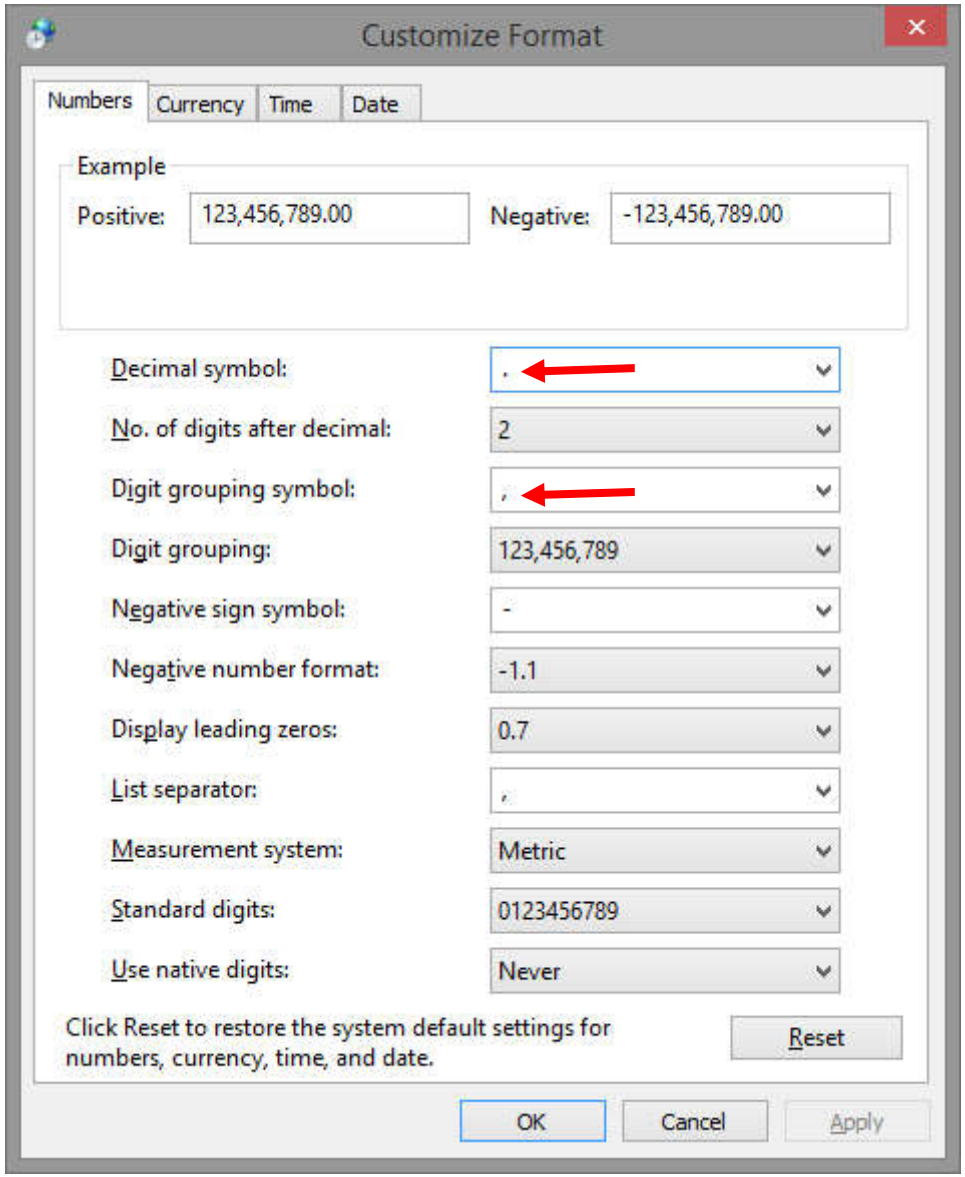
# Μικροσυστοιχίες

Δρ Ιωάννης Μιχαλόπουλος

Σύμβολο Υποδιαστολής;

·

# Αλλαγή locale στο PC



# Numeric Notation?

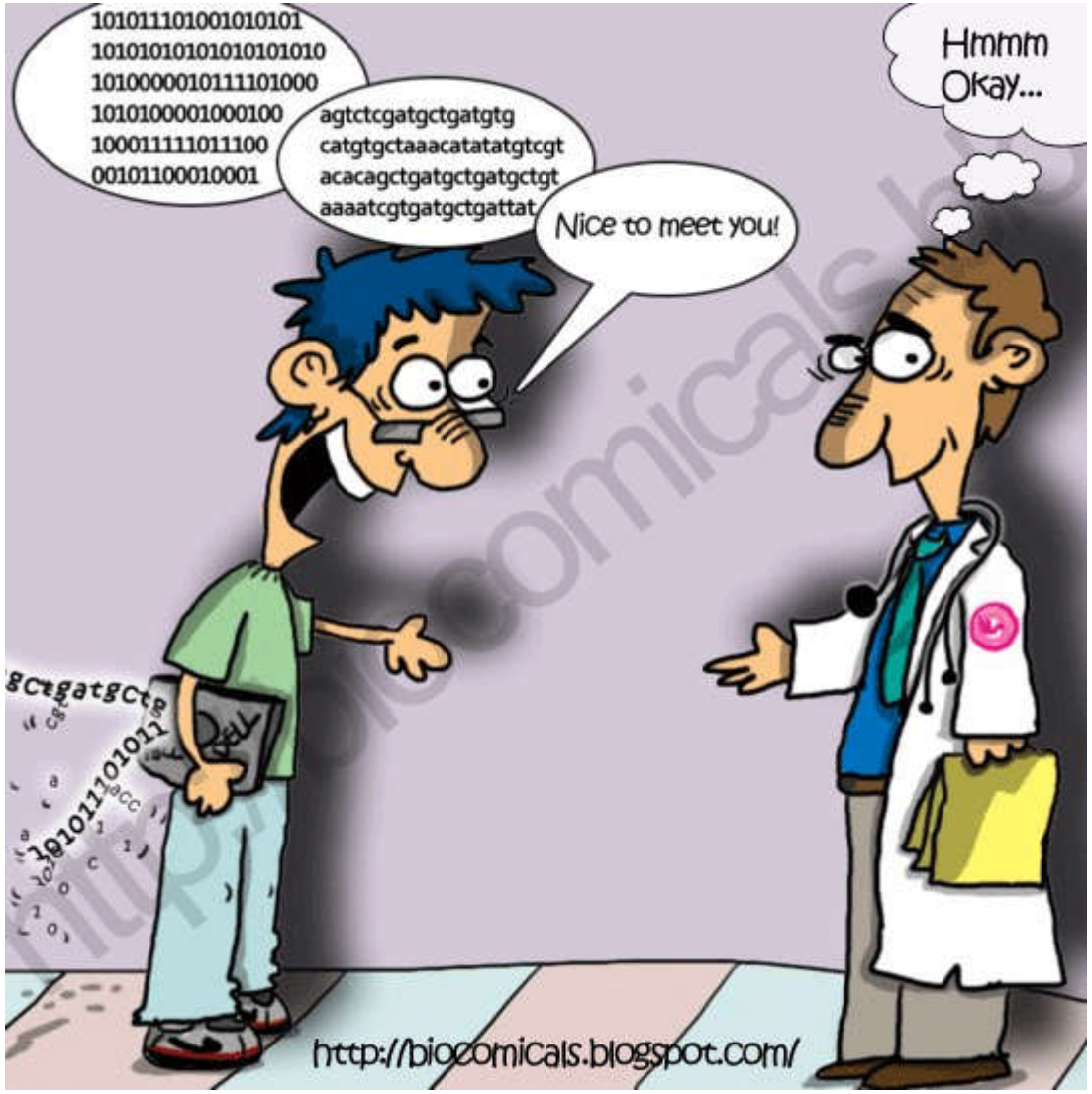
1.32E-16

$1.32 \cdot 10^{-16}$

# Βιοπληροφορική

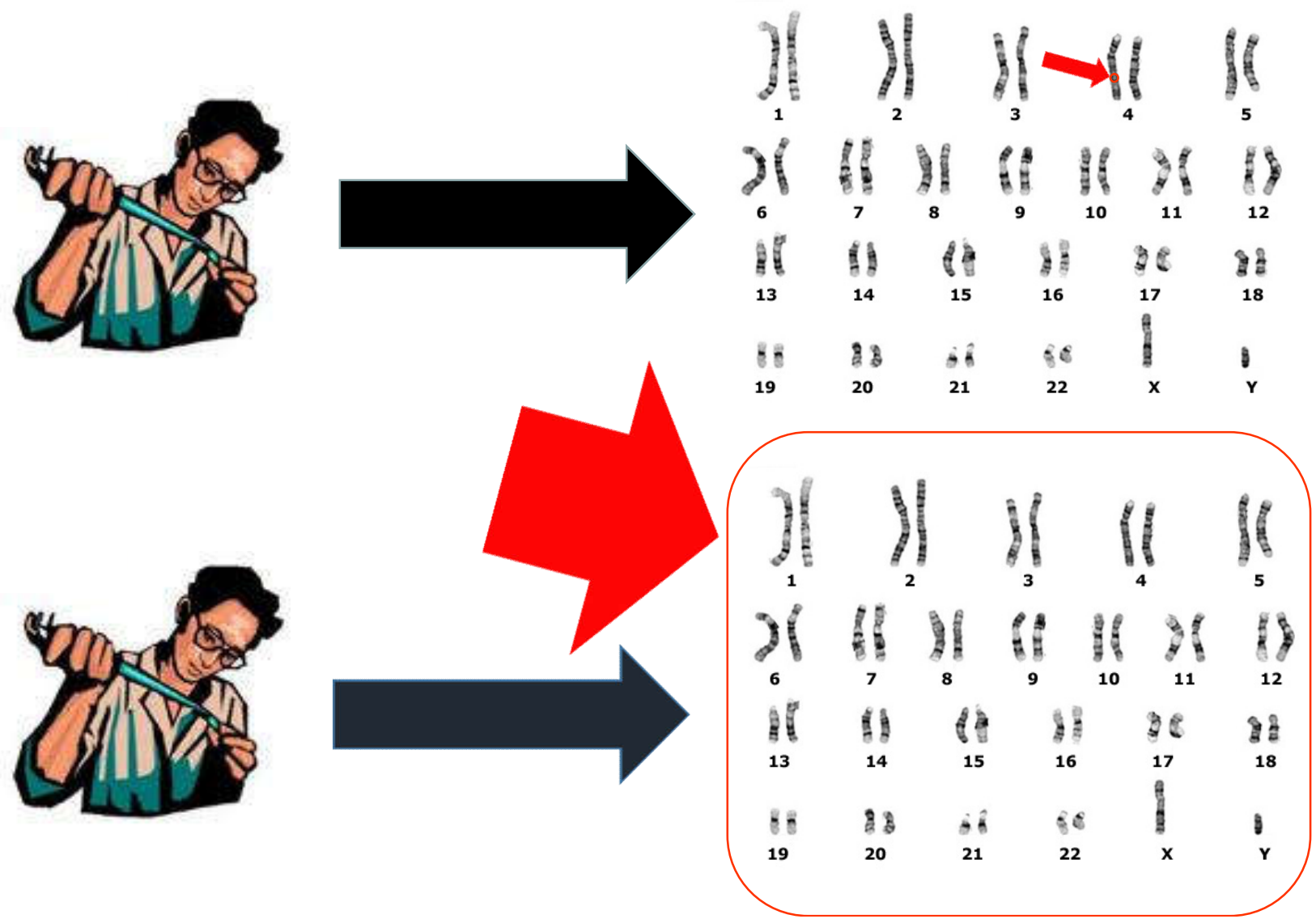

«Η επιστήμη που εφαρμόζει υπολογιστικές μεθόδους (επιστήμη υπολογιστών, εφαρμοσμένα μαθηματικά, στατιστική) με σκοπό την οργάνωση, διαχείριση και κατανόηση της πληροφορίας που σχετίζεται με βιομακρομόρια (DNA, RNA, πρωτεΐνες, πολυσακχαρίτες)»

# Κύριοι στόχοι της Βιοπληροφορικής




- Αποδοτική οργάνωση των βιολογικών δεδομένων και πρόσβαση σε αυτά, καθώς και συσσώρευση νέων δεδομένων
- Ανάπτυξη μεθόδων και υπολογιστικών εργαλείων με στόχο την εξαγωγή πληροφοριών από τα δεδομένα
- Χρήση των εργαλείων αυτών για την ανάλυση και ερμηνεία των δεδομένων με ένα βιολογικά αποδεκτό τρόπο

# Τεχνολογίες Υψηλής Απόδοσης



# Τεχνολογίες Υψηλής Απόδοσης

- Γενωμική
- **Μεταγραφωμική**
- Πρωτεωμική
- Μεταβολομική
- Επιγενωμική
- …

# Μικροσυστοιχίες

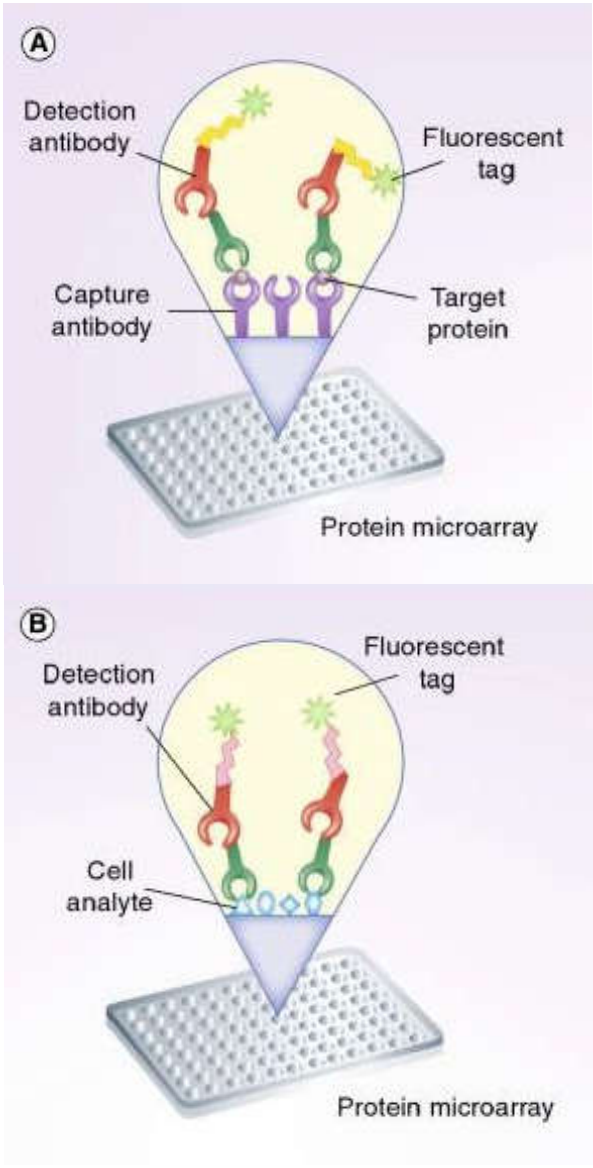
- Ο γενικός όρος περιλαμβάνει διαφόρους τύπους μικροσυστοιχιών:
  - Μικροσυστοιχίες Πρωτεϊνών
  - Μικροσυστοιχίες Ιστών
  - **Μικροσυστοιχίες cDNA**

# Μικροσυστοιχίες Πρωτεϊνών

- Κύτταρα ή ιστοί λύονται χωρίς να αποδιαταχθούν οι πρωτεΐνες του δείγματος
- Το πρωτεϊνικό λύμα προστίθεται σε κάθε πηγάδι μιας συστοιχίας πηγαδιών, όπου λαμβάνει χώρα κάποια παραλλαγή της ELISA, χωριστά σε κάθε πηγάδι

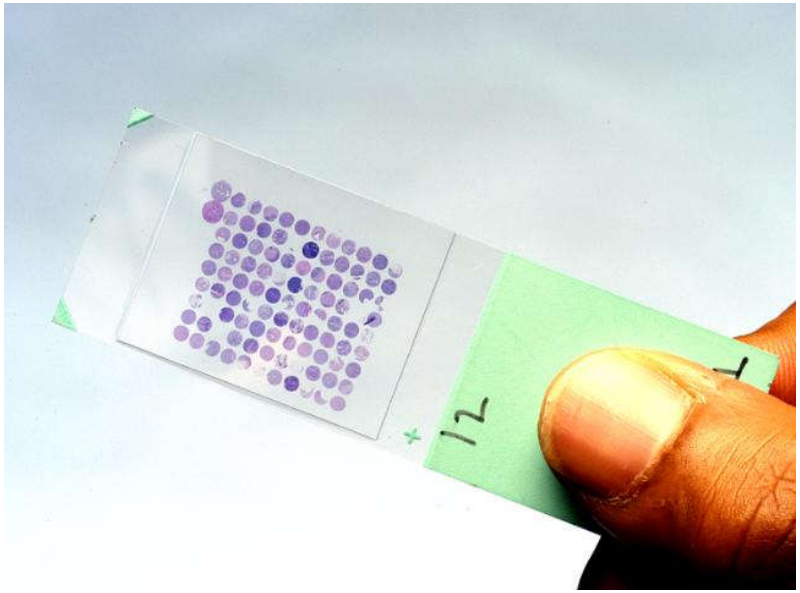
# Μικροσυστοιχίες Πρωτεϊνών

- Πρόσθια Φάση
  - Ακινητοποίηση σε κάθε πηγάδι, ενός πρωτογενούς αντισώματος που αναγνωρίζει μία πρωτεΐνη
  - Προσθήκη πρωτεϊνικού λύματος σε κάθε πηγάδι
  - Προσθήκη ενός διαφορετικού δευτερογενούς αντισώματος που αναγνωρίζει μία πρωτεΐνη σε κάθε πηγάδι
  - Προσθήκη σημασμένου με φθωρίζουσα ουσία, τριτογενούς αντισώματος που αναγνωρίζει το δευτερογενές αντίσωμα, σε κάθε πηγάδι
- Οπίσθια Φάση
  - Ακινητοποίηση σε κάθε πηγάδι των πρωτεϊνών του λύματος
  - Προσθήκη ενός πρωτογενούς αντισώματος που αναγνωρίζει μία πρωτεΐνη, σε κάθε πηγάδι
  - Προσθήκη σημασμένου με φθωρίζουσα ουσία, δευτερογενούς αντισώματος που αναγνωρίζει το πρωτογενές αντίσωμα, σε κάθε πηγάδι



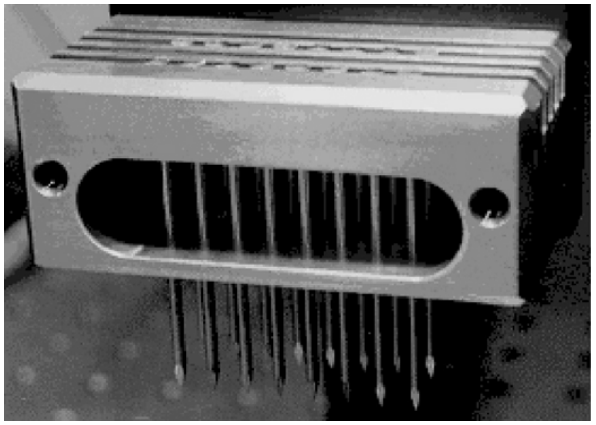
# Μικροσυστοιχίες Ιστών

- Μία κοίλη βελόνα μικρής διαμέτρου χρησιμοποιείται για την αφαίρεση κυλινδρικών δειγμάτων ιστού από περιοχές ενδιαφέροντος κύβων παραφίνης κλινικών βιοψιών ή δειγμάτων όγκων
- Τα κυλινδρικά δείγματα που συλλέγονται, εισάγονται σε έναν κύβο παραφίνης-δέκτη στοιχισμένα σε ίσες ακριβώς αποστάσεις
- Τα τμήματα από τον κύβο-δέκτη, τέμνονται εγκάρσια με μικροτόμο, τοποθετημένο σε αντικειμενοφόρο πλάκα μικροσκοπίου
- Κάθε μπλοκ μικροσυστοιχίας μπορεί να κοπεί σε 100-500 τμήματα, τα οποία μπορούν να υποβληθεί σε ανεξάρτητες ιστολογικές δοκιμές, πχ ανοσοϊστοχημείας, και φθορίζοντος υβριδισμού *in situ*
- Οι μικροσυστοιχίες ιστών είναι ιδιαίτερα χρήσιμες στην ανάλυση δειγμάτων καρκίνου

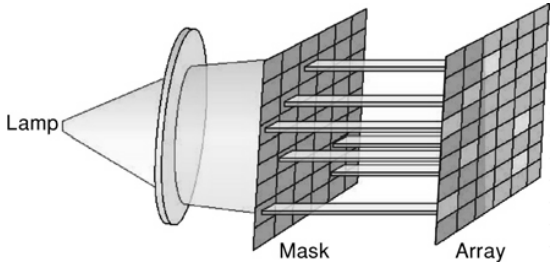


# Μικροσυστοιχίες cDNA

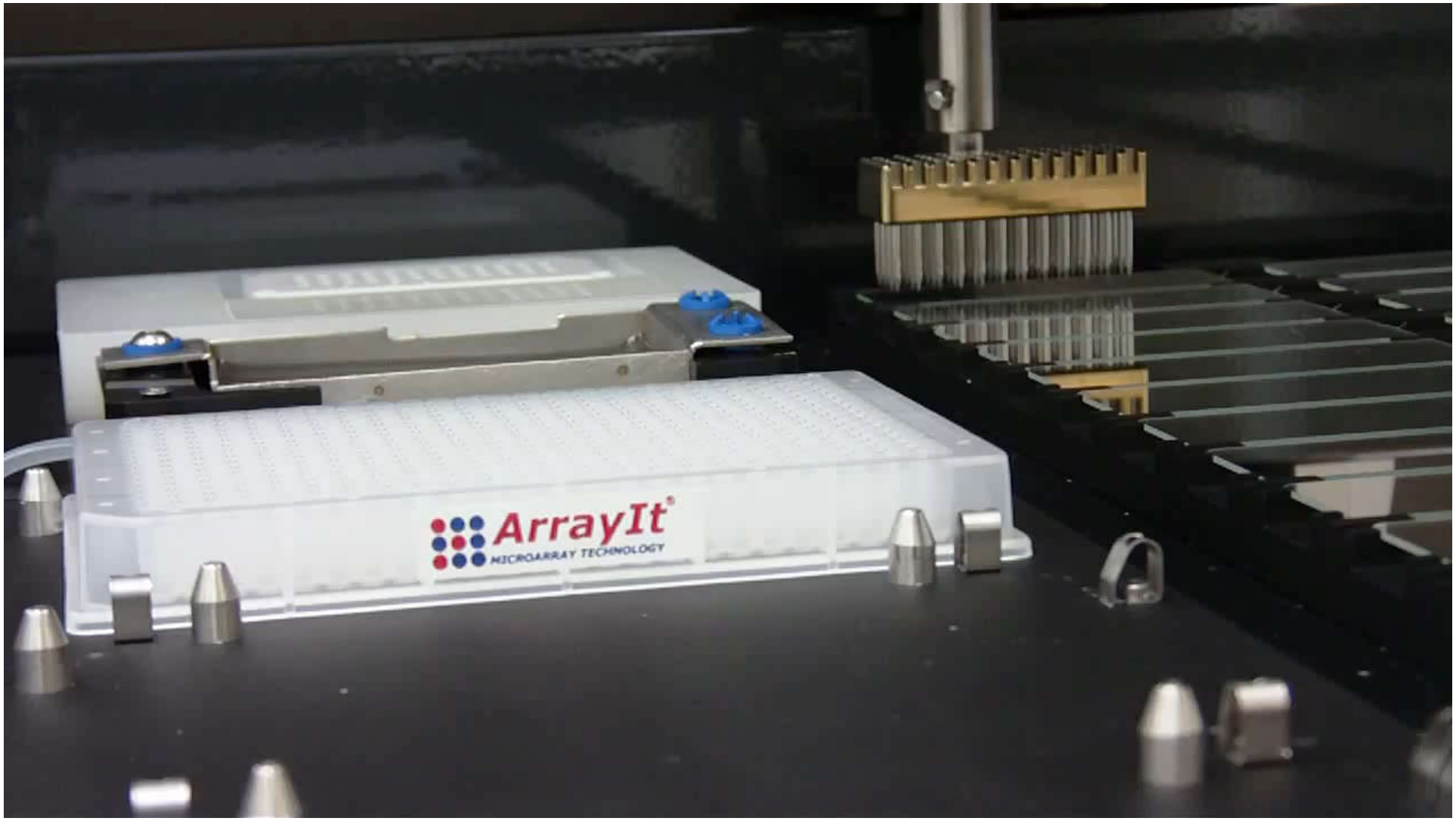
- Ένα πλέγμα κηλίδων DNA (ανιχνευτές) επί ενός υποστρώματος που χρησιμοποιείται για την ανίχνευση συμπληρωματικών αλληλουχιών
- Οι κηλίδες του DNA εναποτίθενται με:
  - Πιεζοηλεκτρισμό (όπως στην εκτύπωση με εκτόξευση μελάνης)
  - Εκτυπωτικές ακίδες
  - **Φωτολιθογραφία (Affymetrix)**
- Το υπόστρωμα μπορεί να είναι:
  - Πλαστικό
  - Γυαλί
  - **Πυρίτιο (Affymetrix)**
  - Σημασμένο DNA ή RNA (Affymetrix) υβριδίζεται στη μικροσυστοιχία
  - Η υβριδοποίηση ανιχνεύεται οπτικά

