ABC report 1

Marko Toplak

June 8, 2008

1 Basic data analysis

Histogram of variations for data for 8 ABC mutants is shown on Fig. 1. One available mutant (A5-) was not included in this preliminary analysis because it is annotated a bit differently. Each mutant was measured at 4 different time points with 3 technical replications. Same genes at different time points were treated as different attributes.

1.1 Mutant distances

Distances between mutant replicates (8 mutants × 3 replications) were measured with Pearson correlation between replicates (distance $= 1 - |\text{correlation}|$). Same genes at different time points were treated as different attributes. Distances are illustrated with MDS scaling and a dendrogram on Fig. 2.

On Fig. 3 we show distances between mutants if replicates are "joined" (expressions averaged). Mean of expressions between replicates of the same mutant was computed.

We have tried to estimate how similar are replicates to each other. On Fig. 4 a histogram of the distribution of distances between all possible triplets of replicates is shown. Vertical lines represent distances between replicates of the same mutant. Distance between a triplet was computed as the sum of distances between all pairs of replicates in the triplet.

2 Analysis with gene sets

2.1 Gene set source

Gene matching used to match genes in data set to genes in gene sets is currently implemented in Orange by look ups to the KEGG database. Therefore we performed a brief survey of Dictyostelium discoideum genes in KEGG. 13458 genes for dicty are in KEGG database. From all

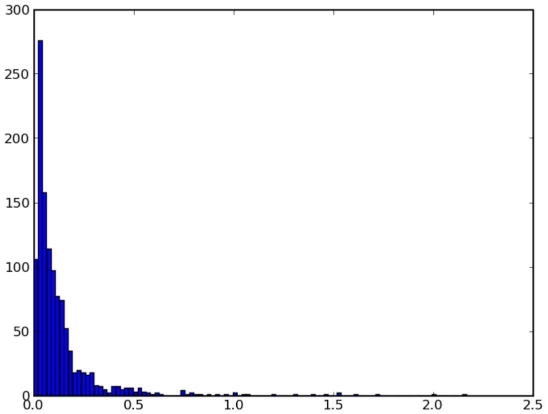

Figure 1: Variations of attributes for 8 ABC mutants (3 technical replications each). Replications are treated separately. One attribute specifies a gene at one time point.

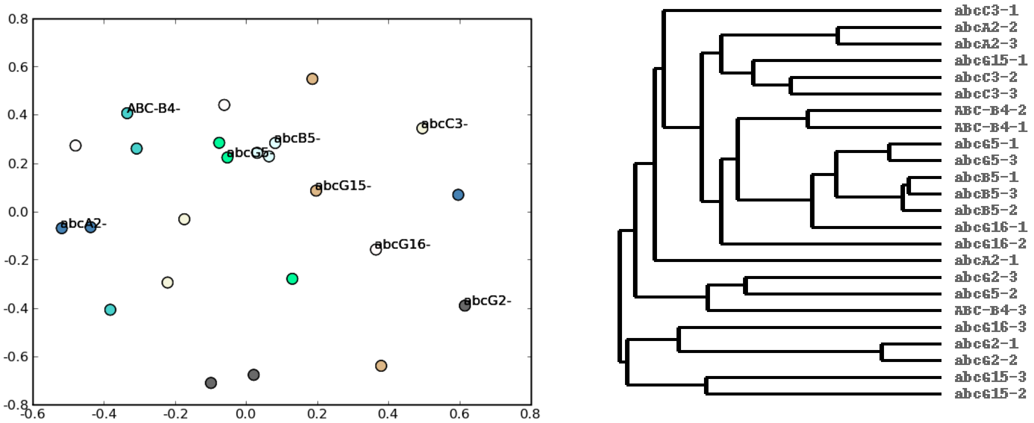

Figure 2: Distances between replicates (MDS, dendrogram).

