Gene expression

Microarray phenotyping in Dictyostelium reveals a regulon of chemotaxis genes

Ezgi O. Booth¹,², Nancy Van Driessche², Olga Zhuchenko³, Adam Kuspa²,³ and Gad Shaulsky²,¹,*
¹Graduate Program in Structural and Computational Biology and Molecular Biophysics, ²Department of Molecular and Human Genetics and ³Department of Biochemistry and Molecular Biology, Baylor College of Medicine, Houston, TX 77030, USA

Received on August 24, 2005; revised on October 14, 2005; accepted on October 17, 2005
Advance Access publication October 18, 2005

ABSTRACT

Motivation: Coordinate regulation of gene expression can provide information on gene function. To begin a large-scale analysis of Dictyostelium gene function, we clustered genes based on their expression in wild-type and mutant strains and analyzed their functions.

Results: We found 17 modes of wild-type gene expression and refined them into 57 submodes considering mutant data. Annotation analyses revealed correlations between co-expression and function and an unexpected correlation between expression and function of genes involved in various aspects of chemotaxis. Co-regulation of chemotaxis genes was also found in published data from neutrophils. To test the predictive power of the analysis, we examined the phenotypes of mutations in seven co-regulated genes that had no published role in chemotaxis. Six mutants exhibited chemotaxis defects, supporting the idea that function can be inferred from co-expression. The clustering and annotation analyses provide a public resource for Dictyostelium functional genomics.

Supplementary information: http://dictygenome.org/supplement/gadi-okyay_2005/
Contact: gadi@bcm.tmc.edu

INTRODUCTION

Dictyostelium cells grow as single cells and develop into multicellular organisms upon starvation (Kessin, 2001). During development, the cells aggregate by chemotaxis to cAMP, differentiate into prespore and prestalk cells, sort into distinct domains and undergo morphological transformations that lead to fruiting bodies consisting of spores and stalks. The morphological changes are accompanied by transcriptional changes that encompass about half the genes in the genome (Van Driessche et al., 2002). One of the questions raised by these findings was the extent to which changes in gene expression reflected developmental function.

Combining large-scale expression data and gene annotation has been applied to gene function analysis in several organisms (Hughes et al., 2000; Stuart et al., 2003; Zhang et al., 2004). In Dictyostelium, transcriptional phenotypes have been used to derive epistatic relationships (Van Driessche et al., 2005) and to explore development, dedifferentiation and spore germination (Katoh et al., 2004; Van Driessche et al., 2002; Xu et al., 2004). The combination of gene expression and annotation analysis was also used to determine gene function. We have predicted that the histidine kinase gene dhkA was involved in dedifferentiation and confirmed the prediction by showing attenuated dedifferentiation in dhkA mutants (Katoh et al., 2004). We also predicted that the myosin gene myoI was involved in spore germination and showed that myoI mutants had germination defects (Xu et al., 2004). Discovery of novel functions was also demonstrated by Iranfar et al. who found novel cell-type specific genes by microarray analysis (Iranfar et al., 2001). Subsequently, they identified 104 prestalk genes by microarray analysis and verified them by in situ hybridization (Maeda et al., 2003).

Dictyostelium cells produce and respond to pulsatile cAMP signals during development. Iranfar et al. treated wild-type cells and acAA mutants with cAMP pulses and tested gene expression with microarrays (Iranfar et al., 2003). They found groups of cAMP-regulated genes and distinguished the pulse-dependent genes from the pulse independent. These experiments suggested a complex regulation of gene expression in Dictyostelium and a possible correlation between gene expression and function (Shaulsky and Loomis, 2002).

Here we analyzed the correlation between expression and function by clustering co-regulated genes in microarray data from wild-type cells (Van Driessche et al., 2002). We then refined the clustering using data from mutant strains (Kibler et al., 2003a,b; Van Driessche et al., 2005) and analyzed the annotations in each cluster (Xu and Shaulsky, 2005). We found co-regulation of genes that encode components of protein complexes such as the ribosome and the cytoskeleton. Additional analysis revealed an unexpected cluster of genes involved in various aspects of chemotaxis, suggesting that other co-regulated genes may also have chemotaxis roles. We verified that prediction by showing chemotaxis defects in strains mutated in the co-regulated genes.

