A very simplified portrait of gene action and interaction…

Assume the genetic material has been “fairly well-mapped”

Sequencing of the material (AATCGGGCTTACTTGG…)

Identification of fragments that have/might have a function; **Genes**
What genes are involved in given conditions or processes?

- Differentiation among cell types (normal vs cancerous, sub-classifications of cancer cells)
- Response to a stimulus or class of stimuli (drug treatments, attacks to the cell)
- Cell cycle, etc.

What are the relationships among genes?

- Genes who act upon one another through their products; regulatory mechanisms
- Genes whose products are involved in similar or related functions, and are therefore likely to be co-regulated, etc.

Reconstructing modular networks among genes: the genome-level “contraction” of more complex pathways involving extra-nuclear and/or extra-cellular signaling and interactions.

To answer these types of questions (and many others!) one needs

- Sequence information
- Product information
- Expression information: how intensively genes are “used”/ “activated” in various settings.

for a wide set of candidate genes…

MICROARRAY TECHNOLOGIES ALLOW US TO COLLECT EXPRESSION INFORMATION FOR THOUSANDS OF GENES SIMULTANEOUSLY
Some intuitive examples on the use of expression information (no statistics involved yet):

What are the genes that are under or over-expressed in a cancerous cell with respect to a normal one?

<table>
<thead>
<tr>
<th>Gene</th>
<th>Normal</th>
<th>Cancerous</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gene 1</td>
<td></td>
<td>Yes</td>
</tr>
<tr>
<td>Gene 2</td>
<td></td>
<td>Yes</td>
</tr>
<tr>
<td>Gene 3</td>
<td>No</td>
<td></td>
</tr>
<tr>
<td>Gene 4</td>
<td>No</td>
<td></td>
</tr>
</tbody>
</table>

What are the genes whose expression patterns allow us to discriminate across various sub-types of cancer cells?

<table>
<thead>
<tr>
<th>Gene</th>
<th>Type A</th>
<th>Type B</th>
<th>Type C</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gene 1</td>
<td>Yes</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Gene 2</td>
<td></td>
<td>Yes, together</td>
<td></td>
</tr>
<tr>
<td>Gene 3</td>
<td></td>
<td>Yes, together</td>
<td></td>
</tr>
<tr>
<td>Gene 4</td>
<td>No</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

In these examples the “conditions” on which expression is observed are cell types. We could have one or several instances (sample of cells, maybe coming from different individuals) for each. After isolating the relevant genes:

Can we predict cell type on the basis of gene expression?
Can we group cell types on the basis of expression similarity?
More intuitive examples on the use of expression information (no statistics involved yet):

What are the genes whose expression patterns are related to the cell cycle?

Gene 1
Genes 2, 3
Gene 4
Gene 5

Yes, variation along the cycle that repeats itself
(early, middle, late phase, respectively?)

No

What are the genes whose expression patterns are altered in response to a given stimulus?

Gene 1
Genes 2, 3
Gene 4
Gene 5

Yes, variation pattern after
(early, middle, late response, respectively?)

No

In these examples the “conditions” on which expression is observed are a sample of times (expression profiles over a time-course). After isolating the relevant genes:
Can we identify characteristic expression patterns?
Can we group relevant genes on the basis of expression similarity?
Can we use expression profiles to infer regulatory relationships?
Microarrays:

How is expression information collected? A super-simplified summary:

1. Slides (chips) with cDNA clones spotted on a 2D grid, are either built in loco (e.g. microarrayer robot at Penn State microarray facility!), or bought – more or less customized. Each spot corresponds to a gene, or more generally a well identified fragment of genetic material.

2. For any one condition, an mRNA population is extracted from a corresponding cell culture, “pored and gently stirred” on a chip after being “color treated”, and allowed to hybridize; that is, to “stick” to the spots on the chip.

3. The chip is scanned with a laser, producing an image in which light intensity on the color channel in use is proportional to expression for each spot. The image is then processed to obtain numerical readings for each spot.

4. Readings corresponding to different conditions are made comparable through normalization procedures, using baseline (control, reference) readings – the green channel readings for two-color arrays.

5. Normalized readings are often further “massaged” before data analysis; for example:
   - Imputing missing values, correcting out-of-range values
   - Centering and standardizing by gene over conditions (in a sense, make readings corresponding to different genes comparable by eliminating from the analysis overall level and variation magnitude of expression)
   - Centering by condition over genes.

Outcome:

\[
\begin{align*}
? & X_{1,1} & \ldots & X_{1,T} & ? \\
? & ? & \ldots & ? & , \quad i = 1 \ldots N \ (genes), \ j = 1 \ldots T \ (conditions) \\
? & X_{N,1} & \ldots & X_{N,T} & ?
\end{align*}
\]
Sources of experimental error:

Spotting, building the chips

Production of mRNA, color treatment, hybridization

Reading, processing the image

Means to evaluate and manage error:

Ex-ante…
- **Replicates:** the same gene (fragment of genetic material) is spotted in more than one location on the 2D grid.
- **Controls:** some locations on the 2D grid are spotted with genes (fragments of genetic material) to which none, or more generally a fixed “systematic” amount, of mRNA should stick in the conditions under study – e.g. completely unrelated genes from another organism.
- **Empty spots:** to control for background or spurious shining in the readings processing.

Ex-post…
- “Robustify” data analysis

We will go back to some of these issues.