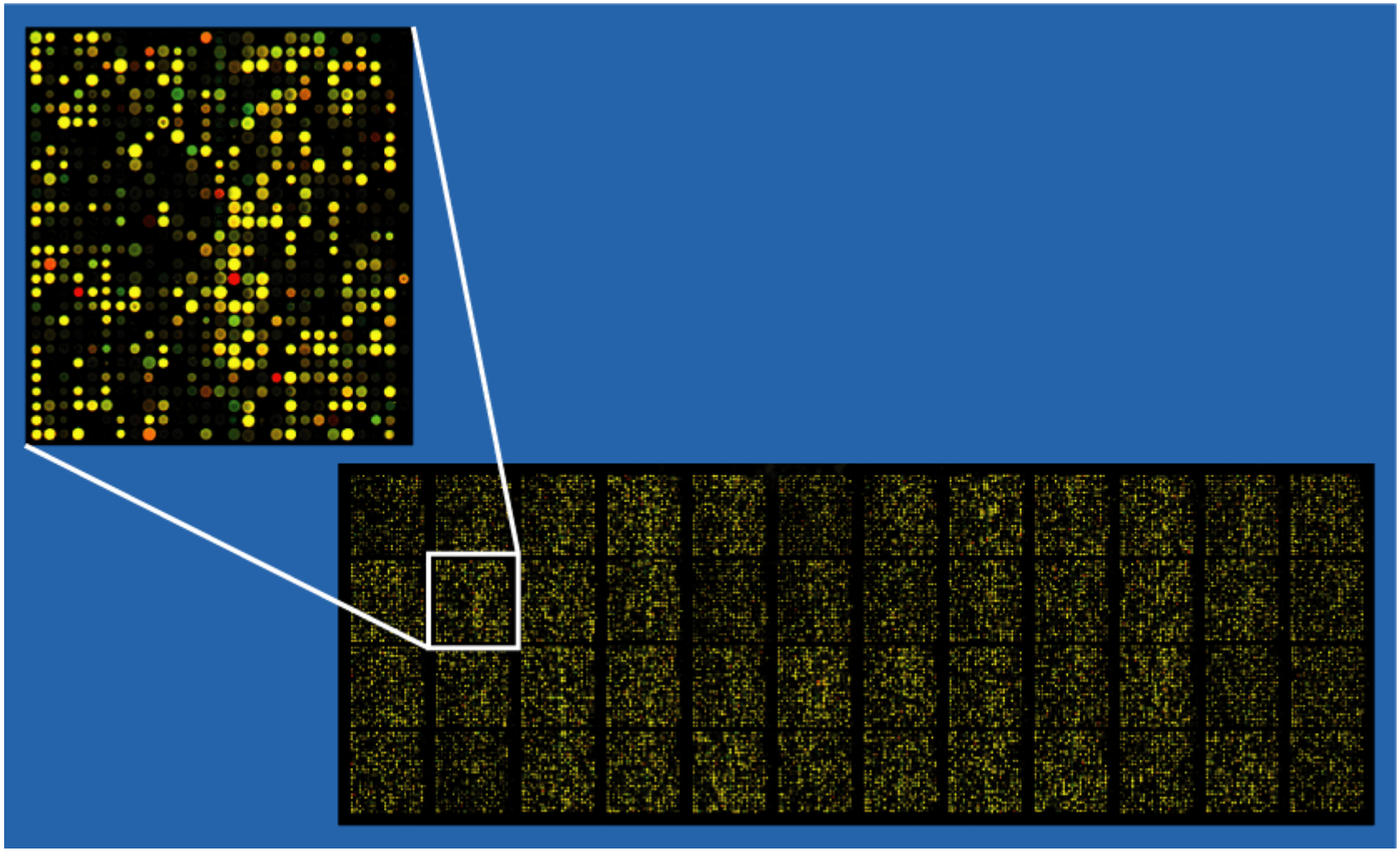


Spotted Array



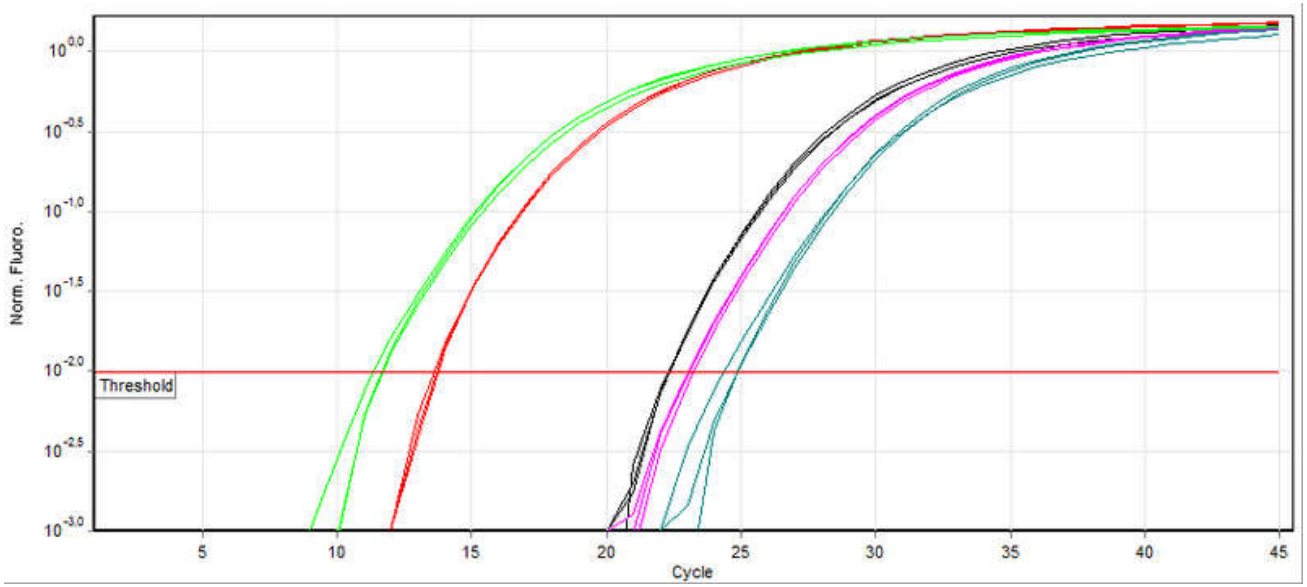

Oligo Array


ArrayIt
MICROARRAY TECHNOLOGY



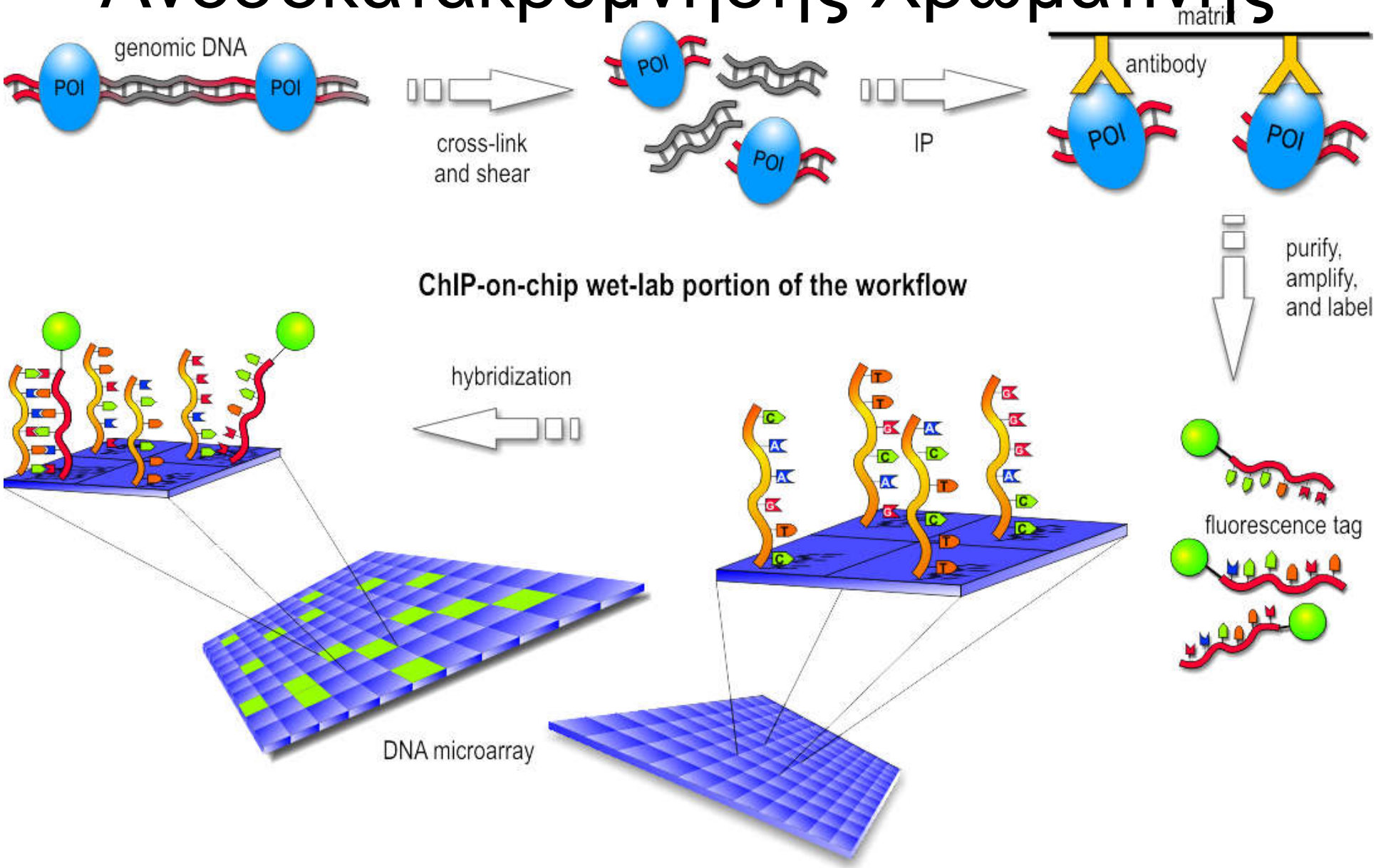
# Μικροσυστοιχίες cDNA

- Μικροσυστοιχίες Ανοσοκατακρύμνησης Χρωματίνης
- Μικροσυστοιχίες Μεθυλίωσης DNA
- Μικροσυστοιχίες Απλών Νουκλεοτιδικών Πολυμορφισμών
- Μικροσυστοιχίες Συγκριτικής Γενωμικής Υβριδοποίησης
- **Μικροσυστοιχίες Έκφρασης**




# Υποχρεωτική επιβεβαίωση με qPCR

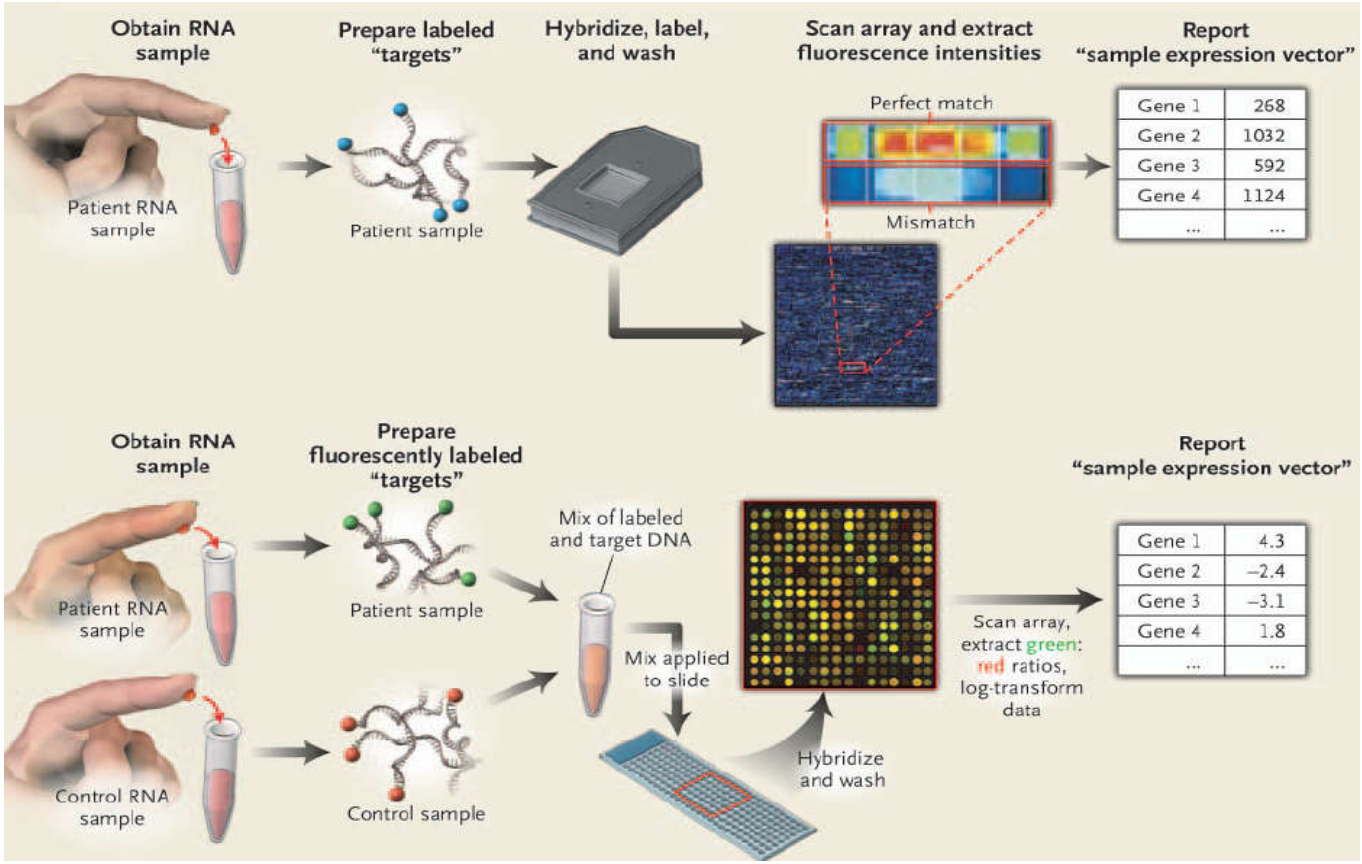
# Μικροσυστοιχίες Ανοσοκατακρύμνησης Χρωματίνης



# Μικροσυστοιχίες Έκφρασης

1-channel microarray

2-channel microarray



# Gene Expression Omnibus
# http://www.ncbi.nlm.nih.gov/geo/

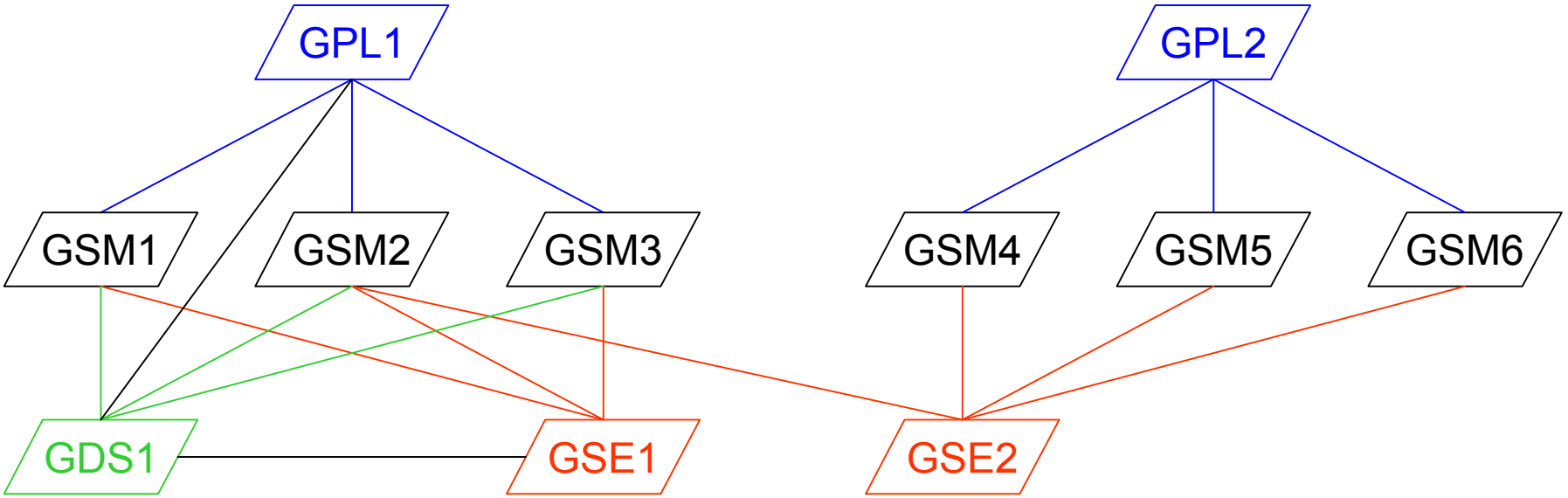
Δημόσιο αποθετήριο δεδομένων γονιδιακής έκφρασης, συμμορφωμένων προς τις απαιτήσεις του MIAME:

- Μικροσυστοιχίες (Affymetrix, Agilent, Illumina, κλπ.)
- Next Generation Sequencing

# Minimal Information About Microarray Experiment (MIAME)

•Πρωτογενή δεδομένα κάθε υβριδισμού (π.χ., αρχεία CEL ή GPR)
•Επεξεργασμένα (κανονικοποιημένα) δεδομένα για το σύνολο των υβριδισμών στην πειραματική μελέτη (π.χ., ο πίνακας δεδομένων γονιδιακής έκφρασης που χρησιμοποιείται για να εξαχθούν τα συμπεράσματα από τη μελέτη)
•Βασικοί σχολιασμοί των δειγμάτων, συμπεριλαμβανομένων των πειραματικών παραγόντων και των τιμών τους (π.χ., ουσία και δόση σε πείραμα απόκρισης σε δόση)
•Πειραματικός σχεδιασμός, συμπεριλαμβανομένων των σχέσεων δειγμάτων και δεδομένων (π.χ., ποιο αρχείο πρωτογενών δεδομένων σχετίζεται με ποιο δείγμα, ποιοι υβριδισμοί είναι τεχνικές και βιολογικές επαναλήψεις)
•Επαρκής σχολιασμός της μικροσυστοιχίας (π.χ. αναγνωριστικά γονιδίου, γενωμικές συντεταγμένες, ακολουθίες ολιγονουκλεοτιδικών ιχνηθετών ή αριθμός καταλόγου εμπορικών μικροσυστοιχιών)
•Βασικά πρωτόκολλα εργαστηριακής επεξεργασίας και επεξεργασίας δεδομένων (π.χ., ποια μέθοδος κανονικοποίησης έχει χρησιμοποιηθεί για τη λήψη των τελικών επεξεργασμένων δεδομένων)

# Δομή αρχειοθέτησης στη GEO




•Κάθε πειραματική **πλατφόρμα** έχει ένα μοναδικό αριθμό **GPL**
•Κάθε **δείγμα** έχει ένα μοναδικό αριθμό **GSM** και ανήκει σε μία πλατφόρμα (GPL)
•Κάθε **σειρά δειγμάτων** έχει ένα μοναδικό αριθμό **GSE** και αποτελεί σύνολο από ένα ή περισσότερα δείγματα (GSMs) που ανήκουν σε μία ή περισσότερες πλατφόρμες (GPLs). Το ίδιο GSM μπορεί να ανήκει σε περισσότερες από μία σειρές (GSEs)
•Κάθε **σύνολο δεδομένων** έχει ένα μοναδικό αριθμό **GDS** και αποτελεί μια επιμελημένη συλλογή περισσοτέρων του ενός δειγμάτων (GSMs) που ανήκουν σε μία πλατφόρμα (GPL) και μία σειρά (GSE)

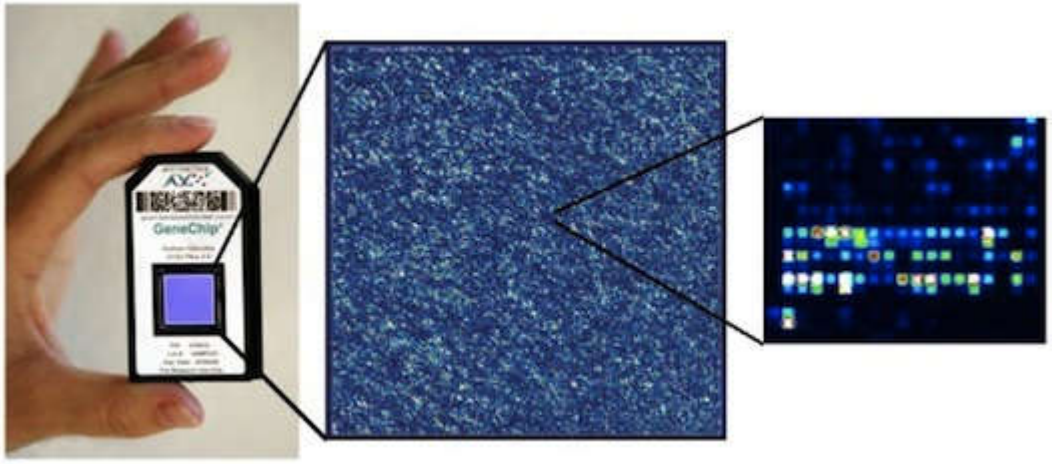
# Διαφορική Έκφραση - Συνέκφραση

- Οι Μικροσυστοιχίες χρησιμοποιούνται κυρίως για την ανίχνευση γονιδίων που εκφράζονται διαφορικά σε δύο ή περισσότερα διακριτά σύνολα δειγμάτων
  - Τα διαφορικώς εκφραζόμενα γονιδίων μπορεί να είναι υπεύθυνα για τις διαφορετικές ιδιότητες των συνόλων των δειγμάτων
- Η ανάλυση της γονιδιακής συνέκφρασης συνδυάζει άσχετα σύνολα δεδομένων μικροσυστοιχιών του ίδιου οργανισμού από διάφορους ιστούς ή αναπτυξιακά στάδια ή κάτω από διαφορετικές πειραματικές συνθήκες
  - Τα συνεκφραζόμενα γονίδια τείνουν να εμπλέκονται σε παρόμοιες βιολογικές διεργασίες

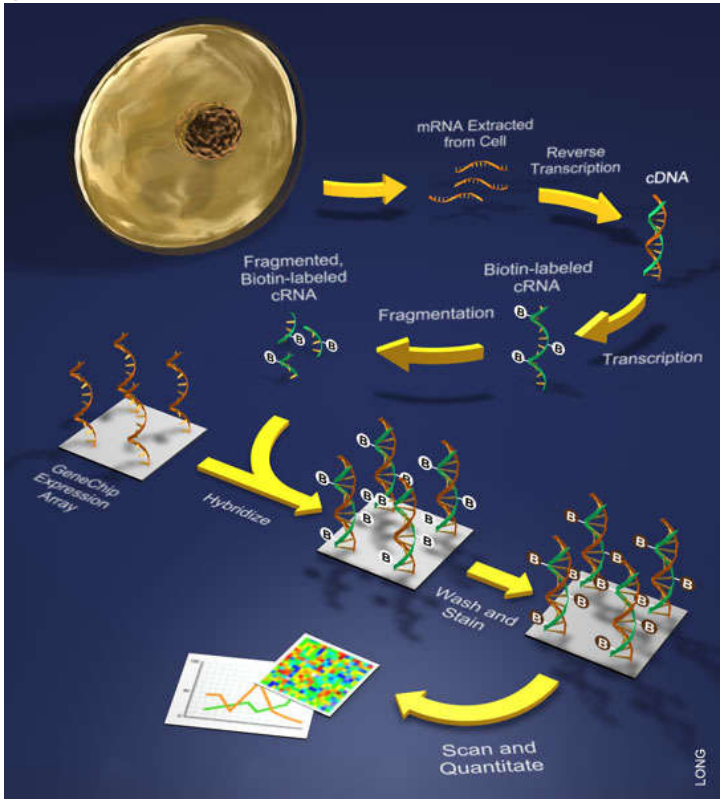
# Μονοκαναλικές Μικροσυστοιχίες

# Affymetrix GeneChip

- Η τεχνολογία GeneChip της Affymetrix παρέχει αποτελεσματική πρόσβαση σε γενετικές πληροφορίες γονιδιακής έκφρασης χρησιμοποιώντας μικρές σε μέγεθος συστοιχίες υψηλής πυκνότητας ολιγονουκλεοτιδικών ανιχνευτών. Η ανάλυση του προφίλ της γονιδιακής έκφρασης επιτρέπει την:
  - παρακολούθηση των αλλαγών στην έκφραση γονιδίων σε κανονικές ή παθολογικές βιολογικές καταστάσεις
  - επιβεβαίωση ή την αναγνώριση νέων φαρμακευτικών στόχων
  - αξιολόγηση τοξικολογικών προφίλ



