QT (Quality Threshold) Clustering

Contents at a glance

I. Definition and Applications 2

II. Overview of the QT Clustering window 2

III. How does the QT Clustering work? 3

IV. Interpreting the Results 4

V. What are some advantages and disadvantages of QT Clustering 4

VI. Most frequently asked questions and answers 5

VII. References 5

Agilent Technologies

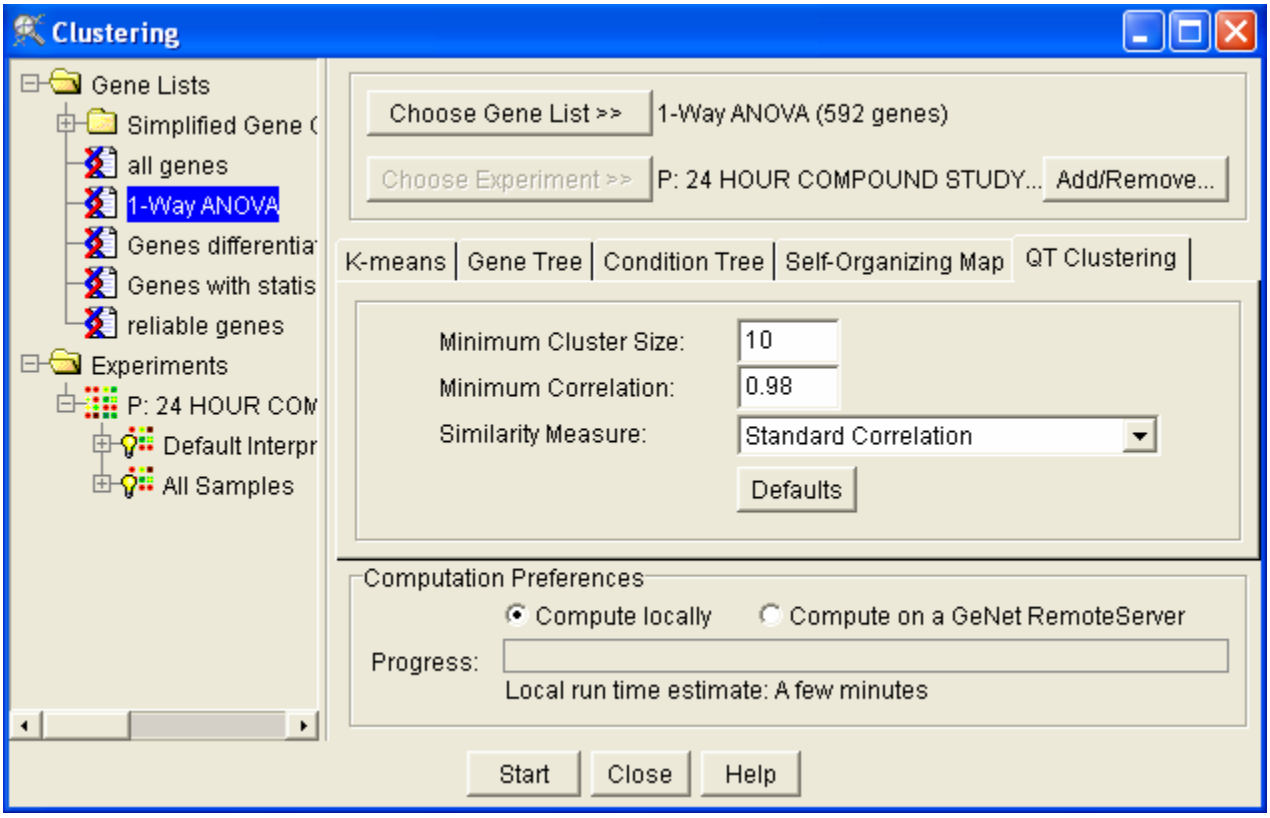
sig_support@agilent.com | Main 866.744.7638

I. Definition and Applications

QT (Quality Threshold) Clustering is an algorithm that groups genes into high quality clusters. Quality is ensured by finding large cluster whose diameter does not exceed a given user-defined diameter threshold. This method prevents dissimilar genes from being forced under the same cluster and ensures that only good quality clusters will be formed.

II. Overview of QT Clustering window

1. Select **Tools** -> **Clustering** -> **QT Clustering**



Choose Gene List: Genes in this selected gene list will be used for clustering.

Choose Experiment: Data from this selected experiment will be used to determine the similarity in genes expression pattern.

Minimum Cluster Size: Minimum number of genes that you would like to have in each cluster.

Minimum Correlation: Minimum correlation that genes within each cluster must have to one another. The diameter is the equivalent of 1 minus the minimum correlation

Similarity Measure: Measure used to determine the similarity of gene expression patterns.

III. How does QT clustering work?

The goal of QT clustering is to form large clusters of genes with similar expression pattern, and to ensure a quality guarantee for each cluster. Quality is defined by the cluster diameter and the minimum number of genes contained in each cluster.

Algorithm:

1. A random gene is chosen from the selected gene list.

2. The algorithm determines which gene has the greatest similarity to this gene. If their total diameter does not exceed the diameter threshold, then these two genes are clustered together.

3. Other genes that minimize the increase in cluster diameter are iteratively added to this cluster. This process continues until no gene can be added to this first candidate cluster without surpassing the diameter threshold.

4. A second candidate gene is chosen.

5. The algorithm determines which gene has the greatest similarity to this second gene. **All genes in the selected gene list are available for consideration to the second candidate cluster.**

6. Other genes from the selected gene list that minimize the increase in cluster diameter are iteratively added to the second candidate cluster. The process continues until no gene can be added to this second candidate cluster without surpassing the diameter threshold.