SYSTEMS AND METHODS

Datasets and preprocessing

Expression datasets (Kibler et al., 2003a, b; Van Driessche et al., 2002, 2005) were weighted based on the number of replications. Table 1 summarizes the genotypes, phenotypes and weights.
Weighted datasets were concatenated into one matrix of 4839 non-redundant genes and 186 instances.

Wild-type data clustering

Unsupervised k-means analysis was used to determine expression patterns in data from wild-type development. The ‘kmeans()’ function in R was used (www.r-project.org) and the clusters’ uniqueness and tightness were evaluated by the F-ratio method (Dudoit et al., 2002). Distances between genes within each cluster, between clusters, and the ratio between the two were calculated as follows:

$$BW = \frac{(SS_B/d_B)}{(SS_W/d_W)},$$

where SS_B is the sum of squared distances between groups, df_B is the degrees of freedom between groups, SS_W is the sum of squared distances within each group, summed across all groups and df_W is the degrees of freedom within the groups.

The concatenated data were used for unsupervised k-means analysis on each wild-type mode. Genes were ordered according to their similarity (Pearson’s correlation) to the mode’s median expression value.

Group significance calculation

We used the hypergeometric distribution [‘Phyper()’ function in R] to determine how unique the distribution of groups of genes was with similar functions in certain modes relative to their distribution amongst all modes. We also used the χ² test in the R function ‘chisq.test().’

Annotation analysis

GO annotation analysis was done with GOAT (Xu and Shaulsky, 2005). The data are presented graphically. Bar lengths represent the ratio between the list frequency (number of genes in list/number of genes at GO level) and the array group frequency (number of genes with specific GO annotation on array/number of all array genes at particular GO level).

cAMP chemotaxis assays

We tested strains: AX4 (Knecht et al., 1986); dimA− (Thompson et al., 2004); yelA− (Osherov et al., 1997); abcC14−, abcB7−, abcB1−, abcE1− (Mode 15); DDB0191712− (Mode 1). Control strains: abcG7−, abcC3−, abcF1−, abcA5− (Mode 17); abcF2− (Mode 12); abcG10− (Mode 14); abcG6− (Mode 3); and abcG5− (Mode 6). Mutants were from the BCM collection (http://dictyensembl.bioch bcm tmc.edu). Cells were grown, collected and spotted on 10 μM cAMP agar plates (Browning et al., 1995; Kibler et al., 2003b). After 22 h, cells were transferred to clear plastic sheets (PV199ED-50, Avery Office Products), dried and stained for 1 h an aqueous solution of 45% methanol, 10% acetic acid and 0.25% Coomassie Brilliant Blue R-250 (BioRad). Cells were washed in 30% methanol, 5% acetic acid/65% water for 2 min and photographed. The diameter of the migrating cells halo was measured. Each strain was assayed 24 times. A one-sided t-test was used to test the null hypothesis that the median cell migration diameter of the mutant was indistinguishable from that of the wild type. P-values < 0.05 reject the null hypothesis.

IMPLEMENTATION

Gene clustering

To identify co-expressed genes, we clustered transcripts with similar expression profiles from developmental time courses (Kibler et al., 2003a,b; Van Driessche et al., 2002, 2005). We first clustered wild-type data using the k-means algorithm. In the algorithm, the number of clusters (k) must be defined, but there is no formal way to optimize k. We optimized it empirically by increasing k from 2 to 30 and measuring distances between groups and between genes within groups (Fig. 1A). The straightforward prediction was that the distance between groups should increase and the distance within groups should decrease smoothly as k increased in the tested range. We hypothesized that if there were an optimal k in that range, the distances would deviate from the smooth tendency. Figure 1A shows an increased distance between the groups for k-values of 16–25 and a decreased distance within the groups for k-values of 16–19. Plotting the data as heat maps (Van Driessche et al., 2002) and inspection of the clusters revealed an optimal value of 17, which yielded groups that we could not readily break or join (data not shown). We used the average gene trajectories of those groups as centers for supervised k-means clustering and obtained 17 clusters (modes) containing 11 to 912 genes (Fig. 1B). Modes 1–6 consist of genes expressed at higher-than-average levels during growth and during the first 6–10 h of development (Fig. 1B). Genes in Mode 7 were transiently upregulated during the first 6–10 h of development. Mode 8–11 genes were upregulated at the end of development. Mode 12–16 genes were upregulated in mid-late development and remained above average thereafter. Mode 17 contained genes that were not grouped with the other modes. The trajectories of many of these genes have been confirmed by northern blots or by comparison to published data (Supplementary Table S3).

