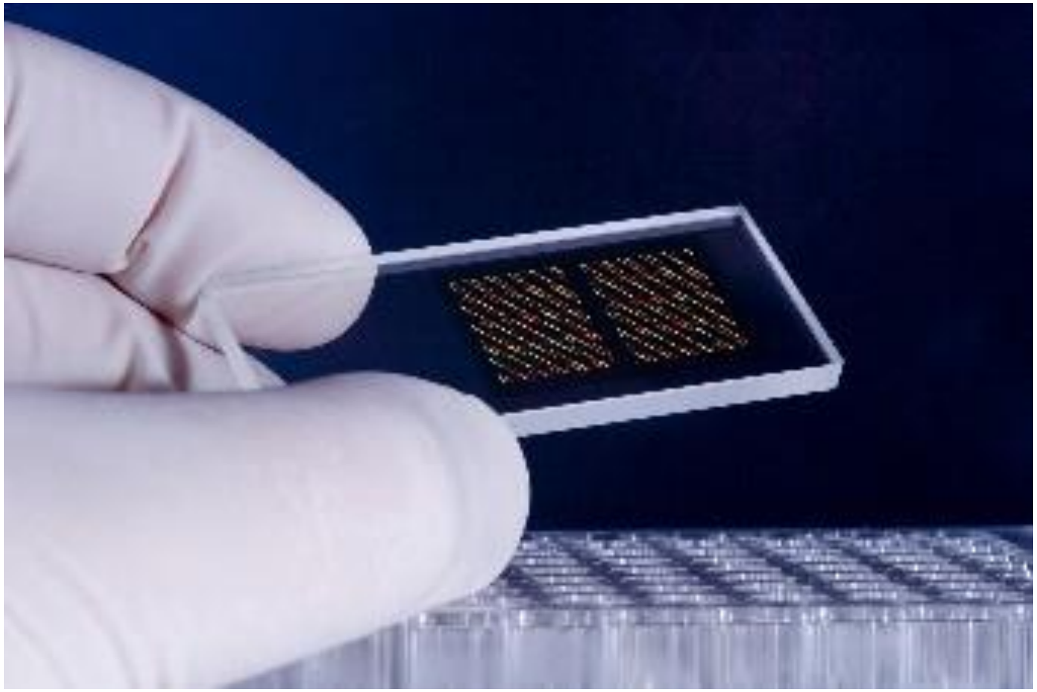
Βιοπληροφορική ΙΙ

Παντελής Μπάγκος
Αναπληρωτής Καθηγητής

Πανεπιστήμιο Θεσσαλίας
Λαμία, 2015

Μικροσυστοιχίες

Γυάλινο πλακίδιο που αποτελείται από συγκεκριμένες αλληλουχίες οι οποίες είναι ειδικές για συγκεκριμένα γονίδια, τους ανιχνευτές (probes), οι οποίοι είναι ακινητοποιημένοι σε μία κουκκίδα (spot) της γυάλινης επιφάνειας του πλακιδίου.



Μικροσυστοιχίες

- Ταυτόχρονη ανάλυση του τρόπου έκφρασης χιλιάδων γονιδίων σε διαφορετικά δείγματα ή σε διαφορετικά στάδια ανάπτυξης

- Σύγκριση έκφρασης σε φυσιολογικές και παθολογικές καταστάσεις

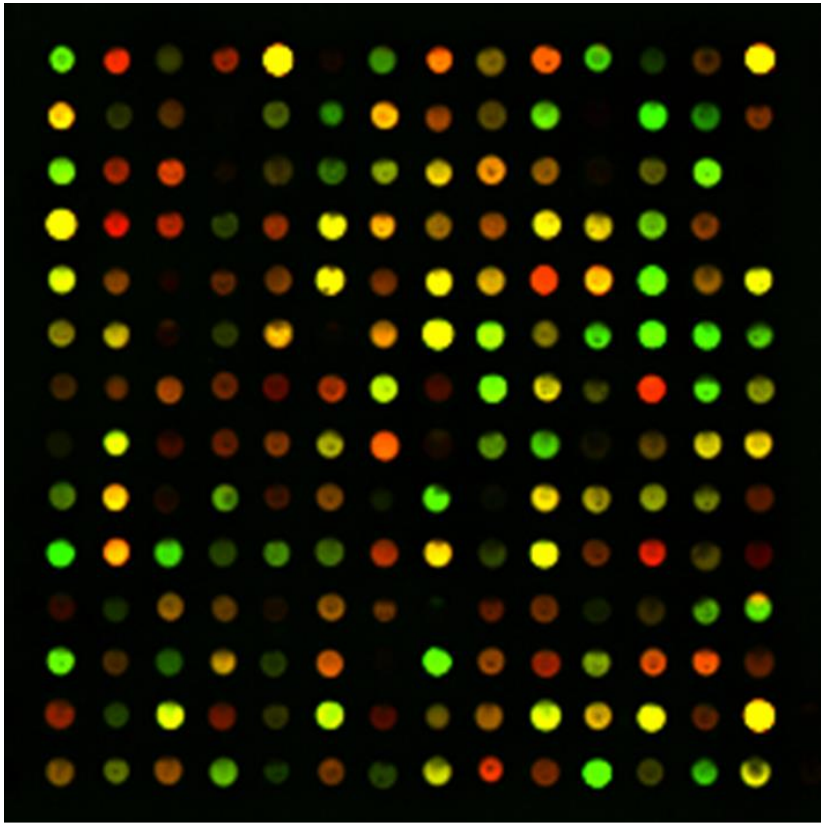
- Ανταπόκριση σε φαρμακευτικές ουσίες ή θεραπείες

- Παρέχουν χρήσιμες πληροφορίες για τη βιολογική λειτουργία ενός οργανισμού, βρίσκοντας ποια γονίδια ενεργοποιούνται ή καταστέλλονται σε διάφορα στάδια ανάπτυξης ή σε απόκριση σε ερεθίσματα του περιβάλλοντος, όπως η απόκριση σε ορμόνες ή σε υψηλή θερμοκρασία

Βασικά βήματα για ένα πείραμα μικροσυστοιχιών

- Διατύπωση του βιολογικού ερωτήματος
- Επιλογή του κατάλληλου τύπου μικροσυστοιχίας (τυπωμένες μικροσυστοιχίες cDNA, τυπωμένες μικροσυστοιχίες ολιγονουκλεοτιδίων, μικροσυστοιχίες που κατασκευάστηκαν με *in situ* σύνθεση ολιγονουκλεοτιδίων)
- Απομόνωση του RNA από τα δείγματα
- Σήμανση των δειγμάτων με φθορίζουσες ουσίες
- Υβριδισμός στην επιφάνεια της μικροσυστοιχίας
- Σάρωση μικροσυστοιχίας στα μήκη κύματος των φθορίζουσων ουσιών και μετρώντας τον αντίστοιχο φθορισμό της κάθε ουσίας
- Χρήση κατάλληλων προγραμμάτων για τη δημιουργία της τελικής εικόνας των μικροσυστοιχιών.

Μικροσυστοιχίες

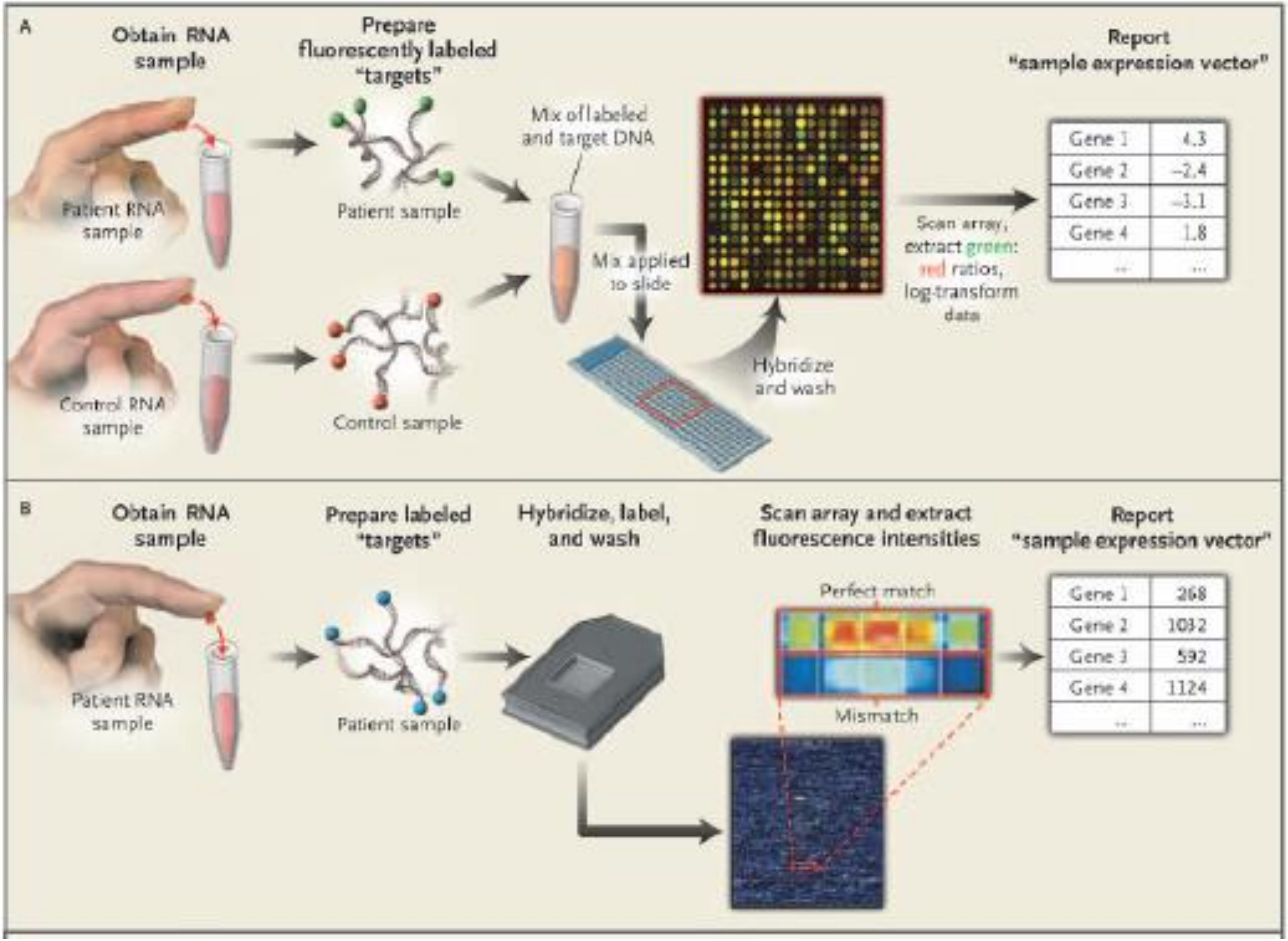


Η συνδυασμένη εικόνα της μικροσυστοιχίας παρέχει ένα βολικό τρόπο ώστε να βρεθούν τα γονίδια τα οποία βρίσκονται σε μεγαλύτερη έκφραση στο δείγμα ελέγχου σε σύγκριση με το δείγμα αναφοράς

Μικροσυστοιχίες

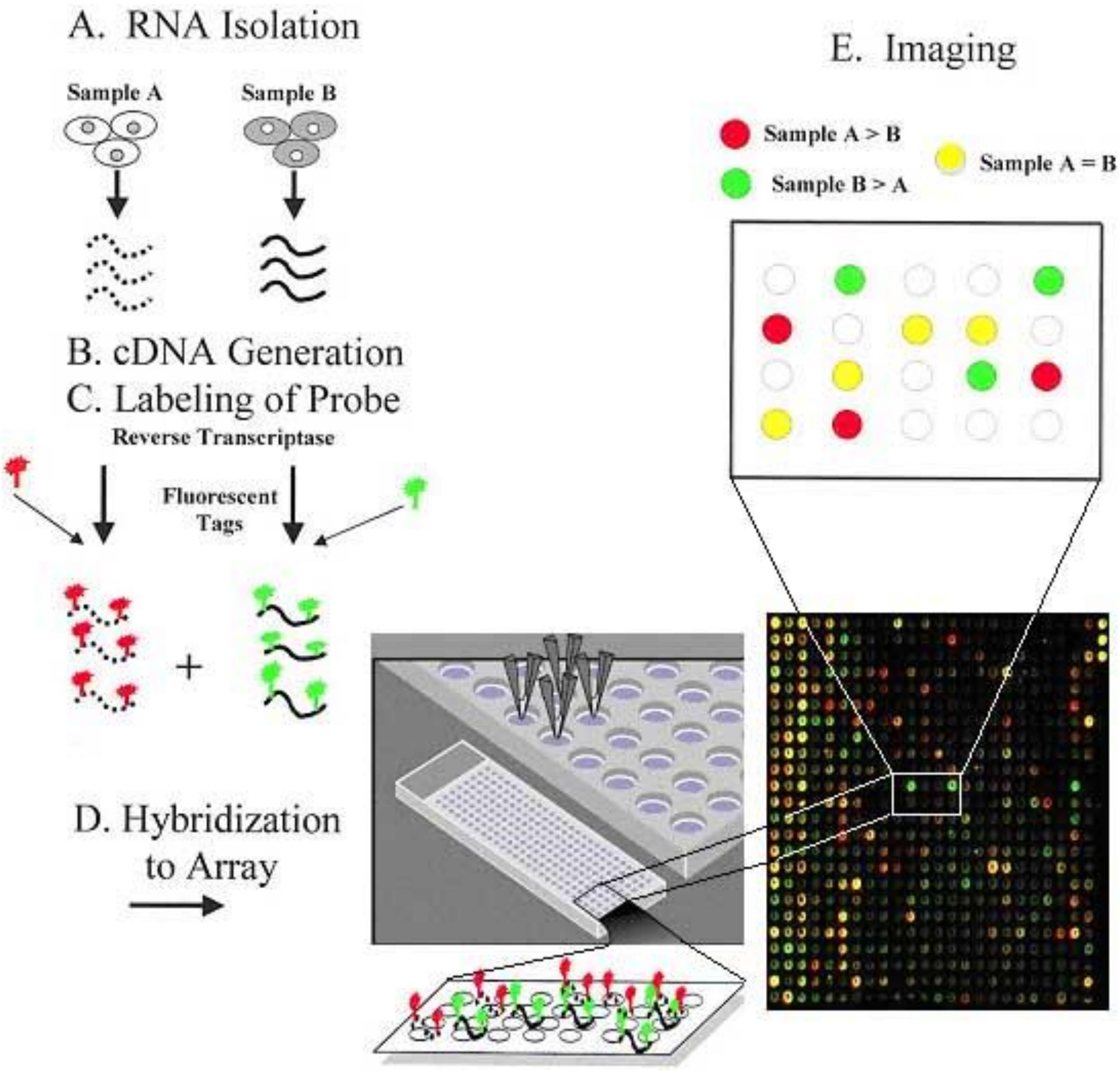
- Μονοχρωματικές μικροσυστοιχίες (Affymetrix): Κάθε δείγμα RNA σημαίνεται με μια χρωστική και τοποθετείται για υβριδισμό σε ένα τσιπ μικροσυστοιχιών.

- Διχρωματικές μικροσυστοιχίες: Δύο δείγματα RNA (ελέγχου – αναφοράς) σημαίνονται με 2 διαφορετικές φθορίζουσες ουσίες και το τοποθετούνται για υβριδισμό στο ίδιο τσιπ μικροσυστοιχιών.

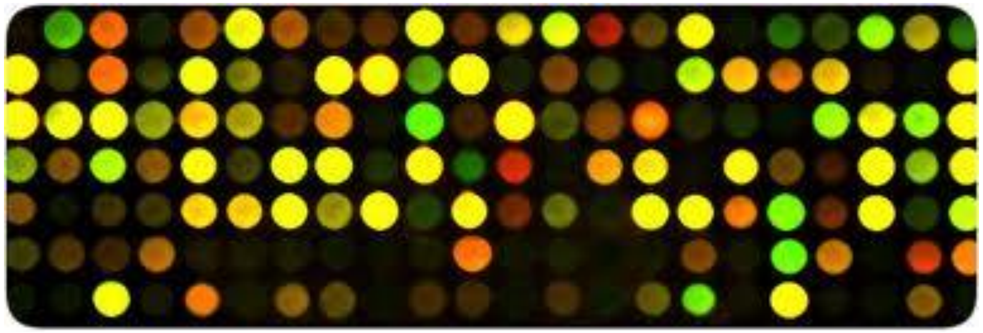

A
Obtain RNA sample
Prepare fluorescently labeled "targets"
Report "sample expression vector"
Patient RNA sample
Patient sample
Mix of labeled and target DNA
Control RNA sample
Control sample
Mix applied to slide
Hybridize and wash
Scan array, extract green: red ratios, log-transform data
B
Obtain RNA sample
Prepare labeled "targets"
Hybridize, label, and wash
Scan array and extract fluorescence intensities
Report "sample expression vector"
Patient RNA sample
Patient sample
Perfect match
Mismatch

Gene 1	4.3
Gene 2	−2.4
Gene 3	−3.1
Gene 4	1.8
...	...

Gene 1	268
Gene 2	1032
Gene 3	592
Gene 4	1124
...	...


A. RNA Isolation
Sample A
Sample B
B. cDNA Generation
C. Labeling of Probe
Reverse Transcriptase
Fluorescent
Tags
+
D. Hybridization
to Array
E. Imaging
Sample A > B
Sample B > A
Sample A = B

Μικροσυστοιχίες



- Με κόκκινο χρώμα εμφανίζεται μια κουκκίδα, αν σε αυτήν η ποσότητα του δείγματος ελέγχου είναι μεγαλύτερο
- Με πράσινο χρώμα εμφανίζεται μια κουκκίδα, αν σε αυτήν η ποσότητα του δείγματος αναφοράς είναι μεγαλύτερο
- Με κίτρινο χρώμα εμφανίζεται μια κουκκίδα, αν σε αυτήν οι ποσότητες του δείγματος ελέγχου και του δείγματος αναφοράς είναι ίσες
- Με μαύρο χρώμα εμφανίζεται μία κουκκίδα αν κανένα δείγμα δεν έχει υβριδοποιηθεί
- Οι υπόλοιπες αποχρώσεις εμφανίζονται για αντίστοιχες ποσότητες των δύο δειγμάτων

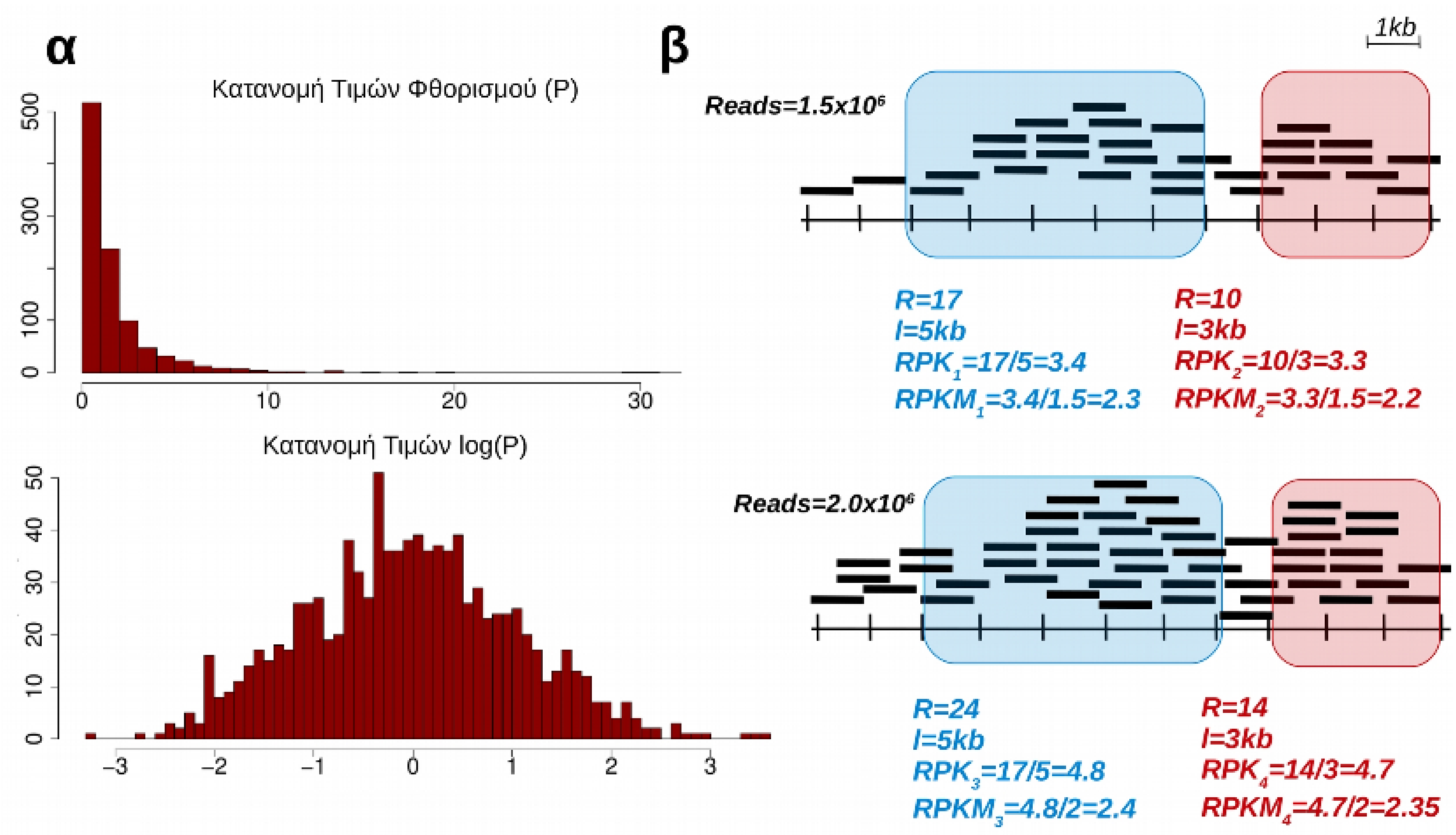
Ποσοτικοποίηση δεδομένων

•Η ένταση του φθορισμού μετατρέπεται σε αριθμητικά δεδομένα και δίνει πληροφορίες σχετικά με την έκφραση των γονιδίων της μικροσυστοιχίας.

