Trios: Transmission Disequilibrium Test (TDT)

		Non-transmitted parental allele	
		A	a
Transmitted parental allele	A	w	x
	a	y	z

- w AA parents (transmit one A, do not transmit other A)
- z aa parents (transmit one a, do not transmit other a)
- x Aa parents that transmit A, do not transmit a
- y Aa parents that transmit a, do not transmit A

Transmission Disequilibrium Test (TDT)

		Non-transmitted parental allele	
		A	a
Transmitted parental allele	A	w	x
	a	y	z

- No variation in w or z (recall homozygous parents non informative)
- $(x-y)^2/(x+y) \sim \chi_1^2$; it's just special case of McNemar's test
- Think of it as testing are there an excess of the A allele in the affected offspring than would happen by Mendel's laws?

TDT test

Transmission Disequilibrium Test (TDT)

		Non-transmitted parental allele	
		A	a
Transmitted parental allele	A	?	78
	a	46	?

Insulin Dependent Diabetes Mellitus (IDDM)

- Example from the text: 94 families, 78 parents transmit allele A, 46 transmit allele a
- $(78-46)^2/(78+46)=8.26$, p-value=0.004

Spielman et al., 1993

FBAT

Generalizations of the original TDT test are called FBATs (Family-Based Association Studies). Compared to case-control association studies, which are more common, a trio from an FBAT design is more informative than 3 individuals in a case-control design, but it is harder to recruit trios or other family structures for a study design, and there can be a lot of missing data, not to mention incorrect data (that can't really be the father....). Consequently, the overall sample size for family-based designs tends to be smaller.

There are proponents of both approaches, but it matters which approach you want to adopt before collecting the data because the study designs are so different. The debate is one that shows up elsewhere in statistics: power versus robustness. The case-control approach tends to be more powerful because it can get larger sample sizes, but advocates for FBATs argue that population based (i.e., not family based) are more robust for population structure. Advocates for case-control designs have in turn argued that you can test for population structure and account for it...

TDT and FBATs

I don't want to take sides in the debate — I just want to point out that there has been a debate (among biostatisticians) in this area. The debate also sounds to me remarkably similar to debates about, for example, nonparametric versus parametric methods, where there is a tradeoff between power and robustness.

I think it is also interesting to think about what data we will have in the future. Suppose 20 or 30 years from now everyone has their entire genome sequenced at birth? In this case it might start to become easy to get family-level data. We might also have every SNP, so we no longer need to worry about SNPs associated with a disease merely being close to the causal SNP. If there's a causal SNP, then it will have been sequenced.

TDT and McNemar's test

Statistically, the TDT test is essentially McNemar's test, a test that comes from the analysis of 2×2 contingency tables.

Often in a contingency table, you find an association between say, a treatment versus a population. For example, you might have a control group given a placebo and a treatment group given a drug. The two groups are independent, and you track whether or not they experience a symptom. A famous example is an aspirin trial where doctors were randomized into placeo versus aspirin groups, and then checked for whether they had a heart attack or not within a given amount of time.

TDT and McNemar's test

For McNemar's test, instead of indepent control and treatment groups, you have correlated observations where the same individual is given both the control and the treatment at different points in time. So you might have a medication for preventing migraines, and you follow patients for 1 week with placebo and 1 week with the drug and check whether or not they experienced a migraine in the week.

	Second week	
Week	Migraine	No migraine
First week	a	b
Second week	c	d

McNemars test is

$$X^2 = \frac{(b-c)^2}{b+c}$$

TDT and McNemar's test

This has an approximate χ^2_1 distribution. Basically, this looks at whether cases in which headache status changed from week to week were different for those on the drug versus those on the placebo. If just as many people had headaches with the drug but not with the placebo as with the placebo but not with the drug, then the test statistic is 0. The diagonals do not contribute to the test statistic.

Gene expression levels

In addition to associations between SNPs and traits such as presence/absence of disease and quantitative single measurements like FEV and height, association studies sometimes look at expression levels for genes.