To refine the clusters, we considered data from 13 mutant strains (Table 1). We concatenated the mutant datasets, performed cluster analysis in each of the 17 modes and found that 14 modes were divided into significant subgroups, resulting in 57 submodes (Supplementary Fig. 1).

Co-expression of cytoskeletal and ribosomal genes during early development

Automated analysis revealed that several modes were enriched in genes with common annotations. For example, Mode 1 contains 191
genes downregulated after ~6–8 h of wild-type development (Fig. 2A). This mode contains many cytoskeletal and ribosomal genes (Fig. 2B). Analysis of the submodes revealed that the cytoskeleton genes were significantly enriched in Mode 1.3 \( (P < 0.002) \) (Supplementary Fig. 1). Therefore, the division into submodes improved the relationship between co-expression and function.

The difference between the Mode 1 submodes was mainly due to altered gene expression in aceA\(^{-}\) and pufA\(^{-}\)pkaC\(^{-}\). Both strains are aggregationless and their microarray phenotypes are more similar to each other than to other aggregationless mutants (Van Driessche et al., 2005). Submodes 1.2 and 1.3 distinguish these mutants (Fig. 2A), but we found no relationship between that distinction and the annotations in the modes. Most Mode 1.3 genes were not downregulated in the aggregationless mutants aceA\(^{-}\), pufA\(^{-}\)pkaC\(^{-}\).

**Fig. 1.** Gene expression modes in wild-type Dictyostelium. Expression data collected with a microarray from developing cells at 2 h intervals (Van Driessche et al., 2002) were subjected to k-means clustering. (A) Optimization: the number of clusters \( (k) \) was increased from 2 to 30. The distance between the groups (left axis) and within the groups (right axis) is plotted. The optimal number of clusters was 17 (arrow). (B) Modes are plotted as heat maps. Yellow indicates higher-than-average and blue indicates lower-than-average expressions. Each row represents a gene and each column represents a developmental time point (0–24 h). Numbers above the charts designate modes and numbers below \( (n = \ldots) \) indicate the number of genes.

**Fig. 2.** Clustering of cytoskeletal and ribosomal protein genes. Mode 1 was refined by analysis of 13 mutant strains. (A) Expression data were plotted as heat maps; each block represents a mutant and a submode as indicated. Within blocks, each row represents a gene, each column represents a developmental time point (0–24 h). Numbers above the charts indicate the time point and, yellow and blue represent higher- and lower-than-average levels of expression, respectively. Genes in the mutant modes are ordered according to the wild-type mode. \( \text{n = \ldots} \) indicates the number of genes in the submodes. (B) Overrepresented annotations of Mode 1 genes: GO tree levels are shown on the left (numbers, ranging 2–9). Bar lengths represent the enrichment fold (scale x-axis). Indented bars are subgroups of bars immediately above, as indicated by the branching pattern. The table indicates the number of genes with a particular annotation in the group (List), on the entire array (Total), the significance \( (P\text{-value}) \) and the annotation. Ribosome and actin cytoskeleton annotations are highlighted.
E. O. Booth et al.

and Mode 15 contained over 48% of the genes. The statistical Modes 1 and 2 contained 10.5 and 5.3% of the genes, respectively, and the genome database (http://dictyBase.org/) for chemotaxis annotation of these genes was incomplete, we searched the literature enrichment in Mode 15 was especially unusual (<i>E. coli</i> is 0.038) and the <i>P</i> &lt; 5 × 10⁻¹⁴).

### Prediction of chemotaxis gene function

The above results suggested that other genes in Modes 1, 2 and 15 might have chemotaxis function. We selected seven genes to test that possibility. We first verified the expression patterns of three genes with quantitative RT–PCR (data not shown). Five of the genes were mutated by the BCM functional genomics project: the ATP Binding Cassette (ABC) genes <i>abcB1</i>, <i>abcB7</i>, <i>abcC14</i>, <i>abcE1</i> and the un-annotated gene DDB0191712, predicted to encode an 8-transmembrane protein (http://dictyBase.org). We also selected the <b2>ZIP</b2>-transcription factor gene <i>dInA</i> (Thompson et al., 2004) and the putative translation initiation factor gene <i>yelA</i> (Osherov et al., 1997). All but DDB0191712 (Mode 1) were in Mode 15. None has been implicated in chemotaxis before and they were selected only based on serendipitous strain availability, so they may be regarded as randomly selected genes from Modes 1 and 15.