•Το σχετικό επίπεδο έκφρασης για κάθε γονίδιο αντιστοιχεί με την ποσότητα του κόκκινου ή του πράσινου φωτός που εκπέμπεται μετά από διέγερση.

•Για να συσχετίσουμε αυτές τις ποσότητες και να εξάγουμε το σχετικό επίπεδο έκφρασης κάθε γονιδίου χρησιμοποιούμε το λόγο έκφρασης

$$T_i = \frac{R_i}{G_i} \qquad T_i' = \log_2(T_i)$$



Εικόνα 7.3: *α) Ιστόγραμμα τιμών φθορισμού που ακολουθούν λογαριθμοκανονική κατανομή (επάνω) και μετά από λήψη των λογαρίθμων τους (κάτω) που μετατρέπει την κατανομή τους σε κανονική β) Γραφική αναπαράσταση του υπολογισμού της τιμής RPKM από δύο πειράματα αλληλούχισης RNA. Οι δύο χρωματισμένες περιοχές περιέχουν διαφορετικό αριθμό μικρο-αναγνώσεων (reads) όμως αυτό είναι αποτέλεσμα του διαφορετικού τους μήκους (5kb έναντι 3kb). Διαίρεση με το μήκος (RPK) δίνει παραπλήσιες τιμές για τις δύο περιοχές στο ίδιο πείραμα. Μεταξύ δύο πειραμάτων με διαφορετικό συνολικό αριθμό αναγνώσεων χρειάζεται μια ακόμα διόρθωση ως προς το συνολικό αριθμό των reads. Ετσι οι τιμές RPKM είναι πολύ παρόμοιες για τις δύο περιοχές και μεταξύ των δύο πειραμάτων.*

Σφάλματα στα πειράματα μικροσυστοιχιών

Τυχαία και συστηματικά σφάλματα συμβαίνουν σε ένα πείραμα μικροσυστοιχιών:

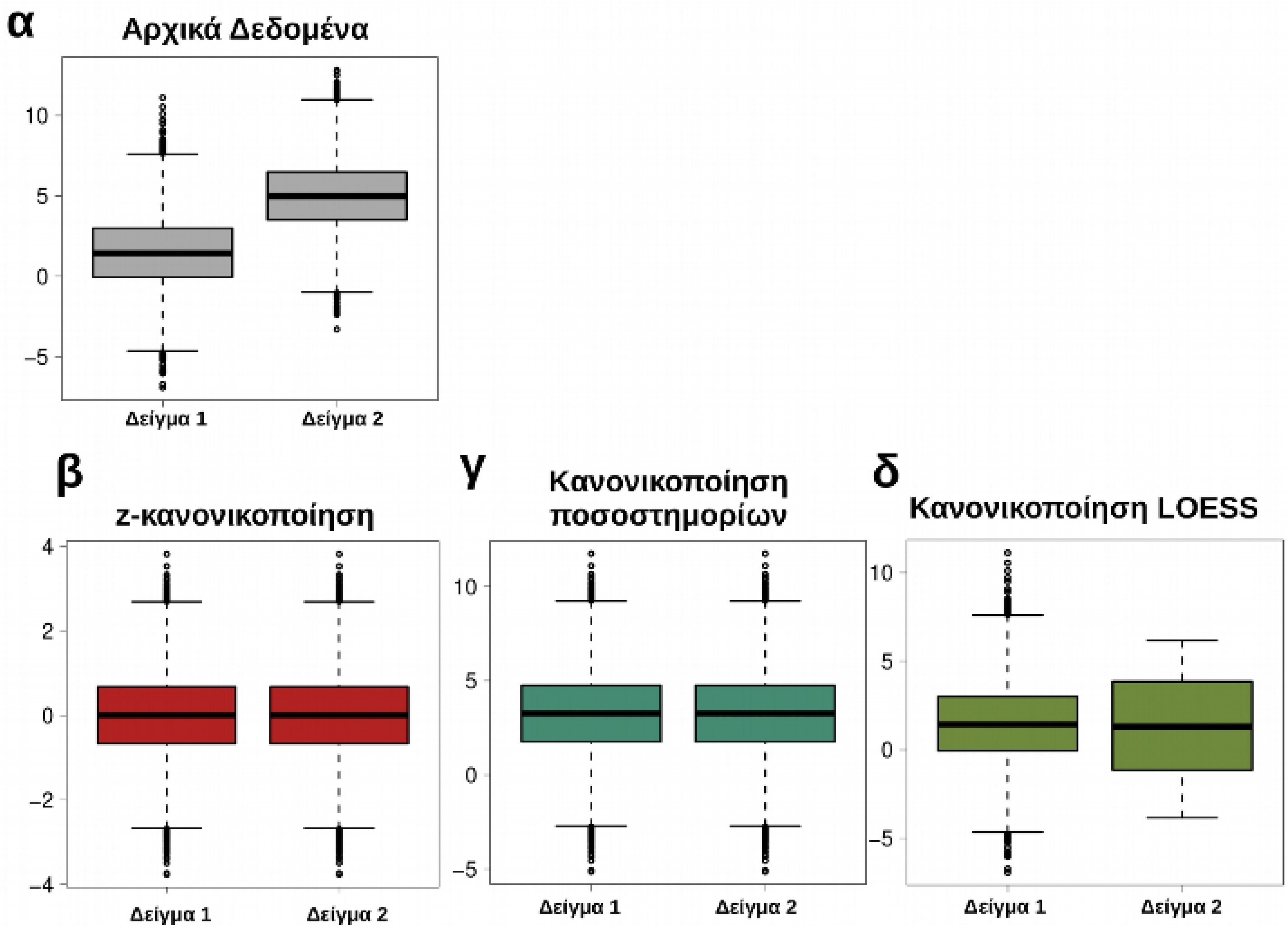
- Χρήση διαφορετικών φθορίζουσων ουσιών
- Χρήση διαφορετικών πλατφορμών
- Διαφορετικές πειραματικές συνθήκες
- Εισαγωγή θορύβου στα δεδομένα από το σαρωτή

image analysis

- Following hybridization, **image analysis** is performed (Yang, Buckley et al. 2001). Pre-filtering/masking method follows and Background Signal adjustment is recommended before scaling. Masking refers to applications of microarray signal correction that account for cross hybridization (Naef and Magnasco 2003), array scratches, scanner improper configuration (Shi, Tong et al. 2005, Timlin 2006), spot light saturation and washing issues (Yauk, Berndt et al. 2005) that may have occurred.

Normalization

- **Normalization** is performed to correct for systematic differences between samples on the same slide, or between slides, which do not represent true biological variation between slides and enables experiments to be combined and/or compared. It focuses on adjusting the individual hybridization intensities in order to balance them appropriately so that meaningful biological comparisons can be made (Quackenbush 2002).
- There are a number of reasons why data must be normalized which include:
  - unequal labeling efficiency,
  - noise of the system and differential expression.
- The decision as to which normalization method is appropriate may depend on the biological nature of the dataset examined. For each microarray technology there is a preferred normalization method (Bolstad, Irizarry et al. 2003, Boes and Neuhauser 2005).
- Typical normalization methods include the rank invariant normalization (Tseng, Oh et al. 2001), quantile (Bolstad, Irizarry et al. 2003), LOWESS/LOESS methods (Tseng, Oh et al. 2001). For many types of commercial arrays, suites of R-BioConductor (Reimers and Carey 2006), based packages are used to do consecutively background adjustment and normalization of data, such as RMA (Robust Multi-Array Average expression measure) (Irizarry, Hobbs et al. 2003) and MAS 5.0 Algorithm (Pepper, Saunders et al. 2007).



Εικόνα 7.4: *Κανονικοποίηση δύο συνόλων τιμών έκφρασης από δύο δείγματα (α) με β) z-κανονικοποίηση που μετατρέπει την κλίμακα σε νέα κλίμακα με κέντρο το 0 γ) κανονικοποίηση ποσοστημορίων που μετατρέπει την κλίμακα σε μια σταθμισμένη κλίμακα με βάση την κατανομή ποσοστημορίων. Τόσο η β) όσο και η γ) διατηρούν τη διασπορά του δείγματος. Η κανονικοποίηση LOESS (δ) αλλάζει τις τιμές στο ένα μόνο δείγμα (εδώ Δείγμα 2) ανάλογα με το πού εφαρμόζεται το μοντέλο. Η πλήρης κανονικοποίηση περιλαμβάνει και την αντίστροφη διαδικασία (κανονικοποίηση του Δείγματος 1 με βάση το 2).*

Κανονικοποίηση

Τρόπος ελαχιστοποίησης των σφαλμάτων στα επίπεδα έκφρασης

- Κανονικοποίηση ολικής έντασης (total intensity normalization)

- Lowess (locally weighted linear regression) κανονικοποίηση

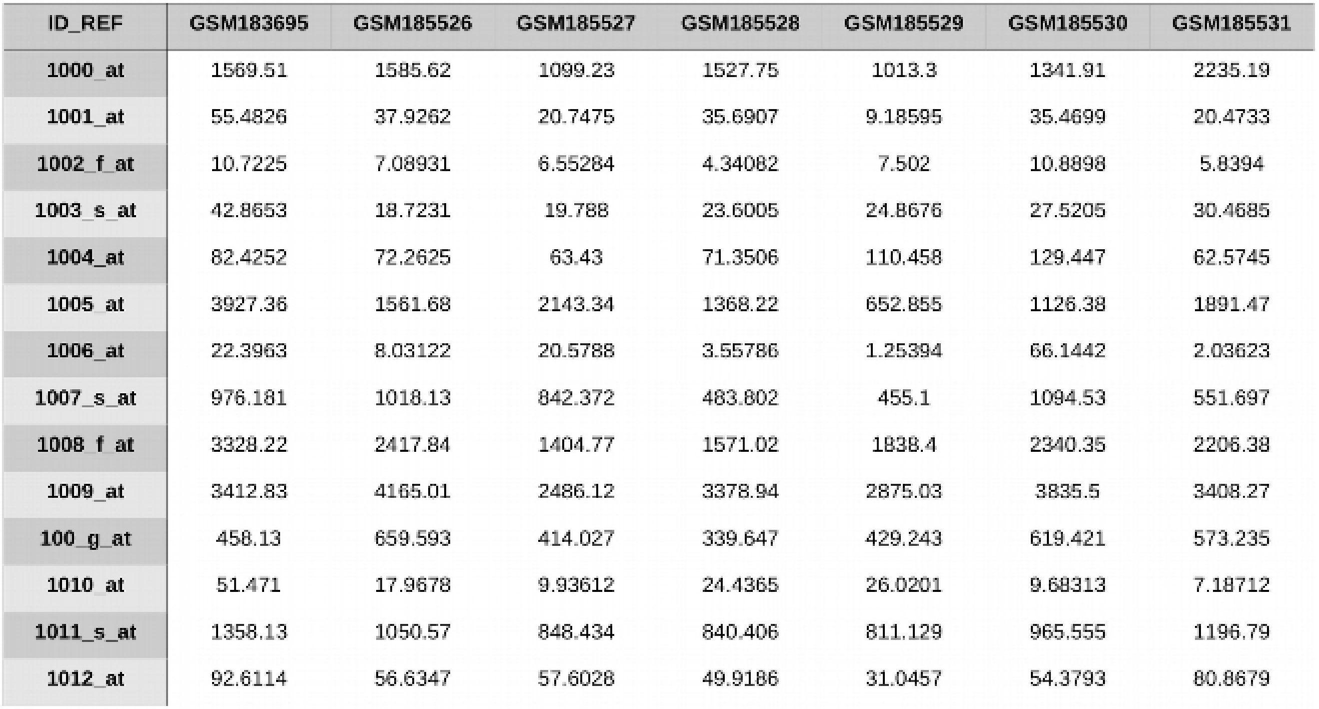
Βάσεις δεδομένων μικροσυστοιχιών

- **GeneExpression Omnibus (GEO):** Βάση δεδομένων του NCBI που παρέχει δεδομένα γονιδιακής έκφρασης
  http://www.ncbi.nlm.nih.gov/geo/

- **Array Express:** Δημόσια βάση δεδομένων μικροσυστοιχιών η οποία διατηρείται στο Ευρωπαϊκό Ινστιτούτο Βιοπληροφορικής EBI **http://www.ebi.ac.uk/arrayexpress/**

- **ONCOMINE:** Βάση δεδομένων που περιέχει πειράματα μικροσυστοιχιών που αφορούν διαφόρους τύπους καρκίνου. Επίσης παρέχει στο χρήστη εργαλεία διαχείρισης των δεδομένων για την αποδοτικότερη εύρεση των επιθυμητών πειραμάτων και γονιδίων **http://www.oncomine.org/**

Δεδομένα μικροσυστοιχιών



ID_REF	GSM183695	GSM185526	GSM185527	GSM185528	GSM185529	GSM185530	GSM185531
1000_at	1569.51	1585.62	1099.23	1527.75	1013.3	1341.91	2235.19
1001_at	55.4826	37.9262	20.7475	35.6907	9.18595	35.4699	20.4733
1002_f_at	10.7225	7.08931	6.55284	4.34082	7.502	10.8898	5.8394
1003_s_at	42.8653	18.7231	19.788	23.6005	24.8676	27.5205	30.4685
1004_at	82.4252	72.2625	63.43	71.3506	110.458	129.447	62.5745
1005_at	3927.36	1561.68	2143.34	1368.22	652.855	1126.38	1891.47
1006_at	22.3963	8.03122	20.5788	3.55786	1.25394	66.1442	2.03623
1007_s_at	976.181	1018.13	842.372	483.802	455.1	1094.53	551.697
1008_f_at	3328.22	2417.84	1404.77	1571.02	1838.4	2340.35	2206.38
1009_at	3412.83	4165.01	2486.12	3378.94	2875.03	3835.5	3408.27
100_g_at	458.13	659.593	414.027	339.647	429.243	619.421	573.235
1010_at	51.471	17.9678	9.93612	24.4365	26.0201	9.68313	7.18712
1011_s_at	1358.13	1050.57	848.434	840.406	811.129	965.555	1196.79
1012_at	92.6114	56.6347	57.6028	49.9186	31.0457	54.3793	80.8679

Εικόνα 7.1: *Οι πρώτες γραμμές του αποτελέσματος ενός πειράματος έκφρασης σε μικροσυστοιχία DNA. Η πρώτη στήλη περιέχει τον κωδικό αριθμό του ανιχνευτή (probe) που μπορεί να αντιστοιχηθεί σε ένα συγκεκριμένο γονίδιο. Οι τιμές που ακολουθούν στις στήλες 2-8 αντιστοιχούν στη μέτρηση φθορισμού για το δεδομένο ανιχνευτή για καθένα από επτά διαφορετικά δείγματα.*

https://repository.kallipos.gr/handle/11419/1585

Ανάλυση Μικροσυστοιχιών

1) Στατιστική ανάλυση για εύρεση γονιδίων που υπέρ ή υποεκφράζονται

2) Ομαδοποίηση (Clustering)

3) Πρόγνωση (Prediction)

Ομαδοποίηση (Clustering)

- Ομαδοποιούνται μαζί γονίδια με βάση τα επίπεδα έκφρασης τους
- Αναπαράσταση των ομάδων αυτών με σκοπό την εύρεση πιθανών σχέσεων μεταξύ των γονιδίων
- Αλγόριθμοι ομαδοποίησης μπορούν να διαχωριστούν σε επιβλεπόμενους (supervised) και μη-επιβλεπόμενους (unsupervised)
- Η απόσταση (distance) μεταξύ δύο γονιδίων χρησιμοποιηται ως είσοδος στους αλγορίθμους ομαδοποίησης:
  - **Ευκλείδεια απόσταση**

    $$d_{AB} = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2}$$

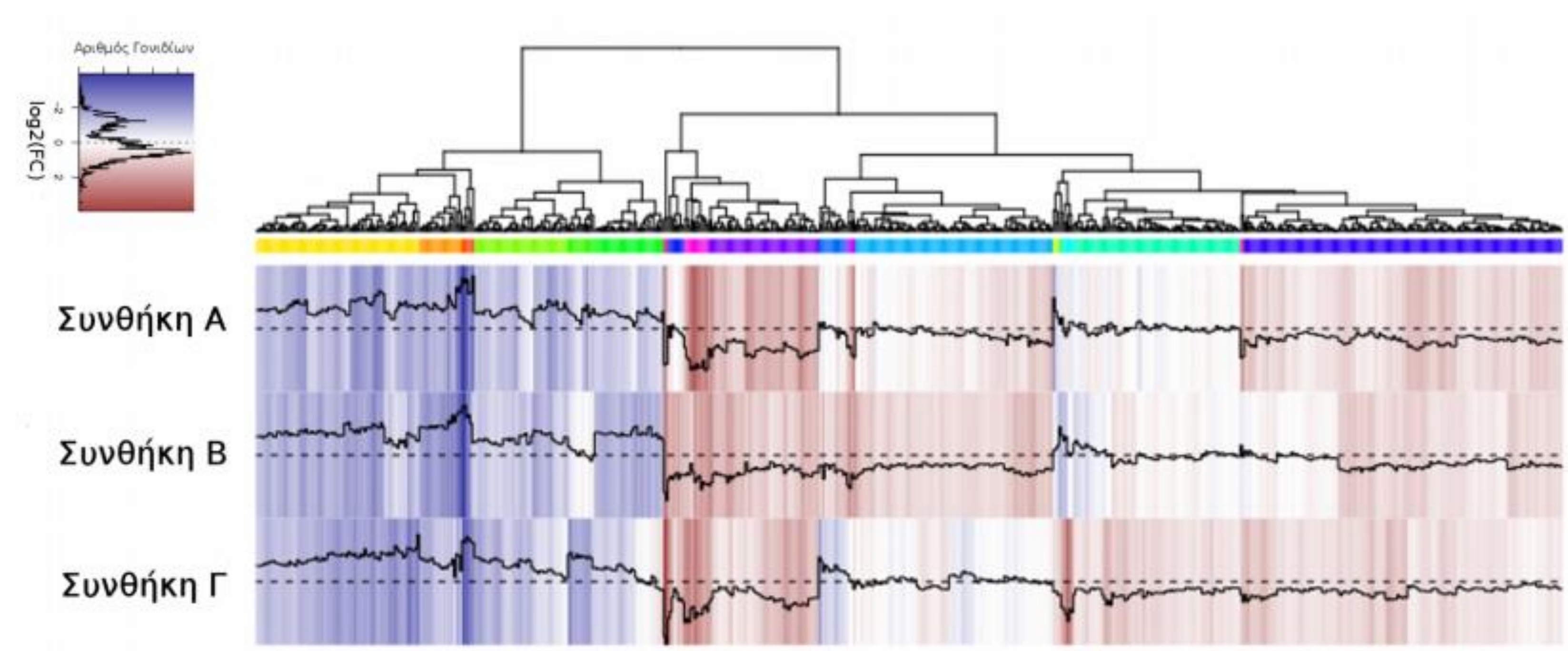
  - **Απόσταση Manhattan**

    $$d_{AB} = \sum_{i=1}^{n} |x_i - y_i|$$

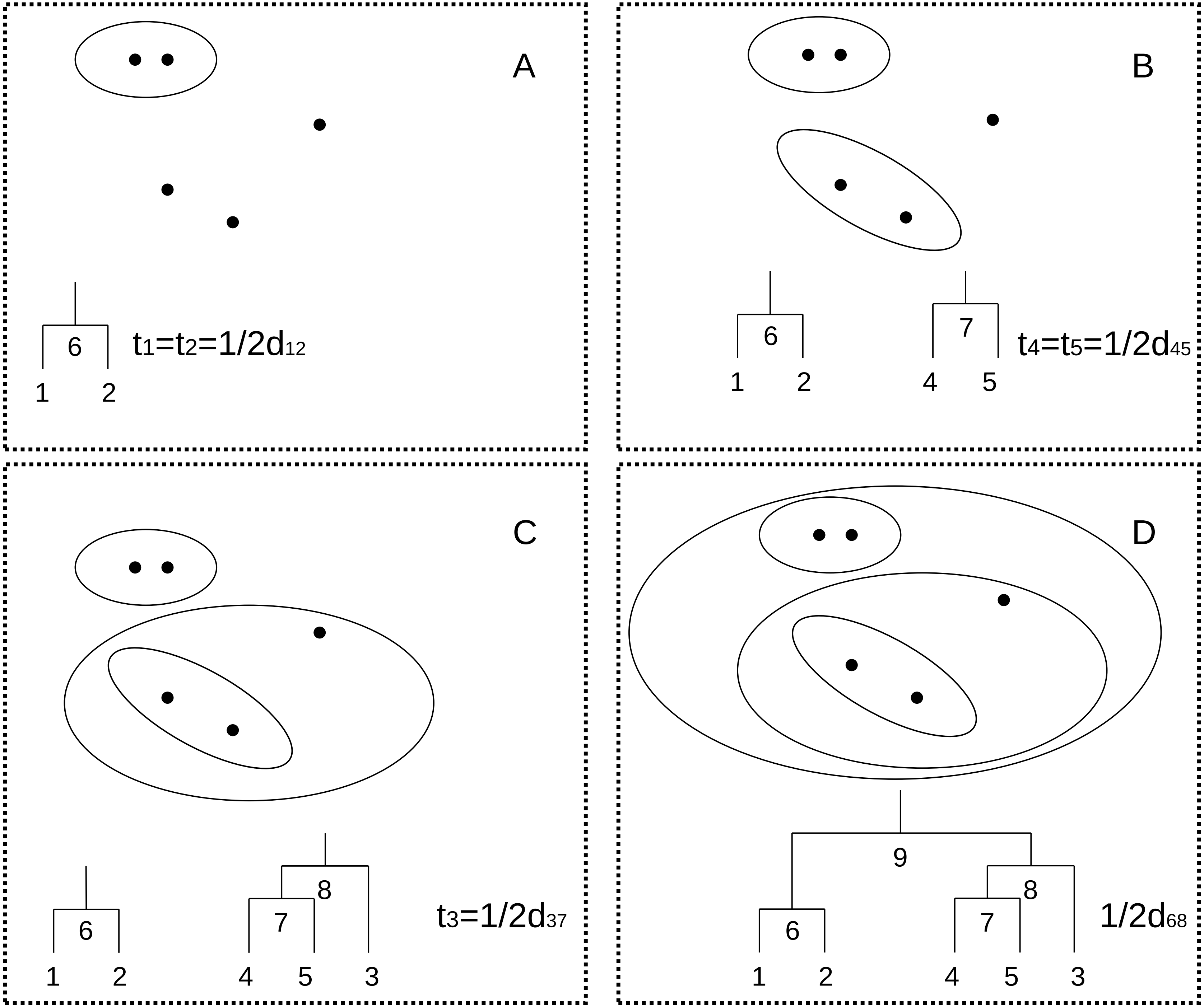
  - **Συντελεστής Συσχέτισης του Pearson**

