Genomic assays and the multiple testing problem

Maria Figueroa, MD
2-2-09
The ‘Omics’ era

High-throughput technologies allow for us to simultaneously query tens of thousands of targets:

- Microarrays
- Proteomics (MS - Protein arrays)
- Massively parallel sequencing

- Increased the amount of biology captured by one experiment
- Significant amount of noise
- Pose specific statistical problems
Basic concepts on microarray technology

- Collection of known ssDNA probes arrayed on a solid surface by covalent attachment to a chemically suitable matrix

- Quantitative and qualitative measurements of nucleic acids

- Rely on the ability of nucleic acids to hybridize to the DNA probes through base pair recognition under specific experimental conditions
Microarray-based experiments: General design

1. Extract and process DNA/RNA
2. Label
3. Competitive hybridization
4. Scan image
5. Extract data
Different type of Biological platforms

- Gene Expression arrays: Changes in gene expression levels

- Array-based Comparative Genomic Hybridization (aCGH): DNA copy number variations

- ChIP-on-chip: Genomic localization of DNA-Protein interactions

- DNA Methylation microarrays: Localization of 5-methyl-Cy

- Genotyping microarrays (SNP microarrays): Single nucleotide polymorphisms
Some statistical considerations

- **Variables far exceed number of samples**
  
  - e.g.: Test clinical response to a new drug for treatment of high blood pressure on 200 pts.
  
  vs. Identify gene expression changes associated with the same drug in 200 pts.

- **Multiple comparisons**
  
  - i.e. in order to identify genes that change in a statistically significant manner with the drug we will need to *test each of the 37,000 genes* on the array in parallel and then select the significant ones
### Multiple comparisons: a practical example

<table>
<thead>
<tr>
<th>Gene 1</th>
<th>Treatment (+)</th>
<th>Treatment (-)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0.701365258</td>
<td>0.847669154</td>
</tr>
<tr>
<td>Gene 2</td>
<td>0.019893544</td>
<td>0.998953774</td>
</tr>
<tr>
<td>Gene 3</td>
<td>0.823234225</td>
<td>0.009390894</td>
</tr>
<tr>
<td>Gene 4</td>
<td>0.831201089</td>
<td>0.672332561</td>
</tr>
<tr>
<td>Gene 5</td>
<td>0.618048089</td>
<td>0.493722217</td>
</tr>
<tr>
<td>Gene 6</td>
<td>0.314244277</td>
<td>0.693208323</td>
</tr>
<tr>
<td>Gene 7</td>
<td>0.834711252</td>
<td>0.975953907</td>
</tr>
<tr>
<td>Gene 8</td>
<td>0.632542712</td>
<td>0.320787292</td>
</tr>
<tr>
<td>Gene 9</td>
<td>0.613812632</td>
<td>0.943127333</td>
</tr>
<tr>
<td>Gene 10</td>
<td>0.326036635</td>
<td>0.138067146</td>
</tr>
<tr>
<td>Gene 11</td>
<td>0.634973714</td>
<td>0.556115332</td>
</tr>
<tr>
<td>Gene 12</td>
<td>0.985398561</td>
<td>0.057188922</td>
</tr>
<tr>
<td>Gene 13</td>
<td>0.122163747</td>
<td>0.363493692</td>
</tr>
<tr>
<td>Gene 14</td>
<td>0.414223175</td>
<td>0.383947257</td>
</tr>
<tr>
<td>Gene 15</td>
<td>0.285974499</td>
<td>0.155930996</td>
</tr>
<tr>
<td>Gene 16</td>
<td>0.672888773</td>
<td>0.772655752</td>
</tr>
<tr>
<td>Gene 17</td>
<td>0.016216298</td>
<td>0.080760328</td>
</tr>
<tr>
<td>Gene 18</td>
<td>0.551922437</td>
<td>0.097837061</td>
</tr>
<tr>
<td>Gene 19</td>
<td>0.889225949</td>
<td>0.629840191</td>
</tr>
<tr>
<td>Gene 20</td>
<td>0.679047253</td>
<td>0.610386551</td>
</tr>
</tbody>
</table>

**1- Gene by gene Two-tailed T test**

**2- Significance of p< 0.05**
Conclusion: Gene 16 is upregulated with the treatment.
But... let's review a few things

✓ p < 0.05: This means we accept the risk of erroneously rejecting the null hypothesis in 5% of the cases i.e. we are willing to accept 5% false positive calls.

