr>
<tr>
<td>13</td>
<td>P1 &amp; P5</td>
<td>45</td>
<td>577</td>
<td>5</td>
<td>83</td>
<td>4</td>
<td>73</td>
<td>1.1438</td>
</tr>
<tr>
<td>14</td>
<td>P2 &amp; P5</td>
<td>48</td>
<td>625</td>
<td>6</td>
<td>89</td>
<td>3</td>
<td>76</td>
<td>1.1424</td>
</tr>
<tr>
<td>15</td>
<td>P4 &amp; P5</td>
<td>49</td>
<td>674</td>
<td>5</td>
<td>94</td>
<td>5</td>
<td>81</td>
<td>1.1395</td>
</tr>
<tr>
<td>16</td>
<td>P9 &amp; P5</td>
<td>60</td>
<td>734</td>
<td>24</td>
<td>118</td>
<td>21</td>
<td>102</td>
<td>1.1608</td>
</tr>
<tr>
<td>17</td>
<td>P1 &amp; P6</td>
<td>51</td>
<td>785</td>
<td>11</td>
<td>129</td>
<td>9</td>
<td>111</td>
<td>1.1643</td>
</tr>
<tr>
<td>18</td>
<td>P2 &amp; P6</td>
<td>51</td>
<td>836</td>
<td>4</td>
<td>133</td>
<td>2</td>
<td>113</td>
<td>1.1591</td>
</tr>
<tr>
<td>19</td>
<td>P4 &amp; P6</td>
<td>47</td>
<td>883</td>
<td>8</td>
<td>141</td>
<td>7</td>
<td>120</td>
<td>1.1597</td>
</tr>
<tr>
<td>20</td>
<td>P9 &amp; P6</td>
<td>57</td>
<td>940</td>
<td>11</td>
<td>152</td>
<td>7</td>
<td>127</td>
<td>1.1617</td>
</tr>
<tr>
<td>21</td>
<td>P1 &amp; P7</td>
<td>48</td>
<td>988</td>
<td>5</td>
<td>157</td>
<td>3</td>
<td>130</td>
<td>1.1589</td>
</tr>
<tr>
<td>22</td>
<td>P2 &amp; P7</td>
<td>44</td>
<td>1032</td>
<td>4</td>
<td>161</td>
<td>3</td>
<td>133</td>
<td>1.1566</td>
</tr>
<tr>
<td>23</td>
<td>P4 &amp; P7</td>
<td>48</td>
<td>1080</td>
<td>1</td>
<td>162</td>
<td>1</td>
<td>134</td>
<td>1.150</td>
</tr>
<tr>
<td>24</td>
<td>P9 &amp; P7</td>
<td>69</td>
<td>1149</td>
<td>15</td>
<td>177</td>
<td>9</td>
<td>143</td>
<td>1.154</td>
</tr>
<tr>
<td>25</td>
<td>P1 &amp; P10</td>
<td>62</td>
<td>1211</td>
<td>14</td>
<td>191</td>
<td>11</td>
<td>154</td>
<td>1.1577</td>
</tr>
<tr>
<td>26</td>
<td>P2 &amp; P10</td>
<td>62</td>
<td>1273</td>
<td>6</td>
<td>197</td>
<td>5</td>
<td>159</td>
<td>1.1548</td>
</tr>
<tr>
<td>27</td>
<td>P4 &amp; P10</td>
<td>69</td>
<td>1342</td>
<td>12</td>
<td>209</td>
<td>7</td>
<td>166</td>
<td>1.1557</td>
</tr>
<tr>
<td>28</td>
<td>P9 &amp; P10</td>
<td>65</td>
<td>1407</td>
<td>14</td>
<td>223</td>
<td>8</td>
<td>174</td>
<td>1.1585</td>
</tr>
</tbody>
</table>

Experimental data for a differential display experiment comparing two developmental stages of the rat inner ear.

2. SYSTEMS AND METHODS

2.1 cDNA preparation for DD
mRNA samples were isolated from organs of Corti of newborn (P0) and of rats on post natal day 14 (P14) using DYNABeads (Dynal, Oslo, Norway) according to the manufacturer’s instructions. cDNA was synthesized by reverse transcription of ∼150 ng of P0 or P14 mRNA with Superscript II (Invitrogen, Karlsruhe, Germany) in supplemented buffer and 10 mM DTT, 0.5 mM dNTPs, 1 pmol oligo(dT)-primer and 0.3 U RNasin (Promega, Madison, WI) for 2 h at 42°C. The integrity of the cDNA was confirmed by a PCR amplifying a fragment of GAPDH cDNA.