7. The algorithm iterates through all genes on the selected gene list and forms a candidate cluster with reference to each gene. In other words, there will be as many candidate clusters as there are genes in the gene list. Once a candidate cluster is formed for each gene, all candidate clusters below the user-specified minimum size are removed from consideration.

8. The largest remaining candidate cluster, with the user-specified minimal number of gene member, is selected and retained as a QT cluster. The genes within this cluster are now removed from consideration. All remaining genes will be used for the next round of QT cluster formation.

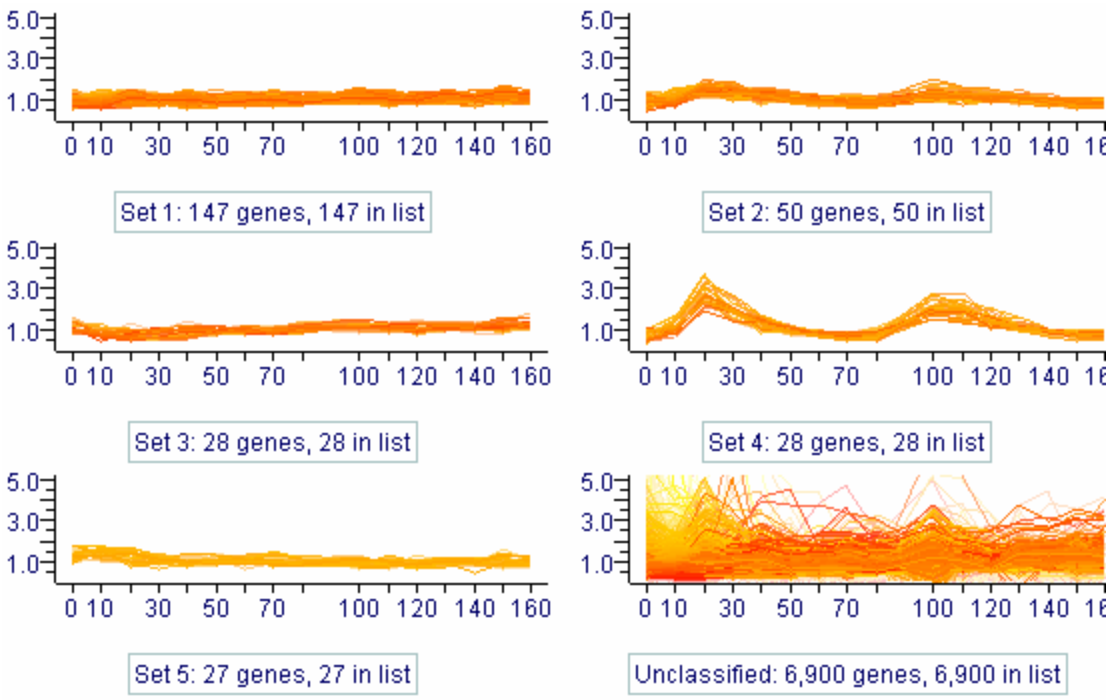
9. The entire process (step 1 to 9) is repeated until the largest remaining candidate cluster has fewer than the user-specified number of genes.

10. The result is a set of non-overlapping QT clusters that meet quality threshold for both size, with respect to number of genes, and similarity, with respect to maximum allowable diameter.

11. Genes that do not belong in any clusters will be grouped under the “unclassified” group.

IV. Interpreting the Results

QT Clusters are displayed according to the cluster size, from the largest to the smallest. Set 1 is the largest cluster, followed by set 2, etc… All sets will have **at least** the user-defined minimum cluster size and the minimum correlation (diameter). For example, all 147 genes in Set 1 below are at least 0.98 correlated to each other. Genes that did not meet the minimum quality are grouped under the "unclassified" category.



III. What are some advantages and disadvantages of QT Clustering?

Advantages:

1. **Quality Guarantee**: only clusters that pass a user-defined quality threshold will be returned.

2. **Number of clusters is not specified a priori:** does not require the user to specify the number of cluster in advanced (unlike self-organizing map or k-means clustering).

3. **All possible clusters are considered:** a candidate cluster is generated with respect to every gene and tested in order of size against quality criteria.

Disadvantages:

1. **Computationally Intensive/Time Consuming:** increasing the **Minimum Cluster Size,** decreasing the **Minimum Correlation,** or increasing the number of genes on the selected gene list can greatly increase the computational time.

Q. Does GeneSpring use the jackknife-correlation method?

A. No. This method is extremely computationally extensive and it is impractical for most researchers working on a desktop computer. The author who devised the QT clustering algorithm (Genome Research, Heyer et al 9:1106-1115) used jackknife-correlation method to cluster 4190 genes, but it took them ~30 minutes to run it on a Sparc Ultra, Unix workstation.

Q. What are the main differences between QT clustering and K-means clustering?

	K-means	QT clustering	Consequence
Need to specify cluster number?	Yes	No	K-means: if users specify too few clusters, genes that are not similar will be forced to group together.
Very computationally intensive?	No	Yes	QT clustering: may be too computationally intensive, depending on available RAM and number of genes in starting gene list, for some desktop computer.
Every gene must be clustered?	Yes	No	K-means: every gene on the selected gene list must belong to a cluster. This could potentially group genes that are not very similar into the same cluster. QT clustering: only cluster with user-specified quality will be formed.

Q. Why does the computation take so long when I decrease the minimum correlation?

Decreasing the minimum correlation will lead to an increase in the diameter size. As a result, the algorithm will undergo more iterations in an attempt to add more genes to each cluster, until the cluster reaches its maximum size allowed.

VII. References

Heyer, L.J., et al. "Exploring Expression Data: Identification and Analysis of Coexpressed Genes". Genome Research, 9:1106-1115 (1999).