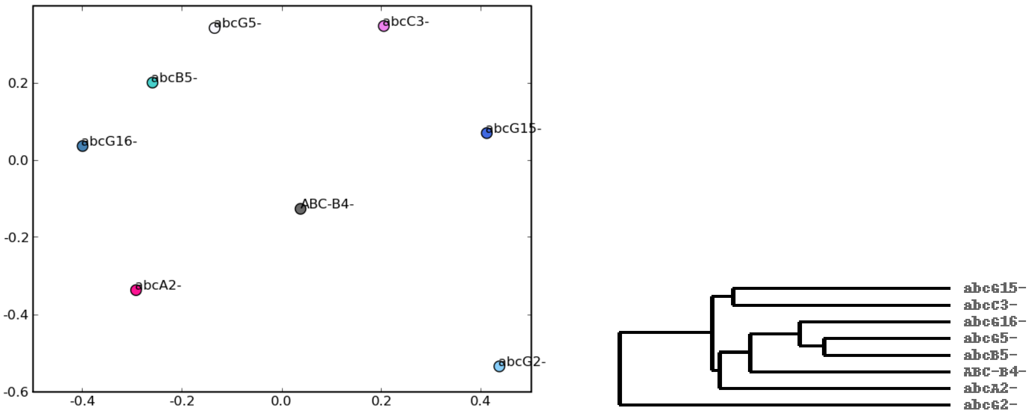

Figure 3: Distances between mutants if replicates are averaged (MDS, dendrogram).

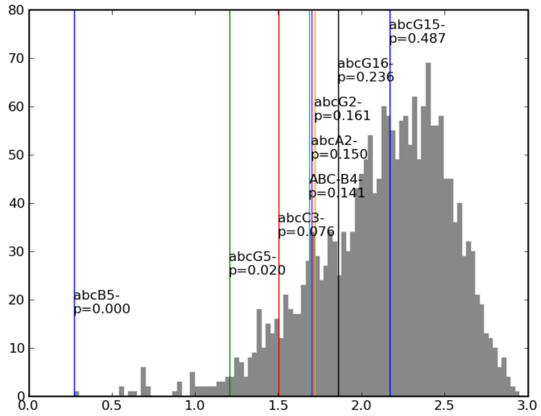

Figure 4: Distances between all possible triplets with distances of replicates of the the same mutant marked by vertical lines.

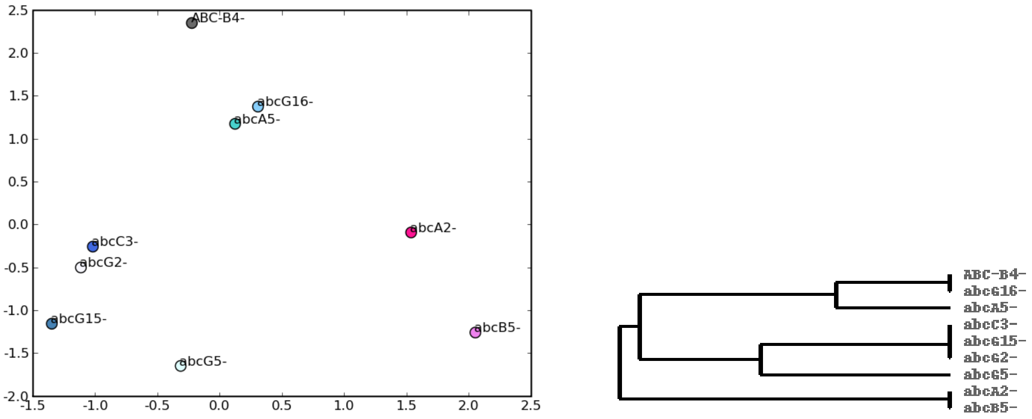

Figure 5: Distances between mutants measured as the number of differentially expressed gene sets with $p < 0.05$. (MDS, dendrogram).

the genes measured by the minichip, 9 can not be unambiguously matched to genes in KEGG database: DDB0252847, DDB0252881, DDB0231422, DDB0266346, DDB0252883, DDB0266410, DDB0252843, DDB0252829, DDB0266477. We have ignored this fact for now, but they can be matched manually if needed.

KEGG contains very few pathways with at least 3 genes measured on minichip, that could be used as gene sets for GSEA. Therefore we also used GO as a source of gene sets. GO contains 6953 dicty genes. Only 384 of them can not be found in KEGG if searching with DDB of gene as specified in GO.

We compiled our collection of gene sets from KEGG pathways and GO groups. We took only sets which contained at least 3 genes also measured on the minichip.

2.2 Measuring distances with GSEA

To compensate for small number of samples we ignored the time component: we grouped minichip only by the mutant, ignoring the time point. That yielded 12 measurements per mutant (3 replications on 4 time points).