We tested chemotaxis in the mutants in an assay where cells are starved on agar containing cAMP. As the cells develop, they generate a gradient by degrading the cAMP in the agar around them. They then move out, up the cAMP gradient, and form a halo whose diameter is a surrogate measure of chemotactic ability (Browning et al., 1995). Five of the mutants showed a marked decrease in chemotaxis (<i>P ≤ 3 × 10⁻⁹</i>), <i>dInA</i> cells showed a small but significant decrease (<i>P = 0.038</i>) and the <i>abcC14</i> cells were essentially indistinguishable from the wild type (Fig. 3). These

### Coordinate expression of genes implicated in various chemotaxis functions

Manual inspection of the data revealed an unexpected clustering of genes involved in various chemotaxis functions. <i>Dictyostelium</i> aggregation is mediated by cAMP chemotaxis and many molecular components that regulate the process are well characterized (Aubry and Firtel, 1999; Kimmel and Parent, 2003; Parent, 2004; Parent and Devreotes, 1999; Postma et al., 2004). Because the electronic annotation of these genes was incomplete, we searched the literature and the genome database (http://dictyBase.org/) for chemotaxis related genes. Fifty-eight of the genes were represented on our microarray (Table 2) and they were enriched in three modes. Modes 1 and 2 contained 10.5 and 5.3% of the genes, respectively, and Mode 15 contained over 48% of the genes. The statistical significance of these results was very high both by the hypergeometric distribution (Table 3) and by a <i>y²</i> test (data not shown). The enrichment in Mode 15 was especially unusual (<i>P ≤ 3 × 10⁻⁹</i>).

### Prediction of chemotaxis gene function

The above results suggested that other genes in Modes 1, 2 and 15 might have chemotaxis function. We selected seven genes to test that possibility. We first verified the expression patterns of three genes with quantitative RT–PCR (data not shown). Five of the genes were mutated by the BCM functional genomics project: the ATP Binding Cassette (ABC) genes <i>abcB1</i>, <i>abcB7</i>, <i>abcC14</i>, <i>abcE1</i> and the un-annotated gene DDB0191712, predicted to encode an 8-transmembrane protein (http://dictyBase.org). We also selected the <b2>ZIP</b2>-transcription factor gene <i>dInA</i> (Thompson et al., 2004) and the putative translation initiation factor gene <i>yelA</i> (Osherov et al., 1997). All but DDB0191712 (Mode 1) were in Mode 15. None has been implicated in chemotaxis before and they were selected only based on serendipitous strain availability, so they may be regarded as randomly selected genes from Modes 1 and 15.

We tested chemotaxis in the mutants in an assay where cells are starved on agar containing cAMP. As the cells develop, they generate a gradient by degrading the cAMP in the agar around them. They then move out, up the cAMP gradient, and form a halo whose diameter is a surrogate measure of chemotactic ability (Browning et al., 1995). Five of the mutants showed a marked decrease in chemotaxis (<i>P ≤ 3 × 10⁻⁹</i>), <i>dInA</i> cells showed a small but significant decrease (<i>P = 0.038</i>) and the <i>abcC14</i> cells were essentially indistinguishable from the wild type (Fig. 3). These

### Coordinate expression of genes implicated in various chemotaxis functions

Manual inspection of the data revealed an unexpected clustering of genes involved in various chemotaxis functions. <i>Dictyostelium</i> aggregation is mediated by cAMP chemotaxis and many molecular components that regulate the process are well characterized (Aubry and Firtel, 1999; Kimmel and Parent, 2003; Parent, 2004; Parent and Devreotes, 1999; Postma et al., 2004). Because the electronic annotation of these genes was incomplete, we searched the literature and the genome database (http://dictyBase.org/) for chemotaxis related genes. Fifty-eight of the genes were represented on our microarray (Table 2) and they were enriched in three modes. Modes 1 and 2 contained 10.5 and 5.3% of the genes, respectively, and Mode 15 contained over 48% of the genes. The statistical significance of these results was very high both by the hypergeometric distribution (Table 3) and by a <i>y²</i> test (data not shown). The enrichment in Mode 15 was especially unusual (<i>P ≤ 3 × 10⁻⁹</i>).
Table 3. Distribution of chemotaxis genes