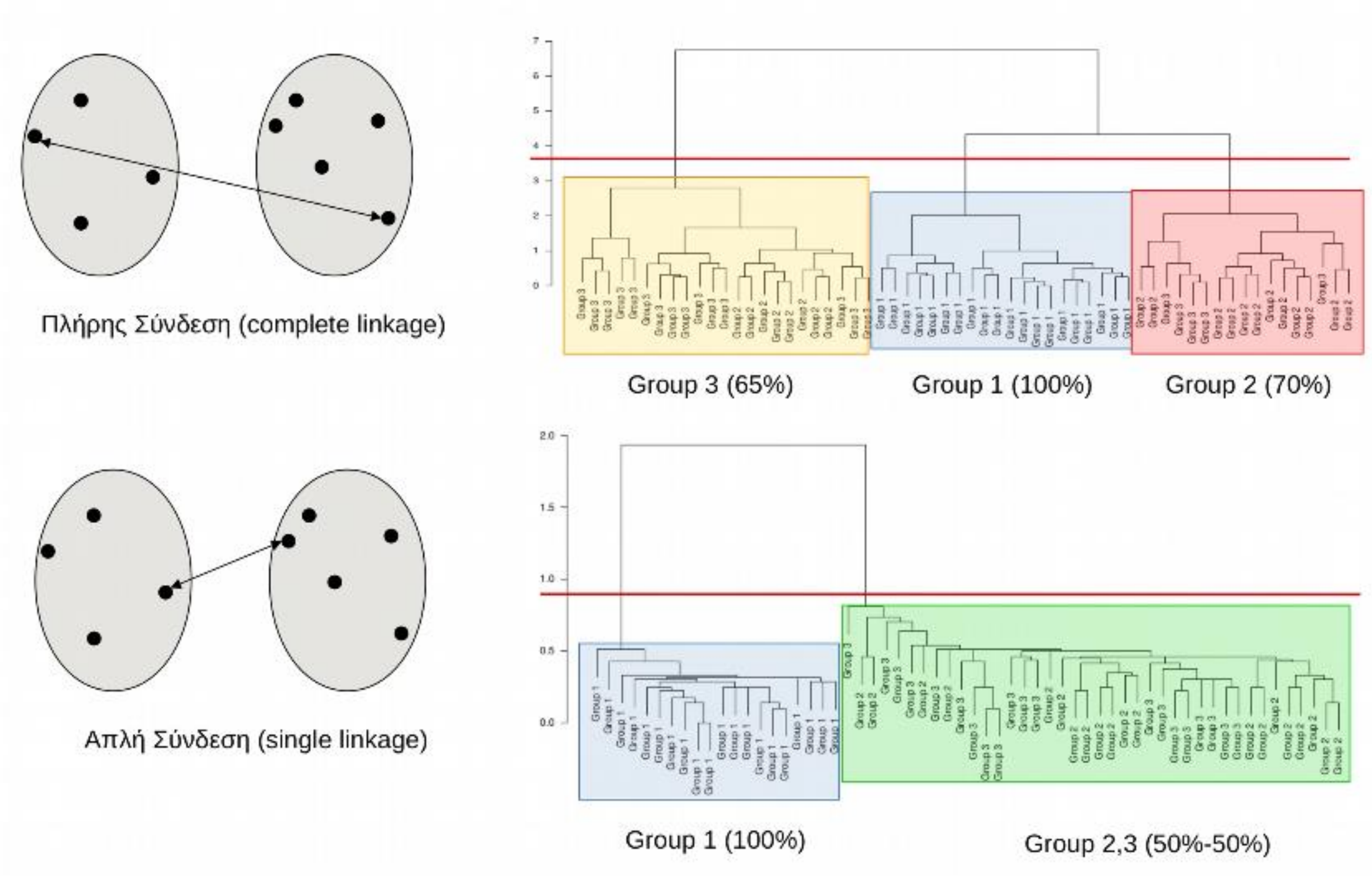
    $$r = \frac{\sum_{i=1}^{n} (x_i - \overline{X})(y_i - \overline{Y})}{\sqrt{\sum_{i=1}^{n} (x_i - \overline{X})^2} \sqrt{\sum_{i=1}^{n} (y_i - \overline{Y})^2}}$$

- **Clustering** analysis tries to group genes or individuals according to their expression levels and leads to a representation that can be helpful for identifying patterns in time and space. Clustering operates in an unsupervised manner, since in such analyses all individuals (usually the patients are treated equally) and the clustering method result in some classification that can be of interest. Some of the methods require that the number of clusters should be defined beforehand, whereas in others, the number of clusters is automatically defined. Several clustering methods exist, the most commonly used for microarray analysis are the Unweighted Pair Group Method with Arithmetic Mean (UPGMA) and hierarchical clustering for tree based representations. Evolutionary tree based algorithms such as Neighbor Joining could also be applied. In the k-Means algorithm the number of clusters should be pre-defined and is also widely used in microarray experiments. One of the preferred clustering algorithms is the self-organizing map (SOM) which is another technique that is particularly well suited for exploratory data analysis. The self-organizing map (SOM) (Tamayo, Slonim et al. 1999) is a method for producing ordered low-dimensional representations of an input data space. Typically such input data is complex and high dimensional with data elements being related to each other in a nonlinear fashion. Most of the aforementioned implementations can be found in BioConductor (Reimers and Carey 2006), Expander (Shamir, Maron-Katz et al. 2005) and Hierarchical Clustering Explorer (HCE).



Εικόνα 7.6: *Θερμικός χάρτης που αναπαριστά τις σχετικές τιμές έκφρασης 650 γονιδίων όπως αυτές μετρήθηκαν σε τρεις διαφορετικές συνθήκες (Α, Β και Γ). Το γαλάζιο αντιστοιχεί σε χαμηλότερη και το κόκκινο σε υψηλότερη έκφραση σε σχέση με την κατάσταση ελέγχου, καθώς στο θερμικό χάρτη εμφανίζονται μόνο σχετικές τιμές έκφρασης. Ο χάρτης συνοδεύεται από ιεραρχική ομαδοποίηση (βλ. Παρακάτω) των γονιδίων με βάση τα πρότυπα έκφρασής τους στις τρεις συνθήκες. Γονίδια που βρίσκονται στον ίδιο κλάδο του δέντρου εμφανίζουν μεγαλύτερη ομοιότητα σε ό,τι αφορά την αυξομείωση των επιπέδων έκφρασης μεταξύ των συνθηκών.*


A
6
1
2
$t_1=t_2=1/2d_{12}$
B
6
1
2
7
4
5
$t_4=t_5=1/2d_{45}$
C
6
1
2
8
7
4
5
3
$t_3=1/2d_{37}$
D
9
6
1
2
8
7
4
5
3
$1/2d_{68}$



Εικόνα 7.10: *Ιεραρχική ομαδοποίηση για 60 από τα 150 γονίδια που αναλύθηκαν με PCA στην προηγούμενη ενότητα με 20 γονίδια να ανήκουν στο καθένα από τα 3 υποσύνολα. Επάνω: Υπολογισμός των αποστάσεων με πλήρη σύνδεση αποδίδει τρεις ομάδες με πολύ καλή συμφωνία με την (εκ των προτέρων γνωστή) αρχική ομαδοποίηση. Κάτω: Υπολογισμός των αποστάσεων με απλή σύνδεση οδηγεί στο σχηματισμό δύο ομάδων χωρίς να μπορεί να διακρίνει μεταξύ των Ομάδων 2 και 3.*

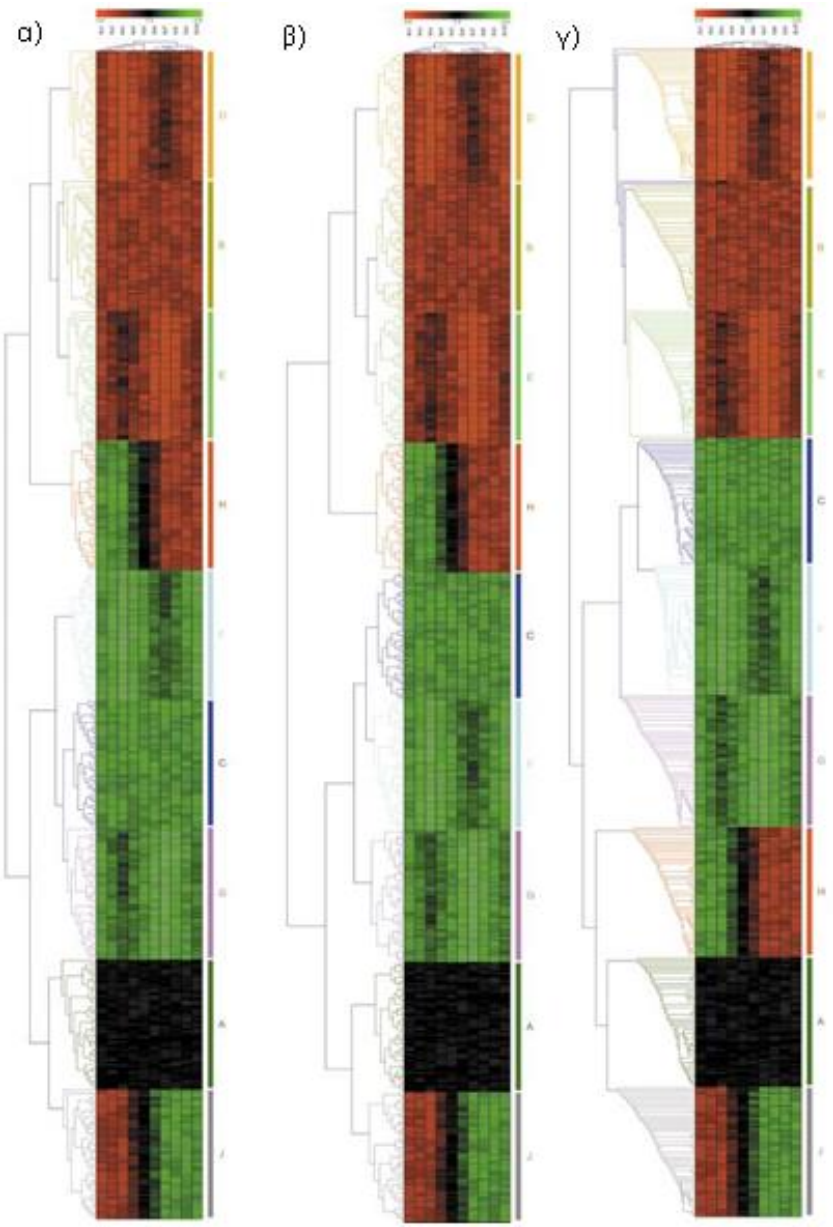
Αλγόριθμοι Ομαδοποίησης

Ιεραρχική ταξινόμηση:

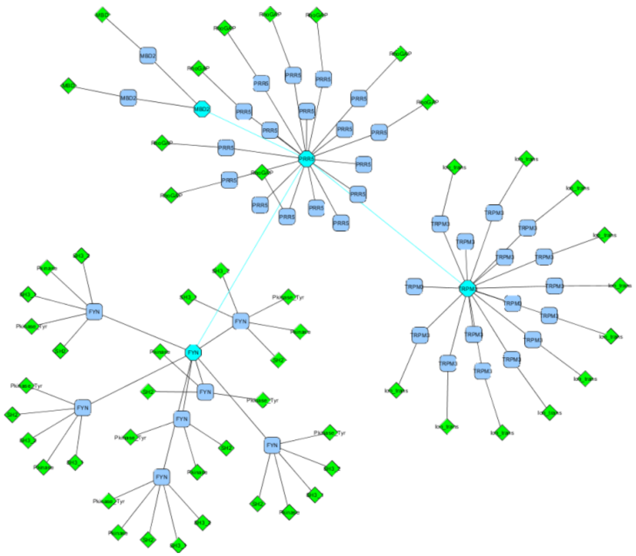
α)Single Linkage Clustering

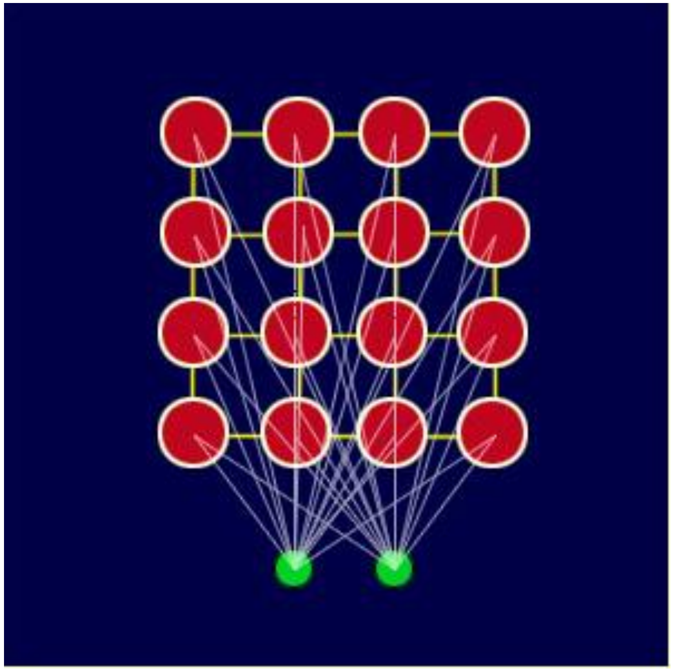
β) Complete Linkage Clustering

γ) Average Linkage Clustering

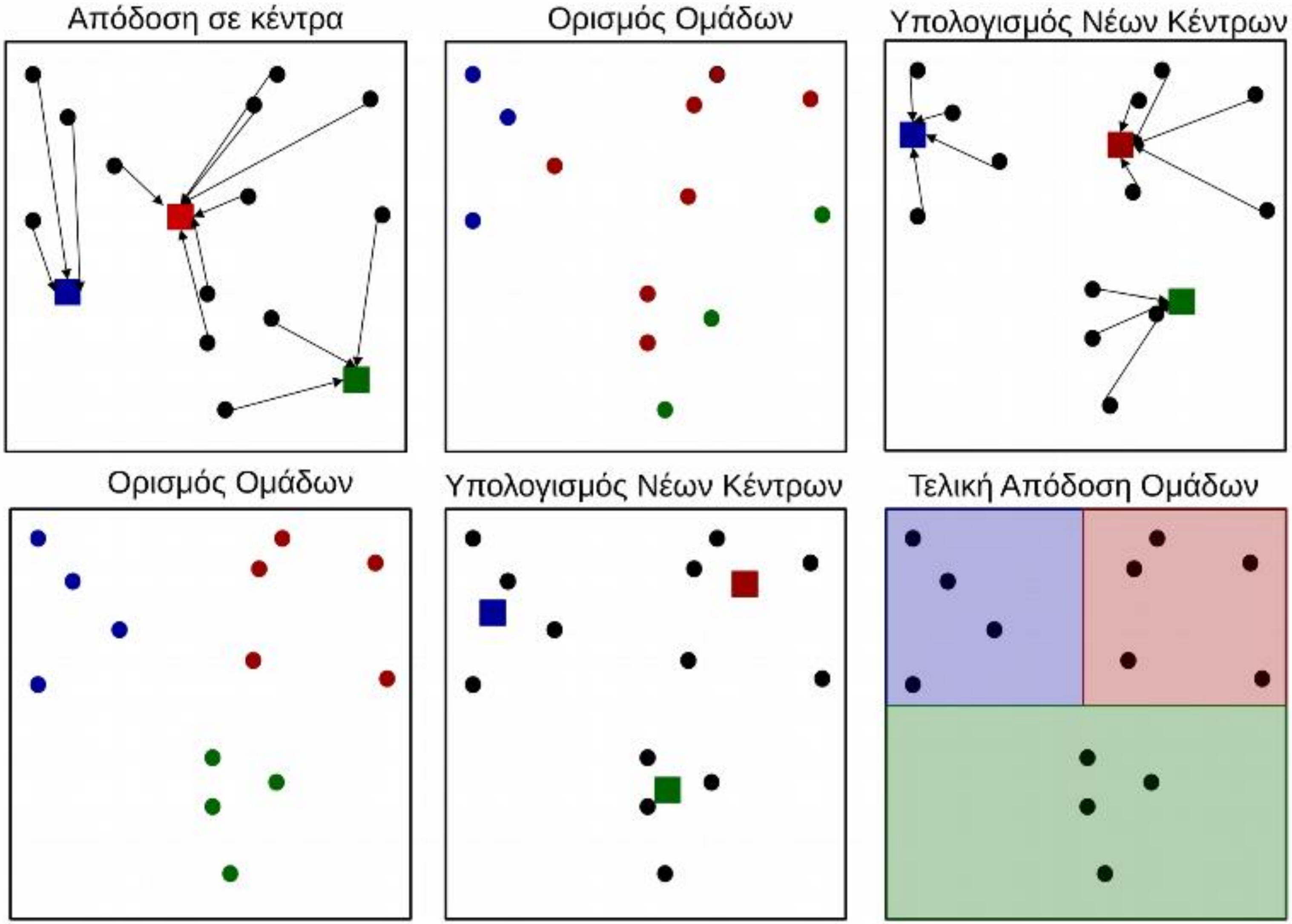


Αλγόριθμοι Ομαδοποίησης





- K-means
- SOMs
- SVM
- PCA
- MCL



Εικόνα 7.11: *Σχηματική αναπαράσταση του αλγορίθμου της ομαδοποίησης k-μέσων.*

Πρόγνωση

- Ενδιαφερόμαστε κυρίως για τη σωστή πρόγνωση (ταξινόμηση) των ασθενών.
- Έχει σημασία σε περιπτώσεις πρόβλεψης της ασθένειας, σαν διαγνωστική δοκιμασία
- Χρησιμοποιούνται οι συνηθισμένες μέθοδοι ταξινόμησης (Νευρωνικά Δίκτυα, SVM, κλπ)
- Πολλές φορές απαιτείται κάποια μέθοδος επιλογής των πιο σημαντικών γονιδίων

- **Classification** refers to class prediction from gene expression patterns. In such a case, we have predefined classes (two or more), for instance healthy individuals vs. diseased ones, and we want to build a classifier that will be able to discriminate them in future applications (Golub, Slonim et al. 1999, Radmacher, McShane et al. 2002), most notably, for screening and diagnostic purposes (Simon, Radmacher et al. 2003). A wide variety of supervised methods taken from the arsenal of machine learning and artificial intelligence have been used for this purpose, including Neural Networks (Khan, Wei et al. 2001), , Support Vector Machines (Furey, Cristianini et al. 2000), Graphical Models (Bura and Pfeiffer 2003), genetic algorithms (Ooi and Tan 2003), nearest neighbour classifiers and many other statistical methods, including shrunken centroids (Tibshirani, Hastie et al. 2002) and Partial Least Squares and Discriminant analysis (Nguyen and Rocke 2002).

feature selection

- Due to the large number of features (genes) given as input to the various classifiers, a subsequent problem is to select the best subset of features that can be used efficiently by the classifier. This problem is known as the **feature selection** problem in machine learning (Guyon and Elisseeff 2003). In addition to the large number of techniques that have already been developed in the machine learning and data mining fields, the advent of microarrays have led to a wealth of newly proposed techniques. Comparison of such methods in gene expression classification can be found in several excellent reviews and evaluation studies (Li, Zhang et al. 2004, Saeys, Inza et al. 2007, Ma and Huang 2008)

Identification of differentially expressed genes

- **Identification of differentially expressed genes,** finally, is the most obvious approach in order to assign biological functions to genes, in cases where there are two or more classes in which individuals can be classified in advance, for example when normal and diseased tissues are compared or the gene expression is studied with respect to a particular treatment. The main aim is to identify which genes are pinpointed by their differential expression levels and see which of them is up-, or down-regulated. Ideally, the identification of DEGs is a simple procedure reduced to a statistical test for the equality of means (e.g. t-test, see below). However, statistically microarrays datasets are characterised by several distinctive features such small number of samples (individuals), large number of variables and large amount of noise, and thus several advanced statistical methods have been proposed in order to overcome these. Moreover, the accumulation of similar datasets from various laboratories has lead to the need of combining these datasets in order to increase the sample size. This approach, which is termed meta-analysis in the medical literature, has been increasingly popular during the last years and several methods exist.

t-test

One sample t-test

$$\bar{X}_1 - \bar{X}_2 = \bar{X}_D \qquad (1)$$

$$t = \frac{\bar{X}_D}{S_D / \sqrt{n}} \qquad (2)$$

$$n = \frac{n_1 n_2}{n_1 + n_2} \qquad (3)$$

Two sample t-test with equal variances

$$t = \frac{\bar{X}_1 - \bar{X}_2}{S_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \qquad (4)$$

$$S_p = \sqrt{\frac{(n_1 - 1) S_1^2 + (n_2 - 1) S_2^2}{n_1 + n_2 - 2}} \qquad (5)$$

Two sample t-test with unequal variances

$$t = \frac{\bar{X}_1 - \bar{X}_2}{S_{\bar{X}_1 - \bar{X}_2}} \qquad (6)$$

$$S_{\bar{X}_1 - \bar{X}_2} = \sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}} \qquad (7)$$

$$d.f. = \frac{\left(S_1^2/n_1 + S_2^2/n_2\right)^2}{\left(S_1^2/n_1\right)^2 \Big/ (n_1 - 1) + \left(S_2^2/n_2\right)^2 \Big/ (n_2 - 1)} \qquad (8)$$