The idea for measuring gene expression is that during normal cell activity, certain genes are turned on or off in the sense that the proteins they code for are either made or not made. Which genes get turned on or off partly depends on the tissue they are in: certain proteins are needed more often in the liver than in the brain, for example.

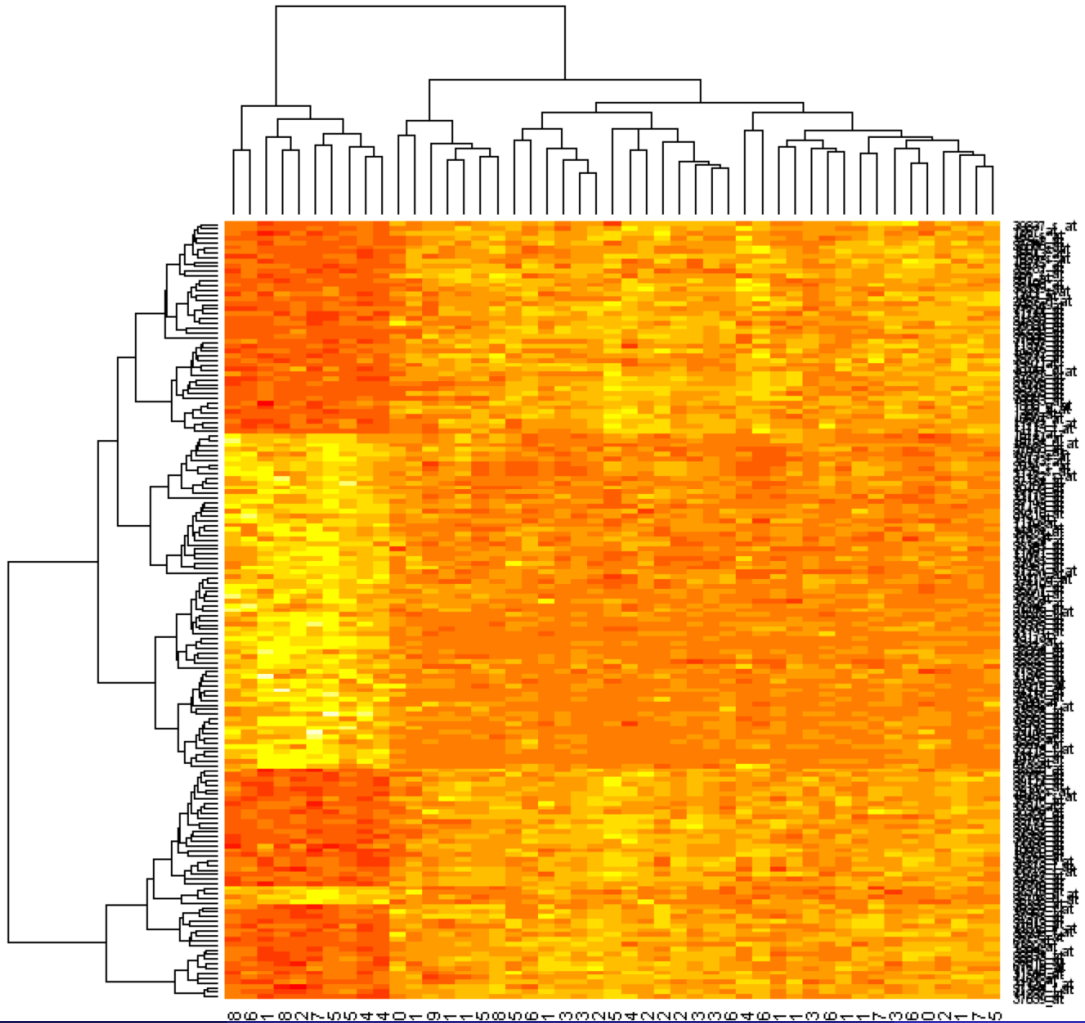
Gene expression levels

Gene expression studies look at levels of protein production or mRNA (a precursor to the protein) in different tissues. mRNA abundance is easier to measure but less direct information about which genes are functioning in the cell. One book (*Computational Genome Analysis* by Deonier et al.) gives the analogy that it is like measuring the prodoctivity of a law office by measuring how many reams of paper it is using.