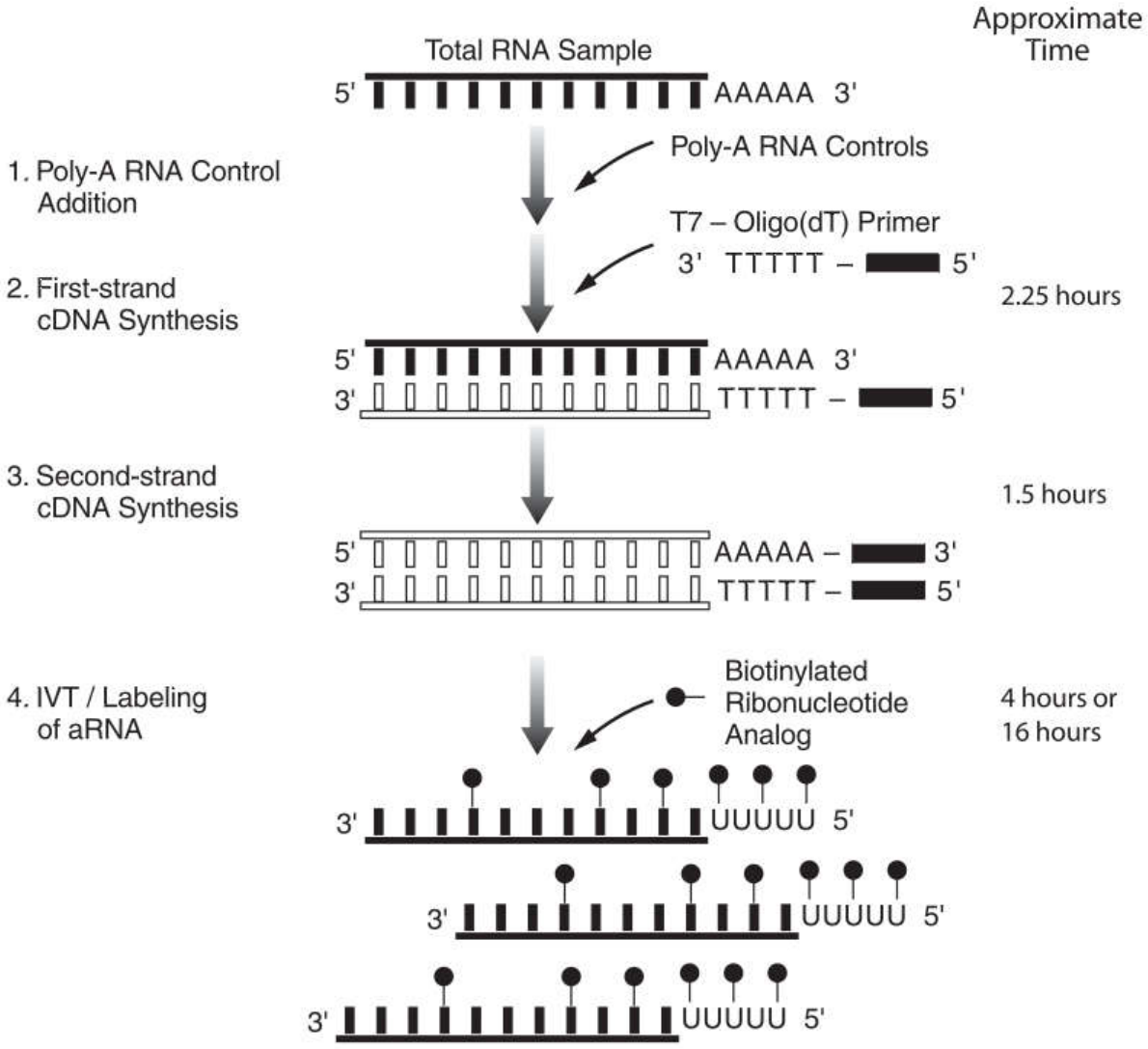
# Affymetrix GeneChip




- Η διαδικασία περιλαμβάνει:
  - Απομόνωση ολικού RNA
  - Αντίστροφη μεταγραφή σε cDNA
  - Μεταγραφή *in vitro* και βιοτινυλίωση
  - Κατακερματισμό
  - Υβριδοποίηση
  - Έκπλυση και φθορίζουσα χρώση
  - Οπτική σάρωση και ποσοτικοποίηση

# Πρωτόκολλα Affymetrix GeneChip

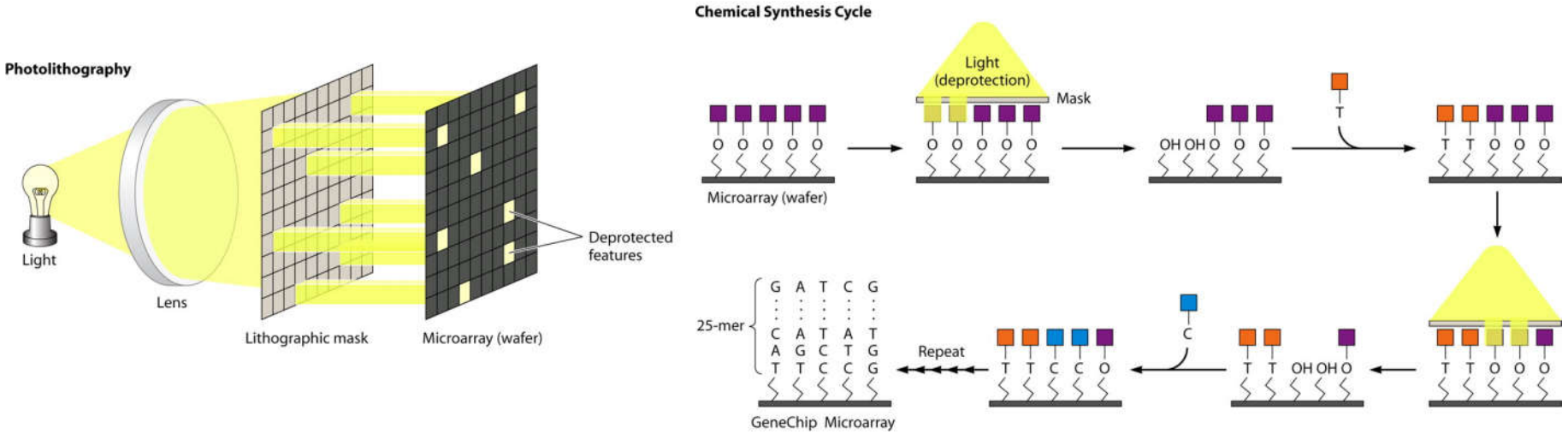
- Απομόνωση ολικού RNA
  - Ολικό RNA απομονώνεται από τα κύτταρα στόχους ή ιστούς χρησιμοποιώντας συμβατικά πρωτόκολλα και αξιολογείται η ποιότητά του
- Σύνθεση Δίκλωνου cDNA
  - Το RNA μετατρέπεται σε cDNA χρησιμοποιώντας εκκινητές oligo-dT που το τέλος τους έχουν έναν υποκινητή T7 για τη σύνθεση του πρώτου κλώνου DNA
  - Στη συνέχεια γίνεται η σύνθεση δεύτερου κλώνου cDNA
- Μεταγραφή *in vitro*
  - Το δίκλωνο cDNA χρησιμοποιείται ως καλούπι σε μια αντίδραση μεταγραφής *in vitro* (IVT) που καταλύεται από πολυμεράση T7 και που περιέχει βιοτινυλιωμένα CTP και UTP εκτός από τα τέσσερα μη τροποποιημένα τριφωσφορικά ριβονουκλεοσίδια
  - Το βιοτινυλιωμένο συμπληρωματικό RNA (cRNA) καθαρίζεται από το μίγμα της αντίδρασης IVT χρησιμοποιώντας στήλες καθαρισμού δείγματος
  - Η ποσότητα και η καθαρότητα του cRNA αξιολογείται φασματοφωτομετρικά
- Κατακερματισμός και υβριδισμός
  - Το βιοτινυλιωμένο cRNA κατακερματίζεται χημικά και το κατακερματισμένο cRNA προστίθεται σε ένα διάλυμα υβριδισμού που περιέχει αρκετά βιοτινυλιωμένα ολιγονουκλεοτίδια ελέγχου ( για έλεγχο ποιότητας), και υβριδοποιείται σε chip μικροσυστοιχιών ολονυκτίως στους 45°C.


Approximate Time
Total RNA Sample
5'
AAAAA 3'
1. Poly-A RNA Control Addition
Poly-A RNA Controls
T7 – Oligo(dT) Primer
3' TTTTT – 5'
2. First-strand cDNA Synthesis
2.25 hours
5'
AAAAA 3'
3'
TTTTT – 5'
3. Second-strand cDNA Synthesis
1.5 hours
5'
AAAAA – 3'
3'
TTTTT – 5'
4. IVT / Labeling of aRNA
Biotinylated Ribonucleotide Analog
4 hours or 16 hours
3'
UUUUU 5'
3'
UUUUU 5'
3'
UUUUU 5'

# Πρωτόκολλα Affymetrix GeneChip

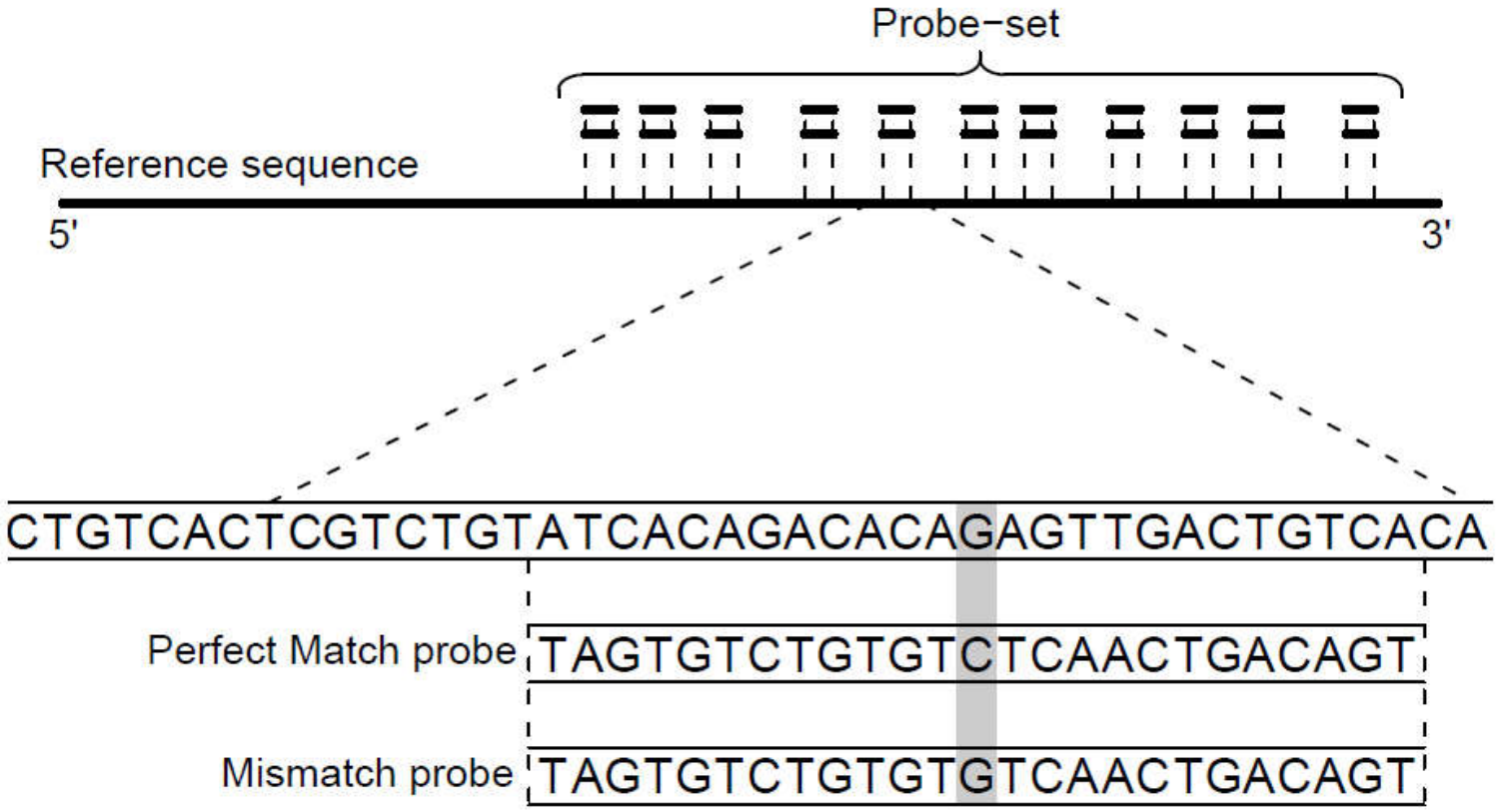
- Έκπλυση, φθορίζουσα χρώση και οπτική σάρωση του τσιπ
  - Τα τσιπ στη συνέχεια μεταφέρονται σε ένα όργανο τεχνολογίας ρευστών που εκτελεί εκπλύσεις για την απομάκρυνση του cRNA που δεν έχει υβριδοποιηθεί σε συμπληρωματικό του ολιγονουκλεοτιδικό ανιχνευτή
  - Το προσδεδεμένο cRNA σημαίνεται με φθορισμό χρησιμοποιώντας στρεπταβιδίνη συζευγμένη με φυκοερυθρίνη (SAPE)
  - Τα πλακίδια στη συνέχεια σαρώνονται: Κάθε cRNA προσδεδεμένο στο συμπληρωματικό του ολιγονουκλεοτίδιο διεγείρεται χρησιμοποιώντας ένα συνεστιακό σαρωτή λέιζερ, και καταγράφονται οι θέσεις και οι εντάσεις των φθοριζόντων εκπομπών
  - Ειδικό λογισμικό χρησιμοποιείται για να μετατρέπει τις πληροφορίες φθορισμού σε δεδομένα σχετικά με τα επίπεδα της γονιδιακής έκφρασης στο αρχικό δείγμα

# Φωτολιθογραφία




- Υπεριώδης ακτινοβολία διέρχεται μέσω της λιθογραφικής μάσκας που δρα ως φίλτρο είτε για να μεταδώσει είτε να μπλοκάρει την ακτινοβολία από τη χημικά προστατευμένη επιφάνεια της μικροσυστοιχίας
- Η διαδοχική εφαρμογή των ειδικών λιθογραφικών μασκών προσδιορίζει τη σειρά της σύνθεσης του ολιγονουκλεοτιδικού ανιχνευτή
- Κύκλος της χημικής σύνθεσης
  - Η υπεριώδης ακτινοβολία αφαιρεί τις προστατευτικές ομάδες (τετράγωνα) από την επιφάνεια του πίνακα, επιτρέποντας την προσθήκη ενός μόνο φωτοχημικά προστατευμένου νουκλεοτιδίου
  - Διαδοχικοί κύκλοι αποπροστασίας με ακτινοβολία, αλλαγής στο μοτίβο φιλτραρίσματος των λιθογραφικών μασκών, και προσθήκης ενός είδους μονονουκλεοτιδίων σχηματίζουν μικροσυστοιχίες με συγκεκριμένα 25μερή ολιγονουκλεοτίδια-ανιχνευτές

# Affymetrix GeneChip




- Οι εκατοντάδες χιλιάδες μονόκλωνοι ανιχνευτές μεγέθους 25 βάσεων, τοποθετούνται ανά ζεύγη, όπου κάθε ζεύγος αποτελείται από έναν ανιχνευτή με τέλεια αντιστοιχία (PM) και έναν με μερική αντιστοιχία (MM)
- Οι ανιχνευτές PM και MM είναι πανομοιότυποι, με εξαίρεση τη μεσαία βάση που είναι συμπληρωματική που είναι συμπληρωματικές
- Κάθε γονίδιο αντιστοιχεί σε ένα σύνολο ανιχνευτών (probe-set) που βρίσκονται σε διαφορετικά σημεία του πλακιδίου

# Affymetrix GeneChip

Παράδειγμα:

- 1415771_at:
  - Description: Mus musculus nucleolin mRNA, complete cds
  - LocusLink: AF318184.1 (Μήκος 2412 bp)
  - Αλληλουχία-στόχος 129 bp

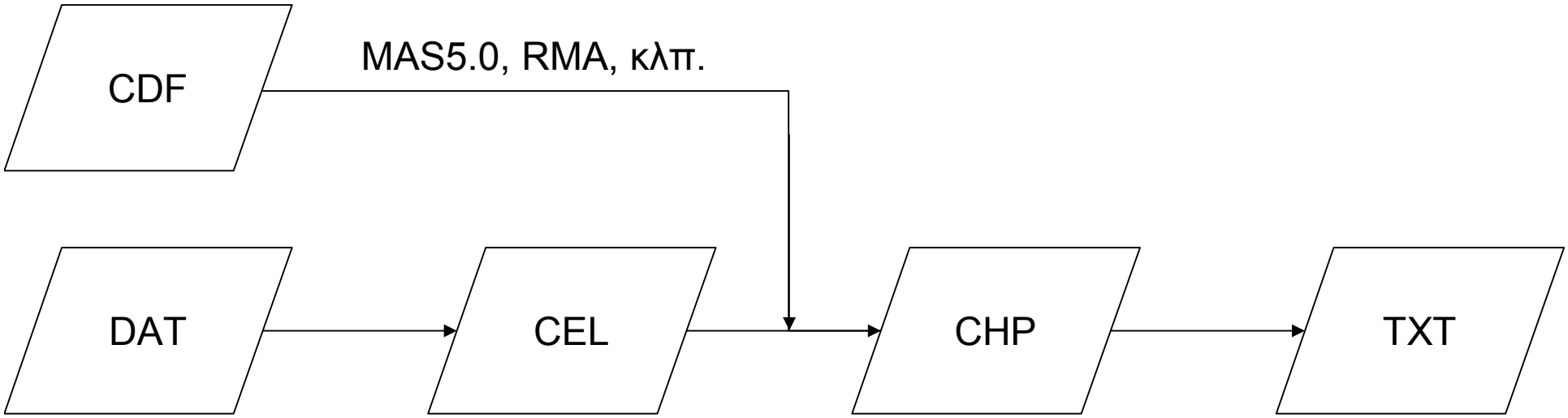
«Πλακόστρωση» της αλληλουχίας-στόχου από 11 ανιχνευτές

**gagaagtcaaccatccaaaactctg**tttgtcaaaggtctgtctgaggataccactgaagagaccttaaaagaatcatttgagggctctgttcgtgcaagaatagtcactgatcgggaaactggttctt
gagaa**gtcaaccatccaaaactctgtttgt**caaaggtctgtctgaggataccactgaagagaccttaaaagaatcatttgagggctctgttcgtgcaagaatagtcactgatcgggaaactggttctt
gagaagtcaaccatccaa**aactctgtttgtcaaaggtctgtct**gaggataccactgaagagaccttaaaagaatcatttgagggctctgttcgtgcaagaatagtcactgatcgggaaactggttctt
gagaagtcaaccatccaaaactctgtttgtca**aaggtctgtctgaggataccactga**agagaccttaaaagaatcatttgagggctctgttcgtgcaagaatagtcactgatcgggaaactggttctt
gagaagtcaaccatccaaaactctgtttgtcaaaggt**ctgtctgaggataccactgaagaga**ccttaaaagaatcatttgagggctctgttcgtgcaagaatagtcactgatcgggaaactggttctt
gagaagtcaaccatccaaaactctgtttgtcaaaggtctgtctgaggat**accactgaagagaccttaaaagaat**catttgagggctctgttcgtgcaagaatagtcactgatcgggaaactggttctt
gagaagtcaaccatccaaaactctgtttgtcaaaggtctgtctgaggataccactgaagagaccttaaa**agaatcatttgagggctctgttcgt**gcaagaatagtcactgatcgggaaactggttctt
gagaagtcaaccatccaaaactctgtttgtcaaaggtctgtctgaggataccactgaagagaccttaaaagaatcat**ttgagggctctgttcgtgcaagaat**agtcactgatcgggaaactggttctt
gagaagtcaaccatccaaaactctgtttgtcaaaggtctgtctgaggataccactgaagagaccttaaaagaatcatttgag**ggctctgttcgtgcaagaatagtca**ctgatcgggaaactggttctt
gagaagtcaaccatccaaaactctgtttgtcaaaggtctgtctgaggataccactgaagagaccttaaaagaatcatttgagggctct**gttcgtgcaagaatagtcactgatc**gggaaactggttctt
gagaagtcaaccatccaaaactctgtttgtcaaaggtctgtctgaggataccactgaagagaccttaaaagaatcatttgagggctctgttcgtgcaagaata**gtcactgatcgggaaactggttctt**

# Affymetrix GeneChip

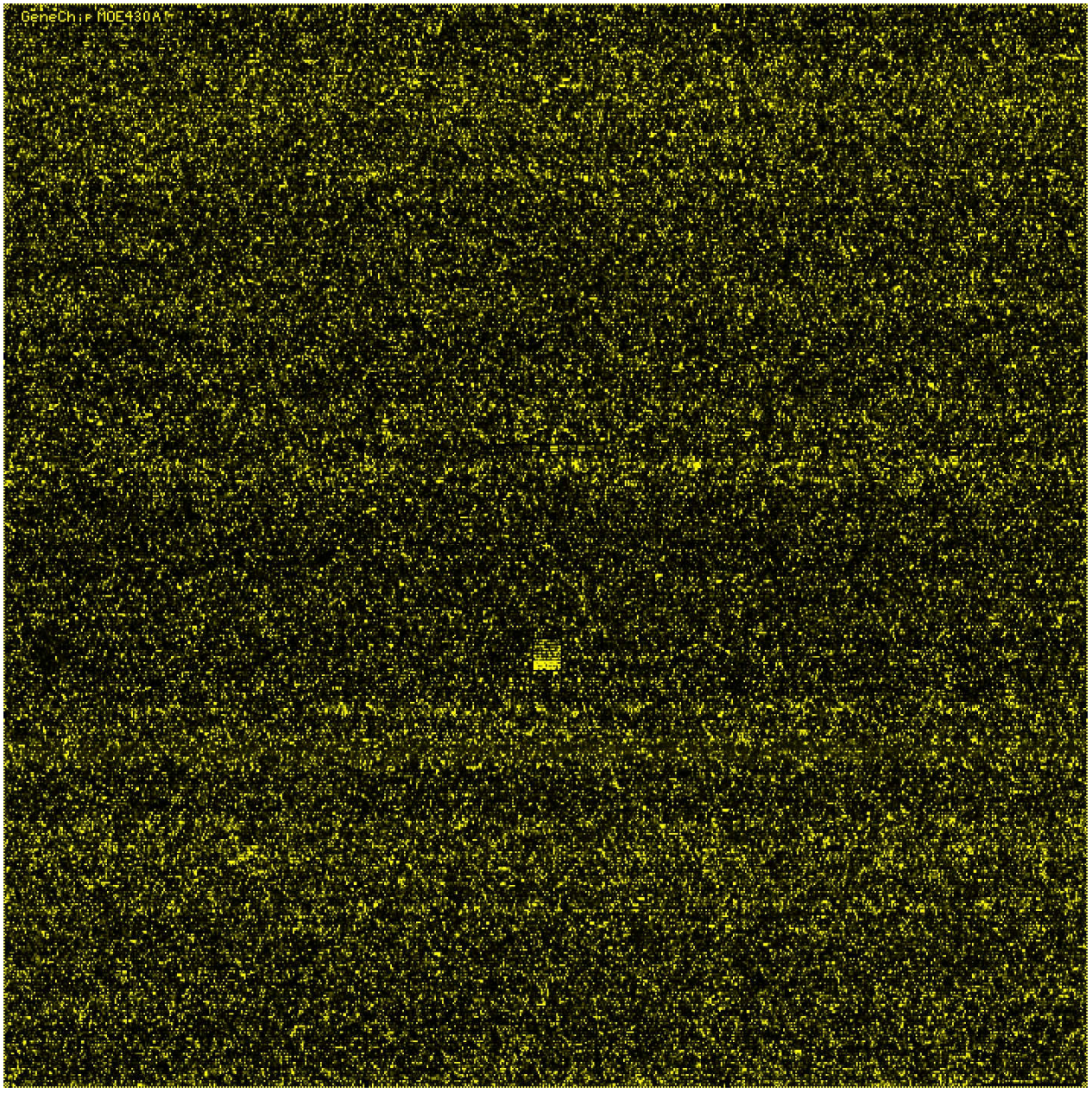
- Το πρόθεμα AFFX έχουν σύνολα ανιχνευτών που είναι για εσωτερικούς ελέγχους της Affymetrix και γενικά δε χρησιμοποιούνται σε αναλύσεις
- Οι καταλήξεις των συνόλων ανιχνευτών στο εξ ορισμού CDF σημαίνουν:
- _at: Υβριδοποιείται με ένα μοναδικό μετάγραφο (antisense)
- _s_at: Όλοι οι ανιχνευτές του συνόλου υβριδοποιούνται με συγκεκριμένο σύνολο αλληλουχιών
- _a_at: Όλοι οι ανιχνευτές του συνόλου υβριδοποιούνται με συγκεκριμένη οικογένια γονιδίων
- _x_at: Τουλάχιστον μερικοί ανιχνευτές υβριδοποιούνται σταυρωτά με άλλες αλληλουχίες-στόχους του πλακιδίου
- …

