POINTS OF SIGNIFICANCE

Clustering

Clustering finds patterns in data—whether they are there or not.

Many biological analyses involve partitioning samples or variables into clusters on the basis of similarity or its converse, distance. For example, in a gene expression study, we might seek subsets of patients with similar expression, or take a complementary approach and identify similarly expressed genes across patients. Clustering is a type of unsupervised learning comprising many different methods[1]. Here we will focus on two common methods: hierarchical clustering[2], which can use any similarity measure, and *k*-means clustering[3], which uses Euclidean or correlation distance.

Fundamentally, all clustering methods apply the same approach. First, we calculate similarity and then use it to group objects (e.g., samples) into clusters. However, the clustering output is useful only if the clusters correspond to the data's biologically relevant features that were not used to define the grouping. To judge clusters' validity, we need external information; clusters are not known in advance. For example, our confidence in the validity of our clusters increases if patients in each cluster share a phenotype, or if genes in each cluster share a sequence motif; but confidence increases only if this information was not used to assess similarity in the first place.

Let's look at how similarity can be calculated. Suppose we have expression profiles for five genes, A–E, across $n = 15$ patients, and we want to compare these profiles to a reference profile (**Fig. 1**). A visual assessment may be misleading. The difference in expression relative to the reference is smaller for gene B than for gene A, for example, and this might make us think that the gene B profile is more similar to the reference than the gene A profile. However,

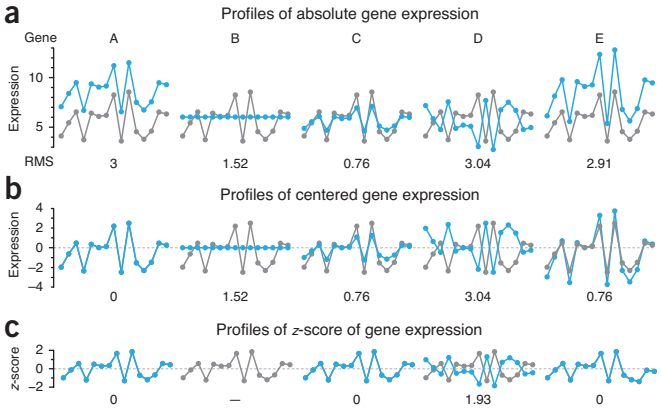

Figure 1 | Similarity measures between expression profiles across $n = 15$ patients (dots) of five putative genes (blue) and a reference (gray). (**a**) Absolute expression profiles of genes A–E generated by various transformations from the reference. Their similarity to the reference is shown as the Euclidian distance expressed as root mean square (r.m.s.). Gene C is most similar to the reference (r.m.s. = 0.76), followed by gene B (r.m.s. = 1.52). (**b**) Profiles from **a** centered on their means and corresponding r.m.s. Gene A and reference profiles now overlap (r.m.s. = 0), and the similarity of gene E to the reference has decreased to be the same as that of gene C (r.m.s. = 0.76). (**c**) Profiles from **a** transformed into *z*-scores. Gene B has no profile because the *z*-score is undefined when no variation is present.

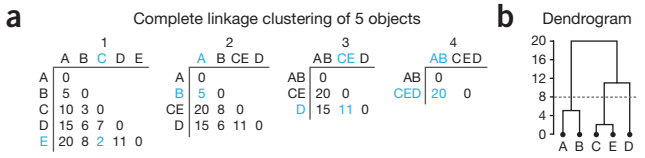

Figure 2 | Complete linkage clustering of five objects. (**a**) Pairwise distances (step 1) are used to merge objects (steps 2–4) where the maximum of all pairwise distances is used. At each merging step, the shortest distance is chosen (blue). (**b**) A dendrogram with a vertical axis showing the distance between merged nodes. To create clusters, one can cut the tree at a fixed height (dashed line).

gene B has a completely different pattern of expression (constant) than that of the reference, while gene A has the same pattern as the reference but with an offset.

While there are many ways to calculate the similarity of two such profiles, including subjective measures, we use the common geometric notion of Euclidian distance expressed as the root mean square (r.m.s.; **Fig. 1a**). This quantity includes a factor of $1/\sqrt{n}$ to avoid dependency solely on n, such as for profiles that differ by only a constant offset. Similarity can be expressed as $|c - \text{r.m.s.}|$ (where c is some constant such as the maximum distance between objects), so that objects with distance c or greater have zero similarity.