Gene expression is often measured with microarrays, which have 2-d arrays with one tiny well per gene. There are a little over 20,000 genes in the human genome, so these experiments usually get 20,000 quantitative gene expression values, possibly with replication, although the number of replications is very low (often just two). We might also get gene expression values on multiple individuals, but this will usually be small compared to the number of genes. Thinking of this as a multivariate data set, we have very observations with a huge number of variables, and many more variables than observations.

Gene expression data

Gene expression levels are often visualized using a figure such as this:



Gene expression data

Typically a diagram like this has rows that represent genes and columns that represent samples (or vice versa). Here redness might indicate higher expression level. Clustering can be done on both the rows and columns to see if certain genes tend to cluster together or certain observations cluster together. We might be interested to see, for example, if cancerous cells have more similar gene expression values to each other than to non-cancerous cells.

For clusters in the genes, a cluster suggests certain genes tend to be co-expressed, e.g., they tend to all have high expression levels in the same tissue. This can help form hypotheses about which genes work together to achieve a certain function. Because certain genes will only be expressed in some tissues, these heat maps can look different in different tissues. Instead of expressing an absolute gene expression level, these heat maps of gene expression might also express relative expression levels (for example, relative expression in liver vs brain or cancer versus normal cells).

Gene expression data

A statistical question in gene expression data is whether gene expression is signficantly different in one group versus another for a given gene, or simply whether it is signficantly different from 0 for a given gene. Because there are so many genes ($\sim$20,000), there is a large potential for false positives. But because sample sizes are typically small for gene expression data, it might be difficult to overcome Bonferroni corrections.

Replication is important in finding results to be significant. Replication can be done by replicating the sample (taking the same slides and retesting for expression levels) and increasing the number of organisms sampled. Replication is difficult for technical reasons. For example, dye has to be applied to the slides to detect mRNA levels, and dye intensity is not constant, so the variability in the dye instensity has to be taken to account. This usually results in a lot of preprocessing of the data before it gets analyzed, and "raw" data isn't analyzed. However, there might be considerable measurement error.

Clustering for gene expression data

Standard clustering techniques from multivariate statistics can be performed for clustering gene expression values. In hierarchical clustering methods, the idea is to create a distance or similarity matrix. These distance matrices could represent Euclidean distances between vectors for each gene, where the number of dimensions of the vector is the number of observations. The Euclidean distances are often standardized by the mean and standard deviation for each variable (i.e., each tissue sample in this case). Alternatively, if you cluster the observations, you could standardize by the mean and standard deviation for the gene. In either case, you end up with individual gene expression values that look like z-scores.

Clustering for gene expression data

To do the clustering (on the genes, say), you first cluster together the first two genes that have the smallest distance. You then proceed iteratively, clustering together the two clusters that have the smallest distance. The second step in the clustering could cluster two new genes, or it could cluster a gene with the first cluster. Which occurs depends on which distance is smallest. This depends on a definition of the distance between two clusters, and there are multiple ways to define this.

Here is an example of yeast data, where the observations are time points – the expression level was tracked as a function of time. For this data, there are 4381 genes and 25 time points. The data is organized with genes in rows just because that is more convenient. I have seen data sets with say, 25,000 columns and only 500 rows, and this is difficult to look at and scroll through. This data is also nicely cleaned with rounded values; typically you'd see many more digits of precision for the expression levels.
`http://www.exploredata.net/Downloads/Gene-Expression-Data-Set`