# Αρχεία Affymetrix

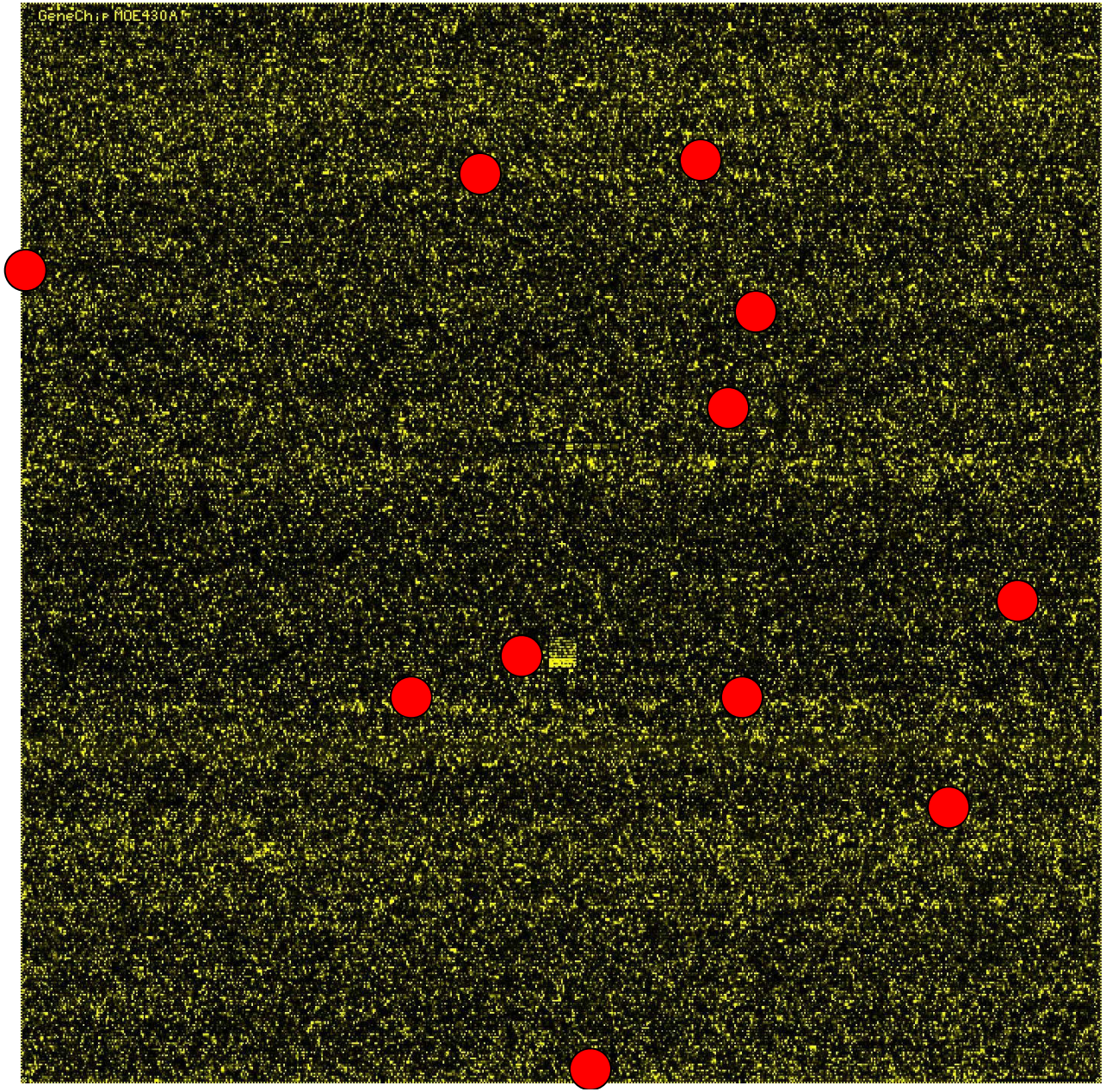



- DAT: Μη επεξεργασμένη οπτική εικόνα του υβριδοποιημένου πλακιδίου
- CDF: Διάταξη του πλακιδίου (από την Affymetrix ή άλλες πηγές)
- CEL: Επεξεργασμένο αρχείο DAT (τιμές έντασης-θέσης)
- CHP: Πειραματικά αποτελέσματα από το συνδυασμό των αρχείων CEL και CDF
- TXT: Τιμές έκφρασης κάθε συνόλου ιχνηθέτη (αρχείο CHP σε διαμόρφωση αρχείου)

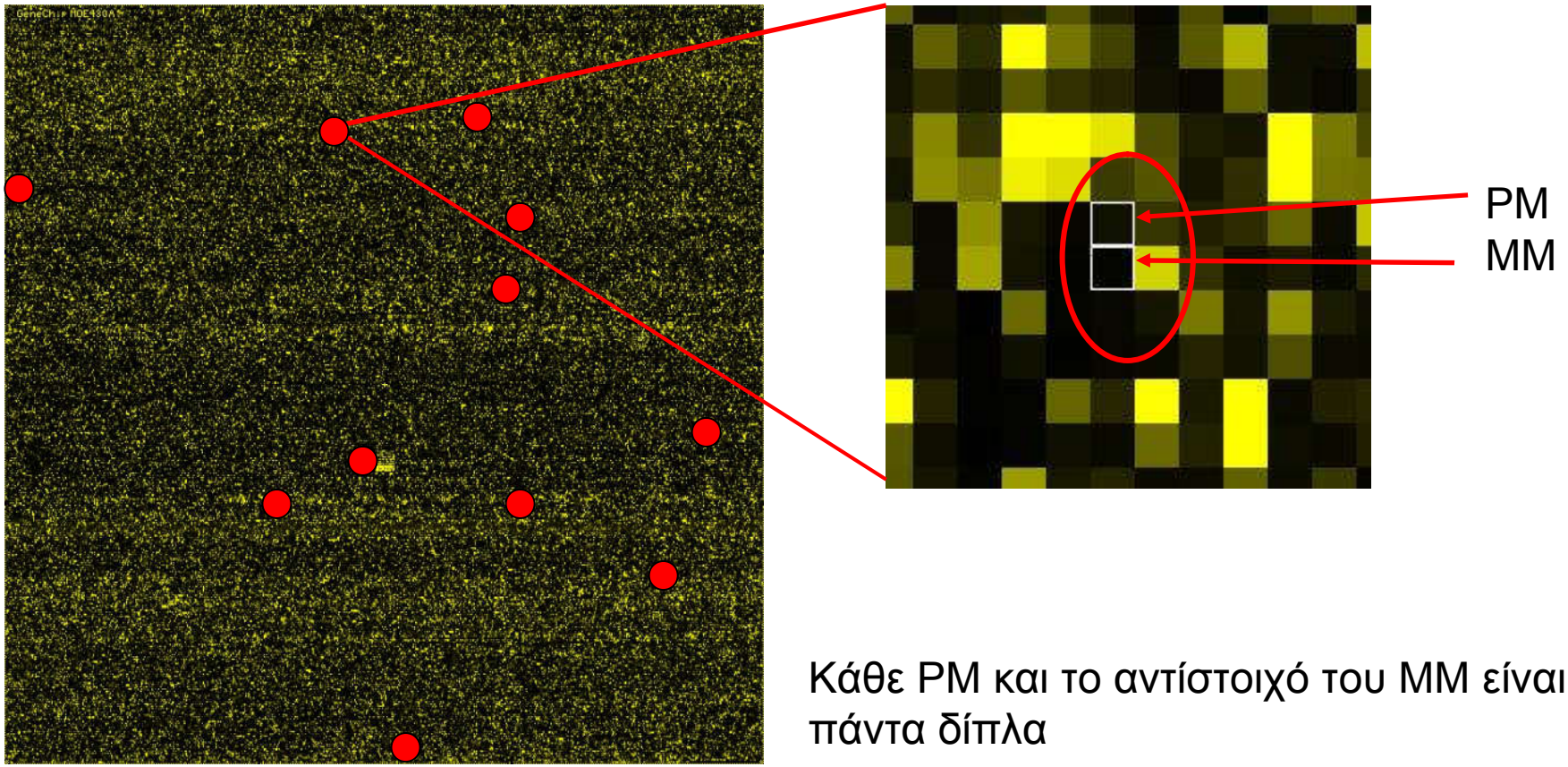
# Ψευδοεικόνα Affymetrix Chip (DAT)



# 1415771_at στο MOE430A

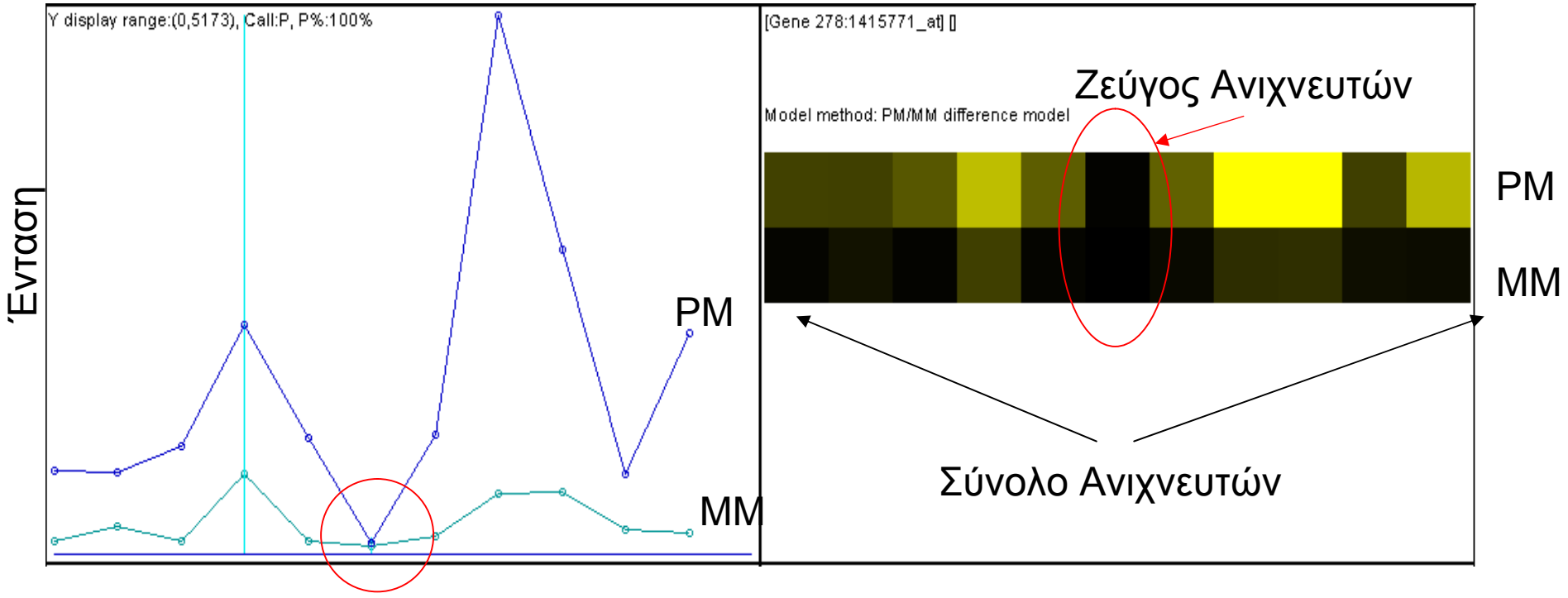


# 1415771_at στο MOE430A




Κάθε PM και το αντίστοιχό του MM είναι πάντα δίπλα

# 1415771_at στο MOE430A



# Ανάλυση Δεδομένων Affymetrix

- Κανονικοποίηση (αφαίρεση μη βιολογικών στοιχείων του σήματος)
  - Αφαίρεση Υποβάθρου
  - Σύνοψη
  - Κλίμακα
  - (Λογαρίθμηση)
- Αλγόριθμοι
  - MAS5.0 (Affymetrix)
  - RMA/GCRMA

# Αλγόριθμος One-Step Tukey

- Υπολογίζουμε τη διάμεσο M των τιμών $x_i$, $i$=1,…,$n$
- Υπολογίζουμε την Απόλυτη Απόκλιση της Διαμέσου S, ως τη διάμεσο των απολύτων τιμών των αποστάσεων από την M
- Η ομοιόμορφη μέτρηση της απόστασης κάθε τιμής $x_i$ από το κέντρο, είναι:

$$u_i = \frac{x_i - M}{cS + \varepsilon}, i = 1,...,n$$

όπου: $c$ (προεπιλογή $c$ = 5) και $\varepsilon$ μια πολύ μικρή τιμή για την αποφυγή της διαιρέσεως δια του μηδέν (προεπιλογή $\varepsilon$ = 0.0001)

- Υπολογίζουμε τα βάρη:

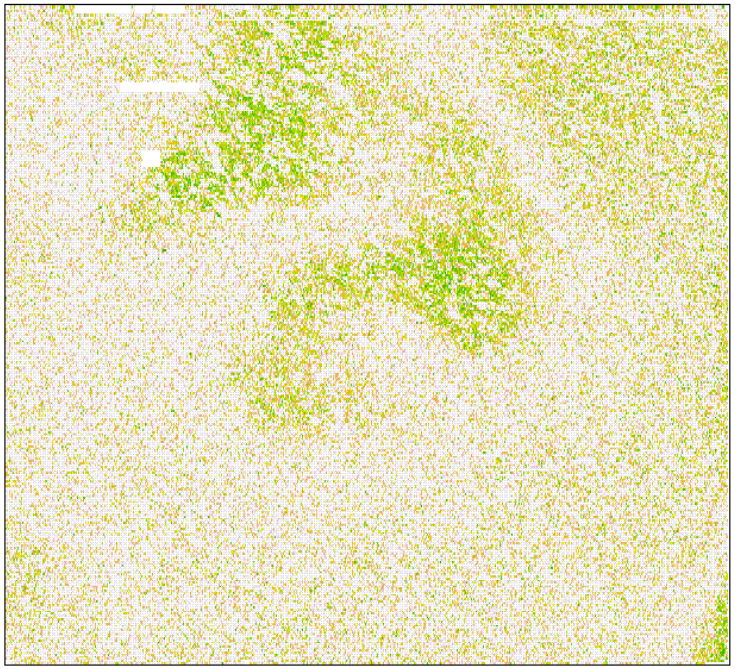
$$w(u_i) = \begin{cases} \left(1 - u_i^{\,2}\right)^2, \left|\mathrm{u_i}\right| \le 1 \\ 0, \left|\mathrm{u_i}\right| > 1 \end{cases}$$

- Υπολογίζουμε τη βεβαρημένη μέση τιμή:

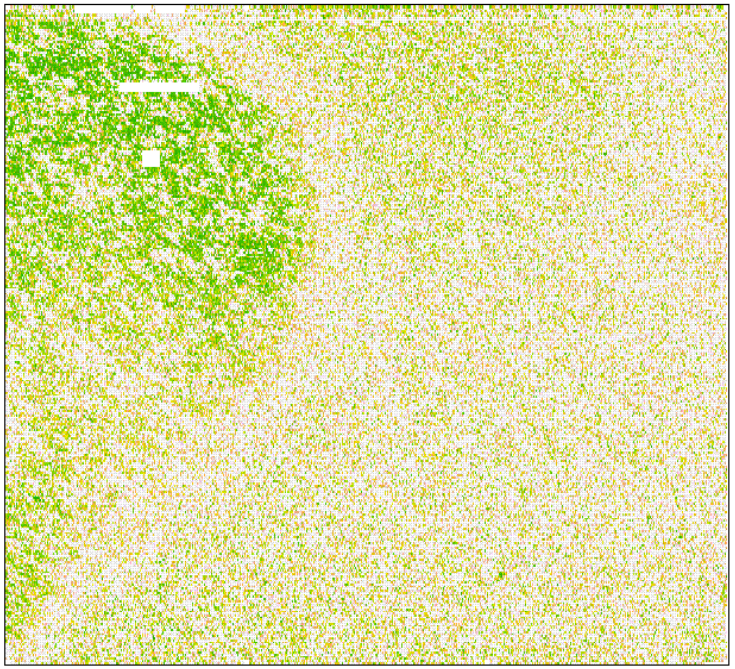
$$T_{bi} = \frac{\sum_{i=1}^{n} w(u_i) x_i}{\sum_{i=1}^{n} w(u_i)}$$

# Affymetrix

2353a99hpp_av08.cel

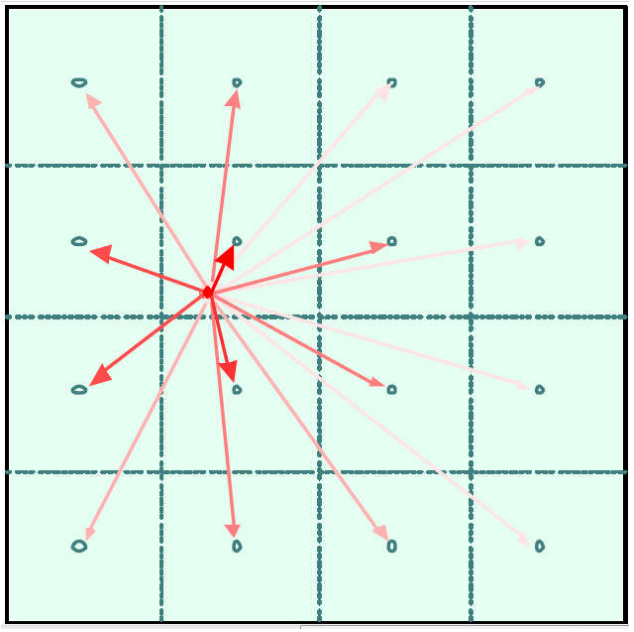


2353p99hpp_av08.cel



# MAS5

**Διόρθωση Υποβάθρου**
Υπολογίζεται μια κατώτατη τιμή υποβάθρου η οποία αφαιρείται από κάθε τιμή έντασης ανιχνευτή.



**Τιμές Ζωνών**

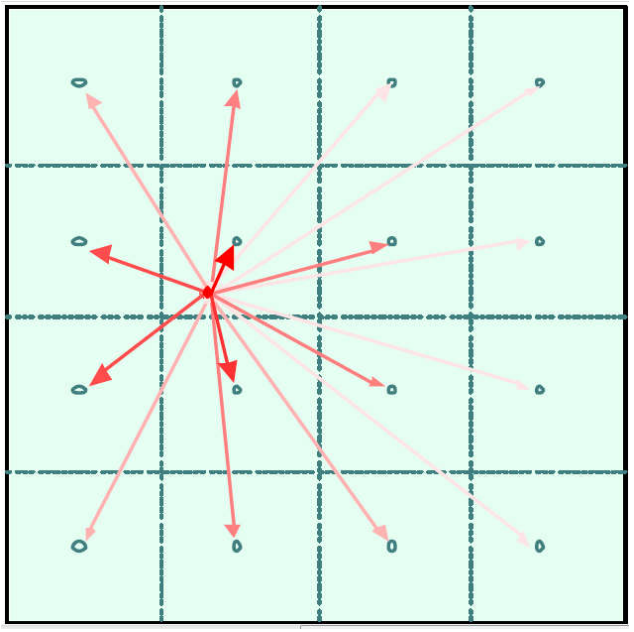
- Για τον υπολογισμό των τιμών υποβάθρου, η μικροσυστοιχία διαιρείται σε ***K*** ορθογώνιες ζώνες ***Z_k*** (όπου ***k*** = 1, ..., ***K***) (προεπιλεγμένη τιμή ***K***=16)
- Οι καλυμμένοι ανιχνευτές και οι ανιχνευτές ελέγχου δεν χρησιμοποιούνται στον υπολογισμό των τιμών υποβάθρου
- Το υπόβαθρο ***b*** της ζώνης ***K*** (***bZ***$_k$) υπολογίζεται ως το 2-οστό ποσοστημόριο των τιμών των εντάσεων των ανιχνευτών της ζώνης αυτής.
- Η τυπική απόκλιση των εντάσεων που είναι χαμηλότερες από την υπολογισμένη τιμή υποβάθρου, χρησιμοποιείται ως εκτίμηση της μεταβλητότητας ***n*** του υποβάθρου της ζώνης ***K*** (***nZ***$_k$).

# MAS5

**Προσαρμογή Εξομάλυνσης**

•Για να υπάρξει μια ομαλή μετάβαση μεταξύ των ζωνών, υπολογίζουμε τις αποστάσεις κάθε ανιχνευτή συντεταγμένων x,y στην μικροσυστοιχία από τα διάφορα κέντρα των ζωνών

•Στη συνέχεια υπολογίζεται ένα σταθμισμένο άθροισμα $W_k(x,y)$ ως το αντίστροφο του αθροίσματος του τετραγώνου της απόστασης από το κέντρο της ζώνης ***K*** και μίας σταθεράς εξομάλυνσης (που προστίθεται για να εξασφαλιστεί ότι ο παρονομαστής δεν θα έχει ποτέ μηδενική τιμή):



$$w_k(x,y) = \frac{1}{d_k^2(x,y) + smooth}$$

•Για κάθε ανιχνευτή με συντεταγμένες x,y η τιμή υποβάθρου ***b*** (σταθερά εξομάλυνσης= 100) υπολογίζεται ως εξής:

$$b(x,y) = \frac{1}{\sum_{k=1}^{K} w_k(x,y)} \sum_{k=1}^{K} w_k(x,y) bZ_k$$

# MAS5

**Διόρθωση Θορύβου**

•Υπολογίζουμε μια προσαρμοσμένη τιμή, που θα ελαττώσει τις εντάσεις βάσει του τοπικού υποβάθρου. Για να γίνει αυτό, πρέπει πρώτα να εξασφαλιστεί ότι οι τιμές δεν θα είναι αρνητικές ή πολύ μικρές. Οι αρνητικές τιμές έντασης είναι προβληματικές στους μετέπειτα υπολογισμούς όπου υπολογίζονται οι λογαριθμικές τιμές.