✓ In our example we did not do *1 comparison* (treated vs. untreated), we in fact did *20 comparisons in parallel*.

✓ Each time we had a 5% error, so if we repeat the test 20 times we are likely to get at least 1 false positive.

✓ Gene 16 may or may not change its expression level with the treatment, but we do not have enough evidence to claim that it does.

Our “example data set” was in fact generated with a random number generator.
### Probabilities of 1 or more false positives by chance

If we set p-value at < 0.05

<table>
<thead>
<tr>
<th># genes tested (N)</th>
<th>False positives incidence</th>
<th>Probability of calling 1 or more false + by chance</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1/20</td>
<td>5%</td>
</tr>
<tr>
<td>2</td>
<td>1/10</td>
<td>10%</td>
</tr>
<tr>
<td>20</td>
<td>1</td>
<td>64%</td>
</tr>
<tr>
<td>100</td>
<td>5</td>
<td>99.4%</td>
</tr>
</tbody>
</table>

\[
100(1-0.95^N)
\]
And on a genomics scale...

⇒ Suppose no genes really changed (e.g. in random samples from the same population)
⇒ ~10,000 genes on an array
⇒ Each gene has a 5% chance of exceeding the threshold at a p-value of 0.05 (Type I error)
⇒ So by chance alone…
    - the p-values for 500 genes should be significant!!
- Most approaches for correcting for multiple comparisons work well for small number of parallel comparisons

- But when tens of thousands of tests are performed most of these are too stringent (e.g. Bonferroni, Sidak, Holm’s)

- The most accepted methods for multiple testing correction in the microarray field are:
  - the False Discovery Rate (FDR) determination (Benjamini-Hochberg)
  - the use of permutations (Westfall-Young, SAM)
## Corrections for multiple comparisons

<table>
<thead>
<tr>
<th>Method</th>
<th>False (±)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bonferroni</td>
<td>False (-)</td>
</tr>
<tr>
<td>Holm’s step down</td>
<td></td>
</tr>
<tr>
<td>Westfall-Young</td>
<td></td>
</tr>
<tr>
<td>Benjamin-Hochberg FDR</td>
<td></td>
</tr>
<tr>
<td>None</td>
<td>False (+)</td>
</tr>
</tbody>
</table>
Permutation-based methods (at least 1000!)

S1 S2 S3 S4 S5 S6 S7 S8 S9 S10 S11 S12

Normal | Tumor

S11 S5 S8 S6 S9 S7 S12 S2 S3 S10 S1 S4

Random 1 | Random 2

Observed significant p-values

Expected significant p-values by chance
Epigenetic Signatures Identify New Clinically Relevant Subtypes and Define Gene Regulatory Patterns in Patients with Acute Myeloid Leukemia (AML)

Maria E Figueroa, Sanne Lugthart, Yushan Li, Claudia Erpelinck-Verschueren, Paul J Christos, Xutao Deng, Fabien Campagne, Madhu Mazumdar, Bob Löwenberg, John Greally, Peter JM Valk, Ruud Delwel, and Ari Melnick.
Gene Expression Depends On Multiple Factors

- DNA Sequence
- DNA Methylation
- Histone Modifications

Genetic expression depends on multiple factors.