Gene expression data

	A	B	C	D	E	F	G
1	time	40	50	60	70	80	90
2	YAL001C	-0.07	-0.23	-0.1	0.03	-0.04	-0.12
3	YAL014C	0.215	0.09	0.025	-0.04	-0.04	-0.02
4	YAL016W	0.15	0.15	0.22	0.29	-0.1	0.15
5	YAL020C	-0.35	-0.28	-0.215	-0.15	0.16	-0.12
6	YAL022C	-0.415	-0.59	-0.58	-0.57	-0.09	-0.34
7	YAL036C	0.54	0.33	0.215	0.1	-0.27	0.45
8	YAL038W	-0.625	-0.6	-0.4	-0.2	-0.13	0.33
9	YAL039C	0.05	-0.24	-0.19	-0.14	-1.22	-0.16
10	YAL040C	0.335	0.05	-0.04	-0.13	0.02	0.04
11	YAL044C	-0.43	-0.46	-0.39	-0.32	-0.66	0.03
12	YAL046C	0.135	0.23	0.125	0.02	0.09	0.16
13	YAL048C	0.005	0.02	-0.05	-0.12	0.1	0.1
14	YAL049C	-0.2	-0.32	-0.32	-0.32	-0.33	-0.26
15	YAL051W	0.155	0.2	0.22999999	0.26	-0.04	0.34

Gene expression data

To create the clusters, we need to define the distance between two clusters. When clustering genes, the distance between clusters can depend on the number of the elements in the clusters. We start with each gene being its own cluster. We call these **singleton** clusters. The distance between two singleton clusters is the Euclidean distance between them. So the distance between gene 1 and gene 2 could be written

$$D_{12} = \sqrt{\sum_{i=1}^{p}(x_{1i} - x_{2i})^2}$$

This generalizes the Euclidean distance between two points in the plane

$$d((x_{11}, x_{21}), (x_{12}, x_{22})) = \sqrt{(x_{11} - x_{12})^2 + (x_{21} - x_{22})^2}$$

Gene expression data

The more complicated question is how to define the distance between two clusters. If cluster C_1 has 3 genes and cluster C_2 has two genes, what is the distance between cluster C_1 and C_2?

There are a number of choices possible, which create different clustering algorithms:

1. $d(C_1, C_2) = \min\{d(x, y) : x \in C_1, y \in C_2\}$ (single linkage)
2. $d(C_1, C_2) = d(\overline{x}, \overline{y})$ where $\overline{x}$ and $\overline{y}$ are the average multivariate observations (called centroids) in the two clusters
3. $d(C_1, C_2) = \frac{1}{|C_1||C_2|} \sum_{x \in C_1} \sum_{y \in C_2} d(x, y)$ (average linkage)
4. $d(C_1, C_2) = \max\{d(x, y) : x \in C_1, y \in C_2\}$ (complete linkage)

Gene expression data

Other clustering methods are possible as well, but those are the most common hierarchical ones. Another common method is called k-means clustering. For this method, the number of clusters k, is specified in advance. For this method, you need an initial allocation of observations to the clusters, which could be done using a hierarchical method, for instance, or you can specify the centroids of the clusters.