2.2 DD-PCR
Primer sequences were derived from Delta Differential Display Kit (BD Biosciences Clontech, Palo Alto, CA). Arbitrary ‘P-primers’ consisted of a consensus sequence of 16 nt (5′-ATTAACCTCACTAA-3′) and a 9 nt arbitrary sequence at the 3′ end:

5′-TGCTGGGA-3′ (P1), 5′-TGCTGCTAG-3′ (P2), 5′-TGCTGGTAG-3′ (P4), 5′-GATCTGAGT-3′ (P5), 5′-TGCTGGTG-3′ (P6), 5′-TGCTGGTAGT-3′ (P7), 5′-TGCTGGAGG-3′ (P9) and 5′-GCACCGTCC-3′ (P10).

The consensus sequence of the ‘T-primers’ consisted of 28 nt (5′-CATTATGCATGATCATCTTTTTTTTTTTTTTTTTTT-3′) followed by 5′-AG-3′ (T3), 5′-CC-3′ (T5) or 5′-GA-3′ (T7) at the 3′ end. These primers were combined for 28 DD-PCRs as indicated in Table 1. The differential display procedure was performed as recommended in the user manual (PT 1173-1)
2.4 Computer simulation

The computer simulated experiments were performed by generating pseudo-random numbers as realizations of the random variables that will be introduced in the next section. Based on the 90 combinations of T-primers (two-base anchored primers with 5'-Linker-7T, VV-3', where V is A, G or C) with 10 different P-primers and 45 combinations of P-primers with each other possible in a commercially available DD kit, we exemplarily simulated a DD experiment consisting of n=135 primer combinations.

3 ALGORITHM

A DD experiment identified genes that were upregulated during the postnatal development of the inner ear. Identity and significance of the differentially expressed genes for inner ear physiology will be published elsewhere (Reisinger et al., 2005), an extract of the autoradiography is available as Supplementary information at the Journal’s web page. In this DD, a total of 1407 cDNA fragments was amplified, and 223 of these bands were found to be differential. Subsequent analysis by quantitative real-time PCR on independent RNA isolates confirmed that 93% of the identified genes were indeed differential, i.e. exhibiting at least 2-fold upregulation during inner ear maturation (data not shown). From these 223 differential bands 174 originated from different mRNA sequences, whereas 49 bands corresponded to redundantly amplified genes. Closer investigation of this redundancy showed no clear correlation between the frequency with which a gene was identified and the expression level quantified by real-time PCR (data not shown), as observed earlier (Wan et al., 1996). Furthermore, even though it has been described that fingerprints using the same two-base anchored oligo(dT) primer exhibit increased redundancy owing to mispriming effects (Liang et al., 1993), only two redundantly occurring genes were amplified with the same two-base anchored oligo(dT)-primer, and both did not originate from mispriming. Therefore, we consider the event of two primers matching a cDNA so that a 250–1500 bp fragment is amplified as a random event with equal probability.

The number of redundant bands increased with increasing numbers of DD-PCRs performed. From this observation we hypothesized that the number of genes identified during the experiment will saturate, leading to the idea of using a detailed analysis of the redundancy to predict the total number of differentially expressed cDNAs and furthermore, the total number of cDNAs in the compared tissues. To this end, we developed a statistical model for the mRNA DD providing reliable answers of the following questions in statistical mean:

(1) What is the total number of differentially expressed genes?
(2) What is the total number of expressed genes?
(3) How does the redundancy, i.e. the number of repeatedly identified genes increase during the series of DD-PCRs?
(4) What is the development of the coverage?
(5) How many primer combinations are required to achieve a certain coverage?