- GENETIC REGULATION
- EPIGENETIC REGULATION
Gene expression profiling has limitations

- Gives only a snapshot of genes transcribed at the time, with no information on their availability for transcription.
- Does not detect epigenetic/copy number changes
- Only genes with high expression levels stand out above the noise level
- Sometimes biologically significant changes are lost within the noise signal
Aberrant DNA methylation is a hallmark of cancer

**Normal**
- Specific distribution of cytosine methylation
- Promoter CpG island hypomethylation
- Methylation of repetitive elements

**Cancer**
- Global hypomethylation
- Promoter CpG island hypermethylation
- Aberrant silencing of certain tumor suppressors
- Aberrant hypomethylation of certain oncogenes
Identifying aberrant epigenetic patterns in AML will:

a) provide critical insight into the biological complexity of the disease

b) help identify new and clinically relevant disease subtypes
Hypotheses

Identifying aberrant epigenetic patterns in AML will:

a) provide critical insight into the biological complexity of the disease

b) help identify new and clinically relevant disease subtypes
The HELP Assay for Genome-wide 5me-Cy detection

Ligation of Linkers

Differential Restriction Digestion

PCR

HpaII Amplifiable fragment array

Khulan et. al. Genome Res, 2006
Validation of HELP data by MassARRAY EpiTyper
Patients’ characteristics

- 344 patients from Erasmus MC
- HOVON trials 04, 29, 32, 42 and 43
- Median follow-up: 18.2 months (0.1-214.5 m.)
- Median age: 48 years (15-77 years)
- Male: 188; Female: 156
- CD34+ bone marrow cells from 8 healthy donors
Methods

Unsupervised analysis
- Hierarchical clustering to explore internal complexity of the data
- Class discovery

Supervised analysis
- Understand the biology associated with each methylation cluster

Survival analysis
- Determine risk associated with methylation clusters
DNA methylation captures AML biology

Slide Redacted
Unpublished Results
DNA methylation captures AML biology

Slide Redacted
Unpublished Results
DNA methylation captures AML biology

Slide Redacted
Unpublished Results
Supervised analysis

- Understand the biology associated with each methylation cluster

Comparison of each cluster to normal CD34 + cells

Identify aberrant DNA methylation signature for each cluster

Pathway and Gene ontology analysis to understand associated biology
Supervised analysis

- Understand the biology associated with each methylation cluster

### Multiple testing

**Problem #1**: Gene 25,626

<table>
<thead>
<tr>
<th></th>
<th>K0 = Normals</th>
<th>K1</th>
<th>(K…)</th>
<th>K16</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gene 1</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Gene 2</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(Gene…)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Gene 25,626</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Problem #2**

- Gene 25,626
Supervised analysis - Understand the biology associated with each methylation cluster

**Methods**

ANOVA x 25,626 + BH correction

Dunnett’s method

<table>
<thead>
<tr>
<th></th>
<th>K0 = Normals</th>
<th>K1</th>
<th>(K…)</th>
<th>K16</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gene 1</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Gene 2</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(gene…)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Gene 25,626</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
1- Select genes with ANOVA $p < 0.0005$ after BH

2- Run Dunnett’s method for pairwise comparisons against a reference group

3- Select for each cluster the genes that are significant at $p < 0.0005$
Slide Redacted
Unpublished Results
Combining statistical and biological significance

- Increases our chances of capturing biologically significant changes
- Still requires that we correct the p value for multiple testing
“If I correct I do not get any significant genes, so I am better off not correcting”

Wrong! If you do not correct, your “significant” genes are probably not significant at all. This is like cheating your own self!

“My hypothesis was wrong because I do not have any significant genes after correction”

This may or may not be the case. You may just have insufficient power in your design to detect small changes. You can:

1- Increase the number of replicates/samples

2- Select a smaller number of genes to begin your analysis with (high variance genes, high SNR) and in this way the stringency of your correction will be reduced
In Summary

- High-throughput methods are very useful in biology.
- However, there is a risk for drawing the wrong conclusions if we are not careful.
- Conventional statistical approaches may not always be the most appropriate for these data sets.
- When selecting an analytical approach we need to remember the nature of the data we are analyzing (high number of correlated genes, lack of normality, etc)
- For multiple testing: B-H FDR and permutation-based methods are acceptable ways of dealing with this
- Nothing can replace experimental validations!!