Προβλήματα

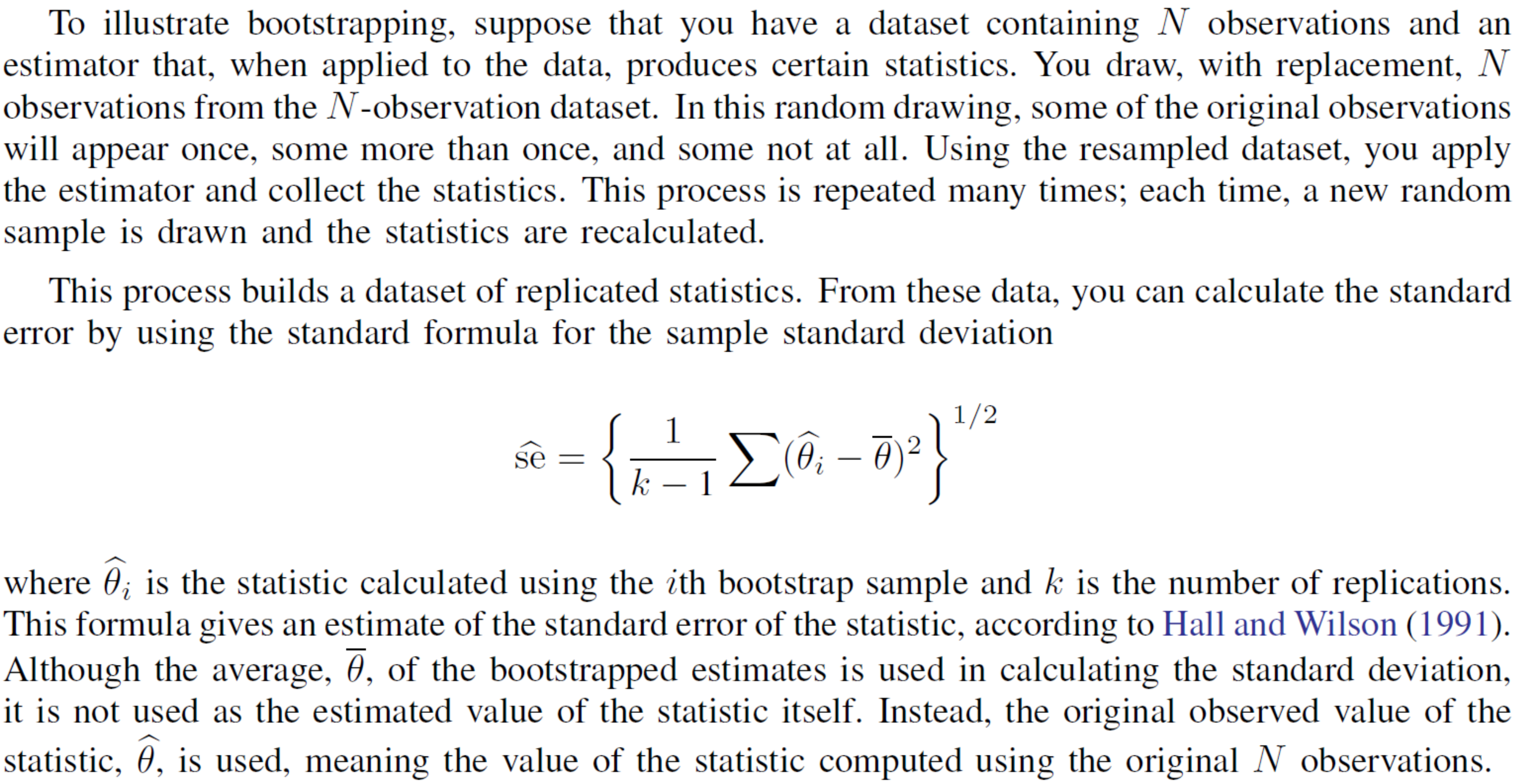
- Το βασικό πρόβλημα με το t-test είναι οτι απαιτεί σχετικά “μεγάλο” μέγεθος δείγματος
- Στα περισσότερα πειράματα μικροσυστοιχιών, έχουμε δείγμα μικρότερο των 20 ατόμων, και καμιά φορά μικρότερο των 10
- Έτσι, οι προϋποθέσεις για την κανονικότητα του πληθυσμού δεν ισχύουν
- Πολλές φορές, ειδικά όταν το δείγμα είναι <5, μπορεί να έχουμε και “περίεργα” μικρή διασπορά που θα μας δημιουργήσει πρόβλημα
- Τέλος, ο μεγάλος αριθμός γονιδίων, μας οδηγεί στο πρόβλημα των πολλαπλών συγκρίσεων

Computationally Intensive methods

- Resampling methods
- Bayesian t-test
- Empirical Bayesian

bootstrap

- The **Bootstrap** (Efron 1982, Efron and Tibshirani 1993) is a statistical method for estimating the sampling distribution of an estimator by sampling with replacement from the original sample. The Bootstrap is an ideal method when no formula the sampling distribution is available or when available formulas make inappropriate assumptions (e.g. small sample size, non-normal distribution).
- The logic behind the bootstrap is that all measures of precision come from a statistic's sampling distribution. When the statistic is estimated on a sample of size n from some population, the sampling distribution tells you the relative frequencies of the values of the statistic. The sampling distribution, in turn, is determined by the distribution of the population and the formula used to estimate the statistic. The accuracy of the bootstrap depends on the number of observations in the original sample and the number of replications.
- A crudely estimated sampling distribution is adequate if you are only going to calculate, for instance, a standard error. A better estimate is needed if you want to construct a 95% confidence interval (and we need to emphasize that there are various methods for constructing a Bootstrap confidence interval from the resampled statistics – the normal approximation method, the bias corrected method, the percentile method and the t-percentile method - see (Efron 1987)).
- Generally, replications on the order of 1,000 produce very good estimates, more may be needed for accurate estimation of p-values, but only 50–200 replications are needed for estimating standard errors (this may have implications for meta-analysis, see below). Various methods have been proposed for estimating the necessary number of replications (Andrews and Buchinsky 2000, Davidson and MacKinnon 2000).
- The Bootstrap has been applied in microarray experiments and empirical evidence suggests that it has good properties, at least for moderate sample sizes (Meuwissen and Goddard 2004). For really small sample sizes (i.e. <10), various modifications to the standard method have been proposed (Neuhauser and Jockel 2006, Jiang and Simon 2007).



To illustrate bootstrapping, suppose that you have a dataset containing N observations and an estimator that, when applied to the data, produces certain statistics. You draw, with replacement, N observations from the N-observation dataset. In this random drawing, some of the original observations will appear once, some more than once, and some not at all. Using the resampled dataset, you apply the estimator and collect the statistics. This process is repeated many times; each time, a new random sample is drawn and the statistics are recalculated.

This process builds a dataset of replicated statistics. From these data, you can calculate the standard error by using the standard formula for the sample standard deviation

$$\widehat{\text{se}} = \left\{ \frac{1}{k-1} \sum (\widehat{\theta}_i - \overline{\theta})^2 \right\}^{1/2}$$

where $\widehat{\theta}_i$ is the statistic calculated using the ith bootstrap sample and k is the number of replications. This formula gives an estimate of the standard error of the statistic, according to Hall and Wilson (1991). Although the average, $\overline{\theta}$, of the bootstrapped estimates is used in calculating the standard deviation, it is not used as the estimated value of the statistic itself. Instead, the original observed value of the statistic, $\widehat{\theta}$, is used, meaning the value of the statistic computed using the original N observations.

http://www.stata.com/manuals13/rbootstrap.pdf

When the `mse` option is specified, the standard error is estimated as

$$\widehat{\mathrm{se}}_{\mathrm{MSE}} = \left\{ \frac{1}{k} \sum_{i=1}^{k} (\widehat{\theta}_i - \widehat{\theta})^2 \right\}^{1/2}$$

Otherwise, the standard error is estimated as

$$\widehat{\mathrm{se}} = \left\{ \frac{1}{k-1} \sum_{i=1}^{k} (\widehat{\theta}_i - \overline{\theta})^2 \right\}^{1/2}$$

where

$$\overline{\theta} = \frac{1}{k} \sum_{i=1}^{k} \widehat{\theta}_i$$

The variance–covariance matrix is similarly computed. The bias is estimated as

$$\widehat{\mathrm{bias}} = \overline{\theta} - \widehat{\theta}$$

The percentile method yields the confidence intervals

$$\left[\theta^*_{\alpha/2},\, \theta^*_{1-\alpha/2}\right]$$

where θ^*_p is the pth quantile (the 100pth percentile) of the bootstrap distribution $(\widehat{\theta}_1, \dots, \widehat{\theta}_k)$.

Let

$$z_0 = \Phi^{-1}\{\#(\widehat{\theta}_i \leq \widehat{\theta})/k\}$$

where $\#(\widehat{\theta}_i \leq \widehat{\theta})$ is the number of elements of the bootstrap distribution that are less than or equal to the observed statistic and Φ is the standard cumulative normal. z_0 is known as the median bias of $\widehat{\theta}$. When the `ties` option is specified, z_0 is estimated as $\#(\widehat{\theta}_i < \widehat{\theta}) + \#(\widehat{\theta}_i = \widehat{\theta})/2$, which is the number of elements of the bootstrap distribution that are less than the observed statistic plus half the number of elements that are equal to the observed statistic.

Let

$$a = \frac{\sum_{i=1}^{n}(\overline{\theta}_{(\cdot)} - \widehat{\theta}_{(i)})^3}{6\left\{\sum_{i=1}^{n}(\overline{\theta}_{(\cdot)} - \widehat{\theta}_{(i)})^2\right\}^{3/2}}$$

where $\widehat{\theta}_{(i)}$ are the leave-one-out (jackknife) estimates of $\widehat{\theta}$ and $\overline{\theta}_{(\cdot)}$ is their mean. This expression is known as the jackknife estimate of acceleration for $\widehat{\theta}$. Let

$$p_1 = \Phi\left\{z_0 + \frac{z_0 - z_{1-\alpha/2}}{1 - a(z_0 - z_{1-\alpha/2})}\right\}$$

$$p_2 = \Phi\left\{z_0 + \frac{z_0 + z_{1-\alpha/2}}{1 - a(z_0 + z_{1-\alpha/2})}\right\}$$

where $z_{1-\alpha/2}$ is the $(1-\alpha/2)$th quantile of the normal distribution. The bias-corrected and accelerated (BC$_a$) method yields confidence intervals

$$\left[\theta^*_{p_1},\, \theta^*_{p_2}\right]$$

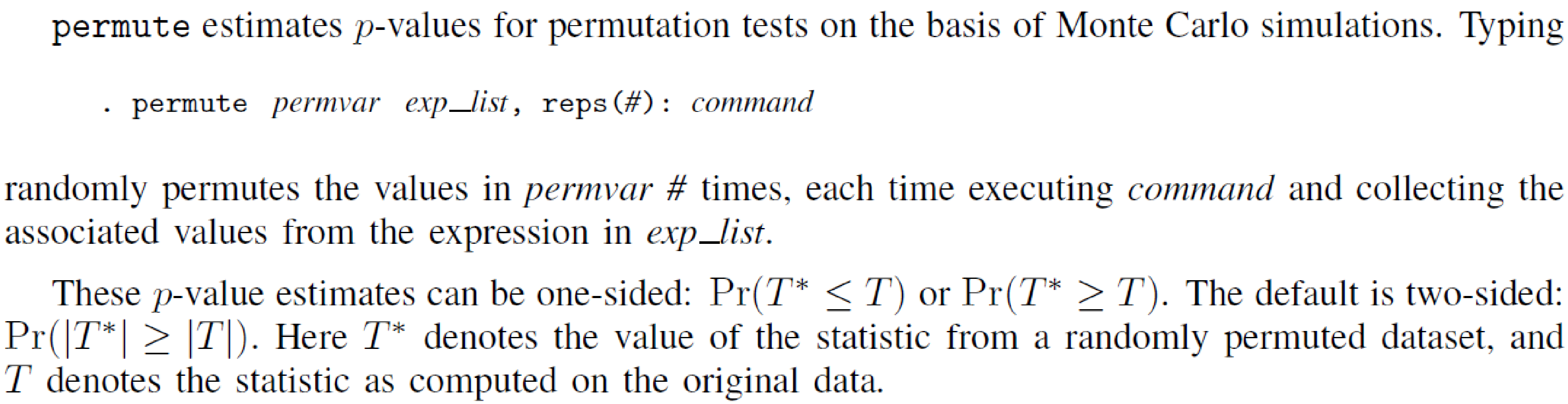
where θ^*_p is the pth quantile of the bootstrap distribution as defined previously. The bias-corrected (but not accelerated) method is a special case of BC$_a$ with $a = 0$.

permutation

- A conceptually different resampling method is the **permutation** test. This is a type of statistical significance test in which the distribution of the test statistic under the null hypothesis is obtained by calculating all possible values of the test statistic following rearrangements of the labels on the observations. If the labels are exchangeable under the null hypothesis, then the resulting tests yield exact significance levels. Confidence intervals can then be derived from the tests.
- The theory has evolved from the works of Ronald Fisher and E. J. G. Pitman in the 1930s (Kaiser 2007). For small samples, all possible permutations can be evaluated, but for sample sizes >15 this is prohibitive. Thus, a random sample of the permutation is used instead, hence the name Monte Carlo permutation.
- An important assumption behind a permutation test is that, under the null hypothesis, the observations are exchangeable. Thus, a consequence of this is that tests of difference in location (like the t-test) require equal variance. In this respect, the permutation t-test shares the same weakness as the classical Student's t-test (the Behrens–Fisher problem).
- Generally, since the permutation computes a p-value by counting the times that the statistic is larger than the observed one, a large number of replications are required (typically of the order of 1,000 or more). Permutation tests have been used for analysis of microarray data (Tsai, Chen et al. 2003). However, when sample sizes are very small, the number of distinct permutations can be severely limited, and pooling the permutation-derived test statistics across all genes has been proposed. However, since the null distribution of the test statistics under permutation is not the same for all genes, this can have a negative impact on both p-value estimation (Yang and Churchill 2007).

Software

- Bootstrap and permutation methods are readily available in major statistical packages like Stata and R. Bootstrap is available with various options using the bootstrap command in Stata and the boot command in R. Permutation can be performed with the permute and permtest (for paired observations) commands in Stata, as well as with the perm command in R. In the Appendix we give examples of performing bootstrap and permutation t-test is Stata.



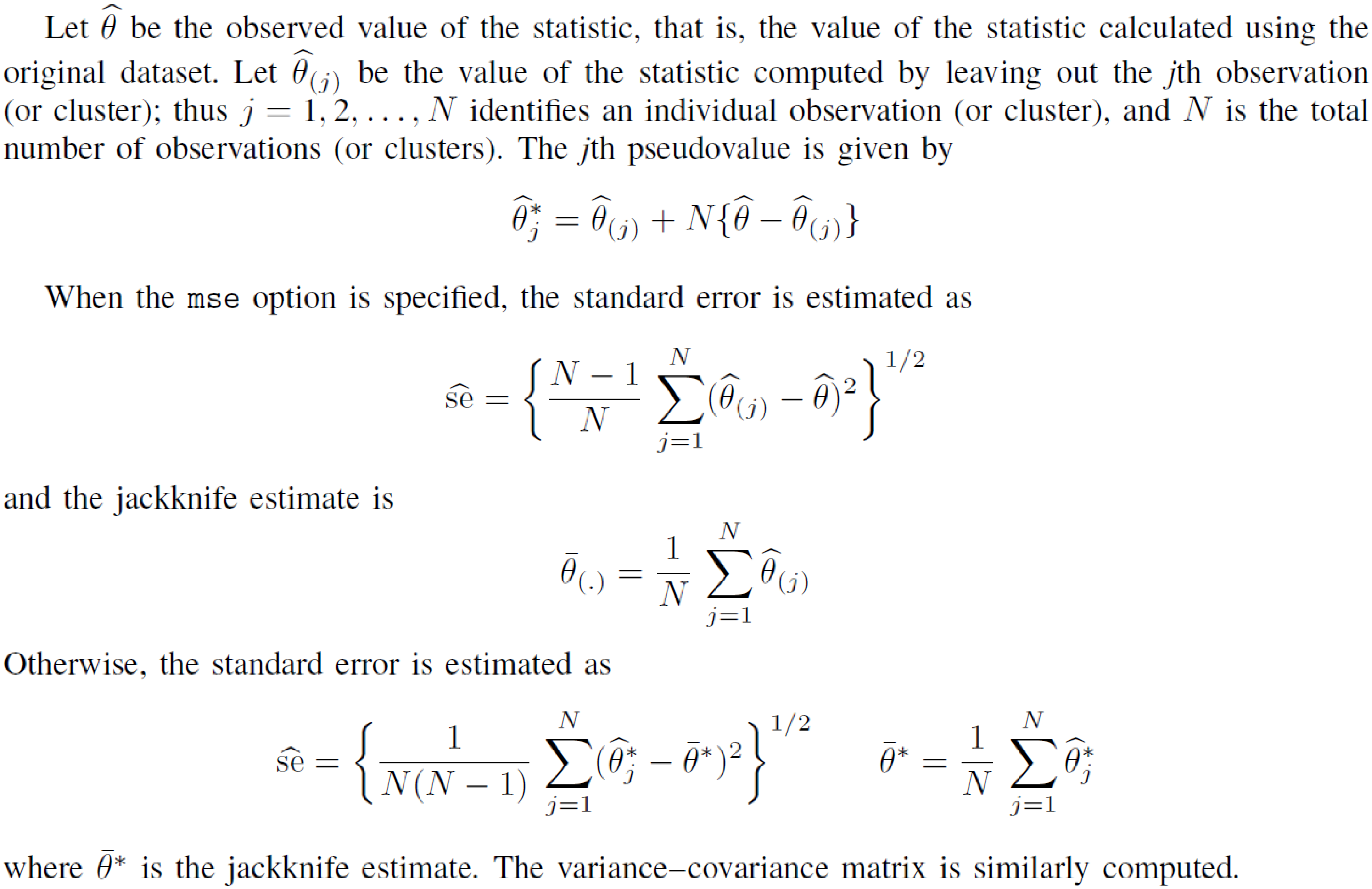
`permute` estimates p-values for permutation tests on the basis of Monte Carlo simulations. Typing

```
. permute permvar exp_list, reps(#): command
```

randomly permutes the values in *permvar* # times, each time executing *command* and collecting the associated values from the expression in *exp_list*.

These p-value estimates can be one-sided: $\Pr(T^* \leq T)$ or $\Pr(T^* \geq T)$. The default is two-sided: $\Pr(|T^*| \geq |T|)$. Here T^* denotes the value of the statistic from a randomly permuted dataset, and T denotes the statistic as computed on the original data.

http://www.stata.com/manuals13/rpermute.pdf



Let $\widehat{\theta}$ be the observed value of the statistic, that is, the value of the statistic calculated using the original dataset. Let $\widehat{\theta}_{(j)}$ be the value of the statistic computed by leaving out the jth observation (or cluster); thus $j = 1, 2, \ldots, N$ identifies an individual observation (or cluster), and N is the total number of observations (or clusters). The jth pseudovalue is given by

$$\widehat{\theta}_j^* = \widehat{\theta}_{(j)} + N\{\widehat{\theta} - \widehat{\theta}_{(j)}\}$$

When the `mse` option is specified, the standard error is estimated as

$$\widehat{\text{se}} = \left\{\frac{N-1}{N}\sum_{j=1}^{N}(\widehat{\theta}_{(j)} - \widehat{\theta})^2\right\}^{1/2}$$

and the jackknife estimate is

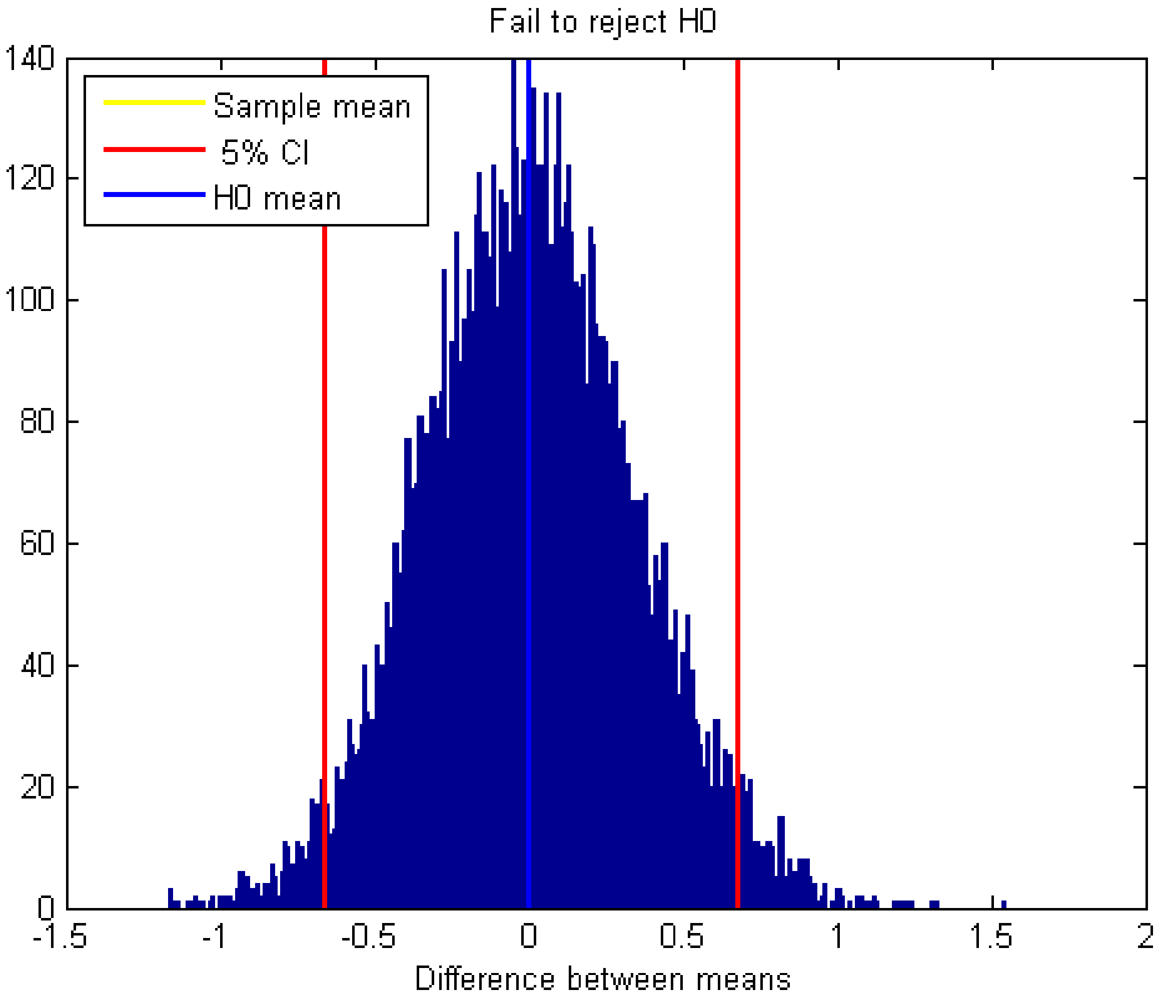
$$\bar{\theta}_{(.)} = \frac{1}{N}\sum_{j=1}^{N}\widehat{\theta}_{(j)}$$