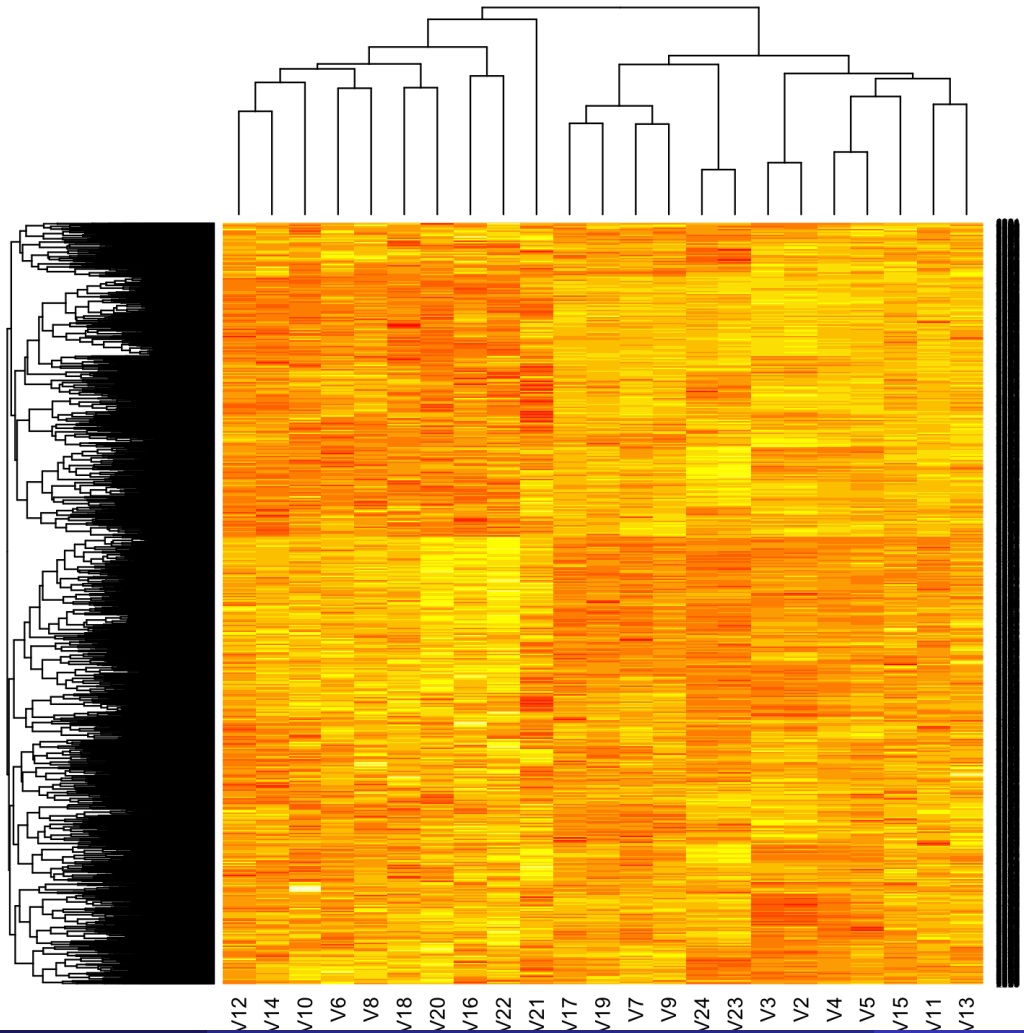
Then you compute the centroid of each cluster. You then compute the distance from each observation to each centroid, and reallocate observations to the cluster where the distance to the nearest centroid is minimized. This leads to recomputing the centroids, and another round of reallocation. The method is iterative and can be stopped after convergence or after a maximum number of iterations has been reached. The method can outperform hierarchical clustering on test data (when the cluster membership is known).

Gene expression data

```
> x <- read.table("spellman.csv",sep=",")
> y <- as.matrix(x[-1,-1]) # x into matrix y, removing time points
(first row) and first column (yeast gene names)
> z <- scale(t(y)) # scale the transposed matrix in order to scale
rows instead of columns
> a <- kmeans(t(z),10)
> t(t(a$cluster))
...
4372   10
4373    6
4374    6
4375    6
4376    6
4377    9
4378    1
4379    4
4380    1
4381    5
```