<table>
<thead>
<tr>
<th>Mode</th>
<th>Count</th>
<th>%</th>
<th>P-value*</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>6</td>
<td>10.3</td>
<td>0.0036</td>
</tr>
<tr>
<td>2</td>
<td>3</td>
<td>5.2</td>
<td>0.0039</td>
</tr>
<tr>
<td>3</td>
<td>4</td>
<td>6.9</td>
<td>0.9335</td>
</tr>
<tr>
<td>4</td>
<td>1</td>
<td>1.7</td>
<td>0.7947</td>
</tr>
<tr>
<td>5</td>
<td>2</td>
<td>3.4</td>
<td>0.9364</td>
</tr>
<tr>
<td>6</td>
<td>4</td>
<td>6.9</td>
<td>0.7892</td>
</tr>
<tr>
<td>7</td>
<td>1</td>
<td>1.7</td>
<td>0.4566</td>
</tr>
<tr>
<td>8</td>
<td>8</td>
<td>13.8</td>
<td>0.2369</td>
</tr>
<tr>
<td>15</td>
<td>28</td>
<td>48.3</td>
<td>5.4E-14</td>
</tr>
<tr>
<td>17</td>
<td>1</td>
<td>1.7</td>
<td>0.9999</td>
</tr>
<tr>
<td>Total</td>
<td>58</td>
<td>100.0</td>
<td></td>
</tr>
</tbody>
</table>

*From the hypergeometric distribution.

Fig. 3. Co-regulation with known chemotaxis genes predicts chemotaxis functions. Mutant cells were deposited on cAMP agar, incubated for 22 h, fixed, stained and photographed (pictures on bottom, bar = 2 mm). The diameter of the halo was measured and plotted as box plots. The gray box indicates the inner quartile (25th–75th percentile of the data). The line across the box indicates the median. Barred dotted lines indicate the upper and lower limits of the data outside the inner quartile. Data were compared with the wild type (WT) using a t-test; P-values are indicated above the boxes. The genotypes and respective transcriptional modes are indicated on the x-axis.

results suggest that six of the tested genes have chemotaxis functions.

As controls, we tested mutants in eight ABC genes from other modes: *abcG7*, *abcC3*, *abcF1*, *abcA5* (Mode 17); *abcF2* (Mode 12); *abcG10* (Mode 14); *abcG6* (Mode 3) and *abcC5* (Mode 6). We compared the chemotaxis ability of the test group (Fig. 3) and the control group to the wild type using the Wilcoxon test and found that the test group was significantly different from the wild type (*P* = 0.04) whereas the control group was not (*P* = 0.14).

DISCUSSION AND CONCLUSIONS

The correlation between gene expression and function is somewhat controversial. On one hand, comprehensive comparisons between expression and function, measured by fitness in yeast, indicated a low correlation (Giaever et al., 2002; Winzeler et al., 1999). On the other hand, experiments in yeast, nematodes and mice described good correlations between groups of co-regulated genes and their GO annotations (Hughes et al., 2000; Stuart et al., 2003; Zhang et al., 2004). A possible explanation is that good correlations are observed in genes whose products function in protein complexes. For example, ribosomal protein genes are co-regulated in all the above organisms. Our findings are consistent with that explanation, because ribosomal protein genes and cytoskeletal genes were among the most significant groups of co-regulated genes. Co-regulation of chemotaxis genes has not been described before as such, but the correlation between expression and function in chemotaxis genes is reasonable because chemotaxis involves the coordinated polymerization and polarization of the cytoskeleton protein complex. The signal transduction mechanisms that regulate cytoskeletal changes also involve protein complexes such as Arp2/3 and SCAR/PIR121 (Parent, 2004).