Otherwise, the standard error is estimated as

$$\widehat{\text{se}} = \left\{\frac{1}{N(N-1)}\sum_{j=1}^{N}(\widehat{\theta}_j^* - \bar{\theta}^*)^2\right\}^{1/2} \qquad \bar{\theta}^* = \frac{1}{N}\sum_{j=1}^{N}\widehat{\theta}_j^*$$

where $\bar{\theta}^*$ is the jackknife estimate. The variance–covariance matrix is similarly computed.

http://www.stata.com/manuals13/rjackknife.pdf


Fail to reject H0
Sample mean
5% CI
H0 mean
140
120
100
80
60
40
20
0
-1.5
-1
-0.5
0
0.5
1
1.5
2
Difference between means

Εφαρμογή t-test σε παρατηρήσεις δυο δειγμάτων

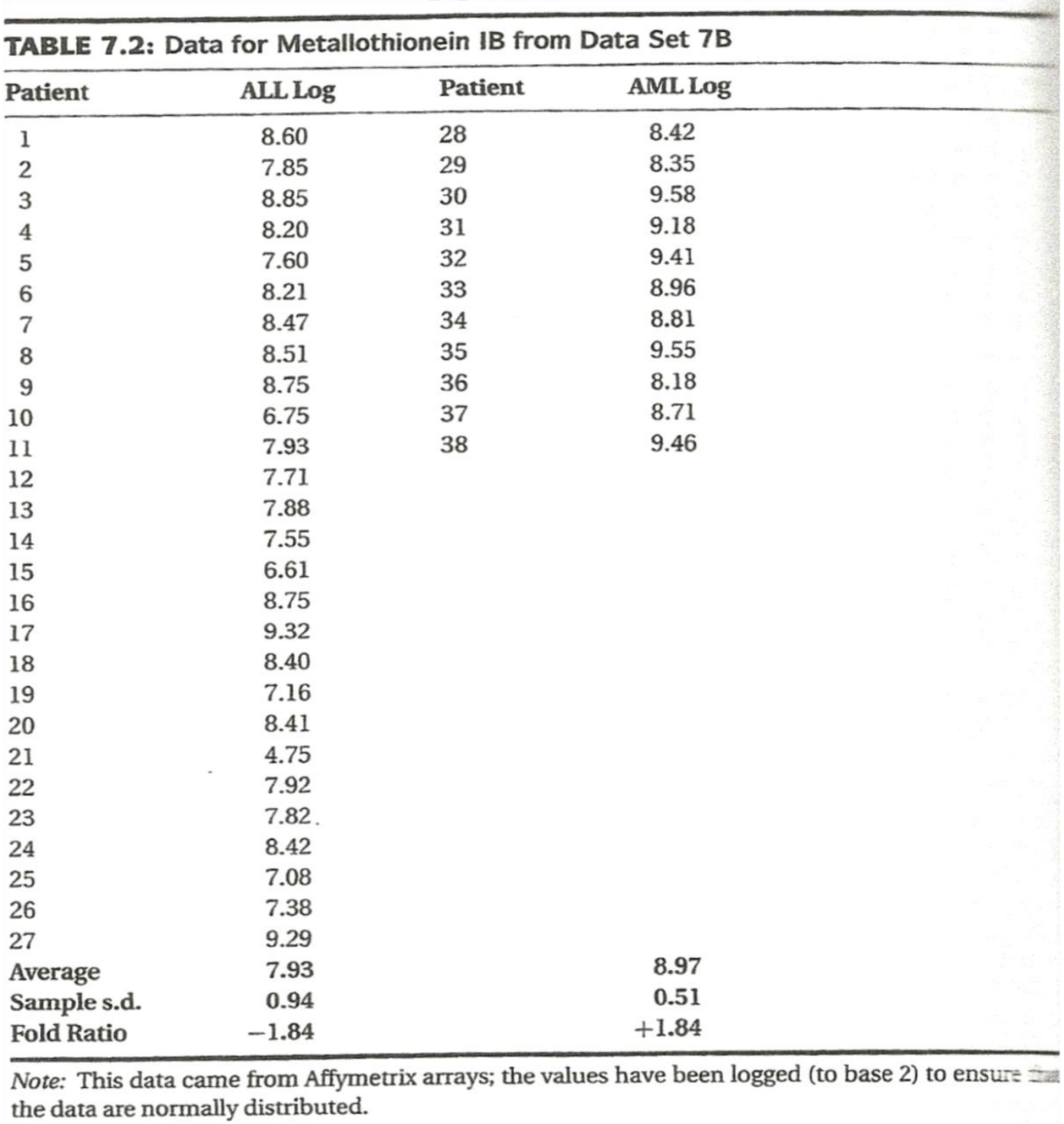
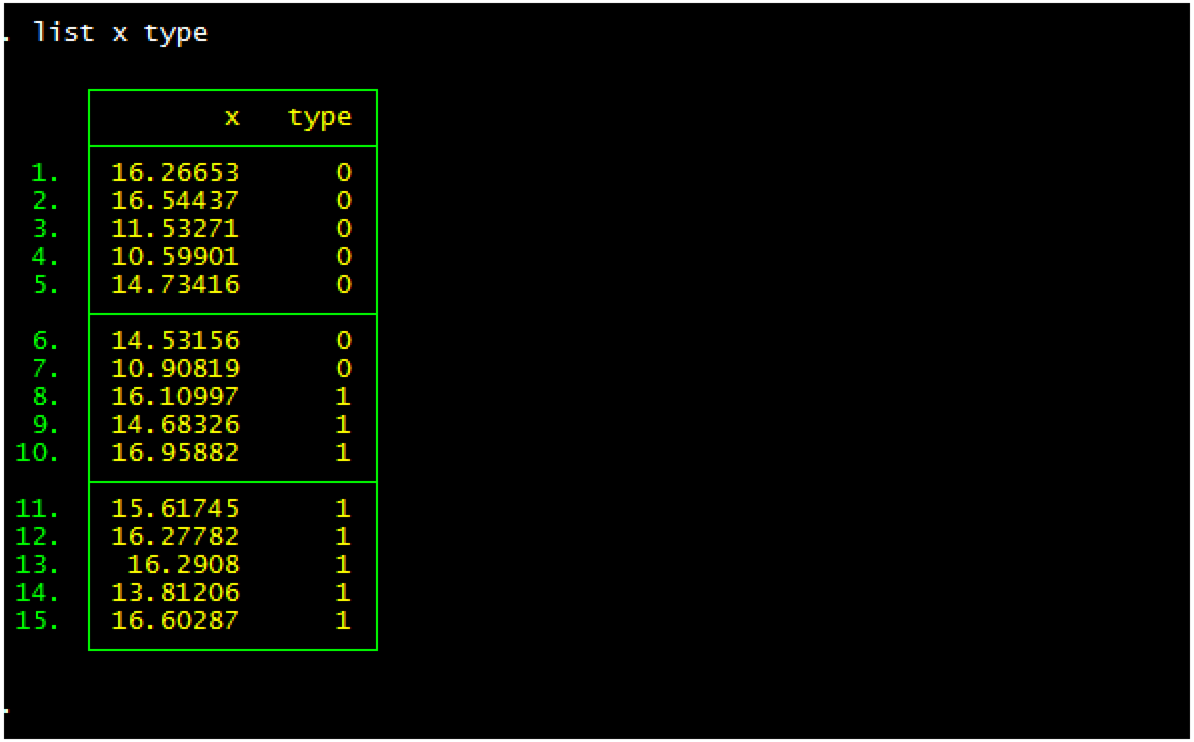


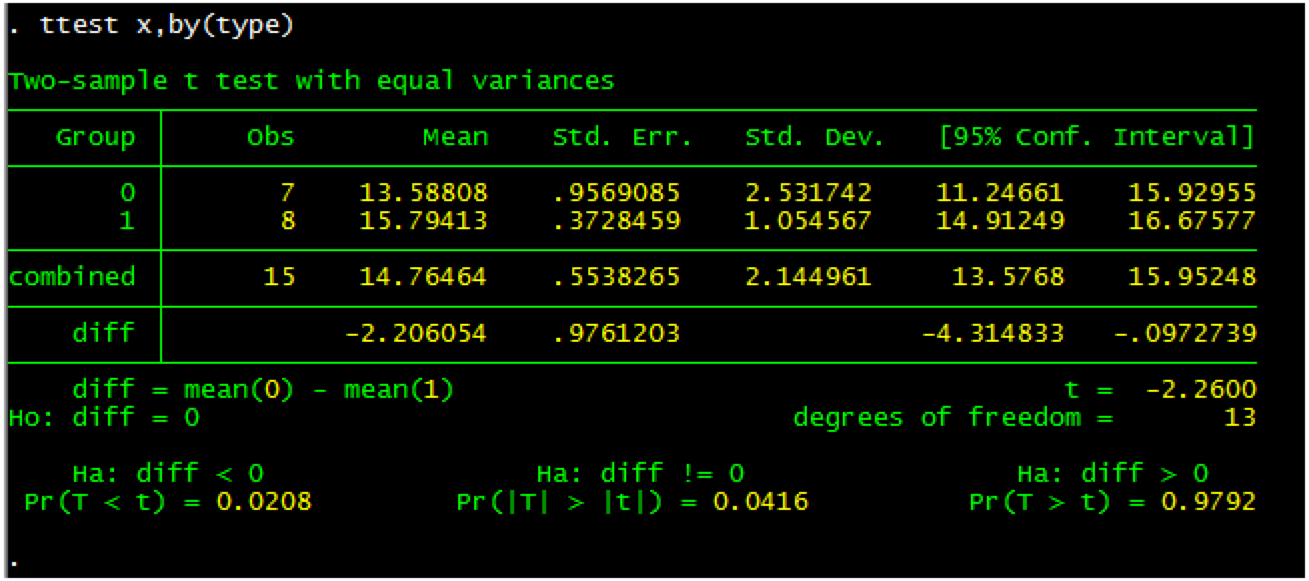
TABLE 7.2: Data for Metallothionein IB from Data Set 7B

Patient	ALL Log	Patient	AML Log
1	8.60	28	8.42
2	7.85	29	8.35
3	8.85	30	9.58
4	8.20	31	9.18
5	7.60	32	9.41
6	8.21	33	8.96
7	8.47	34	8.81
8	8.51	35	9.55
9	8.75	36	8.18
10	6.75	37	8.71
11	7.93	38	9.46
12	7.71		
13	7.88		
14	7.55		
15	6.61		
16	8.75		
17	9.32		
18	8.40		
19	7.16		
20	8.41		
21	4.75		
22	7.92		
23	7.82		
24	8.42		
25	7.08		
26	7.38		
27	9.29		
Average	**7.93**		**8.97**
Sample s.d.	**0.94**		**0.51**
Fold Ratio	**−1.84**		**+1.84**

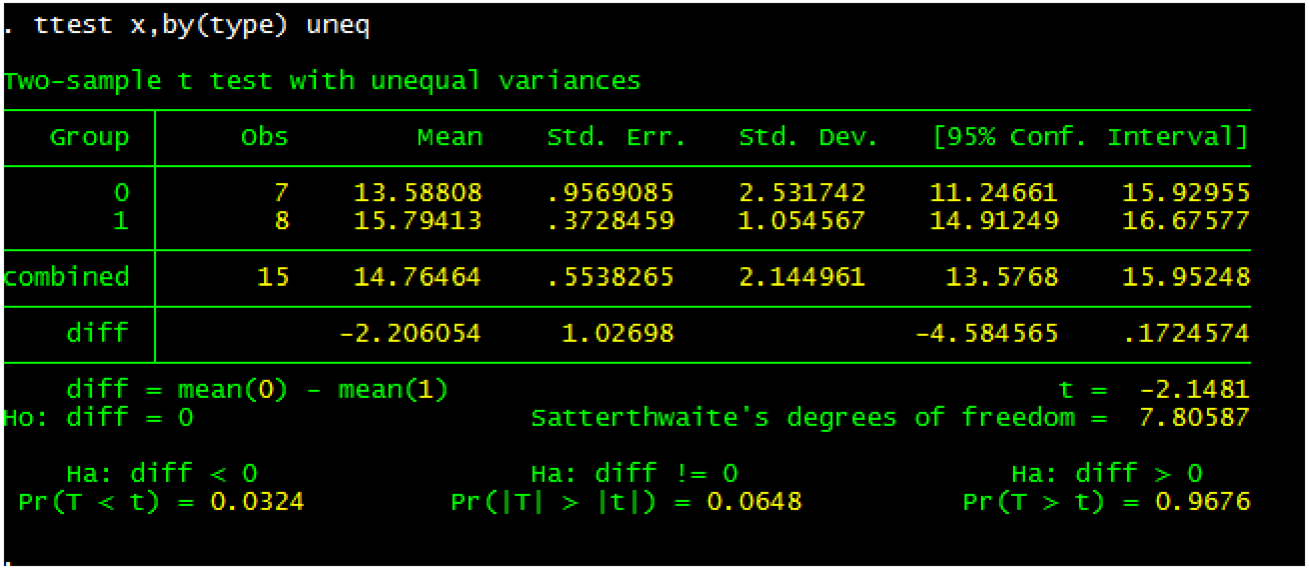
Note: This data came from Affymetrix arrays; the values have been logged (to base 2) to ensure [illegible] the data are normally distributed.



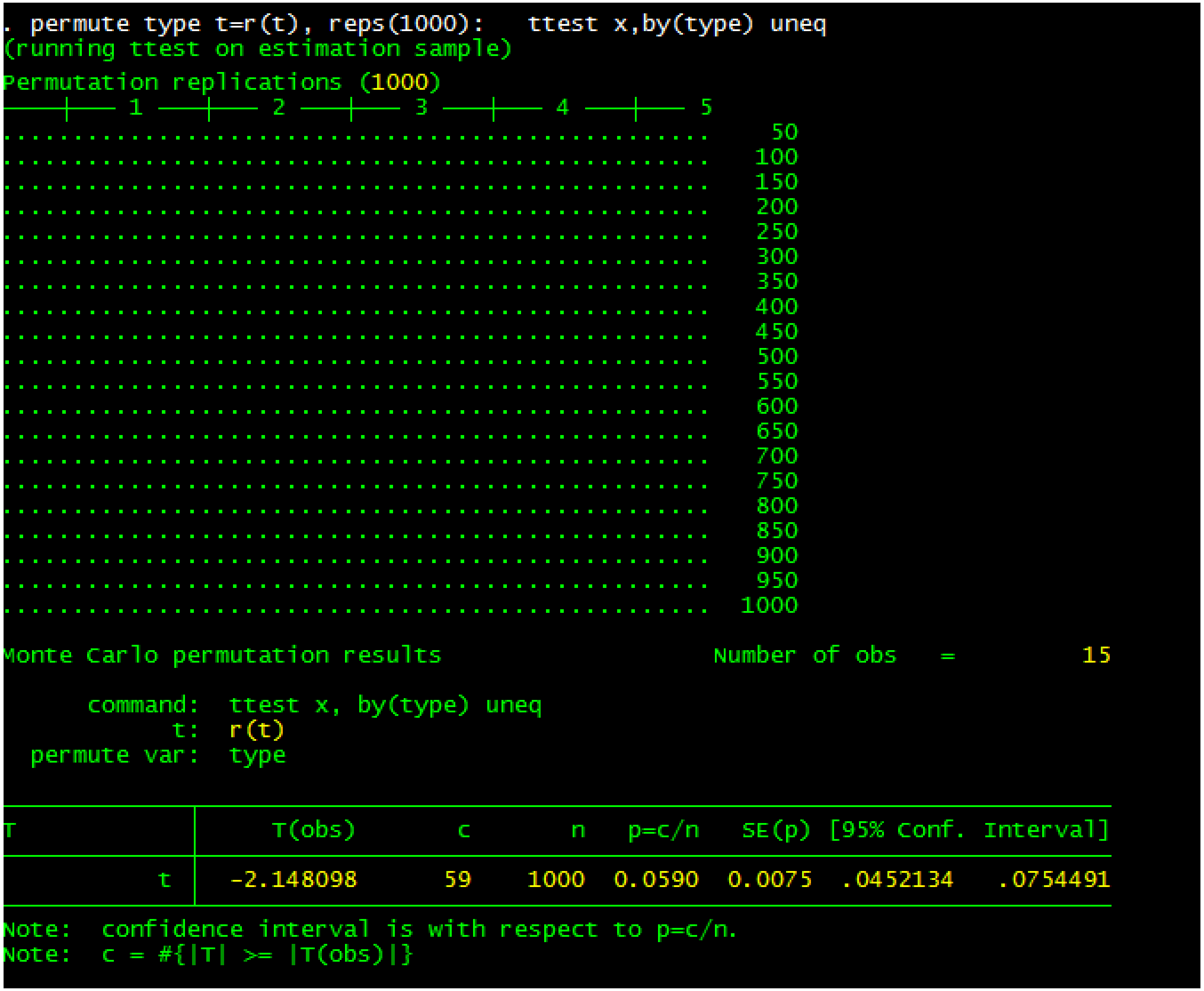
```
. list x type

     +-----------------+
     |        x   type |
     |-----------------|
  1. | 16.26653      0 |
  2. | 16.54437      0 |
  3. | 11.53271      0 |
  4. | 10.59901      0 |
  5. | 14.73416      0 |
     |-----------------|
  6. | 14.53156      0 |
  7. | 10.90819      0 |
  8. | 16.10997      1 |
  9. | 14.68326      1 |
 10. | 16.95882      1 |
     |-----------------|
 11. | 15.61745      1 |
 12. | 16.27782      1 |
 13. |  16.2908      1 |
 14. | 13.81206      1 |
 15. | 16.60287      1 |
     +-----------------+

.
```



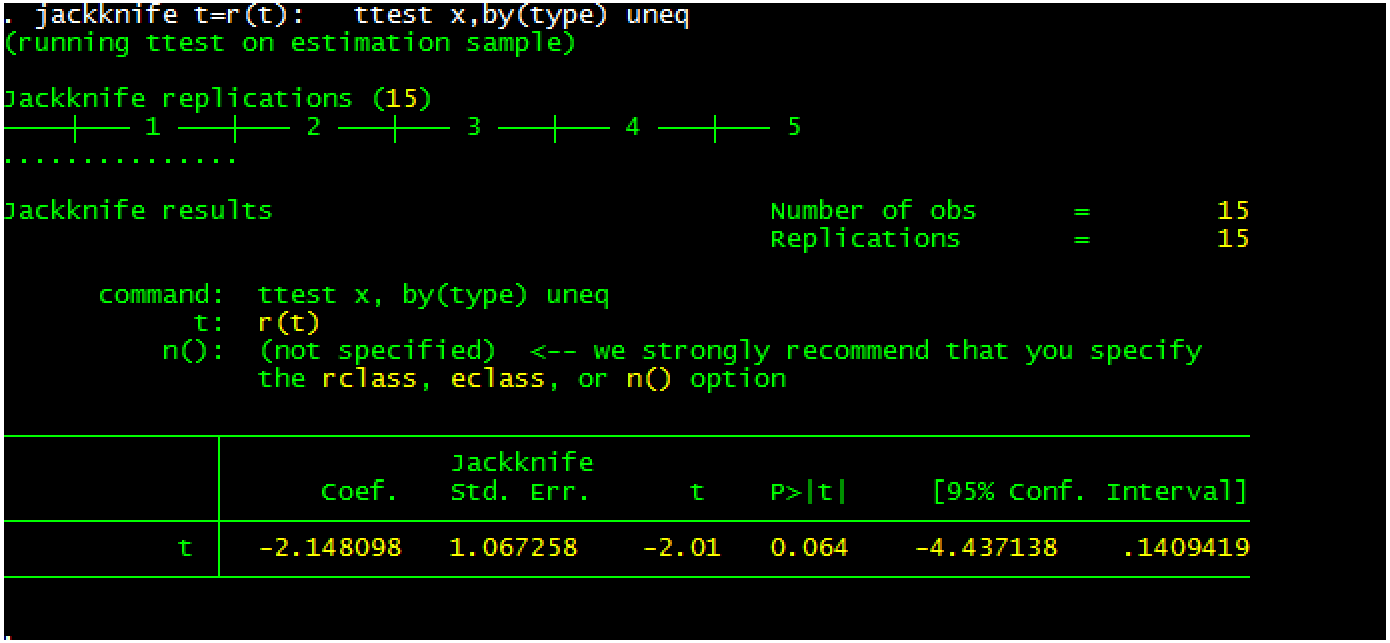
```
. ttest x,by(type)

Two-sample t test with equal variances
------------------------------------------------------------------------------
   Group |     Obs        Mean    Std. Err.   Std. Dev.   [95% Conf. Interval]
---------+--------------------------------------------------------------------
       0 |       7    13.58808    .9569085    2.531742    11.24661    15.92955
       1 |       8    15.79413    .3728459    1.054567    14.91249    16.67577
---------+--------------------------------------------------------------------
combined |      15    14.76464    .5538265    2.144961     13.5768    15.95248
---------+--------------------------------------------------------------------
    diff |           -2.206054    .9761203               -4.314833   -.0972739
------------------------------------------------------------------------------
    diff = mean(0) - mean(1)                                      t =  -2.2600
Ho: diff = 0                                     degrees of freedom =       13

    Ha: diff < 0                 Ha: diff != 0                 Ha: diff > 0
 Pr(T < t) = 0.0208         Pr(|T| > |t|) = 0.0416          Pr(T > t) = 0.9792

.
```