For the following considerations, the total of fingerprints originating from the same primer combination are termed as 'subexperiment'. To describe the statistical model for the mRNA DD, the subsequent notation is introduced:

\[
\begin{align*}
N & \quad \text{the total number of expressed genes}, \\
N_1 & \quad \text{the number of differentially expressed genes}, \\
p & \quad N_1/N \quad \text{the proportion of differentially expressed genes}, \\
n & \quad \text{number of performed subexperiments}. \\
\end{align*}
\]

Generally, three types of bands may be distinguished in any of our DD experiments: redundant differential (\(b_r\)) and non-differential (\(b_o\) and \(b_n\) ones. Every subexperiment in which a subset of cDNA fragments is amplified corresponds to a random choice of bands that will consist of these three possible types. Thus, with \(y_i\) denoting the sample size of the \(i\)-th subexperiment and \(z_{i,j}\) denoting the total number of identified non-redundant differential bands during the first \(i-1\) subexperiments, we can describe the \(i\)-th subexperiment by a probability space given by the sample space

\[
\Omega_i := \{(b_r, b_o, b_n) \in \{0, \ldots, y_i\}, b_r + b_o + b_n = y_i\},
\]

and by the multinomial distribution,

\[
\mathbb{P}(b_r, b_o, b_n) = \left( \frac{y_i}{b_r, b_o, b_n} \right) q_1 q_2 q_3.
\]

The success probabilities \(q_1, q_2, q_3\) are given by the respective proportions of the three types of bands, i.e.

\[
q_1 := \frac{z_{i-1}}{N}, \quad q_2 := \frac{N_1 - z_{i-1}}{N}, \quad q_3 := \frac{N - N_1}{N} = 1 - p.
\]

Hence, \(\mathbb{P}(b_r, b_o, b_n)\) is the probability that the \(i\)-th subexperiment produces \(b_r\) redundant differential bands, \(b_o\) non-redundant differential bands and \(b_n\) non-differential bands.

In the DD, the sample size \(y_i\) is unknown, too. Therefore, we also have to model the sample size in each subexperiment. To that end let the random variable

\(Y_i\) denote the sample size of the \(i\)-th subexperiment, \(i = 1, \ldots, n\).

We assume that the random variables \(Y_1, \ldots, Y_n\) are independent and identically distributed. In total, our statistical model for the DD consists of

\((\Omega, \mathbb{P}): \text{probability space with multinomial distribution}, \quad Y_1, \ldots, Y_n: \text{independent and identically distributed}.)

With this model in hand we can derive other random variables of interest. We denote by

\(X_i\) the number of differential bands in the \(i\)-th subexperiment,

\(D_i\) the total number of differential bands after \(i\) subexperiments,

\(Z_i\) the total number of non-redundant differential bands after \(i\) subexperiments,
the coverage $C_i$ is the proportion of differentially expressed genes that had been identified after $i$ subexperiments, i.e.

$$C_i = \frac{Z_i}{N_i}. \quad (1)$$

This statistical model allows to derive estimators for some parameters of interest (see the Appendix in Supplementary information for details). Denoting by $\bar{y}_i, \ldots, \bar{y}_n$ the experimental sample sizes and by $\bar{y}$ the estimator, we use

$$\bar{y} = \frac{\bar{y}_1 + \cdots + \bar{y}_n}{n} \quad (2)$$

as best unbiased estimator for the sample size. For the parameter $p$ we use the maximum-likelihood estimator, which after $i$ subexperiments is given by

$$\hat{p}_i = \frac{\sum_{j=1}^{i} \bar{x}_i}{\bar{y}_1 + \cdots + \bar{y}_n} = \frac{\sum_{j=1}^{i} \bar{x}_i}{\bar{y}_i}, \quad i = 1, \ldots, n. \quad (3)$$

where $\bar{x}_1, \ldots, \bar{x}_i$ is a sample of $X_1, \ldots, X_i$, i.e. the experimentally observed number of differential bands. Similarly, we get an estimator $\tilde{N}$ for the total number of cDNAs by maximizing the function

$$F(N) := f_1(N)(\bar{u}_i) \times \cdots \times f_n(N)(\bar{u}_n), \quad (4)$$

so that

$$F(\tilde{N}) \geq F(N) \quad \text{for all } N \in \mathbb{N}, \quad (5)$$

where $f_i(N)$ is the binomial density function of $U_i$, i.e.

$$f_i(N)(k) = \left(\frac{\bar{y}_i}{k}\right) \left(\bar{p}_i - \bar{z}_{i-1}/N\right)^k \left(1 - \left(\bar{p}_i - \bar{z}_{i-1}/N\right)\right)^{\bar{y}_i-k}, \quad k \in \{0, \ldots, \bar{y}_i\}. \quad (6)$$