•Για τη διόρθωση θορύβου, για κάθε ανιχνευτή υπολογίζεται μια τοπική τιμή θορύβου n(x, y) ως εξής:

$$n(x,y) = \frac{1}{\sum_{k=1}^{K} w_k(x,y')} \sum_{k=1}^{K} w_k(x,y) nZ_k$$

- Στη συνέχεια, ορίζεται ένα όριο ως κλάσμα της τοπικής τιμής θορύβου, ώστε καμία τιμή να μην προσαρμόζεται κάτω από αυτό το όριο. Δηλαδή, για μια ένταση ανιχνευτή ***I'***(x, y) στις συντεταγμένες (x, y), υπολογίζουμε μια προσαρμοσμένη ένταση:

$$A(x,y) = \max(I'(x,y) - b(x,y), NoiseFrac * n(x,y))$$

$$\text{όπου: } I'(x,y) = \max(I'(x,y), 0.5)$$

Το επιλεγμένο κλάσμα της ολικής διακύμανσης υποβάθρου είναι το *NoiseFrac*. (προεπιλογή *NoiseFrac* = 0.5)

# MAS5

**Ιδανικό ατελές ταίριασμα (IM)**

•Υπολογίζουμε με τον αλγόριθμο one-step biweight (***$T_{bi}$***) το ειδικό υπόστρωμα biweight $SB_i$ στο σύνολο ανιχνευτών ***i*** (όπου ***j*** το ζεύγος ανιχνευτών):

$$SB_i = T_{bi}\left(\log_2\left(PM_{i,j}\right) - \log_2\left(MM_{i,j}\right) : j = 1,\ldots,n_i\right)$$

- Το ιδανικό ατελές ταίριασμα IM για το *j* ζεύγος ανιχνευτών στο σύνολο ανιχνευτών *i* υπολογίζεται ως εξής:

$$IM_{i,j}\begin{cases} MM_{i,j}, & MM_{i,j} < PM_{i,j} \\ \dfrac{PM_{i,j}}{2^{(SB_i)}}, & MM_{i,j} \geq PM_{i,j}\ and\ \mathrm{SB_i} > contrast\tau \\ \dfrac{PM_{i,j}}{2^{\left(\frac{contrast\tau}{1+\left(\frac{contrast\tau - SB_i}{scale\tau}\right)}\right)}}, & MM_{i,j} \geq PM_{i,j}\ and\ SB_i \leq contrast\tau \end{cases}$$

Scaleτ είναι το σημείο αποκοπής που περιγράφει την μεταβλητότητα των ζευγών ανιχνευτών στο σύνολο ανιχνευτών (προεπιλεγμένο *contrastτ*=0.03 και *scaleτ*=10).

# MAS5

**Τιμή ανιχνευτή και λογαριθμική τιμή σήματος**

•Με δεδομένο το ιδανικό ατελές ταίριασμα, ο τύπος για την τιμή του ανιχνευτή (PV):

$$V_{i,j} = \max\left(PM_{i,j} - IM_{i,j}, d\right)$$

προεπιλεγμένο d = $2^{-20}$

•Τώρα υπολογίζουμε την τιμή του ανιχνευτή PV για κάθε ζεύγος ανιχνευτών *j* στο σύνολο ανιχνευτών *i*. Το πλήθος των ζευγών ανιχνευτών σε κάθε σύνολο ανιχνευτών είναι ο αριθμός *n*:

$$PV_{i,j} = \log_2\left(V_{i,j}\right), j = 1,\ldots n_i$$

- Υπολογίζουμε την απόλυτη τιμή έκφρασης για το σύνολο ανιχνευτών ***i*** ως την εκτίμηση one-step biweight των ***i n*** αναπροσαρμοσμένων τιμών ανιχνευτών:

$$SignalLogValue_i = {}'T_{bi}\left(PV_{i,1},\ldots,PV_{i,n_i}\right)$$

# MAS5

**Κλιμακούμενη τιμή Ανιχνευτή**

•Εάν ο αλγορίθμου είναι ρυθμισμένος για κανονικοποίηση σε όλα τα σύνολα ανιχνευτών ή σε επιλεγμένα σύνολα ανιχνευτών, υπολογίζουμε έναν παράγοντα κλιμάκωσης (***sf***):

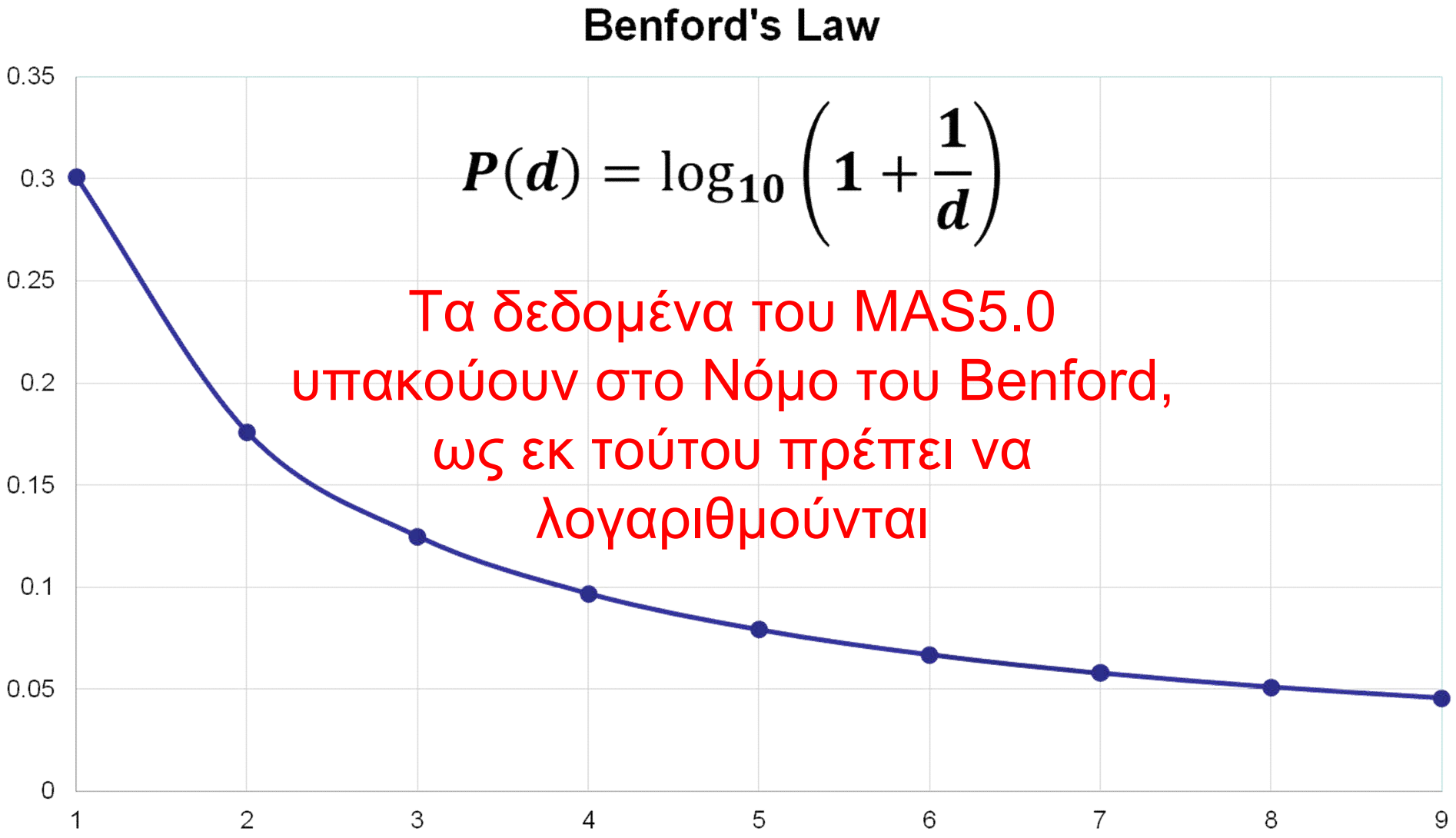
$$sf = \frac{Sc}{TrimMean\left(2^{SignalLogValue_i}, 0.02, 0.98\right)}$$

όπου ***Sc*** είναι το σήμα-στόχος (προεπιλογή ***Sc*** = 500)

- Η αναφερόμενη τιμή του συνόλου ανιχνευτών *i* είναι:

$$ReportedValue(i) = sf \cdot 2^{\left(SignalLogValue_i\right)}$$

# $\log_2$-tranformation



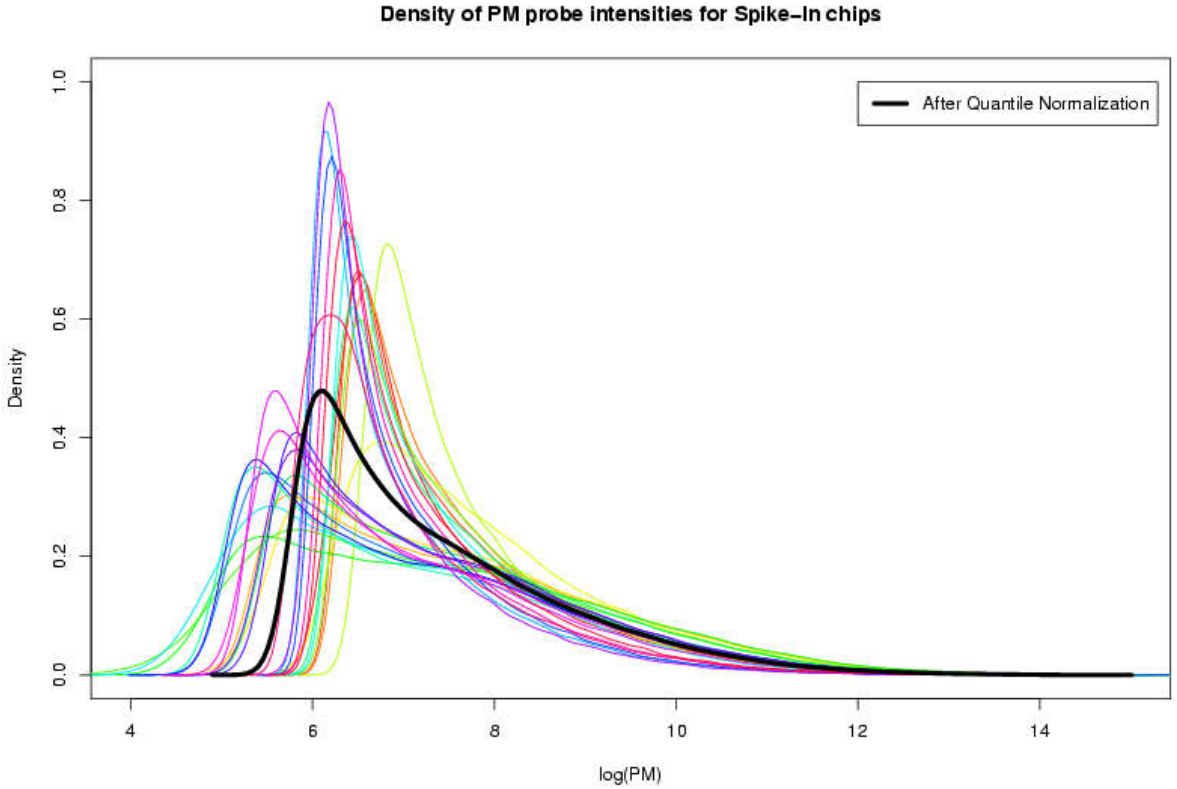
| Probe | Control | Tumour | Difference | Relative Difference | ratio | $Log_2$ ratio |
|---|---|---|---|---|---|---|
| A | 1 | 4 | 3 | +300% | 4 | 2 |
| B | 4 | 1 | -3 | -75% | 0.25 | -2 |
| C | 2 | 8 | 6 | +300% | 4 | 2 |
| D | 8 | 2 | -6 | -75% | 0.25 | -2 |

# MAS5

- Θετικά:
  - Χρησιμοποιείται και για μοναδικά πλακίδια (αν και αντίγραφα οι προτιμώνται)
  - Μπορεί να δώσει p-value για τα δεδομένα έκφρασης
- Αρνητικά
  - Πολλοί διορθωτικοί παράγοντες στον αλγόριθμο
  - Όχι τόσο αναπαραγώγιμα αποτελέσματα
- Διάφορα
  - Πιο διαδεδομένη μέθοδος στα πλακίδια Affymetrix
  - Εξαρτώμενη από τα MM

# Robust Multichip Analysis (RMA)

- Χρησιμοποιείται για >3 πλακίδια (όσο περισσότερο, τόσο καλύτερα)
- Θεωρεί ότι όλα τα πλακίδια έχουν το ίδιο υπόβαθρο και κατανομή τιμών (?)
- Δε χρησιμοποιεί τους ανιχνευτές MM γιατί το PM-MM* οδηγεί σε μεγάλη διασπορά
  - Οι μισοί ανιχνευτές δεν χρησιμοποιούνται, παρόλα αυτά ο αλγόριθμος δίνει καλά αποτελέσματα
- Το να αγνοηθούν οι MM μειώνει την ακρίβεια και αυξάνει την πιστότητα
- Μια παραλλαγή του RMA, το GCRMA, λαμβάνει τα MM υπ' όψιν του


Density of PM probe intensities for Spike–In chips
After Quantile Normalization
Density
0.0
0.2
0.4
0.6
0.8
1.0
4
6
8
10
12
14
log(PM)

# Κανονικοποίηση Ποσοστιμορίου

| | Ω |
|---|---|
| A | 5 |
| B | 7 |
| C | 3 |
| D | 6 |
| E | 2 |

Καθορισμός σειράς

| | Ω |
|---|---|
| A | iii |
| B | v |
| C | ii |
| D | iv |
| E | i |

Αντικατάσταση τιμών

| | Ω |
|---|---|
| A | 7 |
| B | 12 |
| C | 4 |
| D | 8 |
| E | 3 |

Κατανομή Αναφοράς

| | |
|---|---|
| i | 3 |
| ii | 4 |
| iii | 7 |
| iv | 8 |
| v | 12 |

Για να κανονικοποιηθεί μια δοκιμαστική κατανομή σε κατανομή αναφοράς ιδίου μήκους, ταξινομούνται οι δύο κατανομές κατά αύξουσα σειρά. Η υψηλότερη τιμή της δοκιμαστικής κατανομής λαμβάνει στη συνέχεια την αξία της υψηλότερης τιμής της κατανομής αναφοράς, η επόμενη υψηλότερη τιμή της δοκιμαστικής κατανομής την επόμενη υψηλότερη της κατανομής αναφοράς, κοκ, ώσπου η δοκιμαστική κατανομή είναι μια αναδιάταξη της κατανομής αναφοράς.

# Κανονικοποίηση Ποσοστιμορίου

| | X | Y | Z |
|---|---|---|---|
| A | 9 | 8 | 11 |
| B | 1 | 7 | 16 |
| C | 4 | 3 | 6 |
| D | 2 | 11 | 7 |
| E | 5 | 2 | 10 |

Καθορισμός σειράς

| | X | Y | Z |
|---|---|---|---|
| A | v | iv | iv |
| B | i | iii | v |
| C | iii | ii | i |
| D | ii | v | ii |
| E | iv | i | iii |

Αντικατάσταση τιμών

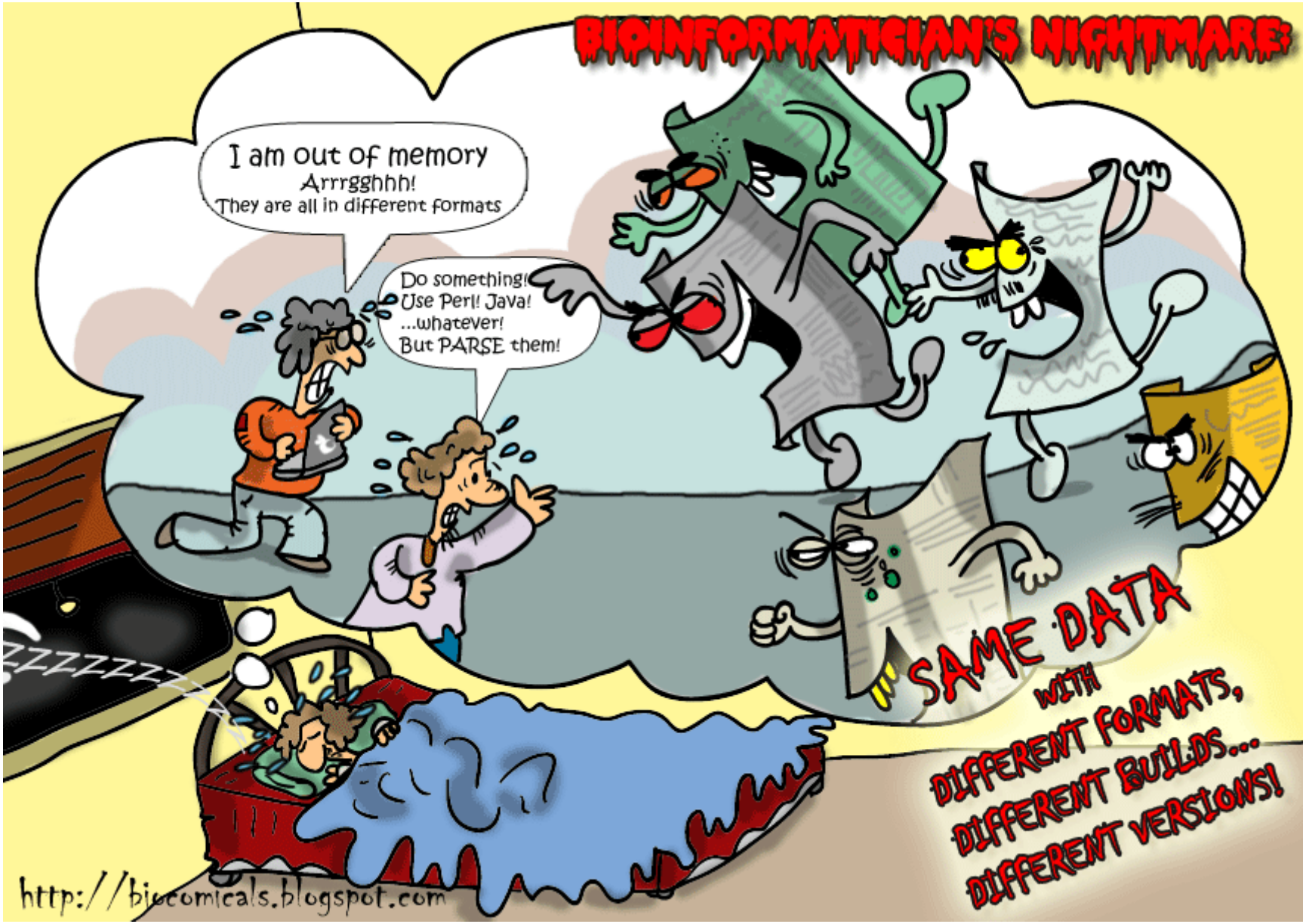
| | X | Y | Z |
|---|---|---|---|
| A | 12 | 8 | 8 |
| B | 3 | 7 | 12 |
| C | 7 | 4 | 3 |
| D | 4 | 12 | 4 |
| E | 8 | 3 | 7 |

Αναδιάταξη

| | X | Y | Z | |
|---|---|---|---|---|
| i | 1 | 2 | 6 | 3 |
| ii | 2 | 3 | 7 | 4 |
| iii | 4 | 7 | 10 | 7 |
| iv | 5 | 8 | 11 | 8 |
| v | 9 | 11 | 16 | 12 |

Δημιουργεί πανομοιότυπες κατανομές ως προς τις στατιστικές τους ιδιότητες

Για να κανονικοποιηθούν δύο ή περισσότερες κατανομές μεταξύ τους, χωρίς κατανομή αναφοράς, οι κατανομές ταξινομούνται κατά αύξουσα σειρά, και μετά ορίζεται ο μέσος όρος (συνήθως, ο αριθμητικός) των κατανομών. Έτσι, η υψηλότερη τιμή σε όλες τις περιπτώσεις γίνεται ο μέσος όρος των υψηλοτέρων τιμών, η δεύτερη υψηλότερη τιμή γίνεται ο μέσος όρος των τιμών της δεύτερη υψηλότερης τιμής, κοκ.


BIOINFORMATICIAN'S NIGHTMARE:
I am out of memory
Arrrgghhh!
They are all in different formats
Do something!
Use Perl! Java!
...whatever!
But PARSE them!
SAME DATA
WITH
DIFFERENT FORMATS,
DIFFERENT BUILDS...
DIFFERENT VERSIONS!
http://biocomicals.blogspot.com

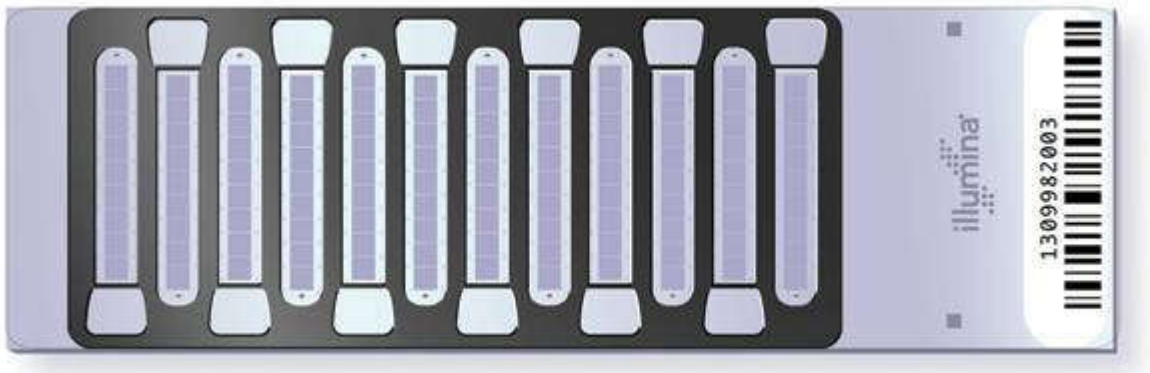
# Πηγές Πολυπλοκότητας

- Διαφορετικές εμπορικές μάρκες και πλατφόρμες της ίδιας μάρκας
  - Affymetrix
    - Affymetrix AtGenome1
    - Affymetrix ATH1
  - Illumina
  - Agilent
- Διαφορετικά Αποθετήρια
  - GEO
  - NASCArrays
- Διαφορετικοί Αλγόριθμοι
  - MAS5.0
  - RMA (Quantile normalisation)

- Διαφορετικά αρχεία περιγραφής πλακιδίων (CDFs)
  - Εξ ορισμού Affymetrix CDF (outdated)
  - Προσαρμοσμένα CDFs
    - Διαφορετικοί κατασκευαστές CDF
      - Διαφορετικές εκδόσες CDF του ίδιου κατασκευαστή (Brainarray)
        - Ετήσιες Αναβαθμίσεις
        - Διαφορετικοί ορισμοί συνόλων ανιχνευτών
          - Πηγές
            - TAIR
            - Entrez
            - Vega
          - Σύνολα Ανιχνευτών
            - Γονίδιο
            - Μετάγραφο

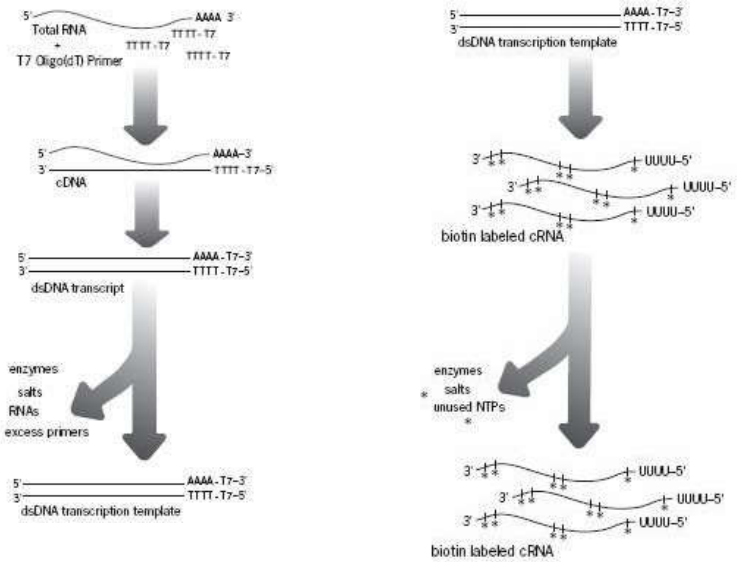
- Λύση
  - Επιλογή της δημοφιλέστερης πλατφόρμας (Affymetrix ATH1)
  - Χρήση της ένωσης των δεδομένων όλων των αποθετηρίων
  - Κανονικοποίηση των πρωτογενών δεδομένων (CEL)
  - Χρήση του τελευταίου Brainarray CDF που βασίζεται πχ στους ορισμούς των γονιδίων κατά TAIR

# Illumina BeadChip



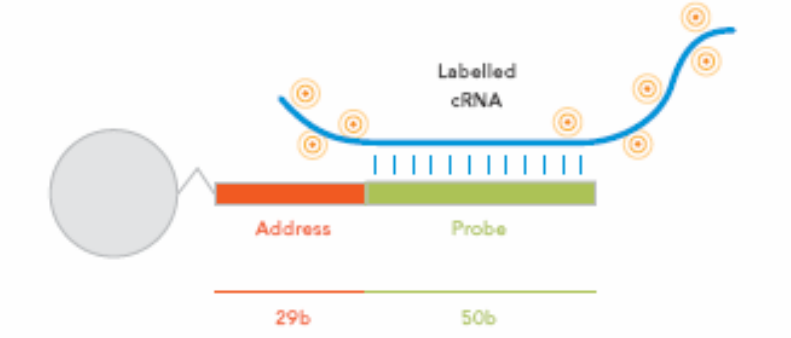

Η Illumina BeadChip είναι μια σχετικά νέα μέθοδος για την εκτέλεση πολλαπλής γονιδιακής ανάλυσης. Το βασικό στοιχείο της τεχνολογίας BeadChip είναι η επισύναψη ολιγονουκλεοτιδίων σε πυριτικά σφαιρίδια. Τα σφαιρίδια στη συνέχεια αποτίθενται τυχαία σε φρεάτια σε ένα υπόστρωμα (για παράδειγμα, ένα γυάλινο πλακίδιο). Η προκύπτουσα συστοιχία αποκωδικοποιείται για να καθοριστεί ποιος συνδυασμός ολιγονουκλεοτιδίου-σφαιριδίου είναι σε ποιο φρεάτιο. Οι αποκωδικοποιημένες συστοιχίες μπορούν να χρησιμοποιηθούν σε μια σειρά εφαρμογών, συμπεριλαμβανομένων της ανάλυσης γονιδιακής έκφρασης, γονότυπου, μεθυλώματος, κλπ.

