♠ Affymetrix expression data

- See Affymetrix tutorial
- Expression summaries: summary expression estimates and other information
- Array normalisation; Hybridisation problems - low levels of intensity

♠ One gene: Sample statistics, summaries and exploration

- Sample $x_1, \dots, x_n$ of expression on one gene
- Sample mean $\bar{x} = n^{-1} \sum_{j=1}^n x_j$
- Sample variances $s = (n-1)^{-1} \sum_{j=1}^n (x_j - \bar{x})^2$
- Sample standard deviation (sd or std) $\sqrt{s}$
- Median, quartiles
- Histograms

♠ Multiple genes: Sample statistics, summaries and exploration

- Gene i, $i = 1, \dots, p$
- Expression level $x_{i,j}$ on array j
- Column vector $\mathbf{x}_j$ of dimension $p \times 1$ – all genes on array j
- $p \times n$ (tall, skinny) matrix $\mathbf{X} = [\mathbf{x}_1, \dots, \mathbf{x}_n]$
- Gene i sample summaries $\bar{x}_i, s_i$
- Sample covariance between genes i, k is $s_{i,k} = (n-1)^{-1} \sum_{j=1}^n (x_{i,j} - \bar{x}_i)(x_{k,j} - \bar{x}_k)$
  note: $s_{i,i} = s_i$
- Sample correlation between genes i, k is $r_{i,k} = s_{i,k}/\sqrt{s_i s_k}$
  note: $r_{i,i} = 1$
- Matrix notation: Sample variance (or covariance) matrix $\mathbf{S} = (s_{i,k})$

$$\mathbf{S} = \begin{pmatrix} s_1 & s_{1,2} & \cdots & s_{1,p} \\ s_{1,2} & s_2 & \cdots & s_{2,p} \\ \vdots & \vdots & \ddots & \vdots \\ s_{1,p} & s_{2,p} & \cdots & s_p \end{pmatrix}$$

- $\mathbf{S}$ is $p \times p$ symmetric, non-negative definite matrix
- Matrix notation: Sample correlation matrix $\mathbf{R} = (r_{i,k})$

$$\mathbf{R} = \begin{pmatrix} r_{1,1} & r_{1,2} & \cdots & r_{1,p} \\ r_{1,2} & r_{2,2} & \cdots & r_{2,p} \\ \vdots & \vdots & \ddots & \vdots \\ r_{1,p} & r_{2,p} & \cdots & r_{p,p} \end{pmatrix} = \begin{pmatrix} 1 & r_{1,2} & \cdots & r_{1,p} \\ r_{1,2} & 1 & \cdots & r_{2,p} \\ \vdots & \vdots & \ddots & \vdots \\ r_{1,p} & r_{2,p} & \cdots & 1 \end{pmatrix}$$

- $\mathbf{R}$ is $p \times p$ symmetric, non-negative definite matrix

♠ Matrix notation

- Suppose $\mathbf{X}$ has variables all standardised to zero mean; that is, $\bar{x}_i = 0$ for all genes $i = 1, \dots, p$ (just subtract a non-zero mean to begin to assure this). Then the variances and covariances simplify, $s_{i,k} = (n-1)^{-1} \sum_{j=1}^n x_{i,j} x_{k,j}$. Then $\mathbf{S} = (n-1)^{-1} \sum_{j=1}^n \mathbf{x}_j \mathbf{x}_j'$ or $\mathbf{S} = (n-1)^{-1} \mathbf{X}\mathbf{X}'$. So correlation patterns among genes are derived from $\mathbf{X}\mathbf{X}'$.

♠ Matlab data input, exploration, summary, graphics