Here $\bar{u}_i = \bar{z}_i - \bar{z}_{i-1}$ denotes the number of new non-redundant bands in the $i$-th subexperiment. Note that the numerical maximization of $F(N)$ will not exhibit a maximum if no redundancy in the experimental data is provided. Using the estimators $\bar{y}, \bar{p}_i, \bar{z}_i$ and $\tilde{N}$ one can derive estimators $\tilde{N}_i, \bar{z}_i, \bar{u}_i, \bar{r}_i$ and $\hat{c}_i$ for the total number of differential cDNAs $N_i$, the total number of non-redundant differential bands $Z_i$, the number of new non-redundant differential bands $U_i$ and the redundancy $R_i$ in the $i$-th subexperiment and finally for the coverage $C_i$ after $i$ subexperiments (for details see the Appendix, Supplementary information).

$$\tilde{N}_i = \frac{\bar{p}_i \cdot \tilde{N}}{\bar{n}}, \quad (6)$$

$$\bar{d}_i = \bar{p}_i \bar{y}, \quad i = 1, \ldots, n, \quad (7)$$

$$\hat{z}_i = \bar{p}_i \tilde{N} \left[1 - \left(1 - \frac{\bar{y}}{N}\right)^{i-1}\right], \quad i = 1, \ldots, n, \quad (8)$$

$$\bar{u}_i = \hat{z}_i - \hat{z}_{i-1} = \bar{p}_i \bar{y} \left[1 - \frac{\bar{y}}{N}\right]^{i-1}, \quad i = 1, \ldots, n, \quad (9)$$

$$\bar{r}_i = \hat{p}_i \bar{y} \left[1 - \left(1 - \frac{\bar{y}}{N}\right)^{i-1}\right] = \frac{\bar{y}}{N} \hat{z}_{i-1}, \quad (10)$$

These estimators provide answers for all questions set out above. This will be demonstrated for an experimental dataset in the following section.

4 IMPLEMENTATION

The numbers of bands in total, differential and non-redundant bands in each of the 28 DD-PCRs as well as the applied primers are listed in Table 1. Note that in the first 12 subexperiments an arbitrary P-primer was combined with a T-primer (see Systems and Methods), whereas in the following subexperiments two P-primers were combined (RAP-PCRs). Using the data in Table 1, estimators for all parameters may be calculated.

First, according to Equation (2) the mean sample size is given as

$$\bar{y} = \frac{\bar{y}_1 + \cdots + \bar{y}_n}{n} = \frac{1407}{28} = 50.25$$

with a standard deviation of 9.05 (see Appendix in Supplementary information). The sample sizes differed slightly between the classical DD experiments (1–12) with $\bar{y}_{DD} = 44.33$ and the RAP-PCRs (13–28) with $\bar{y}_{RAP} = 54.56$. In the following considerations, the mean value of all 28 subexperiments was used. Next, the maximum-likelihood estimator for $p$ is calculated after any subexperiment according to Equation (3) and given in the last column of Table 1. Figure 1A illustrates the convergence of the graph for $p$ with an increasing number of subexperiments. The most accurate estimator for $p$ is the last one,

$$\bar{p} := \bar{p}_{28} = 0.158$$

with a 90% confidence interval (CI) of [0.142, 0.176] (see Appendix in Supplementary information). As shown in Figure 1A, the accuracy of the values for $\bar{p}$ increases with an increasing number of subexperiments. More precisely, the length of the confidence interval for $\bar{p}$ is proportional to $1/\sqrt{r}$ with $r$ being the summed sample size.

According to Equation (4), to calculate $\tilde{N}$ the function $F(N)$ has to be maximized, which is the product of the densities $f_i(N)(\bar{u}_i)$. Therefore, the data of Table 1 are implemented in $f_i(N)(k)$. For instance, for every $N \in \mathbb{N}$ we have

$$f_1(N)(\bar{u}_3) = \left(\frac{4^1}{4}\right) \left(\frac{\bar{p} - \frac{10}{N}}{4}\right) \left(1 - \left(\frac{\bar{p} - \frac{10}{N}}{4}\right)\right)^{41-4}. \quad (11)$$