# Πρωτόκολλο Μικροσυστοιχιών Έκφρασης Illumina



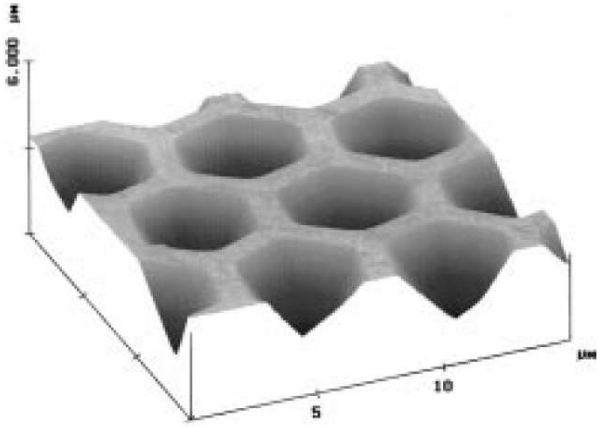

- Υβριδισμός ενός ολιγονουκλεοτιδίου oligo-dT προς το κλάσμα polyA του ολικού RNA. Το ολιγονουκλεοτίδιο φέρει επίσης την ακολουθία ενός υποκινητή της ιικής πολυμεράσης RNA T7.
- Επέκταση του cDNA και στη συνέχεια, σύνθεση της συμπληρωματικής αλυσίδας για την παραγωγή δίκλωνου cDNA.
- Πρόσθεση πολυμεράσης RNA T7 και νουκλεοτιδίων για να ενισχυθεί γραμμικά το RNA. Το νεοπαραγόμενο cRNA ενσωματώνει UTP σεσημασμένο με βιοτίνη.
- Υβριδοποίηση του σεσημασμένου με βιοτίνη cRNA στο BeadChip.
- Χρώση του BeadChip με συνδεδεμένη σε στρεπταβιδίνη, Κυανίνη 3 (Cy3).
- Σάρωση υψηλής ανάλυσης σε σαρωτή Illumina BeadStation.

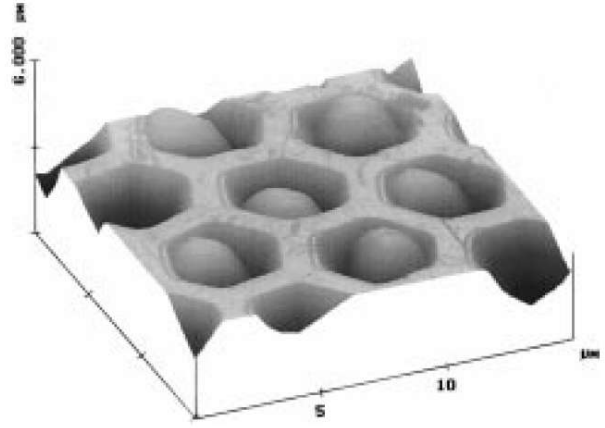
# Illumina BeadChip

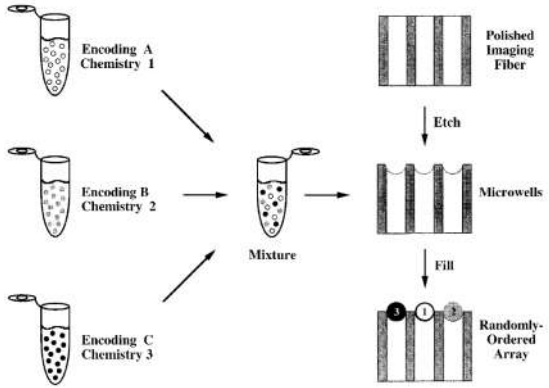



Η ανάλυση γονιδιακής έκφρασης εκτελείται χρησιμοποιώντας ένα ολιγονουκλεοτίδιο 79 βάσεων που έχει δύο τμήματα: Το τμήμα 5' του ολιγονουκλεοτιδίου, 50 βάσεων, έχει σχεδιαστεί για να υβριδοποιείται με αλληλουχίες που διατίθενται σε κοινόχρηστα αποθετήρια δεδομένων. Είναι το τμήμα που προσδένεται στον σημασμένο στόχο που προέρχεται από το κλάσμα poly(A) του συνολικού RNA. Το τμήμα 3' του ολιγονουκλεοτιδίου, 29 βάσεων, είναι η «διεύθυνση». Η διεύθυνση είναι μία μοναδική αλληλουχία που δημιουργήθηκε από την Illumina ειδικά για να επιτρέπει την αναμφισβήτητη ταυτοποίηση του ολιγονουκλεοτιδίου αφού έχει αποτεθεί επί της συστοιχίας.
Κάθε ολιγονουκλεοτίδιο συντίθεται σε μία μεγάλη παρτίδα χρησιμοποιώντας τυπικές τεχνολογίες. Τα ολιγονουκλεοτίδια στη συνέχεια επισυνάπτονται στην επιφάνεια πυριτικού σφαιριδίου 3μm. Σε κάθε σφαιρίδιο επισυνάπτονται εκατοντάδες χιλιάδες αντίγραφα μίας μόνο ολιγονουκλεοτιδικής αλληλουχίας.
Οι συστοιχίες μπορούν να έχουν μέχρι και 44.000 μοναδικά ολιγονουκλεοτίδια.

# Illumina BeadChip

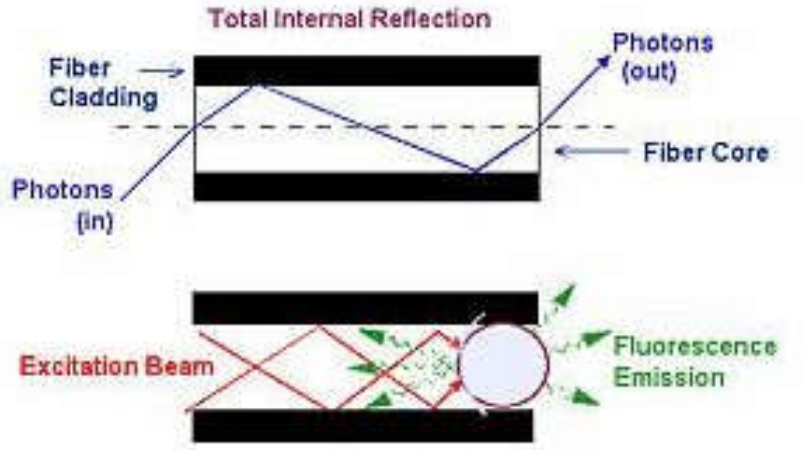


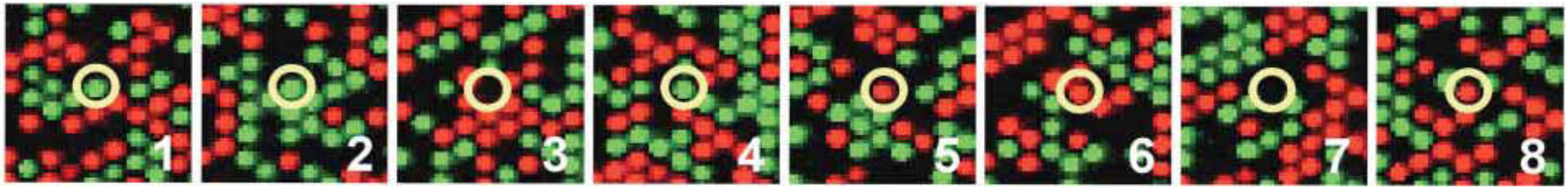

Τυπικές λιθογραφικές τεχνικές χρησιμοποιούνται για να δημιουργήσουν ένα κυψελοειδές πρότυπο φρεατίων στην επιφάνεια του γυάλινου πλακιδίου. Κάθε φρεάτιο μπορεί να χωρέσει μόνο ένα σφαιρίδιο. Τα σφαιρίδια μιας δεδομένης συστοιχίας αναμειγνύονται σε ίσες ποσότητες και εναποτίθενται στην επιφάνεια. Τα σφαιρίδια καταλαμβάνουν τα φρεάτια με τυχαία κατανομή. Κάθε ολιγονουκλεοτίδιο αντιπροσωπεύεται, κατά μέσο όρο, από περίπου 20 σφαιρίδια σε κάθε συστοιχία. Η ταυτότητα του κάθε σφαιριδίου προσδιορίζεται με αποκωδικοποίηση της αλληλουχίας διεύθυνσης. Έτσι, παράγεται ένα μοναδικό για κάθε συστοιχία αρχείο διάταξης, που στη συνέχεια θα χρησιμοποιηθεί για την αποκωδικοποίηση των δεδομένων κατά τη σάρωση της συστοιχίας.

# Illumina BeadChip



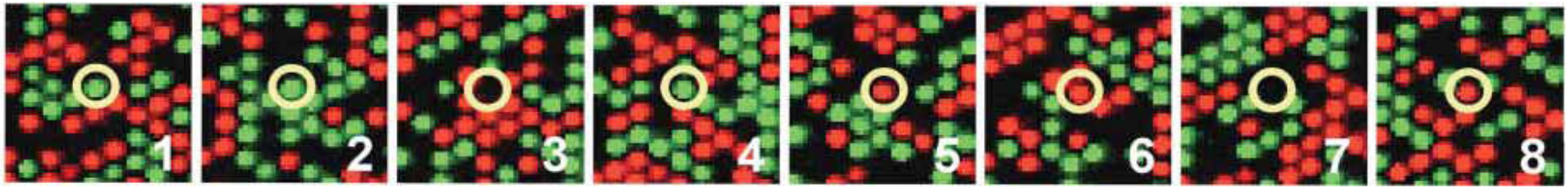

Ο φωτισμός των ανιχνευτών γίνεται μέσω δέσμης οπτικών ινών. Κάθε ίνα καλύπτεται από ένα μόνο σφαιρίδιο. Φως διέγερσης διασχίζει τη δέσμη και διεγείρει τις φθορίζουσες ουσίες που συνδέονται με τα σφαιρίδια. Στη συνέχεια, το εκπεμπόμενο φως συλλέγεται στην επιφάνεια της ίνας.

# Αποκωδικοποίηση Μικροσυστοιχιών Illumina



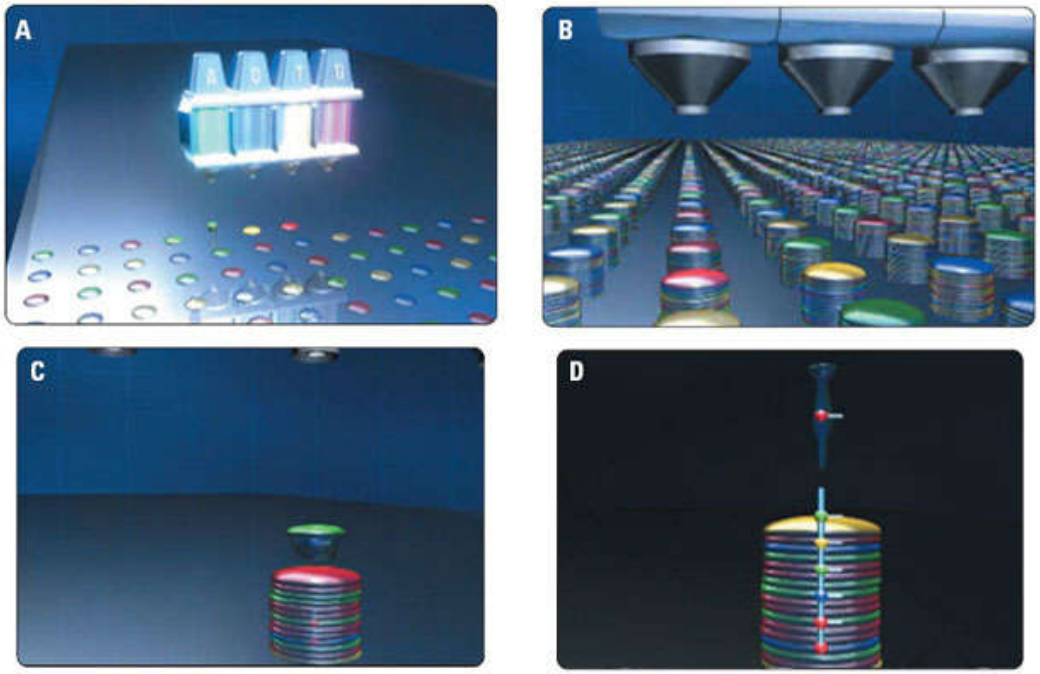

Μόλις τα σφαιρίδια αποτεθούν στην επιφάνεια του πλακιδίου, είναι απαραίτητο να ταυτοποιηθεί ποιος ανιχνευτής μεταγράφου είναι σε ποιο φρεάτιο. Αυτό γίνεται χρησιμοποιώντας το τμήμα διεύθυνσης του ολιγονουκλεοτιδίου. Η αποκωδικοποίηση των μικροσυστοιχιών περιλαμβάνει διαδοχική υβριδοποίηση διαφορικά σημασμένων ανιχνευτών. Η διαφορική σήμανση χρησιμοποιεί τρεις καταστάσεις: Σήμανση με καρβοξυφλουορεσκίνη (FAM) (πράσινο), κυανίνη 3 (Cy3) (κόκκινο), και μη σήμανση (κενό). Κατά τη διάρκεια κάθε κύκλου της διαδικασίας, ένα σφαιρίδιο είναι πράσινο, κόκκινο, ή κενό.

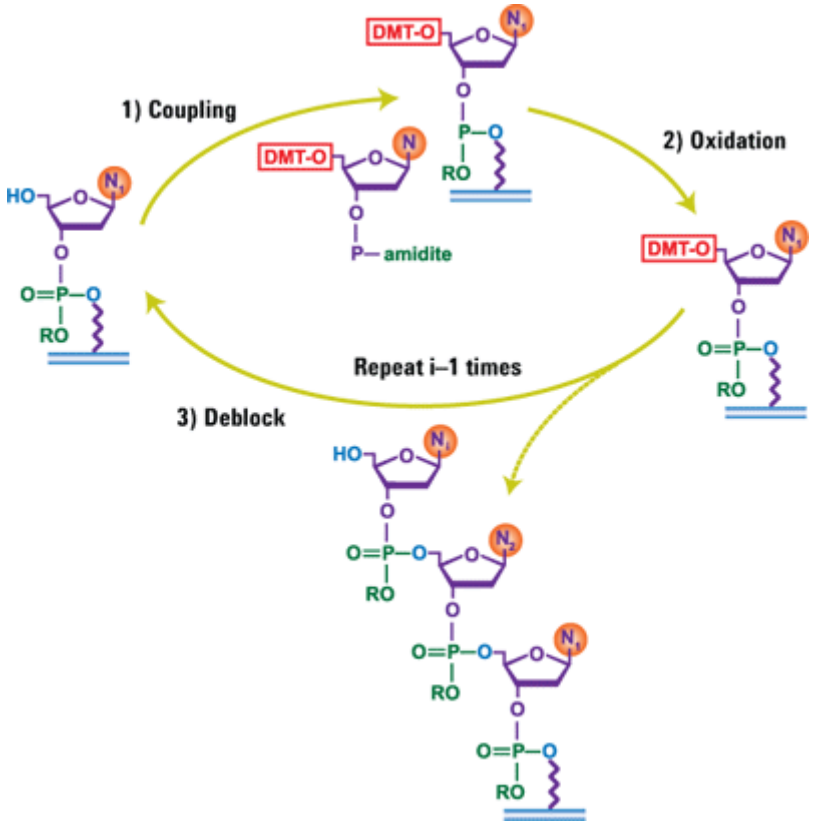
# Αποκωδικοποίηση Μικροσυστοιχιών Illumina




Τα αποκωδικοποιώντα σεσημασμένα ολιγονουκλεοτίδια υβριδοποιούνται στις μικροσυστοιχίες σε υψηλές συγκεντρώσεις, το οποίο επιτρέπει ταχεία υβριδοποίηση, ακολουθούμενη από έκπλυση για την απομάκρυνση μη-ειδικού σήματος και του υποβάθρου. Αν ένας αριθμός αντιστοιχεί σε κάθε κατάσταση - 0 στο κενό, 1 πράσινο, και 2 στο κόκκινο, τότε κάθε κύκλος της διαδικασίας δημιουργεί τριαδικό ψηφίο. Αν κοιτάξουμε ένα υποθετικό ανιχνευτή, τότε ο πρώτος γύρος μπορεί να είναι κόκκινο (πρώτο ψηφίο 2). Ο δεύτερος γύρος είναι πράσινο (ψηφίο 1), έτσι ώστε ο αριθμός είναι τώρα 21. Ο τρίτος γύρος κόκκινο, έτσι ώστε ο αριθμός είναι τώρα 212. Κάθε γύρος προσθέτει απλά ένα νέο ψηφίο στον αριθμό. Αυτό συνεχίζεται μέχρις ότου υπάρξουν επαρκή ψηφία για τον μοναδικό προσδιορισμό κάθε ανιχνευτή. Με ένα ψηφίο μπορεί να ταυτοποιήσουμε τρεις ανιχνευτές (0, 1, ή 2), με δύο ψηφία μπορούμε να προσδιορίσουμε 9 ανιχνευτές (00, 01, 02, 10, 11, 12, 21, 22, 23), με ένα τριψήφιο μπορούμε να εντοπίσουμε 27, κοκ. Η επανυβριδοποίηση συνεχίζεται έως ότου υπάρξουν επαρκή δεδομένα για να προσδιοριστεί σαφώς η ταυτότητα κάθε σφαιριδίου.