Practically, similarity in expression should be based on varying regulation and not absolute abundance. To emphasize regulation, we can center the expression values by subtracting the profile's mean from each of its expression values (**Fig. 1b**). To focus on the pattern rather than magnitude of regulation, one can first convert profiles to *z*-scores, which give the variation from the mean in units of s.d. (**Fig. 1c**). The r.m.s. between *z*-score profiles is $1 - r$, where r is the correlation of the profiles—those perfectly correlated have r.m.s. = 0. Distance may be defined as $1 - |r|$ to cluster genes with opposing regulation, such as gene D, which is perfectly negatively correlated with the reference (**Fig. 1c**). When using correlation distance, it is common to filter out samples with very low variance, where the pattern may be due to chance.

Once we have the similarity between objects, we group them into clusters. In hierarchical clustering, the nodes start off as objects and are then iteratively merged on the basis of pairwise distance (**Fig. 2a**). There are many ways of calculating this distance, but the most common methods are complete linkage clustering and single linkage clustering, which return the maximum or minimum, respectively, of all pairwise distances of objects between nodes. The clustering is typically depicted by a dendrogram, where the height of the branches is either the step at which the nodes were merged or the distance between them (**Fig. 2b**). Clusters are formed by partitioning of the dendrogram—for example, by cutting it at a fixed height and considering each of the resulting subtrees as a cluster. Membership in clusters depends on both the cutoff and similarity measures (**Fig. 3**). Alternatively, clusters can be made with selective cuts informed by underlying biology to find visually pleasing groups.

When comparing two dendrograms, take into account that the order of branches in a dendrogram is arbitrary. Nodes that are near each other (e.g., profiles D5 and E5 in **Fig. 3b**) may lose their spatial adjacency with a single branch flip.

In contrast to hierarchical clustering, *k*-means clustering requires that we first choose the number of clusters, *k*. In **Figure 4a** we illustrate this process using $k = 3$ and a simulated two-dimensional data set with points randomly placed in three adjoining areas (gray

circles). The algorithm begins by selecting *k* data points as 'centroids' (open circles). In our case these are randomly selected points from the data set, but they may also be randomly generated within the range of the data. In the first step, the similarity between each point and each centroid is computed—typically on the basis of Euclidian or correlation distance—and points are grouped with the nearest centroid (colored lines). Subsequently, the position of each centroid is recalculated on the basis of the objects assigned to it, and object assignment is repeated with the new centroid positions. These steps are repeated until the centroids and clusters no longer change.

Cluster quality may be checked using the within-cluster similarity (ideally, high) and between-cluster similarity (ideally, low). Unless the clusters are well separated, with high within-cluster similarity and low between-cluster similarity, different clustering methods will create different clustering of the same data even when the same measure of similarity is used. Within-cluster similarity tends to be favored as a quality measure because clusters may have arbitrary boundaries and may not be well separated[4].

When the method does not always converge to the same solution, such as for *k*-means (whose output depends on the choice of centroids), within-cluster similarity can be used to rate different solutions (**Fig. 4b**). For example, in our simulation of 10,000 trials, the most frequently seen solution (**Fig. 4c**, $d = 39.0$) is not the one with the lowest distance (**Fig. 4a**, $d = 38.4$). When clusters are spatially compact balls around the center of the node (**Fig. 4c**), *k*-means behaves like complete linkage clustering. Solutions in which some or all clusters are stringy (**Fig. 4e–g**) are similar to output of single linkage clustering. In our example data set, the clusters are not well separated, and this is reflected in solutions that do not reconstruct the original grouping of points (**Fig. 4e–g**).

Determining the number of clusters is a difficult problem. Typically the number of clusters is not known in advance, and so exploration of the cluster quality as a function of the number of clusters may be performed. Another approach is to plot the objects in each cluster (**Fig. 3**) to determine how variable the objects (e.g., expression profiles) are and whether there are clusters that have very similar profiles and so should be merged.