Subsequently, the maximization is carried out numerically. Figure 1B shows the graph of $F(N)$. The numerical optimization leads to a maximum of $f(N)$ for

$$\tilde{N} = 2806 \quad (90\% \text{CI: } [2456, 3228])$$

Figure 1C shows stabilization of $\tilde{N}$ with an increasing number of subexperiments. Hence, approximately 2800 different cDNA transcripts may be amplified under standard PCR conditions from the rat organ of Corti. Using Equation (6) leads to an estimated number of upregulated genes of

$$\tilde{N}_1 = \frac{\bar{p} \cdot \tilde{N}}{\bar{n}} = 0.158 \times 2806 \approx 445 \quad (90\% \text{ CI: } [380, \pm 510]).$$

Combining the estimators for $\bar{p}, \tilde{N}$ and $\tilde{N}_1$ and Equations (7) and (8), the expected number of differential bands as well as the number
Experiments correlate with the calculated course for calculated estimators. Accuracy of random numbers for performed computer simulated experiments by generating pseudo-experimental data. To verify the predictions of the model, we (subexperiments 1–28) in Figure 2B, here complemented by the be computed, as illustrated in Figure 2A, and at an expanded scale (Fig. 2) represent a computed DD experiment using the same estimators for \(^p\) and \(N\). The ratio of the sum of differential and non-differential bands is the maximum-likelihood estimator for \(p\). Accuracy of \(p\) increases with an increasing number of subexperiments. (B) The numerical maximization of the likelihood function for \(N\) after 28 subexperiments exhibits a maximum at 2800, corresponding to the total number of cDNAs (available under standard PCR conditions). (C) Training of the model with the first 15–27 subexperiments only leads to the plotted estimations for \(N\).

![Graphs showing prediction curves](image)

**Fig. 1.** Estimation of the parameters \(p\) and \(N\). (A) The ratio of the sum of differential and non-differential bands is the maximum-likelihood estimator for \(p\). (B) The numerical maximization of the likelihood function for \(N\) after 28 subexperiments exhibits a maximum at 2800, corresponding to the total number of cDNAs (available under standard PCR conditions). (C) Training of the model with the first 15–27 subexperiments only leads to the plotted estimations for \(N\).

The model predictions were further evaluated by training the model with the first 14–27 subexperiments and predicting the number of differential and non-redundant differential bands after 28 experiments. The number of training experiments is rather small, the predictions for \(d_{28}\) and \(z_{28}\) are in fair agreement with the experimentally obtained values. Another validation method that shows an even better match is obtained by performing a sort of resampling, also shown in Figure 3. For this purpose, 20 sets of subexperiments were chosen randomly without replacement and the resulting predictions for \(d_{28}\) and \(z_{28}\) were averaged. For example, the averaged predictions of 20 sets comprised 20 subexperiments each leading to a total number of expected differential genes to 226, compared with 223 experimentally obtained differential bands. Similarly, the averaged predictions for \(z_{28}\) using...
assumes that the sample size for the same experimental conditions results in at least 1 redundant band may be expected. For our experimental conditions this will give more accurate and reliable predictions than other mathematical models. The statistical model allows further predictions on the coverage of DD experiments: according to Equation (11), the coverage after \( n \) subexperiments, gray dots. For the resampling (black), a randomly chosen subset of subexperiments is used for prediction and the results are averaged, shown with error intervals (see Appendix in Supplementary information). The predictions improve with an increasing number of considered subexperiments, as indicated by the error intervals. Predictions are in agreement with the experimentally observed number of differential bands (223) and non-redundant bands (174).

The statistical model allows predictions for further DD-PCRs. The resulting model enables one to estimate the total number of genes expressed in the examined tissue, the number of differentially expressed genes and the coverage in dependence of the number of applied primer combinations. The model is strongly supported by our experimental data, the validation of the predictions by the resampling procedure and the results of computer simulated experiments. These validations also show a reasonable robustness of the predictions toward small changes in the dataset.