Gene expression data

The heatmap automatically does the clustering for you as well

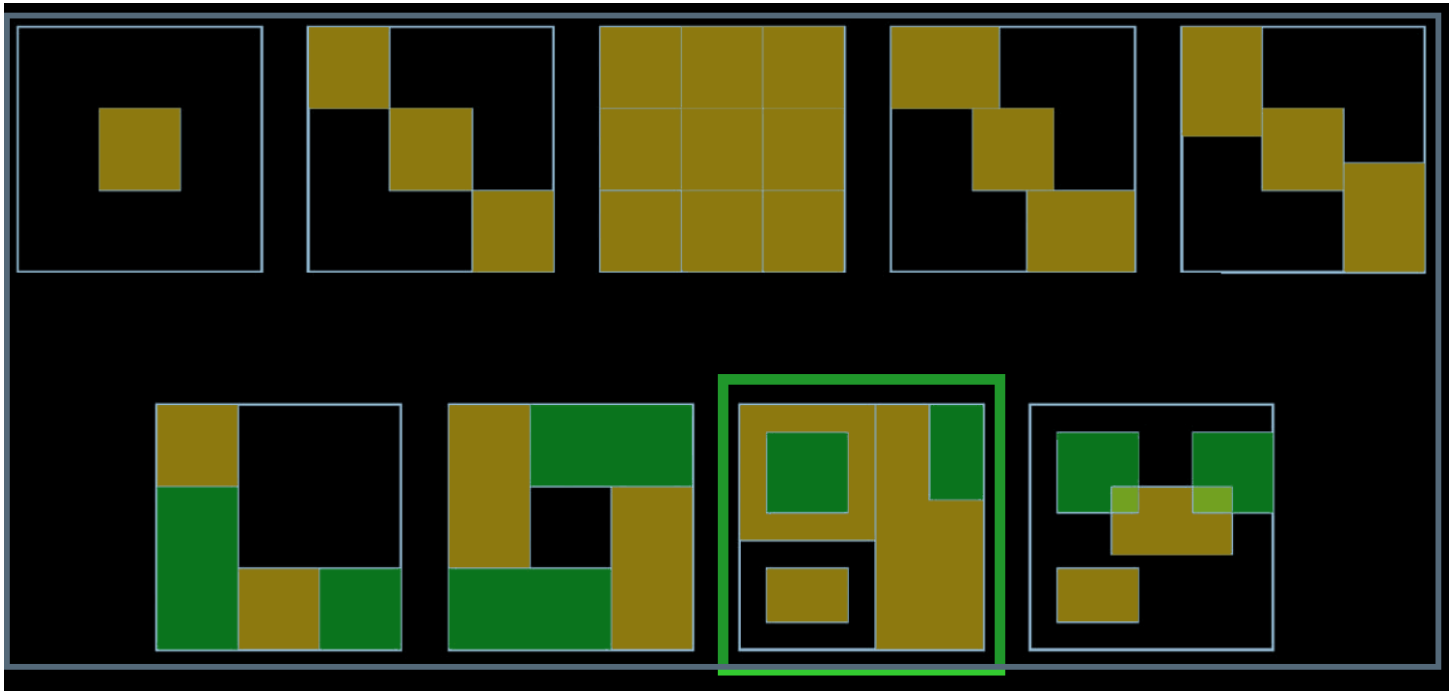


Gene expression data

Another type of clustering is biclustering. Here the idea is to find a subset of rows and subset of columns simultaneously such that they form a block which is considered a cluster. There are many types of biclustering depending on whether you want blocks (biclusters) to be non-overlapping, overlapping in only the rows (for example if genes contribute to more than one type of sample), overlapping only in the columns (if samples can belong to more than one type), and whether biclusters can be nested hierarchically.

Gene expression data

The heatmap automatically does the clustering for you as well



Gene expression data

As mentioned earlier, you can do SNP associations with gene expression data. For humans, if you have say 1 million SNPs and 20K genes, you get 20 billion associations to test. Obviously this can take a while, especially if you had say, 1000 individuals in your study. In stead of blindly searching all possible SNPs against all possible gene expressions, you could limit yourself to SNPs within a certain distance of a gene (say within the same chromosome or within 50 nucleotides.

Homework 3 (Due Wednesday, 22 April)

1. Do problem 7.6 from the book
2. Do problem 7.7 from the book
3. Find a journal article in a refereed journal that uses a survival analysis. Give a 1-page summary of the article including information about the type of survival analysis (or analyses) used, the sample size, number of variables, and type(s) of censoring in the data, and the conclusions they reached. You might search in Google Scholar or Web of Knowledge (or Web of Science) (you can link to this from the webpage for Centennial Library). You could search on particle journals, such as *Biostatistics*, *PLoS Clinical Trials*, *American Journal of Epidemiology*, *New England Journal of Medicine*, *Journal of the American Medical Association*, *Cancer*, *American Journal of Transplantation*, *Journal of Vascular Surgery*, etc. You could also search for particular topics like Cox Proportional Hazards model, log-rank test, etc.