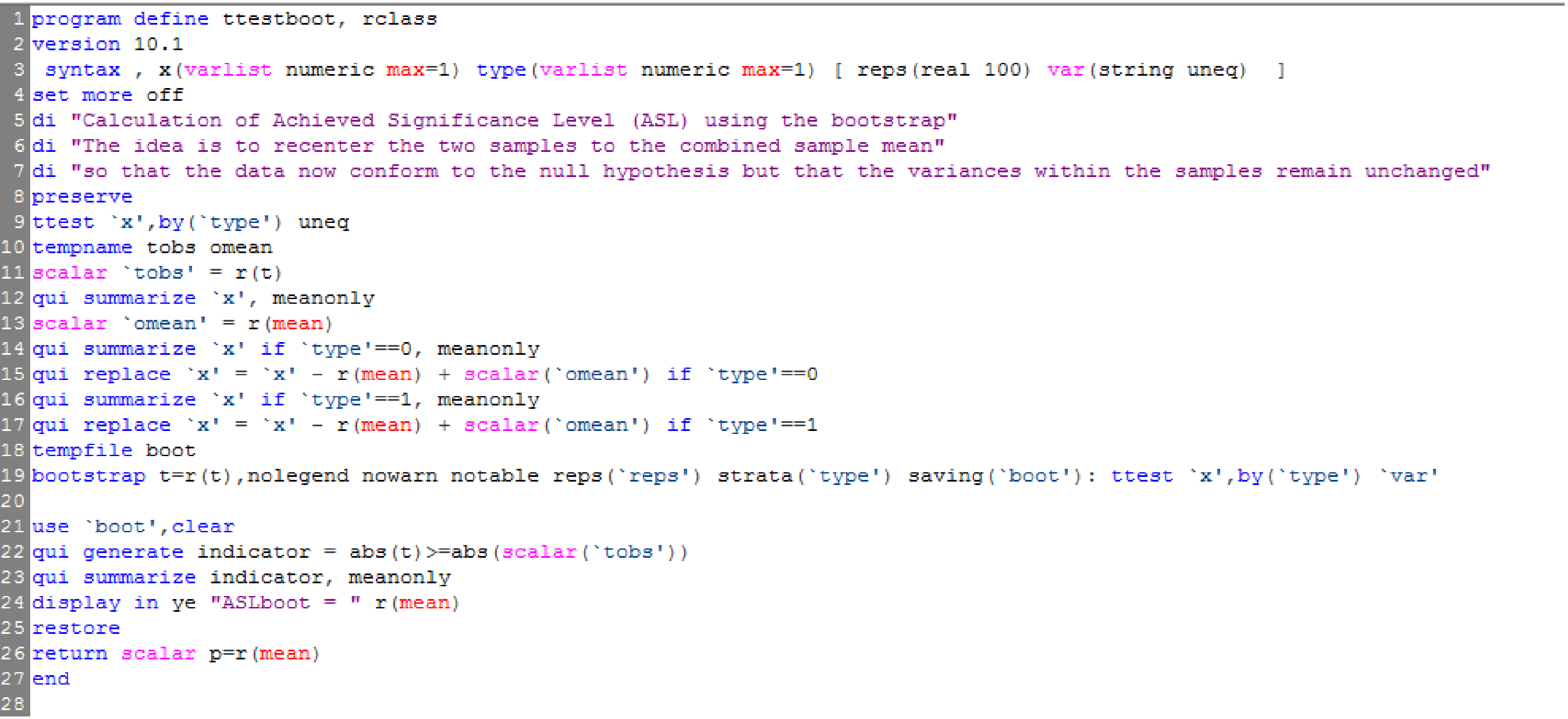
```
. ttest x,by(type) uneq

Two-sample t test with unequal variances
------------------------------------------------------------------------------
   Group |     Obs        Mean    Std. Err.   Std. Dev.   [95% Conf. Interval]
---------+--------------------------------------------------------------------
       0 |       7    13.58808    .9569085    2.531742    11.24661    15.92955
       1 |       8    15.79413    .3728459    1.054567    14.91249    16.67577
---------+--------------------------------------------------------------------
combined |      15    14.76464    .5538265    2.144961     13.5768    15.95248
---------+--------------------------------------------------------------------
    diff |           -2.206054     1.02698               -4.584565    .1724574
------------------------------------------------------------------------------
    diff = mean(0) - mean(1)                                      t =  -2.1481
Ho: diff = 0                     Satterthwaite's degrees of freedom =  7.80587

    Ha: diff < 0                 Ha: diff != 0                 Ha: diff > 0
 Pr(T < t) = 0.0324         Pr(|T| > |t|) = 0.0648          Pr(T > t) = 0.9676

.
```



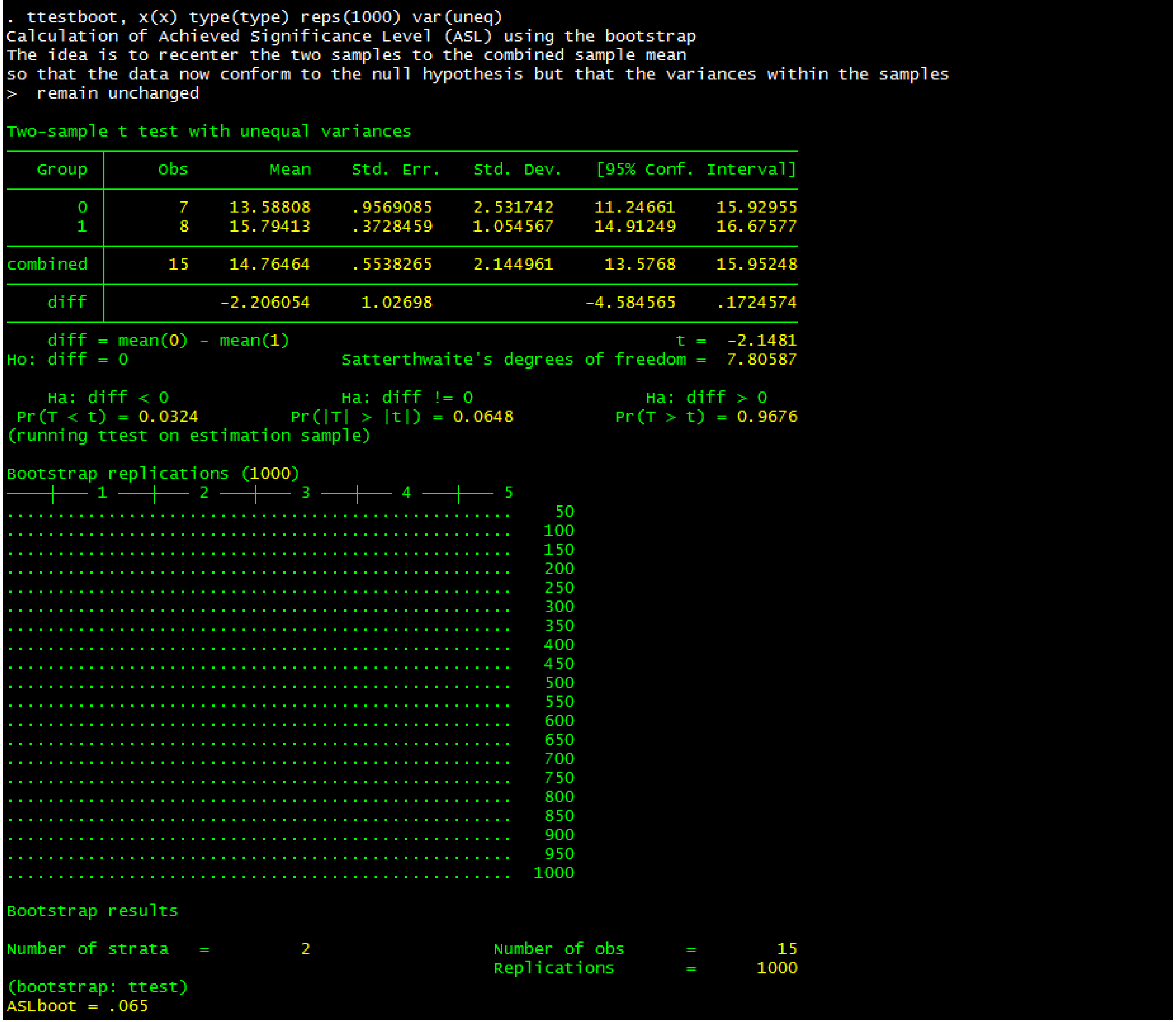
```
. permute type t=r(t), reps(1000):   ttest x,by(type) uneq
(running ttest on estimation sample)

Permutation replications (1000)
----+--- 1 ---+--- 2 ---+--- 3 ---+--- 4 ---+--- 5
..................................................    50
..................................................   100
..................................................   150
..................................................   200
..................................................   250
..................................................   300
..................................................   350
..................................................   400
..................................................   450
..................................................   500
..................................................   550
..................................................   600
..................................................   650
..................................................   700
..................................................   750
..................................................   800
..................................................   850
..................................................   900
..................................................   950
..................................................  1000

Monte Carlo permutation results                 Number of obs    =         15

      command:  ttest x, by(type) uneq
            t:  r(t)
  permute var:  type

------------------------------------------------------------------------------
T            |     T(obs)       c       n   p=c/n   SE(p) [95% Conf. Interval]
-------------+----------------------------------------------------------------
           t |  -2.148098      59    1000  0.0590  0.0075  .0452134   .0754491
------------------------------------------------------------------------------
Note:  confidence interval is with respect to p=c/n.
Note:  c = #{|T| >= |T(obs)|}
```



```
. jackknife t=r(t):   ttest x,by(type) uneq
(running ttest on estimation sample)

Jackknife replications (15)
----+--- 1 ---+--- 2 ---+--- 3 ---+--- 4 ---+--- 5
...............

Jackknife results                               Number of obs      =        15
                                                Replications       =        15

      command:  ttest x, by(type) uneq
            t:  r(t)
          n():  (not specified)  <-- we strongly recommend that you specify
                the rclass, eclass, or n() option

------------------------------------------------------------------------------
             |              Jackknife
             |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
           t |  -2.148098   1.067258    -2.01   0.064    -4.437138    .1409419
------------------------------------------------------------------------------

.
```



```
program define ttestboot, rclass
version 10.1
 syntax , x(varlist numeric max=1) type(varlist numeric max=1) [ reps(real 100) var(string uneq)  ]
set more off
di "Calculation of Achieved Significance Level (ASL) using the bootstrap"
di "The idea is to recenter the two samples to the combined sample mean"
di "so that the data now conform to the null hypothesis but that the variances within the samples remain unchanged"
preserve
ttest `x',by(`type') uneq
tempname tobs omean
scalar `tobs' = r(t)
qui summarize `x', meanonly
scalar `omean' = r(mean)
qui summarize `x' if `type'==0, meanonly
qui replace `x' = `x' - r(mean) + scalar(`omean') if `type'==0
qui summarize `x' if `type'==1, meanonly
qui replace `x' = `x' - r(mean) + scalar(`omean') if `type'==1
tempfile boot
bootstrap t=r(t),nolegend nowarn notable reps(`reps') strata(`type') saving(`boot'): ttest `x',by(`type') `var'

use `boot',clear
qui generate indicator = abs(t)>=abs(scalar(`tobs'))
qui summarize indicator, meanonly
display in ye "ASLboot = " r(mean)
restore
return scalar p=r(mean)
end

```



```
. ttestboot, x(x) type(type) reps(1000) var(uneq)
Calculation of Achieved Significance Level (ASL) using the bootstrap
The idea is to recenter the two samples to the combined sample mean
so that the data now conform to the null hypothesis but that the variances within the samples
>  remain unchanged

Two-sample t test with unequal variances
------------------------------------------------------------------------------
   Group |     Obs        Mean    Std. Err.   Std. Dev.   [95% Conf. Interval]
---------+--------------------------------------------------------------------
       0 |       7    13.58808    .9569085    2.531742    11.24661    15.92955
       1 |       8    15.79413    .3728459    1.054567    14.91249    16.67577
---------+--------------------------------------------------------------------
combined |      15    14.76464    .5538265    2.144961     13.5768    15.95248
---------+--------------------------------------------------------------------
    diff |           -2.206054     1.02698               -4.584565    .1724574
------------------------------------------------------------------------------
    diff = mean(0) - mean(1)                                      t =  -2.1481
Ho: diff = 0                     Satterthwaite's degrees of freedom =  7.80587

    Ha: diff < 0                 Ha: diff != 0                 Ha: diff > 0
 Pr(T < t) = 0.0324         Pr(|T| > |t|) = 0.0648          Pr(T > t) = 0.9676
(running ttest on estimation sample)

Bootstrap replications (1000)
----+--- 1 ---+--- 2 ---+--- 3 ---+--- 4 ---+--- 5
..................................................    50
..................................................   100
..................................................   150
..................................................   200
..................................................   250
..................................................   300
..................................................   350
..................................................   400
..................................................   450
..................................................   500
..................................................   550
..................................................   600
..................................................   650
..................................................   700
..................................................   750
..................................................   800
..................................................   850
..................................................   900
..................................................   950
..................................................  1000

Bootstrap results

Number of strata   =         2                  Number of obs      =        15
                                                Replications       =      1000
(bootstrap: ttest)
ASLboot = .065
```

Εφαρμογή t-test σε ζευγαρωτές παρατηρήσεις

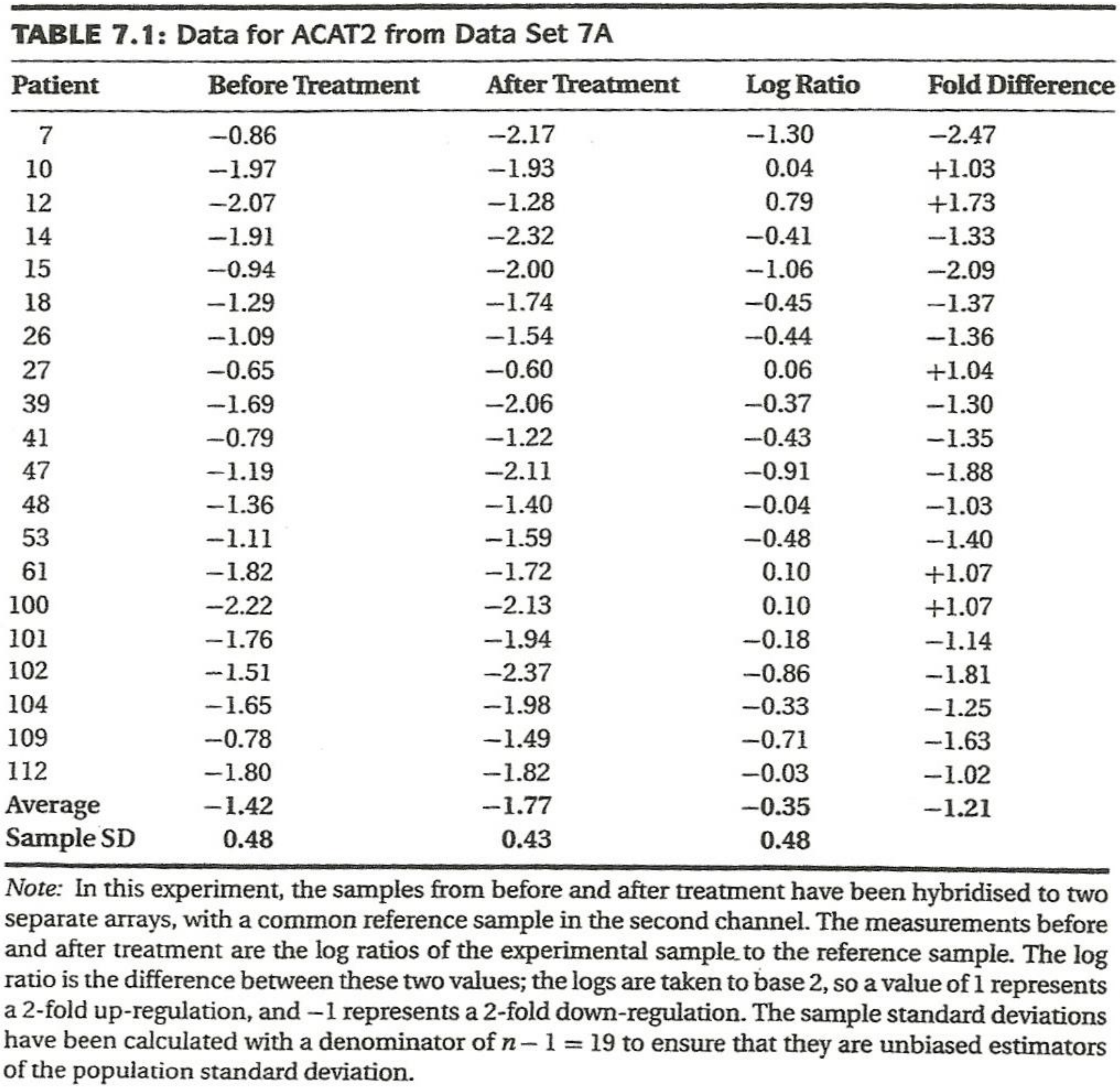
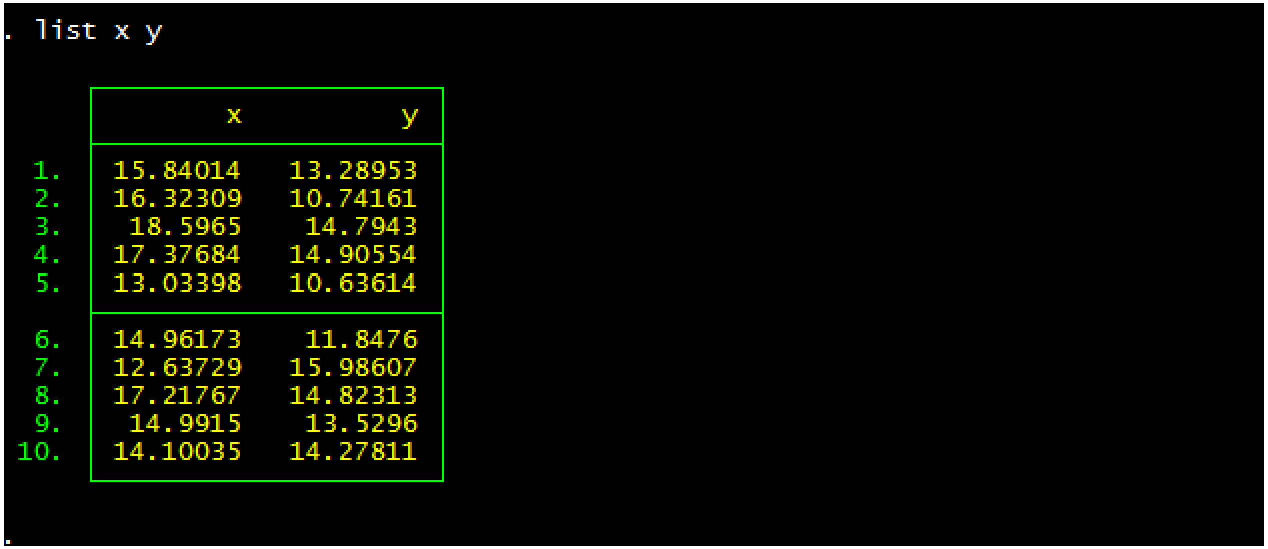


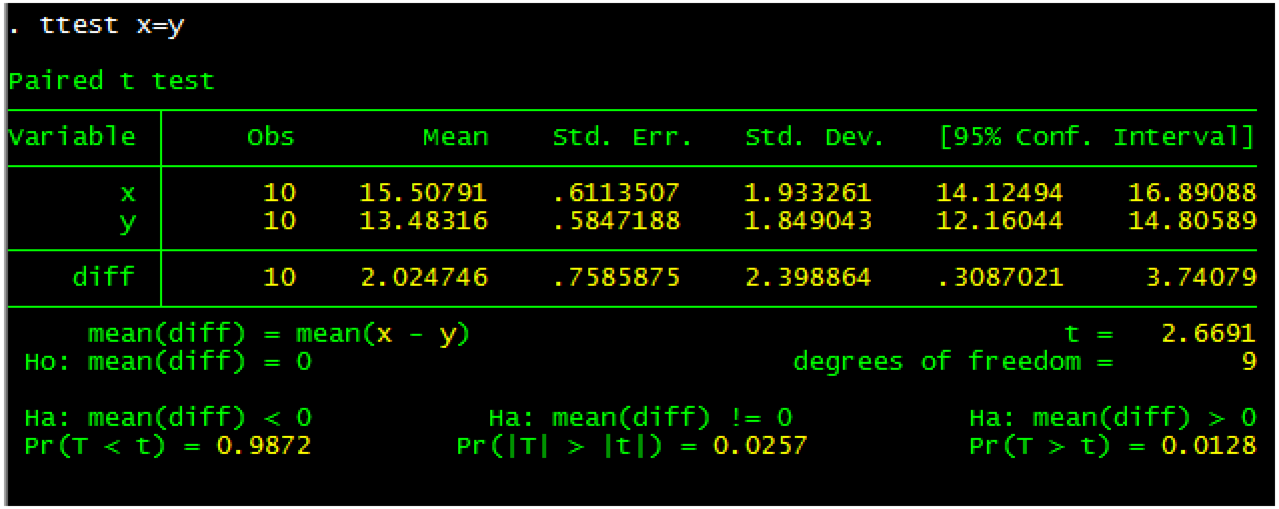
TABLE 7.1: Data for ACAT2 from Data Set 7A

Patient	Before Treatment	After Treatment	Log Ratio	Fold Difference
7	−0.86	−2.17	−1.30	−2.47
10	−1.97	−1.93	0.04	+1.03
12	−2.07	−1.28	0.79	+1.73
14	−1.91	−2.32	−0.41	−1.33
15	−0.94	−2.00	−1.06	−2.09
18	−1.29	−1.74	−0.45	−1.37
26	−1.09	−1.54	−0.44	−1.36
27	−0.65	−0.60	0.06	+1.04
39	−1.69	−2.06	−0.37	−1.30
41	−0.79	−1.22	−0.43	−1.35
47	−1.19	−2.11	−0.91	−1.88
48	−1.36	−1.40	−0.04	−1.03
53	−1.11	−1.59	−0.48	−1.40
61	−1.82	−1.72	0.10	+1.07
100	−2.22	−2.13	0.10	+1.07
101	−1.76	−1.94	−0.18	−1.14
102	−1.51	−2.37	−0.86	−1.81
104	−1.65	−1.98	−0.33	−1.25
109	−0.78	−1.49	−0.71	−1.63
112	−1.80	−1.82	−0.03	−1.02
Average	−1.42	−1.77	−0.35	−1.21
Sample SD	0.48	0.43	0.48	