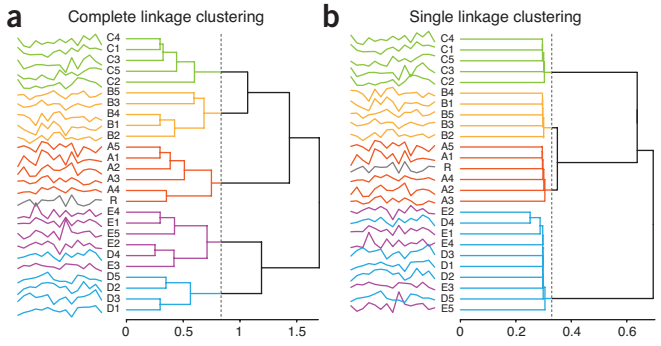

Figure 3 | Dendrograms of hierarchical clustering of gene expression profiles based on correlation distance. The data were generated by creating core profiles A1, B1, C1, D1, and E1 with correlation values of 0.7, 0.5, 0, -0.5, and -0.7 (respectively) with the reference profile R from **Figure 1**. For each core profile (e.g., A1), four additional highly correlated random profiles were generated (e.g., A2–A5). Profiles are colored by group and clusters formed by cutting at a fixed height (dashed line). (**a**) Complete linkage clustering tends to create balanced dendrograms by first clustering objects into small nodes and then clustering the nodes. (**b**) Single linkage clustering tends to create stringy dendrograms by first creating a few nodes and then adding objects to them one at a time.

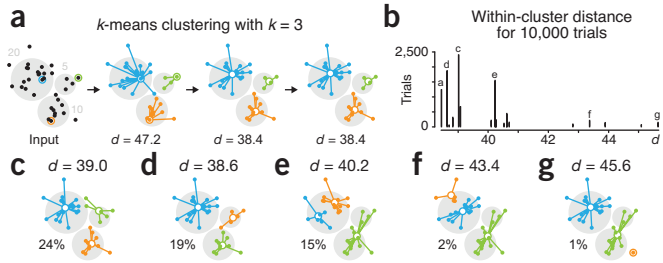

Figure 4 | Simulation of 10,000 trials of *k*-means clustering with $k = 3$ of 35 points (black), of which 20, 10, and 5 were centered on each of the gray circles, respectively, and spatially distributed normally within the circle with s.d. half of the circle radius. Centroids are indicated by colored hollow points; initial centroids were randomly selected points from the data set. (**a**) Evolution of a trial that results in the lowest total within-cluster distance, $d = 38.4$. With each iteration, d generally drops. Points are shown connected to and colored by their assigned centroid. (**b**) Histogram of the total within-cluster distance for 10,000 trials. The lowest $d = 38.4$ solution (**a**) was found in 1,236 (12%) of trials. Bar labels indicate figure panels in which the solution is shown. (**c**,**d**) Two most common solutions, their d and frequency observed. (**e**,**f**) Examples of solutions whose clusters do not follow the original grouping of points. (**g**) Solution with largest d.

Clustering methods always find clusters, even if there are no natural clusters in the data. Ultimately, clusters should be judged by the criterion of utility for biological discovery. We should always ask whether the objects in a cluster have traits in common that were not used to inform the clustering. This question can be used to address the quality of clustering—if we cluster gene expression according to multiple conditions, we might deem more reliable those clusters whose samples share a condition. Alternatively, clusters can help researchers explore data and generate hypotheses—a cluster of objects without obvious similarity might suggest the existence of gene networks, subclasses of diseases, or geographic genetic variability. In these cases, care must be taken to provide additional information to substantiate any claims about the clusters that have been found.

Naomi Altman & Martin Krzywinski

COMPETING FINANCIAL INTERESTS
The authors declare no competing financial interests.

1. Everitt, B. *Cluster Analysis* (Heinemann Educational, 1974).
2. Reynolds, A., Richards, G., de la Iglesia, B. & Rayward-Smith, V. *J. Math. Model. Algorithms* **5**, 475–504 (2006).
3. Hartigan, J.A. & Wong, M.A. *J. R. Stat. Soc. Ser. C Appl. Stat.* **28**, 100–108 (1979).
4. Bryan, J. *J. Multivar. Anal.* **90**, 44–66 (2004).

Naomi Altman is a Professor of Statistics at The Pennsylvania State University. Martin Krzywinski is a staff scientist at Canada's Michael Smith Genome Sciences Centre.