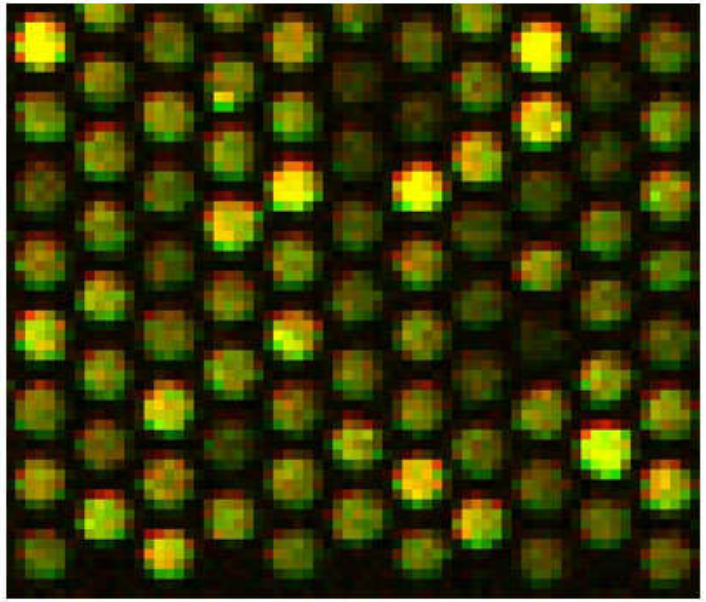
# Agilent



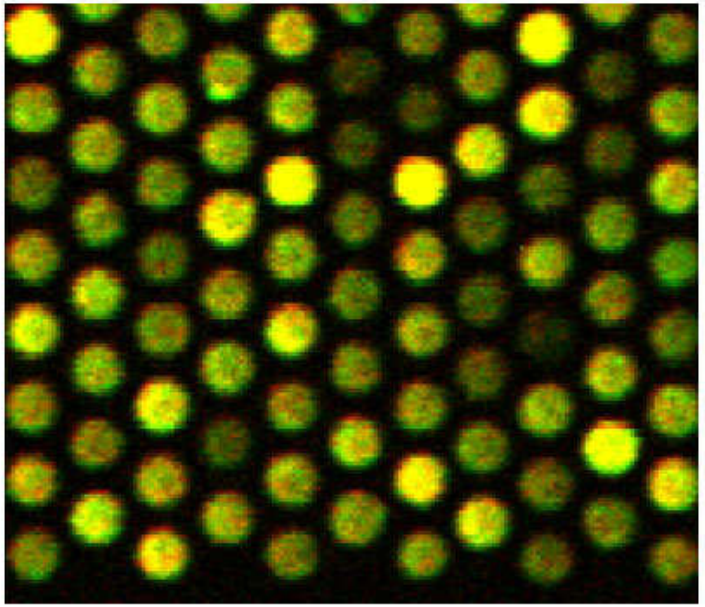




http://www.genomics.agilent.com/article.jsp?pageId=2011&_requestid=719064

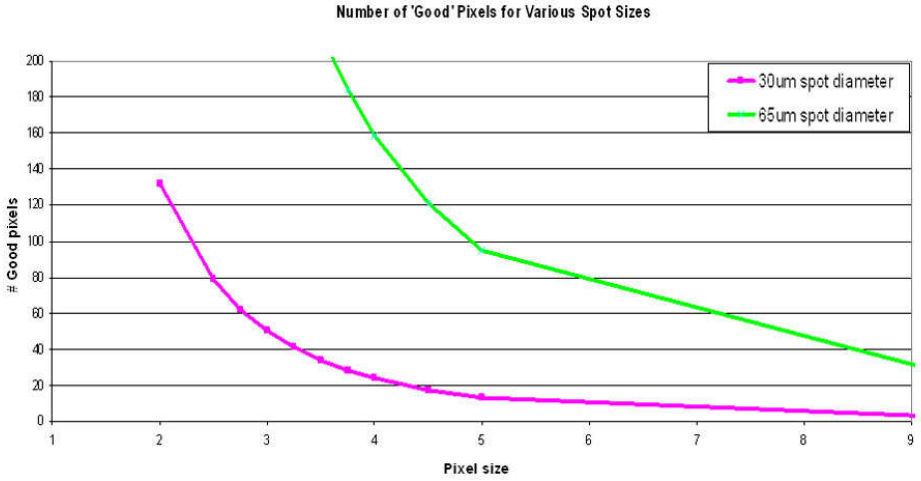
# Agilent

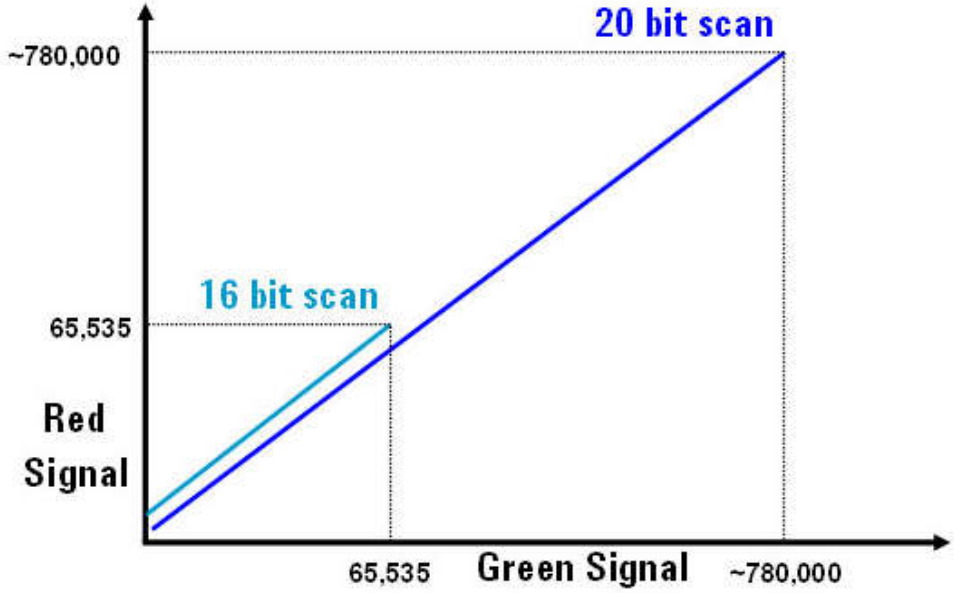


5 μm



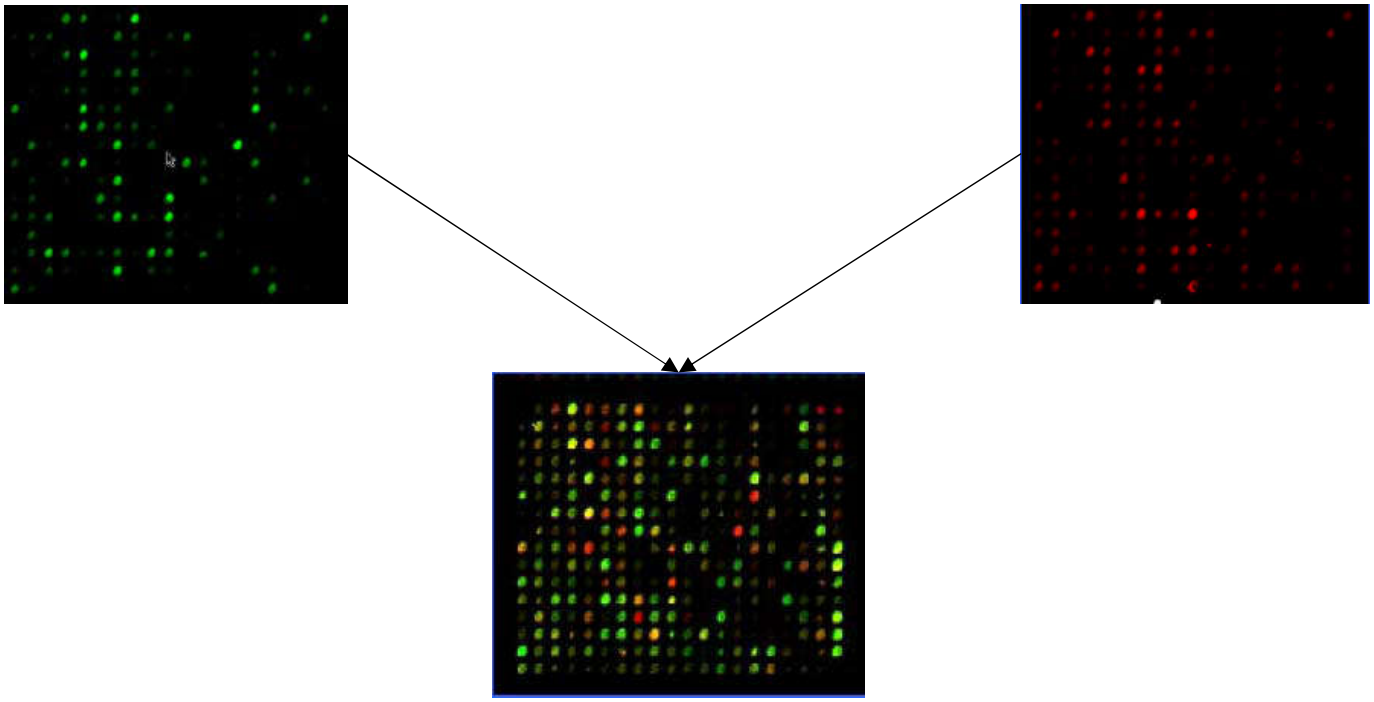
2 μm



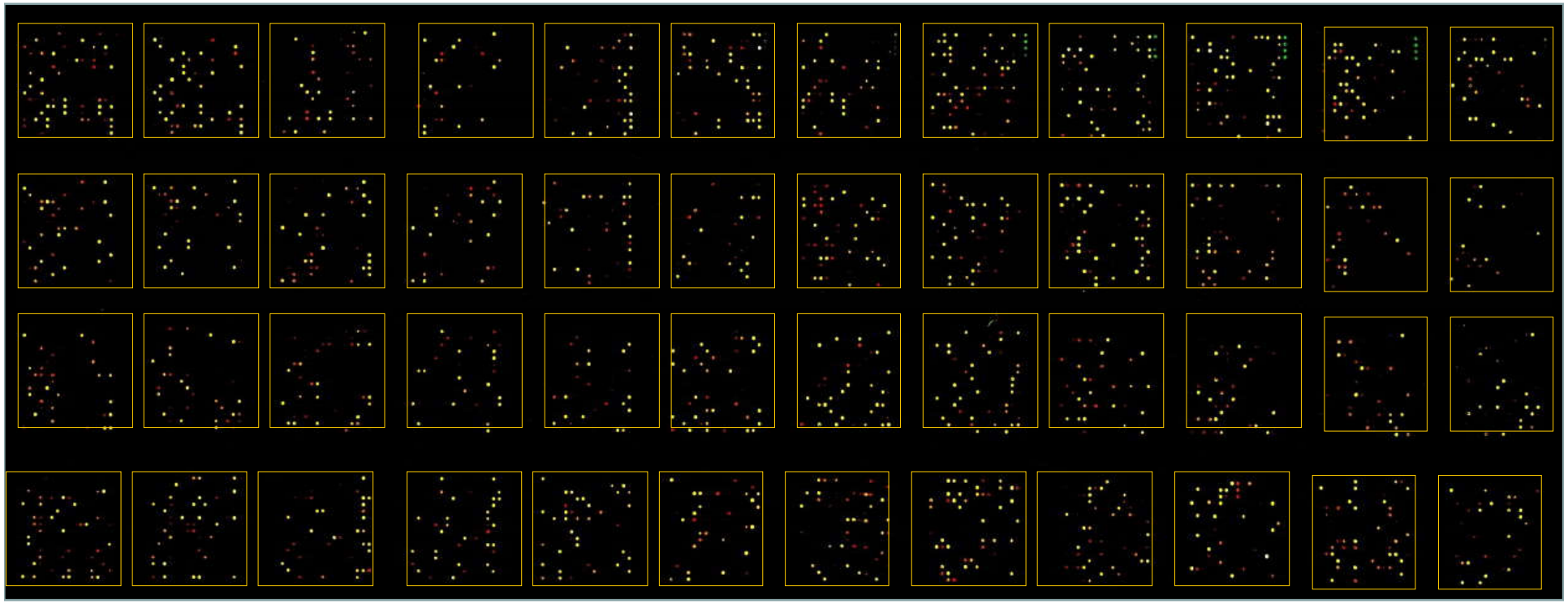



# Δικαναλικές Μικροσυστοιχίες

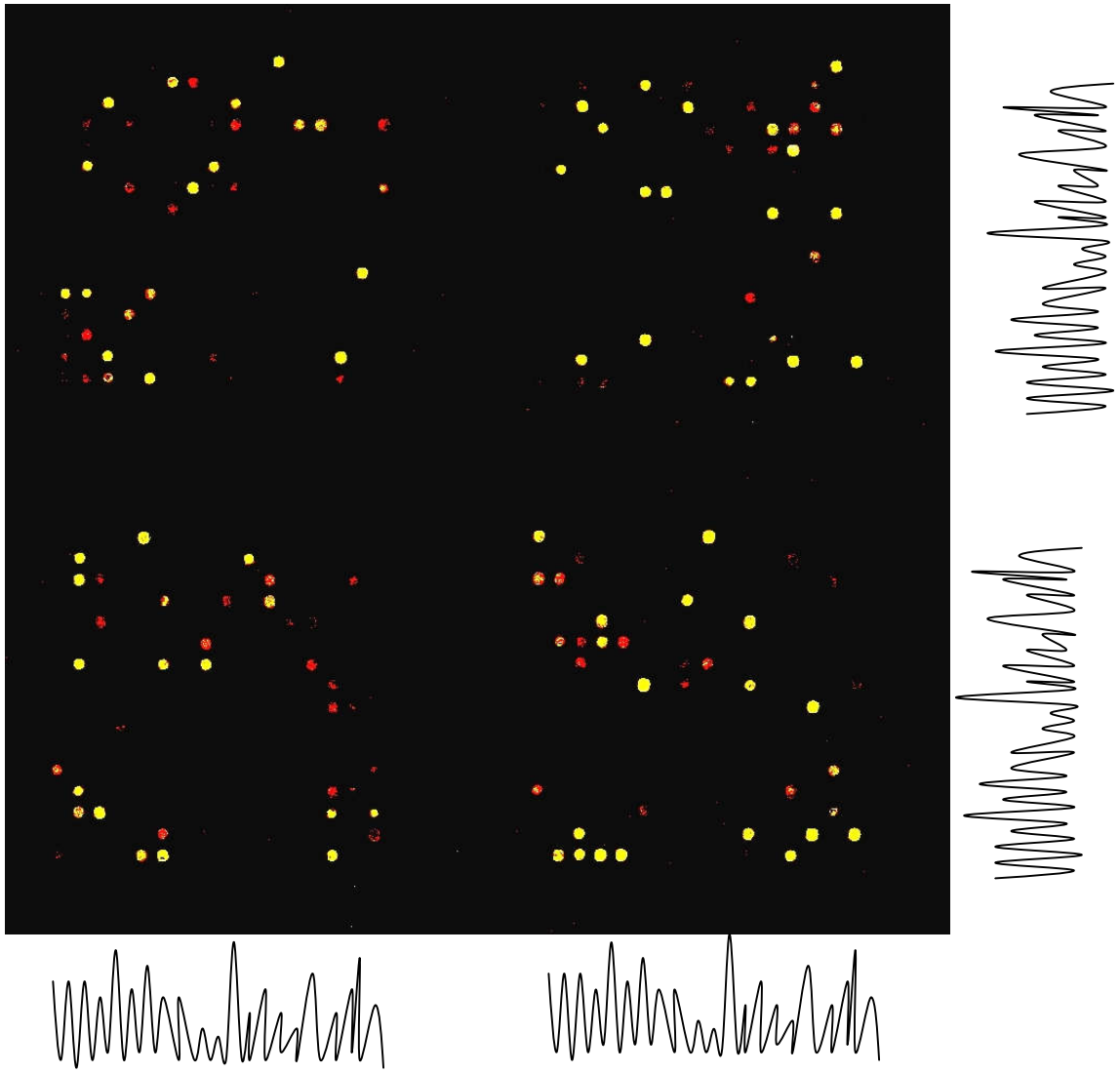
# Δικαναλικές Μικροσυστοιχίες



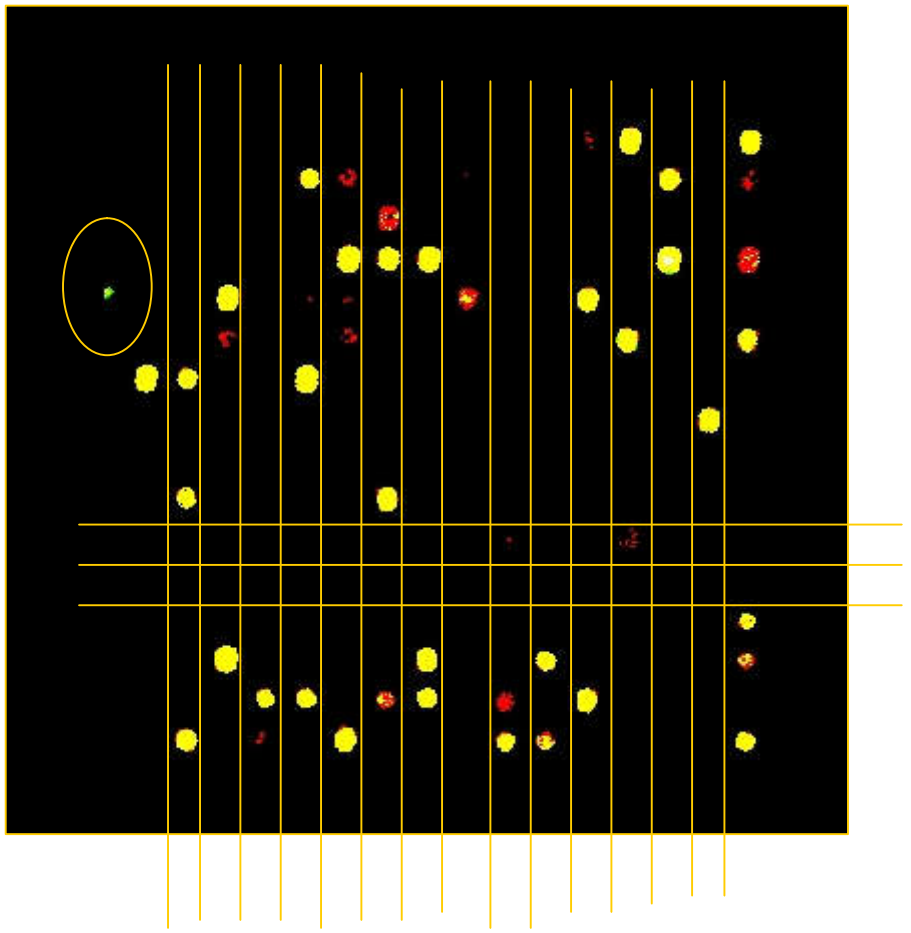
# Εύρεση πλεγμάτων



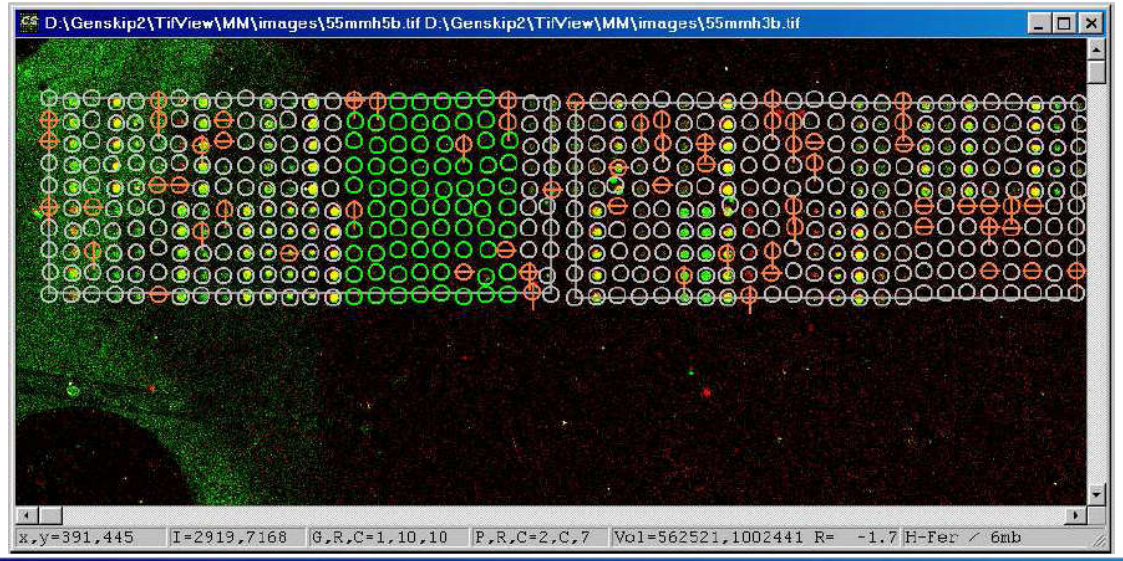
# Εύρεση κηλίδων

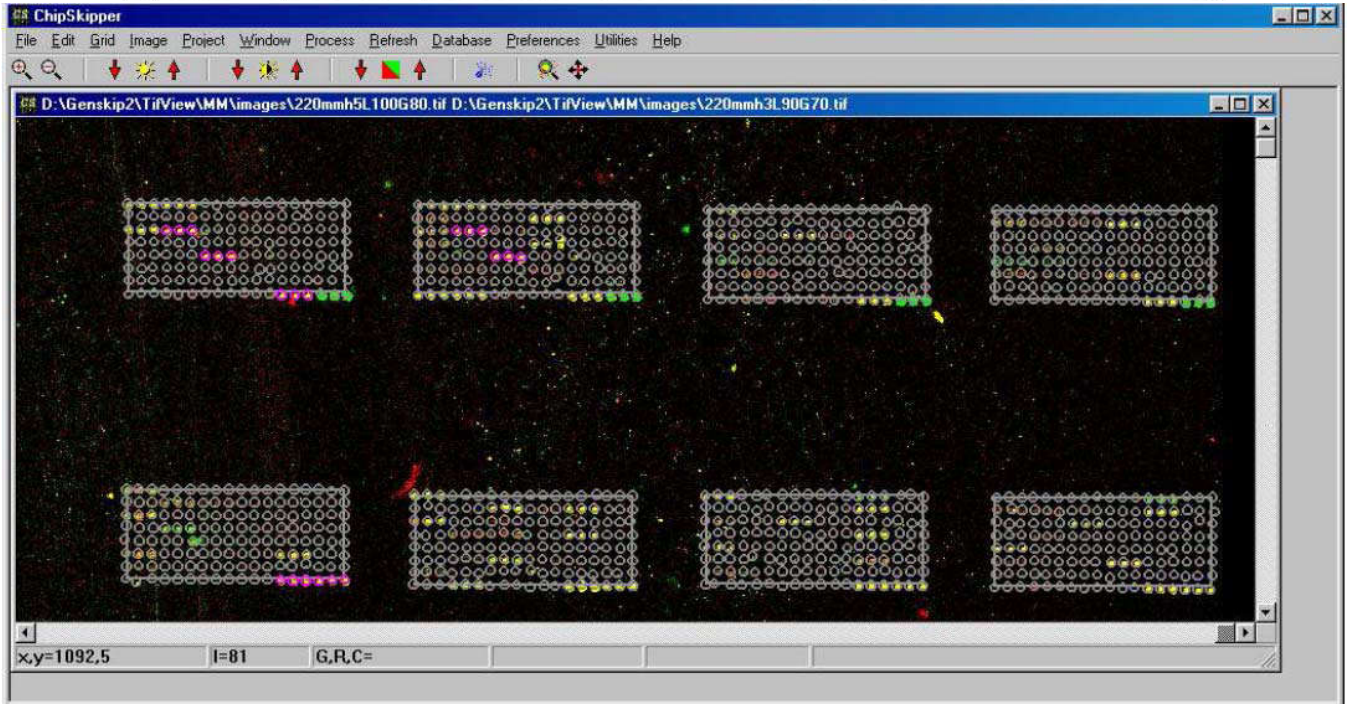


# Εύρεση κηλίδων

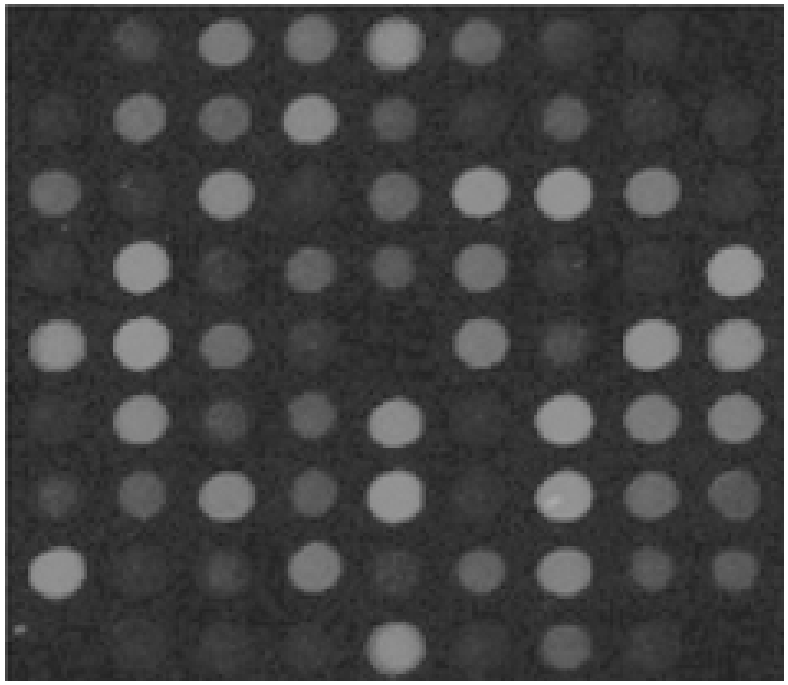


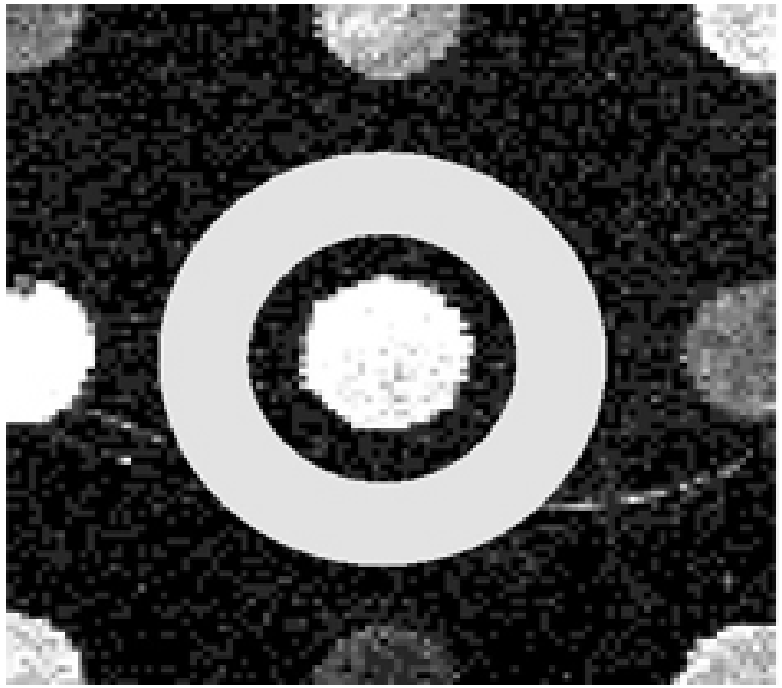
# Εύρεση κηλίδων



# Διόρθωση Υποβάθρου



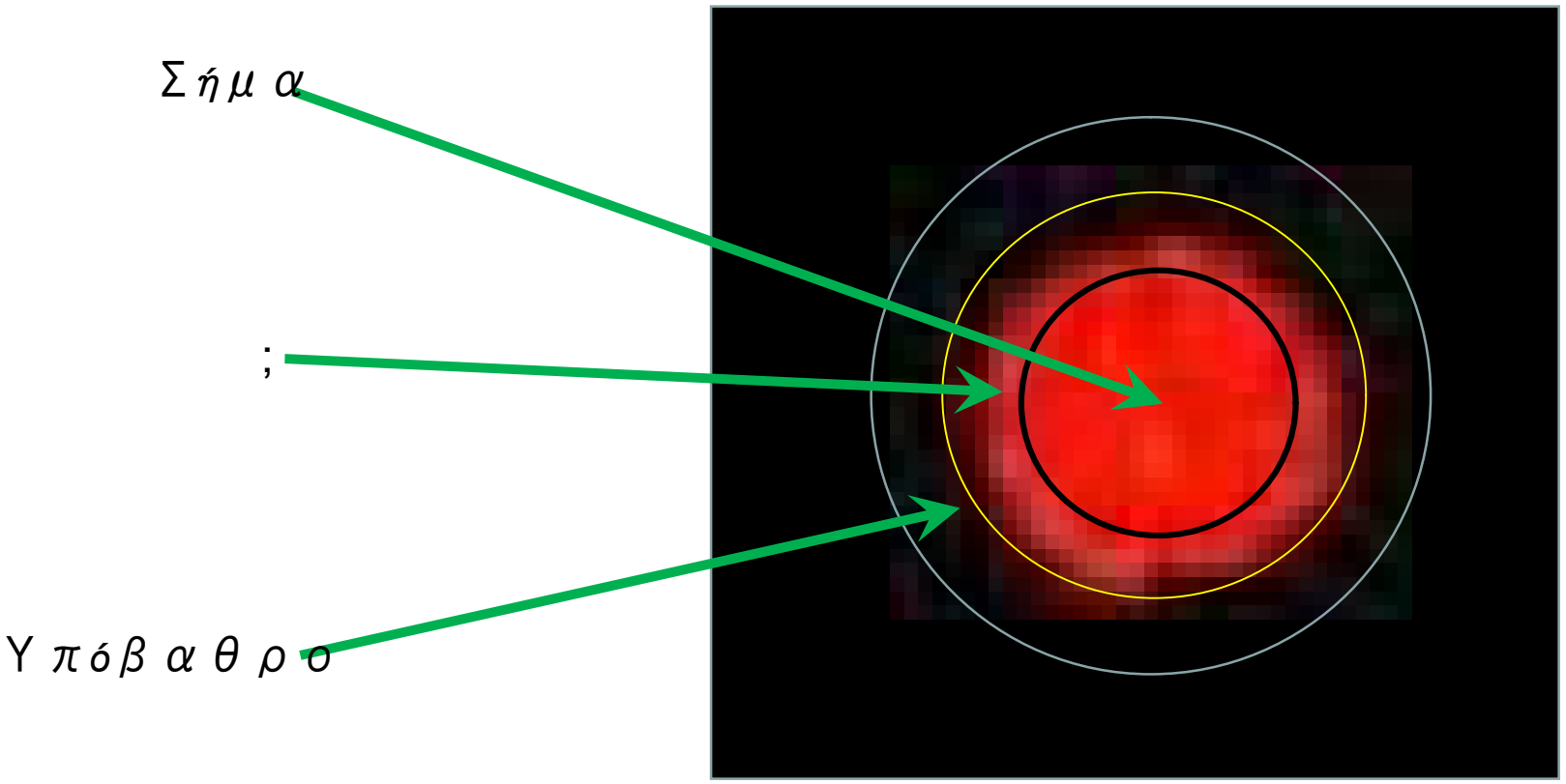


$$signal = \max\left( mv, signal_f - signal_b\left(\frac{s_f}{s_b}\right)\right)$$