Note: In this experiment, the samples from before and after treatment have been hybridised to two separate arrays, with a common reference sample in the second channel. The measurements before and after treatment are the log ratios of the experimental sample to the reference sample. The log ratio is the difference between these two values; the logs are taken to base 2, so a value of 1 represents a 2-fold up-regulation, and −1 represents a 2-fold down-regulation. The sample standard deviations have been calculated with a denominator of $n - 1 = 19$ to ensure that they are unbiased estimators of the population standard deviation.



```
. list x y

     +---------------------+
     |        x          y |
     |---------------------|
  1. | 15.84014   13.28953 |
  2. | 16.32309   10.74161 |
  3. |  18.5965    14.7943 |
  4. | 17.37684   14.90554 |
  5. | 13.03398   10.63614 |
     |---------------------|
  6. | 14.96173    11.8476 |
  7. | 12.63729   15.98607 |
  8. | 17.21767   14.82313 |
  9. |  14.9915    13.5296 |
 10. | 14.10035   14.27811 |
     +---------------------+

.
```



```
. ttest x=y

Paired t test
------------------------------------------------------------------------------
Variable |     Obs        Mean    Std. Err.   Std. Dev.   [95% Conf. Interval]
---------+--------------------------------------------------------------------
       x |      10    15.50791    .6113507    1.933261    14.12494    16.89088
       y |      10    13.48316    .5847188    1.849043    12.16044    14.80589
---------+--------------------------------------------------------------------
    diff |      10    2.024746    .7585875    2.398864    .3087021     3.74079
------------------------------------------------------------------------------
     mean(diff) = mean(x - y)                                      t =   2.6691
 Ho: mean(diff) = 0                               degrees of freedom =        9

 Ha: mean(diff) < 0           Ha: mean(diff) != 0           Ha: mean(diff) > 0
 Pr(T < t) = 0.9872         Pr(|T| > |t|) = 0.0257          Pr(T > t) = 0.0128
```

Bayesian methods

- The **bayesian** framework provides an intuitively appealing framework for dealing with most of the problems encountered in analysis of gene expression data. The t-test being one of the simplest and widely used methods has been into the centre of research for years and several bayesian counterparts of the t-test have been proposed, whereas some of them were developed specifically to address problems in microarray research.
- The various methods that have been proposed share some common features but also show marked differences according to various criteria, especially when it comes to definition of the prior distribution for the hyperparameters.
- Moreover, some of the methods are oriented toward hypothesis testing by relying on the Bayes Factor to compare the null against the alternative hypothesis (Gottardo, Pannucci et al. 2003, Gönen, Johnson et al. 2005, Rouder, Speckman et al. 2009, Wang and Liu 2015), whereas others are oriented towards parameter estimation and compute credible intervals for the parameters of interest, usually the difference of the means (Wetzels, Raaijmakers et al. 2009, Kruschke 2013).
- A convenient property of the t-test is the fact that its simplicity allows in many cases a closed form expression to be derived, especially for the Bayes Factor (Gottardo, Pannucci et al. 2003, Gönen, Johnson et al. 2005, Rouder, Speckman et al. 2009, Wang and Liu 2015), whereas other methods rely on MCMC to sample from the posterior distribution (Wetzels, Raaijmakers et al. 2009, Kruschke 2013).
- Another important feature of the bayesian methods is the fact that within the bayesian framework, one cannot only incorporate the problem of uncertainty and small sample size, but also the problem of multiple testing, a feature very helpful in microarray analysis (Gottardo, Pannucci et al. 2003, Fox and Dimmic 2006, Gonen 2010)

Software

- Concerning the above-mentioned methods, there are several software implementations available. For instance, the Bayes Factor method of Rouder and coworkers (Rouder, Speckman et al. 2009), which is known as the Jeffreys–Zellner–Siow (JZS) t-test, is available as a web-calculator (http://pcl.missouri.edu/bayesfactor) as well as an R package (https://cran.r-project.org/web/packages/BayesFactor/index.html).
- The Savage–Dickey (SD) t-test, proposed by Wetzels and coworkers (Wetzels, Raaijmakers et al. 2009), is inspired by the JZS t-test and retains its key concepts but is applicable to a wider range of statistical problems (i.e. allows researchers to test order restrictions and applies to two-sample situations in which the different groups do not share the same variance), is also available as an R package that uses WinBUGS (http://www.ruudwetzels.com/sdtest). Finally, there is the BEST (Bayesian Estimation Supersedes the t-test) package, which provides a Bayesian alternative to a t-test, providing much richer information about than a simple p value (i.e. complete distributions of credible values for the effect size, group means and their difference, standard deviations and their difference, and the normality of the data) (Kruschke 2013). The BEST package is available for R in http://www.indiana.edu/~kruschke/BEST/. There is also available an online calculator (http://sumsar.net/best_online/), whereas the method is also incorporated in the Bayesian First Aid package (https://github.com/rasmusab/bayesian_first_aid) that aims to provide easy to use Bayesian alternatives to the most widely used estimation commands.

Penalised t-test

- As we already noted, the ordinary t-test is not ideal for many microarray experiments because a large t-statistic can be driven by an unrealistically small value for *S*2. Genes with small sample variances, possibly as a result of very small sample size, have a good a chance of giving a large t-statistic even if they are not DE. A broad class of methods have been presented in order to alleviate such problems. These methods are usually called **penalized**, **moderated** or **regularized t-tests**. Most of these methods have been presented with an empirical Bayesian justification (hence, they share a lot of common features with the Bayesian methods), whereas other consist more of ad-hoc rules.
- In any case all of them apply some kind of modification to the denominator of the t-test by increasing the variance (Kooperberg, Aragaki et al. 2005).
- Thus, they all have the same interpretation as an ordinary t-statistic except that the standard errors have been moderated across genes, effectively borrowing information from the ensemble of genes to aid with inference about each individual gene.
- Baldi and Long were among the first to discuss Bayesian methods for the t-test in the context of microarray experiments (Baldi and Long 2001, Kayala and Baldi 2012). However, even though they developed a full Bayesian probabilistic framework for microarray data analysis, they finally chose to use in their web-server, Cyber-T (http://cybert.ics.uci.edu/) an empirical Bayes regularized t-test with variance equal to:

$$S^2_{Cyber-T} = \frac{\nu_0 \sigma_0^2 + (n-1) S^2}{\nu_0 + n - 2}$$

cont.

- The parameter v0 represents the degree of confidence in the background variance σ02 versus the empirical variance. In Cyber-T, the value of v0 can be set by the user by clicking on the corresponding button. The smaller n, the larger v0 ought to be. A simple rule of thumb is to assume that $K > 2$ points are needed to properly estimate the standard deviation and keep n + v0 = K . This allows for a flexible treatment of situations in which the number n of available data points varies from gene to gene. The default value is $K = 10$. In essence, using this approach the empirical variance is modulated by v0 «pseudo-observations» associated with a background variance σ02. For σ0, one could use the standard deviation of the entire dataset or, depending on the situation, of particular categories of genes. Cyber-T uses however a flexible approach under which the background standard deviation is estimated by pooling together all the neighboring genes contained in a window of size w (the default is $w = 101$, corresponding to 50 genes immediately above and below the gene under consideration). As we already mentioned, Cyber-T is available as a web-server as well as an R function (http://cybert.ics.uci.edu/).

Other similar methods

- Another empirical Bayes methods is the method of Lönnstedt and Speed (Lönnstedt and Speed 2002) which uses the moderated variance:

$$S_{LS}^2 = a + S^2$$

- where the penalty *a* is estimated from the mean and standard deviation of the sample variances *S*. Smyth later (Smyth 2005) generalized the approach from Lönnstedt and Speed in the well-known limma (linear models for microarray data) method which uses:

$$S_{limma}^2 = \frac{\nu_0 \sigma_0^2 + nS^2}{\nu_0 + n}$$

- Here, *d*0 and *s*0 are estimated from the data with the method of moments using an empirical bayes approach. The limma method is one of the most widely used methods for analysing DE genes, and there is available as Bioconductor package in R (http://bioinf.wehi.edu.au/limma). Tusher et al (Tusher, Tibshirani et al. 2001) and Efron et al (Efron, Tibshirani et al. 2001) also used a penalized t-statistics of the form

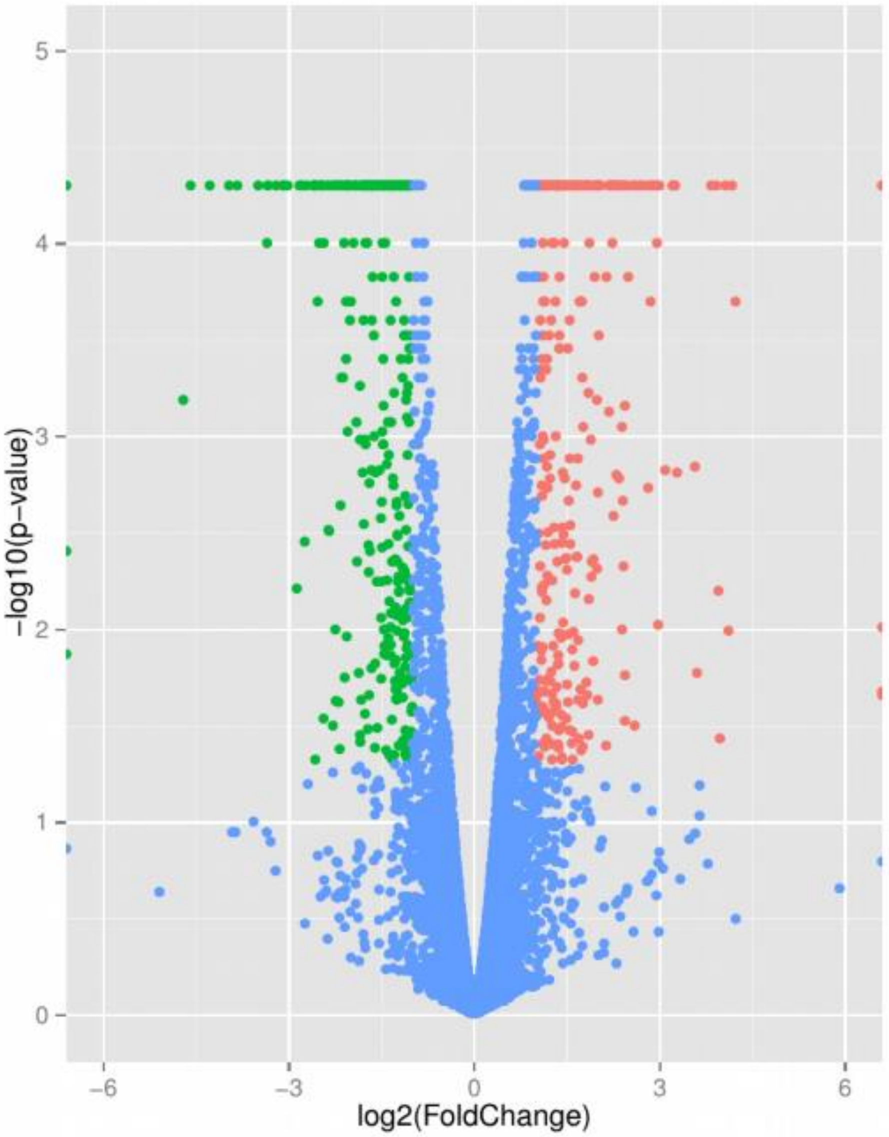
$$S_{SAM} = a + S$$

- This differs slightly from the previous statistics in that the penalty *a* is applied to the sample standard deviation *S* rather than to the sample variance *S*2. Tusher et al (Tusher, Tibshirani et al. 2001) in the so-called «*Significance Analysis of Microarrays*» (SAM) method, choose *a* to minimize the coefficient of variation of the absolute t-values while Efron et al (Efron, Tibshirani et al. 2001), choose *a* to be the 90th percentile of the *S* values. These choices are driven by empirical rather than theoretical considerations. SAM is one of the oldest and widely-used methods and it is available as Excel plugin at http://statweb.stanford.edu/~tibs/SAM/, as well as part of several R packages (samr, ema).

Other Alternatives

- As we already mentioned, the earliest microarray publications judged differential expression purely in terms of fold-change with 2-fold typically considered a worthwhile cutoff. However, fold-change cutoffs do not take variability into account or guarantee reproducibility. Moreover, the FC-based ranking is deficient because a gene with larger variances has a higher probability of having a larger statistic. The moderated t-tests on the other hand, allow for borrowing information across genes and show better performance, providing statistical estimates of statistical significance and the same time giving results more in line with fold-change rankings. However, even these modern statistical tests permit genes with arbitrarily small fold-changes to be considered statistically significant due to the t-statistic possibly having a very small denominator.

- Hence, it has become increasingly common in the literature to require that differentially expressed genes satisfy both *p*-value and fold-change criteria simultaneously. Some authors required genes to satisfy a modest level of statistical significance and then rank significant genes by fold-change with an arbitrary cutoff. Others, first apply a fold-change cutoff and then rank genes by their *p*-value, whereas others declare genes to be differentially expressed if they simultaneously show a fold-change larger than a cutoff and also satisfy criterion for *p* –value. Such combination criteria typically find more biologically meaningful sets of genes than *p*-values alone and in some cases give much better agreement between platforms than *p*-value alone.



Εικόνα 7.5: *Διάγραμμα "κρατήρα ηφαιστείου", (volcano plot) από ένα πείραμα μέτρησης διαφορικής γονιδιακής έκφρασης. Κάθε σημείο αντιστοιχεί σε ένα γονίδιο με τη θέση στον οριζόντιο άξονα να αντιστοιχεί στο δυαδικό λογάριθμο του λόγου διαφορικής έκφρασης και τη θέση στον κάθετο άξονα να αντιστοιχεί στον αρνητικό δεκαδικό λογάριθμο της τιμής p-value. Με πράσινο και κόκκινο φαίνονται τα στατιστικά σημαντικά υπο- και υπερ-εκφραζόμενα γονίδια (για τιμές κατωφλίων |log2FC|>=1.5 και p-value<=0.05).*

TREAT

- A method that tried to impose statistical formalism to these approaches is TREAT (*t*-tests relative to a threshold). This method is an extension of the empirical Bayes moderated *t*-statistic presented by Smyth (limma), and can be used to test whether the true differential expression is greater than a given threshold value. By including the fold-change threshold of interest in a formal hypothesis test, the methods achieve reliable *p*-values for finding genes with differential expression that is biologically meaningful (McCarthy and Smyth 2009). The method has shown very good properties in both real as well simulated data.

WAD

- Similar considerations have lead to the development of the weighted average difference method (WAD) for ranking DEGs (Kadota, Nakai et al. 2008). The authors observed that some top-ranked genes which are falsely detected as "differentially expressed" tend to exhibit lower expression levels. This interferes with the chance of detecting the "true" DEGs because the relative error is higher at lower signal intensities. WAD uses the average difference and relative average signal intensity so that highly expressed genes are highly ranked on the average for the different conditions:

$$WAD = \left(\bar{X}_1 - \bar{X}_2\right) \frac{\bar{X} - \min_p\left(\bar{X}\right)}{\max_p\left(\bar{X}\right) - \min_p\left(\bar{X}\right)}$$

Μετα-Ανάλυση

- Παρουσία θορύβου στα αποτελέσματα
- Μη επαναλήψιμα αποτελέσματα μεταξύ των πειραμάτων

↓

- Στατιστικό εργαλείο που επεξεργάζεται τα δεδομένα και τα αποτελέσματα μελετών που ερευνούν το ίδιο ερώτημα

- Παρέχει ένα τελικό συμπέρασμα το οποίο προέρχεται από μια σύνθεση ανεξάρτητων συνόλων δεδομένων

Normand, S. L. (1999). "Meta-analysis: formulating, evaluating, combining, and reporting." Stat Med 18(3): 321-59

- Meta-analysis is the statistical procedure for combining data from multiple studies. When the treatment effect (or effect size) is consistent from one study to the next, meta-analysis can be used to identify this common effect. When the effect varies from one study to the next/ meta-analysis may be used to identify the reason for the variation. Decisions about the utility of an intervention or the validity of a hypothesis cannot be based on the results of single study, due to the fact that the results typically vary from one study to the next. Rather, a mechanism is needed to synthesize data across studies. Meta-analysis applies objective formulas and can be used with any number of studies.

- ***Issue 1: Selection of Appropriate Microarray Datasets***
- The first, and most critical, step in an experimental study is to clearly state objectives. Meta-analysis enables the identification of differentially expressed genes among multiple samples in order to improve classification within and across platforms, to detect redundancy across diverse datasets, to identify differentially co-expressed genes, and infer networks of genetic interactions. The second step of meta-analysis is to set eligibility criteria, either biological (e.g., tissue type, disease) or technical (e.g., one-channel versus two-channel detection, density of microarrays, technological paltform). Based on these criteria, literature searches are preformed, using appropriate key terms, to retrieve relevant studies. These studies can be complemented by microarray data available in public databases that conform to the MIAME (Minimum Information About a Microarray Experiment) guidelines (Brazma, Hingamp et al. 2001) .

- ***Issue 2: Data Acquisition from Studies***
- The genes found to be differentially expressed in a given study constitute the published gene lists (PGLs) which are either included in the main text or provided as supplementary material. The gene expression data matrices (GEDM) contain preprocessed expression values of every probeset and sample for one gene. The published GEDM cannot be used directly as input for meta-analysis because of the different algorithms used for processing raw data in the original studies, which may generate heterogeneous, non-comparable results.

- ***Issue 3: Preprocessing of Datasets from Diverse Platforms***
- To enable consistent analysis of all datasets, bias introduced by the preprocessing algorithms should be eliminated. To this end, feature-level extraction output (FLEO) files, such as CEL files, should be obtained and converted to GEDM suitable for meta-analysis. Multiple studies from the same platform should be preprocessed using a single algorithm. In case the studies are conducted on different platforms, it is recommended to be preprocessed with comparable algorithms in order to be combinable.

- ***Issue 4: Promiscuous Hybridization between Probes and Genes***
- The datasets are annotated using UniGene or RefSeq gene identifiers, collectively referred to as GeneIDs. Multiple probes can hybridize with the same GeneID, as UniGene represents a cluster of sequences that correspond to a unique gene. Conversely, one non-specific probe can cross-hybridize with multiple GeneIDs due to imperfect specificity. There are also probes with inadequate sequence information that cannot hybridize with any GeneID. One approach to resolve the "many to many" relationships between probes and genes is to include in the meta-analysis only probes that are associated with a single gene, and exclude the promiscuous probes that are associated with more than one gene. In this way, however, important information can be lost. Averaging the expression profiles prior to meta-analysis is not recommended either, given that probe binding affinity differences affect the gene expression measurements. Therefore, it is recommended to apply descriptive statistics, thereby reducing the "many-to-many" into "one-to-one" relationship between probe and GeneID for each study.

- ***Issue 5: Choosing a Meta-Analysis Technique***
- The choice of meta-analysis technique depends on the type of response (e.g., binary, continuous, survival). In this article, we focus on the two-class comparison of microarrays where the objective is to identify genes expressed differentially between two wellknown conditions. There are three generic ways of combining information in such a situation: using effect sizes, using p-values and using ranks.

Statistical methods

- The statistical methods for meta-analysis of differentially expressed genes can be divided in three categories: the methods that rely on some **effect size**, the methods that **combine p-values** and the methods that **combine ranks**.

Μετα-ανάλυση Μικροσυστοιχιών

Μέθοδοι μετα-ανάλυσης:

- **t-test**

$$d_i = \frac{\overline{x}_{1i} - \overline{x}_{2i}}{sd_i} \qquad sd_i = \sqrt{\frac{(n_{1i}-1)sd_{1i}^2 + (n_{2i}-1)sd_{2i}^2}{n_{1i}+n_{2i}-2}}$$

- **Rank Product (Γινόμενο των βαθμών κατάταξης)**

$$RP_g = (\Pi_i \Pi_k r_{gik})\frac{1}{k}$$

- **Συνδυασμός των p-values**

$$s_i = -2\sum_{k=1}^{K} \log(p_{ik})$$

Hong, F. and R. Breitling (2008). "A comparison of meta-analysis methods for detecting differentially expressed genes in microarray experiments." Bioinformatics 24(3): 374-82.