*Dictyostelium* and neutrophils share many chemotaxis properties (Parent, 2004). To test whether chemotaxis genes were also co-regulated in neutrophils, we compared our data to two other studies. One study examined gene expression during inflammation in wild-type mice and in mutants lacking macrophages and neutrophils (Cooper et al., 2005). We found that 19 out of common 35 chemotaxis genes were co-regulated (Pearson’s correlation coefficient > 0.5) in the wild-type mouse data (data not shown). Another study examined expression of human neutrophil genes upon migration to inflammation sites (Theiligard-Monch et al., 2004). Comparing their results with ours revealed a 25% overlap between the co-regulated genes (data not shown). The common regulation of chemotaxis genes in protozoa and mammals suggests evolutionary conservation.

Figure 4 summarizes the correlation between expression and function of the chemotaxis genes on the microarray and their interactions in *Dictyostelium*, following the model proposed in (Parent, 2004). Cell polarization involves localization of PI3K to the front and PTEN to the back of the cell, both activities regulate the levels of phosphoinositides (PIP2/3). The Arp2/3 complex controls actin polymerization in the front. Rac and RacGEF activities, mediated by WASP and SCAR, regulate Arp2/3 in response to changes in PIP2/3 levels, but some of the components that participate in the process are not defined in *Dictyostelium* (Parent, 2004). Based on co-regulation and the similarity between the *Dictyostelium* and neutrophil chemotaxis pathways, we propose that secG, *gefA*, *gefE*, *gefQ*, *wasA*, *ACP3*, *GAPa* and *racA* encode the yet characterized components in this process (Fig. 4).

The cAMP-dependent protein kinase (PKA) controls gene expression and development during chemotaxis (Loomis, 1998). Participation of the PKA pathway is implicated by the co-regulation of the catalytic subunit *pkaC*, the regulatory subunit *pkaR*, and the protein degradation gene *culA* and *erkB* and *regA* are also involved (Maeda et al., 2004), but they are co-regulated with...
Fig. 4. Co-regulation of chemotaxis genes. Schematic view of a polarized chemotactic cell (top = front) after Parent (2004). Chemotaxis components are shown and their transcriptional modes indicated by color (legend in box) or in parentheses next to the gene name. CRAC and ggpA are not represented on the microarray. Rectangles represent gene products. Dashed ellipses represent small molecules.

Mode 6 and Mode 8 genes, respectively, and not with Modes 1, 2 or 15 (Fig. 4). References to the involvement of all the Figure 4 genes in chemotaxis are provided in Supplementary Table 2.

Extracellular cAMP induces signal transduction mechanisms through the receptors encoded by carA and carC. Receptor stimulation activates the adenylyl cyclase encoded by accA and the guanylyl cyclase encoded by sgcA. cGMP and its binding proteins encoded by gbpC and gbpD affect the activity of myosin II and is regulated by the phosphodiesterase PDE3. In the rear, we propose that the p21-activated kinase pakC functions with the known kinase pakA to regulate myosin II assembly. Also co-expressed are the myosin heavy chain gene mhkA and three myosin regulators: the clathrin adaptor apm2, the myosin heavy-chain (mhkA) and light-chain (mhkA) kinases (Parent, 2004; Van Haastert and Devreotes, 2004). The most interesting aspect of our finding is perhaps the co-regulation of genes that are involved in widely different aspects of chemotaxis, including fundamental cell motility, cytoskeletal organization, signal transduction and intercellular signaling.

One of the applications of common regulation and annotation is the discovery of gene function. We tested seven genes that were co-regulated with known chemotaxis genes but not known to have chemotaxis functions and found that six had chemotaxis defects. The halo assay we used is imperfect because it cannot distinguish between cell motility, chemical sensing, cAMP degradation, etc. Therefore the conclusions about chemotaxis defects are tentative and would require more detailed analyses. Nevertheless, our results provide an important proof of principle for the discovery of function from the clustered microarray data. Therefore, the data and analyses in Supplementary Figure 1 serve as a public resource for functional genomics in *Dictyostelium* and could facilitate the discovery of additional gene functions.

ACKNOWLEDGEMENTS
This work was supported by NIH/NICHD Grant P01 HD39691. E.O.B. and N.V.D. were supported in part by training fellowships from the W. M. Keck Foundation of the Gulf Coast Consortia through the Keck Center for Computational and Structural Biology.

Conflict of Interest: none declared.

REFERENCES