Gene Set Enrichment Analysis (GSEA) ranks gene sets according to their differential expression between two groups of samples. We ran GSEA for each pair of mutants. Only gene sets with at least 5% of their genes in data set (on minichip) were considered. We defined the distance between two mutants as the number of gene sets with P-values (computed by GSEA) lower than 0.05. Because some gene sets were composed of the same or almost the same genes, we only counted “different” gene sets: a gene set was added to the list of “different” gene sets only if there was no gene set in the list, for which Jaccard index between it and the candidate gene set would be higher than 0.8. We show MDS and dendrogram of computed distances on Fig. 6.

2.3 Predicting some groups from Sect. 2.2

We have tried using groups from Sect. 2.2 for prediction. We used data for replicates of 8 mutants as in Sect. 1 – each attribute represented gene at one time point. For each group we labelled replicates with 2 labels: one, if mutant belonged to the group, and another, if it did not. We report AUC scores with leave-one-out sampling and their P-values according to 100 random groups of the same size.

```
['abcC3-', 'abcG15-', 'abcG2-'] AUC =  0.866666666667 pval =  0.07
['abcC3-', 'abcG15-', 'abcG2-', 'abcG5-'] AUC =  0.583333333333 pval =  0.68
['ABC-B4-', 'abcG16-'] AUC =  0.509259259259 pval =  0.76
['abcA2-', 'abcB5-'] AUC =  0.805555555556 pval =  0.17
```

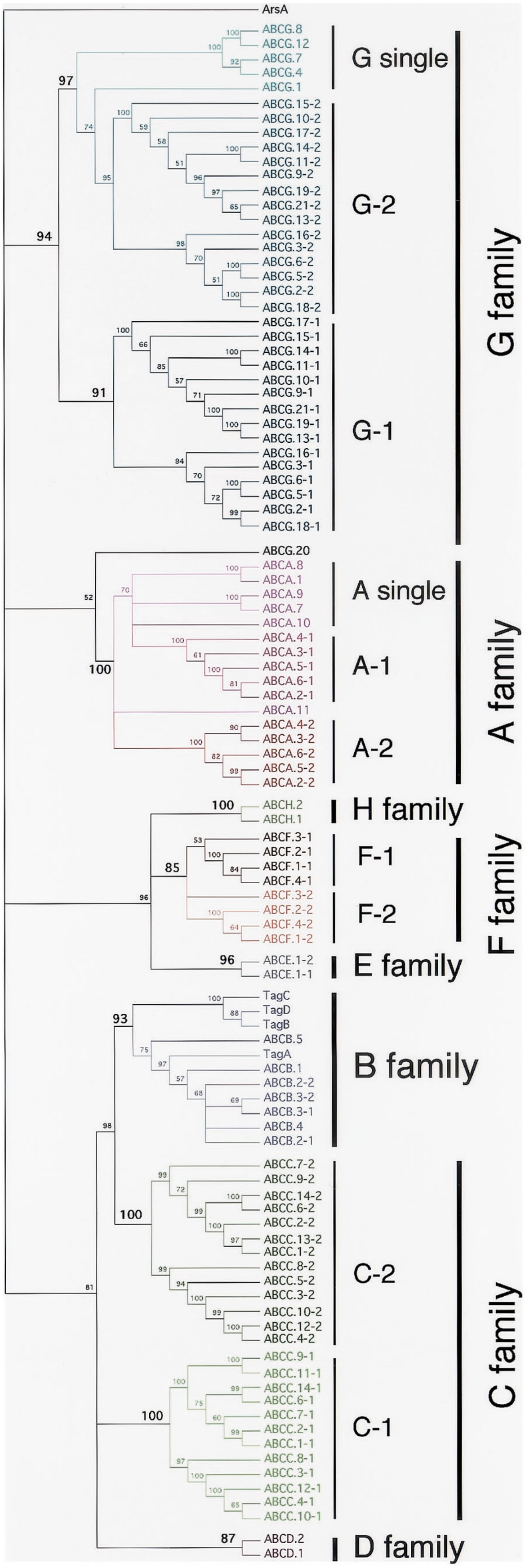

Figure 6: ABC hierarchy from Anjard and Loomis.