Effect size

- The first approach is a standard approach for meta-analysis using fixed or random effects. In principle any suitable effect size can be used, but in practice most authors, for a number of reasons, advocate the standardized mean difference:

$$d_i = \frac{X_{1i} - X_{2i}}{S_{pi}}$$

- Where *X*1*i* and *X*2*i* are the means of the two groups under comparison in the *i*th study, and *Spi* is the pooled standard deviation given by:

$$S_{pi} = \sqrt{\frac{\left(n_{1i} - 1\right) S_{1i}^2 + \left(n_{1i} - 1\right) S_{1i}^2}{n_{1i} + n_{1i} - 2}}$$

- The sample estimate of the standardized mean difference is often called Cohen's d in research synthesis. It turns out that *d* has a slight bias, tending to overestimate the absolute value in small samples. This bias can be removed by a simple correction that yields an unbiased estimate, with the unbiased estimate sometimes called Hedges' *g*. To convert from *d* to Hedges' *g* we use a correction factor, which is called *J*. There is an exact formula for *J*, but in common practice researchers use an approximation given by *gi =Jdi= di* −3*di*/(4*ni* − 9). The estimated variance of *d* is given by

$$\mathrm{var}(d_i) = s_i^2 = \left(\frac{1}{n_{1i}} + \frac{1}{n_{2i}} \right) + \frac{d_i^2}{2\left(n_{1i} + n_{2i}\right)}$$

- When *g* is used, var(*g*)=*J*2var(*d*). In any case, it is straightforward to obtain a pooled estimate of *d* (or *g*):

$$d = \frac{\sum_{i=1}^{k} w_i d_i}{\sum_{i=1}^{k} w_i}$$

- This estimate is the well-known inverse-variance estimate used in meta-analysis with (Petiti 1994, Normand 1999). The above method assumes homogeneity of the effect across studies, an assumption that may be untenable. In case of between-studies heterogeneity, we hypothesize that the true effect varies from study to study and an additive component of the between studies variance (*τ*2) needs to be estimated (random-effects model). The most commonly used method for estimating *τ*2 is the non-iterative method of moments proposed by DerSimonian and Laird (DerSimonian and Laird 1986), even though there are several alternatives including iterative procedures (Thompson and Sharp 1999). In case *τ*2=0, the random-effects and the fixed-effects estimates coincide. In the random-effects case, the weights are calculated by

$$w_i = \left(\tau^2 + s_i^2\right)^{-1}$$

and subsequently Eq. (19) is applied in order to obtain the overall estimate

Προβλήματα

- Τα ίδια με την απλή ανάλυση
- Χρειαζόμαστε πάλι κάποια βελτιωμένη μέθοδο (bootstrap, permutation, empirical Bayes)
  - metaMA (https://cran.r-project.org/web/packages/metaMA/index.html)
  - GeneMeta
  - metaArray
  - MetaDE
- Full Bayesian methods
  - http://people.math.umass.edu/~conlon/research/BayesPoolMicro/

Ranks

- Another class of methods for meta-analysis consists of methods that combine ranks. There are several different approaches, but they all share the biological common sense that if the same gene is repeatedly at the top of the list ordered by up- or down-regulated genes in replicate experiments, the gene will be more likely to be regarded as differentially expressed. The Rank Product (RankProd) method, which we already described in the context of single study, uses the fold-change to rank genes and calculates the products of ranks across individuals and studies (Hong, Breitling et al. 2006). A similar method uses the Rank Sum instead, but all the other calculations are identical. The RankProd software is available at: https://www.bioconductor.org/packages/release/bioc/html/RankProd.html.

cont

- A related method termed METRADISC (Meta-analysis of Rank Discovery Dataset), is based on the same idea, but it is more general (Zintzaras and Ioannidis 2008, Zintzaras and Ioannidis 2012). The ranking within each study can be performed with any available method (FC, t-test, p-value etc) and then the average rank of a particular gene, for each study, can be calculated. The overall mean can be with or without weights, and in the former case the situation resembles the traditional methods for meta-analysis. The between-study heterogeneity of the study-specific ranks can also be computed. The METRADISC software is available in R (http://www.inside-r.org/node/155959) and as a standalone application (http://biomath.med.uth.gr/). The methods that use ranks are quite robust and can incorporate studies using different methods. However, the overall effect cannot be calculated and statistical inferences are based on Monte Carlo permutation tests, which may be time-consuming

- The rank-based methods offer several advantages traditional approaches, including the biologically intuitive of fold-change (FC) criterion, fewer assumptions under the model, and robustness with noisy data and/or low numbers of replicates. The approach overcomes the heterogeneity among multiple datasets and naturally combines them to achieve increased sensitivity and reliability. It is worth pointing out that these methods do not require the simultaneous normalization of multiple datasets using the same technique, which solves a frequently encountered dilemma in microarray meta-analysis pre-processing step. Moreover, the rank-based methods transform the actual expression values into ranks, and thus they can integrate datasets produced by a wide variety of platforms (Affymetrix oligonucleotide arrays, two-color cDNA arrays and so on). As matter of fact, the rank-based methods are quite general and thus can also be used for different types of data, such as proteomics or genetic association data.

Combination of p-values

- Another class of methods that is popular in meta-analysis of microarray studies (Hess and Iyer 2007) is related to the combination of p-values. It is widely accepted that Fisher's original work on combining of p-values (Fisher 1946) was the origin of meta-analysis (Jones 1995). Fisher noted that since p-values from k independent samples are uniform random variables, the sum of their logarithm will follow a *χ*2 distribution with 2*k* degrees of freedom:

$$U = -2\sum_{i=1}^{k}\log\left(p_i\right) = -2\log\left(\prod_{i=1}^{k} p_i\right)$$

Other approaches

- Edgington suggested using the sum of the p-values in order to obtain a pooled estimate (Edgington 1972)

$$p = \frac{\left(\sum_{i=1}^{k} p_i\right)^k}{k!}$$

- Later, the same author suggested using a contrast (Edgington 1972)

$$\bar{p} = \frac{\sum_{i=1}^{k} p_i}{k}$$

in which case $U = (0.5 - \bar{p})\sqrt{12}$ follows a $N(0,1)$

TPM

- A more sophisticated method was presented by Zaykin and coworkers, the so called truncated product method (TPM). Their procedure was to take the product of only those *p*-values less than some specified cut-off value (*τ*) and to evaluate the probability of such a product, or a smaller value, under the overall hypothesis that all *k* hypotheses are true

$$W = \prod_{i=1}^{k} \left(p_i \right)^{I\left(p_i \le \tau \right)}$$

$$P\left(W \le w\right) = \sum_{i=1}^{k} \binom{k}{r} \left(1-\tau\right)^{k-r} \left(w \sum_{s=0}^{r-1} \frac{\left(r \log \tau - \log w\right)^{s}}{s!} I\left(w \le \tau^{r}\right) + \tau^{r} I\left(w > \tau^{r}\right) \right)$$

Stouffer

- Nevertheless, combination of *p*-values although appealing and easily implemented presents serious problems relative to combining effect sizes. For example, there are problems when the p-values are testing different null hypotheses. Moreover, the method does not consider the direction of the association and thus all the p-values has to be one-sided, otherwise up-regulated and down-regulated genes need to be combined separately. Finally, the methods cannot quantify the magnitude of the association (the effect size), and most importantly does not allow for between studies heterogeneity.

$$\overline{Z} = \frac{\sum_{i=1}^{k} Z_i}{\sqrt{k}} \qquad \overline{Z} = \frac{\sum_{i=1}^{k} \sqrt{w_i} Z_i}{\sqrt{\sum_{i=1}^{k} w_i^2}}$$

Multiple Comparisons

- Any time you reject a null hypothesis because a *P* value is less than your critical value, it's possible that you're wrong; the null hypothesis might really be true, and your significant result might be due to chance. A *P* value of 0.05 means that there's a 5% chance of getting your observed result, *if* the null hypothesis were true. It does *not* mean that there's a 5% chance that the null hypothesis is true.
- For example, if you do 100 statistical tests, and for all of them the null hypothesis is actually true, you'd expect about 5 of the tests to be significant at the *P*<0.05 level, just due to chance. In that case, you'd have about 5 statistically significant results, all of which were false positives. The cost, in time, effort and perhaps money, could be quite high if you based important conclusions on these false positives, and it would at least be embarrassing for you once other people did further research and found that you'd been mistaken

- This problem, that when you do multiple statistical tests, some fraction will be false positives, has received increasing attention in the last few years. This is important for such techniques as the use of microarrays, which make it possible to measure RNA quantities for tens of thousands of genes at once; brain scanning, in which blood flow can be estimated in 100,000 or more three-dimensional bits of brain; and evolutionary genomics, where the sequences of every gene in the genome of two or more species can be compared. There is no universally accepted approach for dealing with the problem of multiple comparisons; it is an area of active research, both in the mathematical details and broader epistomological questions.

- The classic approach to the multiple comparison problem is to control the familywise error rate (FWER). Instead of setting the critical *P* level for significance, or alpha, to 0.05, you use a lower critical value. If the null hypothesis is true for all of the tests, the probability of getting *one* result that is significant at this new, lower critical value is 0.05. In other words, if all the null hypotheses are true, the probability that the family of tests includes one or more false positives due to chance is 0.05.
- The most common way to control the familywise error rate is with the Bonferroni correction. You find the critical value (alpha) for an individual test by dividing the familywise error rate (usually 0.05) by the number of tests. Thus if you are doing 100 statistical tests, the critical value for an individual test would be 0.05/100=0.0005, and you would only consider individual tests with P<0.0005 to be significan

- The Bonferroni correction is appropriate when a single false positive in a set of tests would be a problem. It is mainly useful when there are a fairly small number of multiple comparisons and you're looking for one or two that might be significant. However, if you have a large number of multiple comparisons and you're looking for many that might be significant, the Bonferroni correction may lead to a very high rate of false negatives. For example, let's say you're comparing the expression level of 20,000 genes between liver cancer tissue and normal liver tissue. Based on previous studies, you are hoping to find dozens or hundreds of genes with different expression levels. If you use the Bonferroni correction, a *P* value would have to be less than 0.05/20000=0.0000025 to be significant. Only genes with huge differences in expression will have a *P* value that low, and could miss out on a lot of important differences just because you wanted to be sure that your results did not include a single false negative.

- An alternative approach is to control the **false discovery rate (FDR)**. This is the proportion of "discoveries" (significant results) that are actually false positives. For example, let's say you're using microarrays to compare expression levels for 20,000 genes between liver tumors and normal liver cells. You're going to do additional experiments on any genes that show a significant difference between the normal and tumor cells, and you're willing to accept up to 10% of the genes with significant results being false positives; you'll find out they're false positives when you do the followup experiments. In this case, you would set your false discovery rate to 10%.

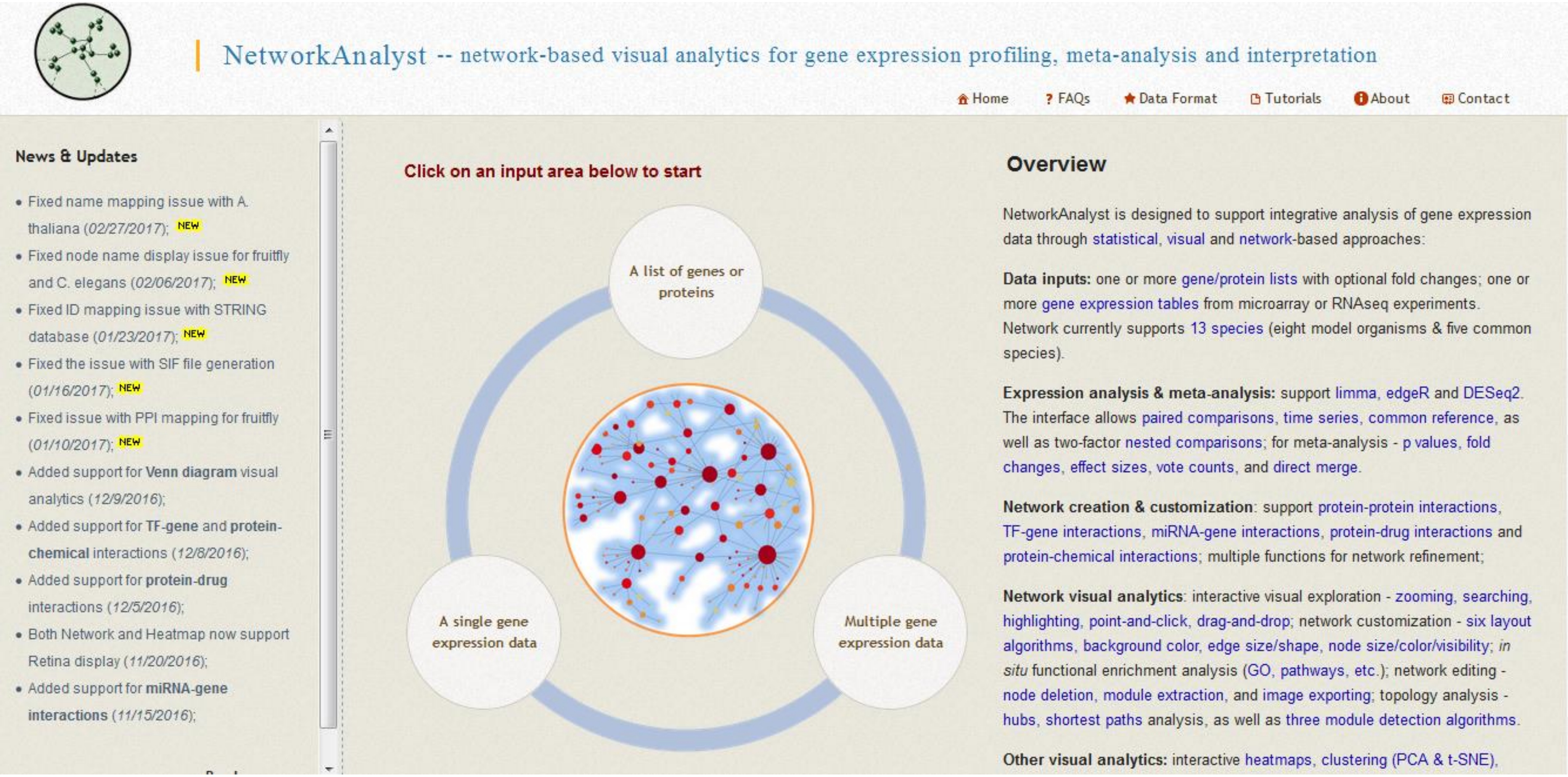
- One good technique for controlling the false discovery rate was briefly mentioned by Simes (1986) and developed in detail by Benjamini and Hochberg (1995). Put the individual *P* values in order, from smallest to largest. The smallest *P* value has a rank of *i*=1, then next smallest has *i*=2, etc. Compare each individual *P* value to its Benjamini-Hochberg critical value, (*i/m*)*Q*, where *i* is the rank, *m* is the total number of tests, and *Q* is the false discovery rate you choose. The largest *P* value that has *P*<(i/m)Q is significant, and *all* of the *P* values smaller than it are also significant, even the ones that aren't less than their Benjamini-Hochberg critical value.

Στατιστική Ανάλυση Μικροσυστοιχιών

- Παράδειγμα: Ας υποθέσουμε ότι εξετάζονται 10000 γονίδια τότε με p-value<0.05, 500 γονίδια αναμένεται να βρεθούν στατιστικά σημαντικά κατά τύχη (by chance)

- Ανάγκη χρησιμοποίησης των μεθόδων διόρθωσης για πολλαπλές συγκρίσεις
  - Bonferroni: $p_{cor(i)} = p_{(i)} * n$
  - Sidak: $p_{cor(i)} = 1 - (1 - p_{(i)})^{\frac{1}{n}}$
  - Holm: $p_{cor(i)} = (n - i) * p_{(i)}$
  - Holland: $p_{cor(i)} = (n - i + 1) * p_{(i)}$
  - FDR: $p_{cor(i)} = \frac{n}{n - i} * p_{(i)}$

NetworkAnalyst


NetworkAnalyst -- network-based visual analytics for gene expression profiling, meta-analysis and interpretation
Home
? FAQs
Data Format
Tutorials
About
Contact
A list of genes or proteins
A single gene expression data
Multiple gene expression data

News & Updates

- Fixed name mapping issue with A. thaliana (*02/27/2017*); NEW
- Fixed node name display issue for fruitfly and C. elegans (*02/06/2017*); NEW
- Fixed ID mapping issue with STRING database (*01/23/2017*); NEW
- Fixed the issue with SIF file generation (*01/16/2017*); NEW
- Fixed issue with PPI mapping for fruitfly (*01/10/2017*); NEW
- Added support for **Venn diagram** visual analytics (*12/9/2016*);
- Added support for **TF-gene** and **protein-chemical** interactions (*12/8/2016*);
- Added support for **protein-drug** interactions (*12/5/2016*);
- Both Network and Heatmap now support Retina display (*11/20/2016*);
- Added support for **miRNA-gene interactions** (*11/15/2016*);

Click on an input area below to start

Overview

NetworkAnalyst is designed to support integrative analysis of gene expression data through statistical, visual and network-based approaches:

Data inputs: one or more gene/protein lists with optional fold changes; one or more gene expression tables from microarray or RNAseq experiments. Network currently supports 13 species (eight model organisms & five common species).

Expression analysis & meta-analysis: support limma, edgeR and DESeq2. The interface allows paired comparisons, time series, common reference, as well as two-factor nested comparisons; for meta-analysis - p values, fold changes, effect sizes, vote counts, and direct merge.

Network creation & customization: support protein-protein interactions, TF-gene interactions, miRNA-gene interactions, protein-drug interactions and protein-chemical interactions; multiple functions for network refinement;

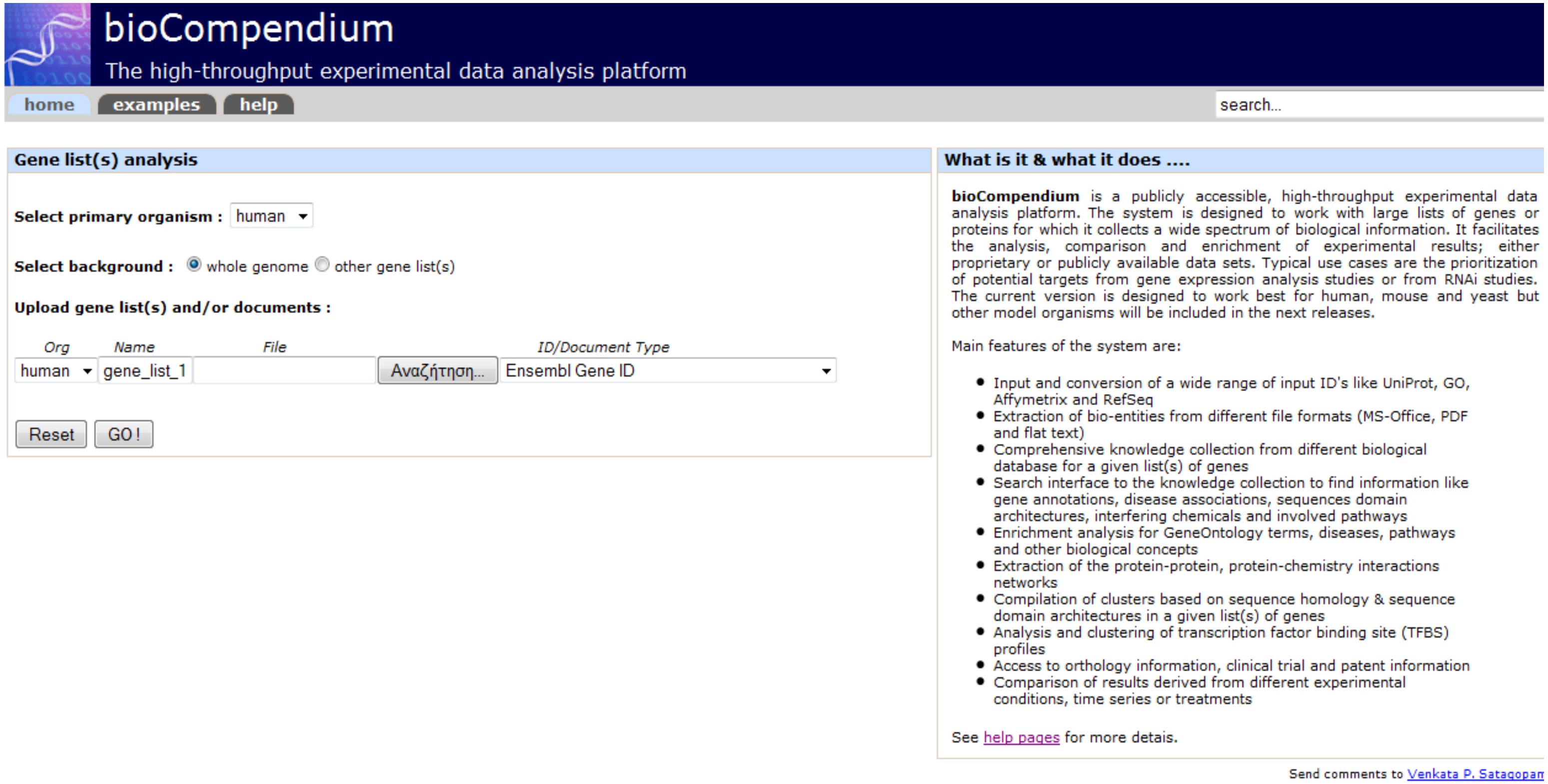
Network visual analytics: interactive visual exploration - zooming, searching, highlighting, point-and-click, drag-and-drop; network customization - six layout algorithms, background color, edge size/shape, node size/color/visibility; *in situ* functional enrichment analysis (GO, pathways, etc.); network editing - node deletion, module extraction, and image exporting; topology analysis - hubs, shortest paths analysis, as well as three module detection algorithms.

Other visual analytics: interactive heatmaps, clustering (PCA & t-SNE),

http://www.networkanalyst.ca

Μετά το clustering και τη μετα-ανάλυση?

- Χρήση λογισμικών για εύρεσης κοινών χαρακτηριστικών μεταξύ ομάδων γονιδίων
- Δημιουργία γονιδιακών υπογραφών με σκοπό την πρόβλεψη ασθενειών



bioCompendium

The high-throughput experimental data analysis platform

Gene list(s) analysis

Select primary organism : human

Select background : ● whole genome ○ other gene list(s)

Upload gene list(s) and/or documents :

Org	Name	File	ID/Document Type
human	gene_list_1	Αναζήτηση...	Ensembl Gene ID

What is it & what it does ....

bioCompendium is a publicly accessible, high-throughput experimental data analysis platform. The system is designed to work with large lists of genes or proteins for which it collects a wide spectrum of biological information. It facilitates the analysis, comparison and enrichment of experimental results; either proprietary or publicly available data sets. Typical use cases are the prioritization of potential targets from gene expression analysis studies or from RNAi studies. The current version is designed to work best for human, mouse and yeast but other model organisms will be included in the next releases.

Main features of the system are:

- Input and conversion of a wide range of input ID's like UniProt, GO, Affymetrix and RefSeq
- Extraction of bio-entities from different file formats (MS-Office, PDF and flat text)
- Comprehensive knowledge collection from different biological database for a given list(s) of genes
- Search interface to the knowledge collection to find information like gene annotations, disease associations, sequences domain architectures, interfering chemicals and involved pathways
- Enrichment analysis for GeneOntology terms, diseases, pathways and other biological concepts
- Extraction of the protein-protein, protein-chemistry interactions networks
- Compilation of clusters based on sequence homology & sequence domain architectures in a given list(s) of genes
- Analysis and clustering of transcription factor binding site (TFBS) profiles
- Access to orthology information, clinical trial and patent information
- Comparison of results derived from different experimental conditions, time series or treatments

See help pages for more detais.

Send comments to Venkata P. Satagopam