Previous mathematical models relied on theoretical predictions of the hybridization of arbitrary P-primers and T-primers with cDNA molecules. The first estimations on how many primer combinations are required to cover the cellular mRNAs (Liang and Pardee, 1992; Bauer et al., 1993, 1994) based on the assumption that higher organisms express \( \sim 100,000 \) genes and individual cells make use of 15% of these. This assumption might easily be corrected using the most recent approximations for the number of expressed genes from the International Human Genome Sequencing Consortium (IHGSC, 2004). However, their predictions on the DD coverage underestimate the required primer combinations as they do not account for the possibility that more than one cDNA fragment of the same gene might be amplified. A more detailed study of the statistics of primer binding of Yang and Liang (2004) predicted 160 arbitrary primers combined with 3 one-base anchored T-primers will amplify a cDNA of 600 bp length with a probability of 93%. The resulting 480 primer combinations for 93% coverage are obviously higher than the 167 primer combinations leading to 93% coverage estimated from Equation (12) in our model (using \( \gamma_{DD} = 44.33 \)). This difference might be explained by the fact that hybridization-based models rely on rough estimations of parameter values that are at least in part not accessible to experimental tests. Moreover, the predictions vary over a broad range if small changes in parameter values as length of the cDNAs or number of base matches between arbitrary primers and cDNA are assumed. Since all parameters in our model are calculated from experimental data, this model is supposed to give more accurate and reliable predictions than other mathematical models.
models describing DD experiments. Furthermore, our statistical model is the first suitable model for experiments using combinations of internally binding arbitrary primers (RAP-PCRs). Another possible application are DD experiments comparing more than two samples, as these can be regarded as a set of one-by-one comparisons (e.g. bands that arise from only one sample but not from all others), and thus each set can be analyzed separately. Moreover, the model predictions are not affected by the use of total RNA instead of mRNA, since most mRNA preparations are not completely free of rRNA. Taken together, the major benefit of the mathematical model proposed here relies on the parameters that are estimated from experimental data, making this model widely applicable. When adopted to a new dataset, the reliability of the model predictions can be verified by the provided validation methods (as in Figs 1A, C and 3).

The usefulness of this model is demonstrated by implementation on a DD comparing two developmental stages of the organ of Corti, a rather small and homogeneous tissue of neuroepithelial origin. The considerable proportion of 15% of upregulated genes reflects intense morphological and physiological maturation processes that take place during inner ear development (Pujol et al., 1997). The calculated 2800 genes predict ~11–14% of the protein encoding genes of higher organisms to be expressed in the mammalian organ of Corti tissue (IHGSC, 2004). For two reasons, however, we expect that the total number of expressed genes in the organ of Corti is somewhat higher: first, a few cDNA fragments were amplified with only one (mostly arbitrary) primer binding in sense and antisense direction. Hence, the use of the same primer in different combinations can lead to a redundant amplification that is not truly independent, resulting in a small overestimation of redundancy. Consequently, the most accurate application for the model proposed here would be an experiment using every primer only once. Second, it should be kept in mind that not every DNA sequence is successfully amplified in a standard PCR, and an efficient amplification of sequences with high GC content requires the supplementation of DMSO or other hydrogen bond weakeners (Arezi et al., 2003). As our DD was performed under standard PCR conditions, we conclude that a series of cDNA sequences, difficult to amplify, is not accounted for the here proposed estimation on the total number of expressed genes in the inner ear and therefore, for the number of differentially expressed genes. However, according to the fact that such sequences will not be amplified with an increasing number of primer combinations, the coverage calculated from our statistical model respects only genes available under standard PCR conditions.

The statistical model proposed in this study is applicable to any DD experiment after the analysis of a sample of subexperiments as soon as substantial redundancy is observed. Hence for any examined tissue, a DD experiment will be suitable to determine the number of cDNA molecules that are amplified under standard PCR conditions and thus the number of expressed genes might be estimated. Although, to get reliable results it is necessary to avoid redundancy originating from mispriming of the one- or two-base anchored oligo(dT) primers described in Liang et al. (1993). This mispriming can be overcome if the reverse transcription reaction (RT) is performed with non-anchored oligo(dT) primers, and anchored primers are used during the PCR only (Rothschild et al., 1997; Frost and Guggenheim, 1999). The increased priming specificity might result from the higher annealing temperature for PCR than during RT.

ACKNOWLEDGEMENTS

We thank Dr Bernd Fakler for helpful discussions and for critically reading the manuscript. This work was supported by DFG-Graduiertenkolleg 843.

Conflict of Interest: none declared.

SUPPLEMENTARY DATA

Autoradiography of the DD and Mathematical Appendix are available on Bioinformatics online.

REFERENCES


Reisinger,E. et al. (2005) Gene expression associated with the onset of hearing detected by mRNA differential display in mammalian organ of Corti, in press.


