Statistics for genomics

Jeff Leek

@jtleek

www.jtleek.com

http://www.ncbi.nlm.nih.gov/pubmed/18271621

An example

OPEN ACCESS Freely available online

PLOS MEDICINE

An Erythroid Differentiation Signature Predicts Response to Lenalidomide in Myelodysplastic Syndrome

Benjamin L. Ebert[1,2,3], Naomi Galili[4], Pablo Tamayo[1], Jocelyn Bosco[1,2], Raymond Mak[1,2], Jennifer Pretz[1,2], Shyam Tanguturi[1], Christine Ladd-Acosta[1], Richard Stone[2,3], Todd R. Golub[1,2,5,6], Azra Raza[4*]

1 Broad Institute of Harvard and Massachusetts Institute of Technology, Cambridge, Massachusetts, United States of America, **2** Dana-Farber Cancer Institute, Harvard Medical School, Boston, Massachusetts, United States of America, **3** Brigham and Women's Hospital, Department of Medicine, Boston, Massachusetts, United States of America, **4** St. Vincent's Comprehensive Cancer Center, New York, New York, United States of America, **5** Childrens's Hospital, Boston, Massachusetts, United States of America, **6** Howard Hughes Medical Institute, Chevy Chase, Maryland, United States of America

1. Normalization

   Key Concepts: Borrowing information

2. Differential Expression

   Key Concepts: Permutation, multiple comparisons

3. Gene Set Enrichment

   Key Concepts: Finding biological patterns

Normalization ensures that differences in intensities are not just due to technology, e.g. reagent, sequencing, imaging, batch effects, quality control artifacts

http://www.genomebiology.com/2015/16/1/117

Raw data

2	4	4	5
5	14	4	7
4	8	6	9
3	8	5	8
3	9	3	5

***Order* values within each sample (or column)**

2	4	3	5
3	8	4	5
3	8	4	7
4	9	5	8
5	14	6	9

***Average* across rows and substitute value with average**

3.5	3.5	3.5	3.5
5.0	5.0	5.0	5.0
5.5	5.5	5.5	5.5
6.5	6.5	6.5	6.5
8.5	8.5	8.5	8.5

***Re-order* averaged values in original order**

3.5	3.5	5.0	5.0
8.5	8.5	5.5	5.5
6.5	5.0	8.5	8.5
5.0	5.5	6.5	6.5
5.5	6.5	3.5	3.5

After normalization/summarization

Response	R	R	•••	NR	NR
	Patient 1	Patient 2	•••	Patient n-1	Patient n
Gene 1	-1.64	-0.42	•••	-1.39	-0.38
Gene 2	-3.12	-3.60	•••	-3.80	-2.82
⋮	⋮	⋮	•••	⋮	⋮
⋮	⋮	⋮	•••	⋮	⋮
			•••		

Association analysis (differential expression analysis) is the search for features, like genes. that show “significant” differences between groups of patients or across phenotypes.

Form a statistic

$$S = \frac{\mu_R - \mu_{NR}}{\sigma_R + \sigma_{NR}}$$

μ_R - average responder expression
μ_{NR} - average non-responder expression
σ_R - standard deviation of responders
σ_{NR} - standard deviation of non-responders

Multiple comparisons

- Family wise error rate:

$$\Pr(\#\text{ False Positives} \geq 1)$$

- False discovery rate:

$$E\left[\frac{\#\text{False Positives}}{\#\text{ Of Discoveries}}\right]$$