Ελάχιστη τιμή πχ $mv$=1

(μέχρι 2% $signal_f < signal_b$ )

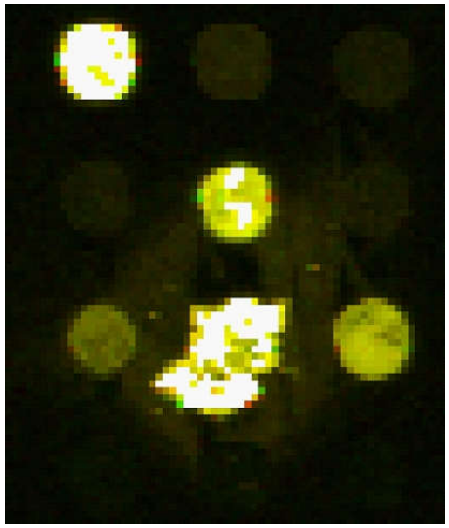
# Διόρθωση Υποβάθρου

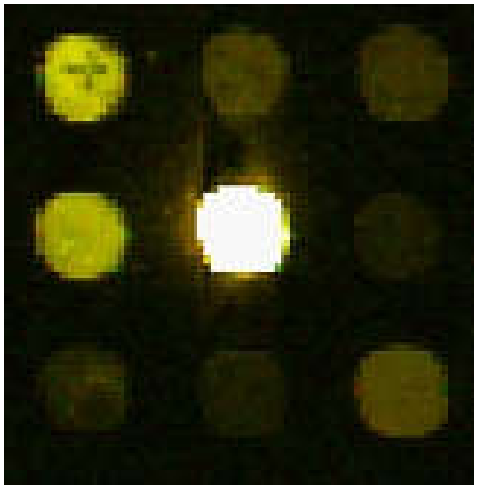


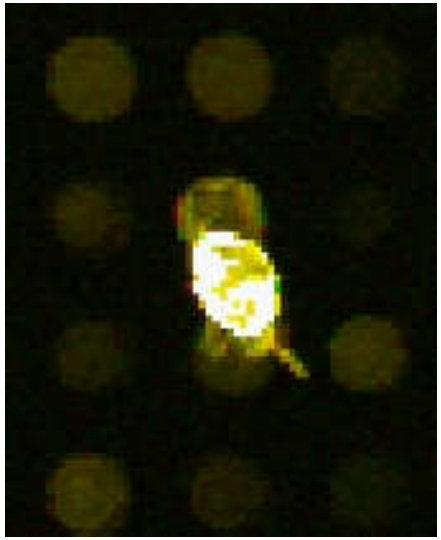

Εξελιγμένες μέθοδοι διόρθωσης υποβάθρου

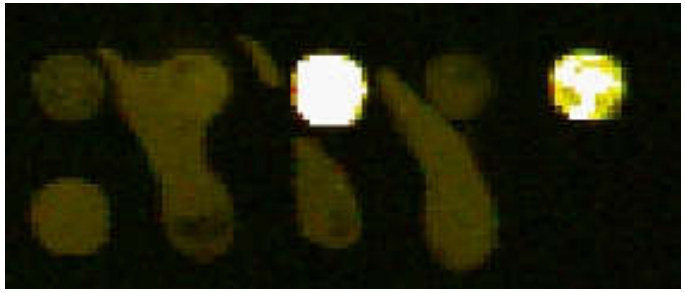
| | |
|---|---|
| Edwards | log-linear |
| Smyth | normexp |
| Kooperberg | Bayesian |

# Προβλήματα









# Διόρθωση Υποβάθρου NormExp+Offset

$$\mathrm{R}_f = \mathrm{R}_b + B + S$$

όπου, $\mathrm{R}_f$ η μετρήσιμη ένταση σήματος, $\mathrm{R}_b$ η μετρήσιμη ένταση υποβάθρου, *B* τα υπόλοιπα υποβάθρου, που έχει εκθετική κατανομή με μέση τιμή *α* και δεν καταγράφεται από το $\mathrm{R}_b$, και *S* το πραγματικό σήμα της έκφρασης, που έχει κανονική κατανομή με μέση τιμή *μ* και διακύμανση $\sigma^2$. Αν *X* η παρατηρούμενη ένταση:

$$X = \mathrm{R}_f - \mathrm{R}_b$$

τότε:

$$X = B + S$$

Ως εκ τούτου:

$$\mathbb{E}(S|X = x) = \mu_{S \cdot X} + \frac{\sigma^2 \phi(0; \mu_{S \cdot X}, \sigma^2)}{1 - \Phi(0; \mu_{S \cdot X}, \sigma^2)}$$

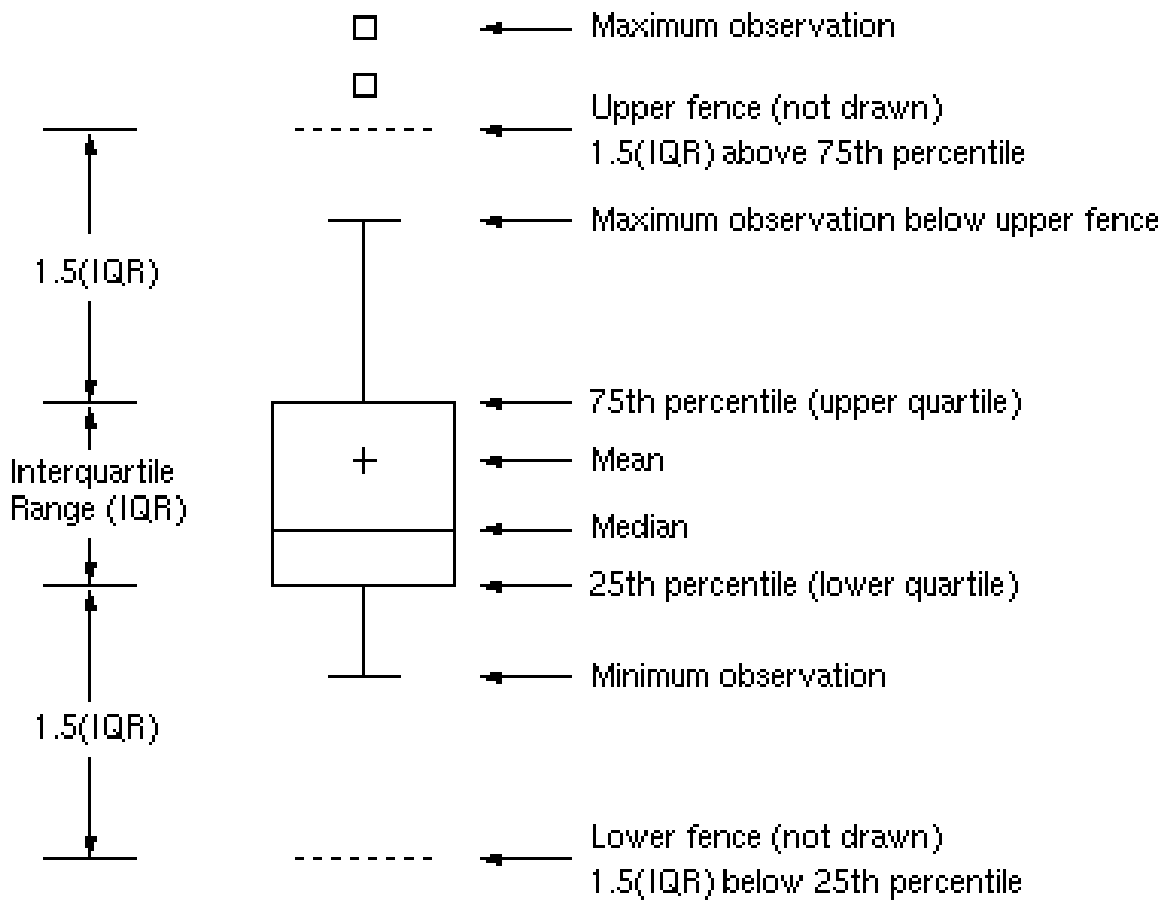
όπου, $\mathbb{E}(S|X = x)$ η εκτίμηση του σήματος με δεδομένη την παρατηρούμενη ένταση, $\phi(0; \mu_{S \cdot X}, \sigma^2)$ η συνάρτηση πυκνότητας κανονικής κατανομής, $\Phi(0; \mu_{S \cdot X}, \sigma^2)$ η συνάρτηση κανονικής κατανομής, και $\mu_{S \cdot X}$:

$$\mu_{S \cdot X} = x - \mu - {\sigma^2}/{\alpha}$$

$$M = \log_2[(R + k)/(G + k)]$$

όπου $k$ το offset, για να μη βγουν αρνητικές τιμές.

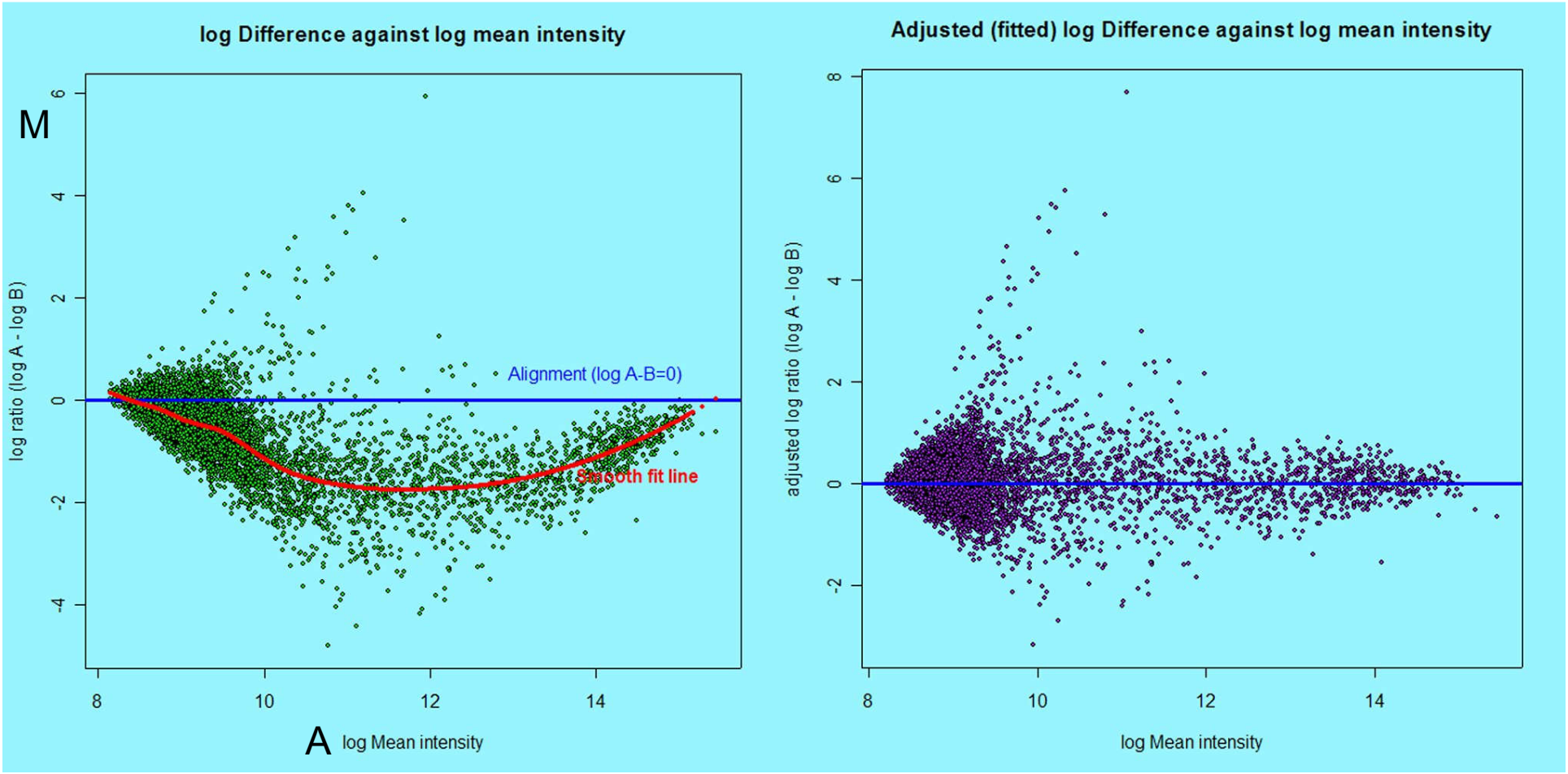
# Box-Plot


Maximum observation
Upper fence (not drawn)
1.5(IQR) above 75th percentile
Maximum observation below upper fence
1.5(IQR)
75th percentile (upper quartile)
Mean
Interquartile
Range (IQR)
Median
25th percentile (lower quartile)
Minimum observation
1.5(IQR)
Lower fence (not drawn)
1.5(IQR) below 25th percentile

# Κανονικοποίηση Loess

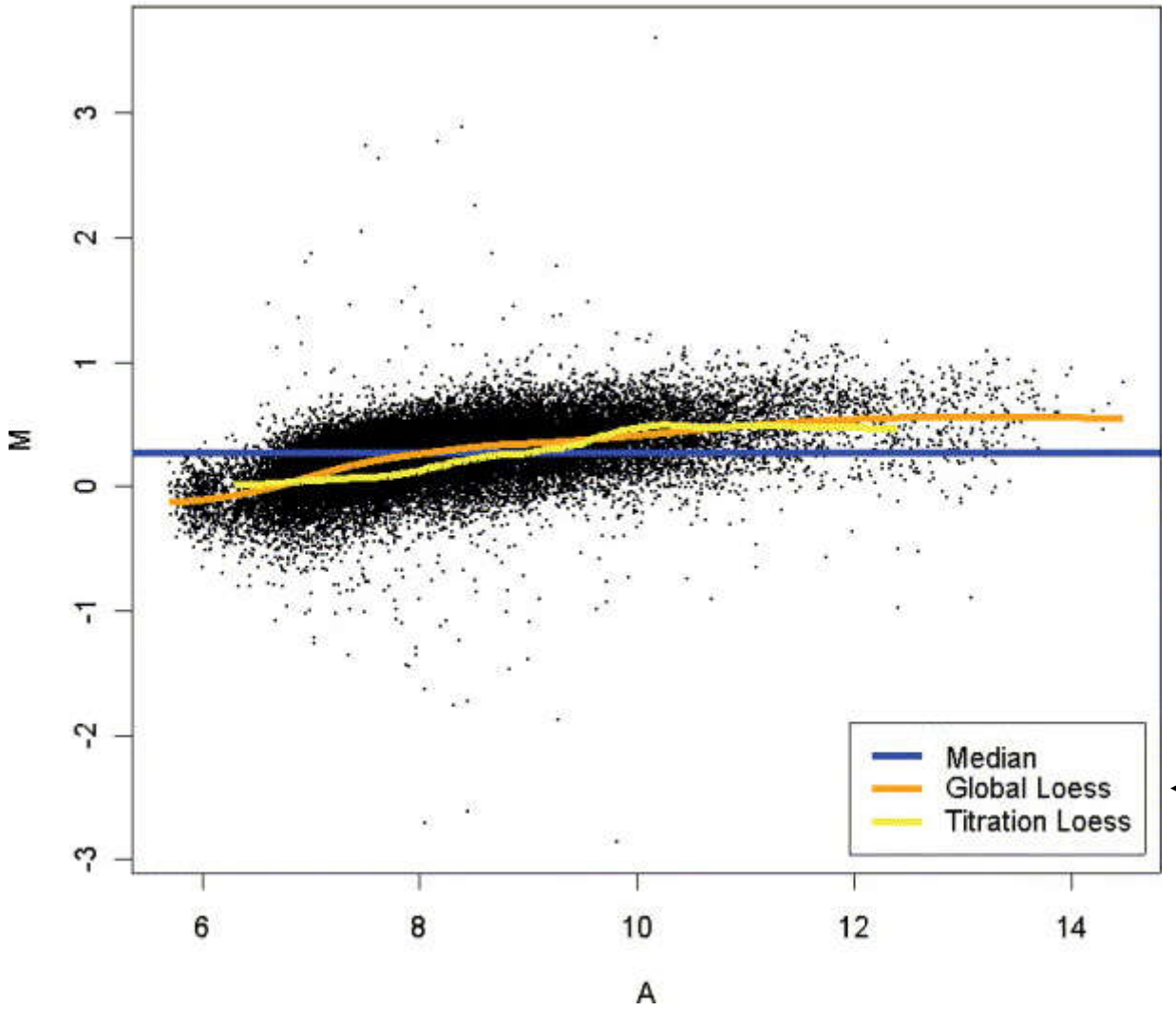
$$M = \log(Cy5) - \log(Cy3) = \log\left(\frac{Cy5}{Cy3}\right)$$

$$A = \frac{1}{2}\left(\log(Cy5) + \log(Cy3)\right) = \log\sqrt{Cy5 \cdot Cy3}$$



# Global Loess

$$N = M - loess(A)$$



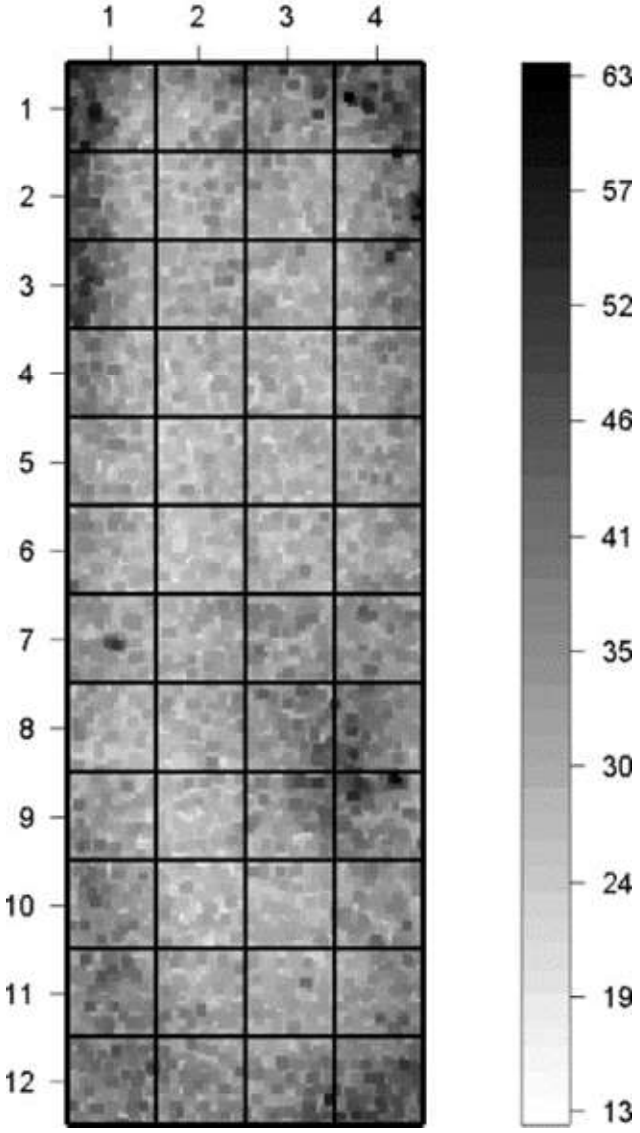

$loess(A)$

# Κανονικοποίηση εντός συστοιχιών

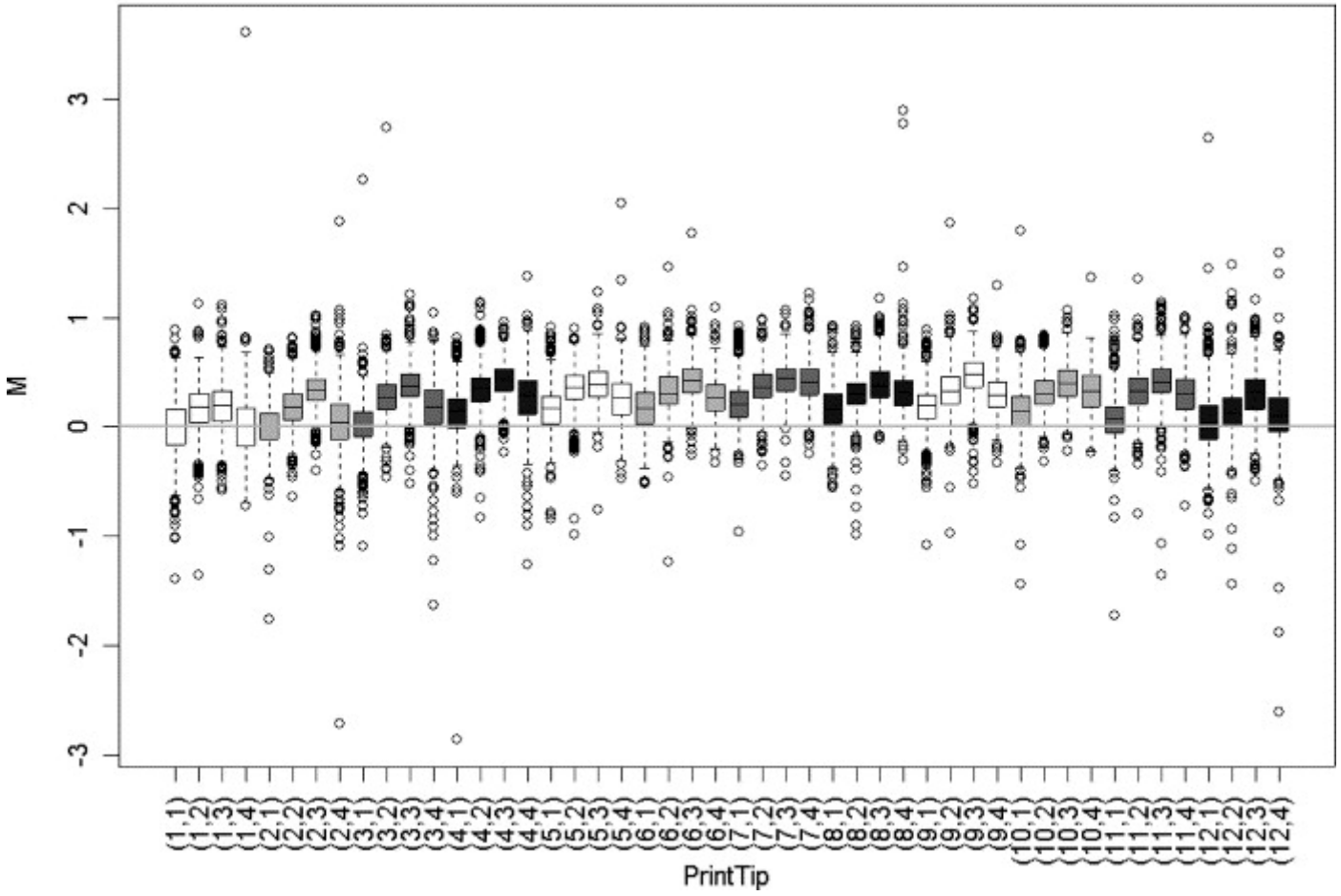
- 4x12 πλέγματα
- 16x12 ανιχνευτές ανά πλέγμα

# Print-tip Loess

$G_b$

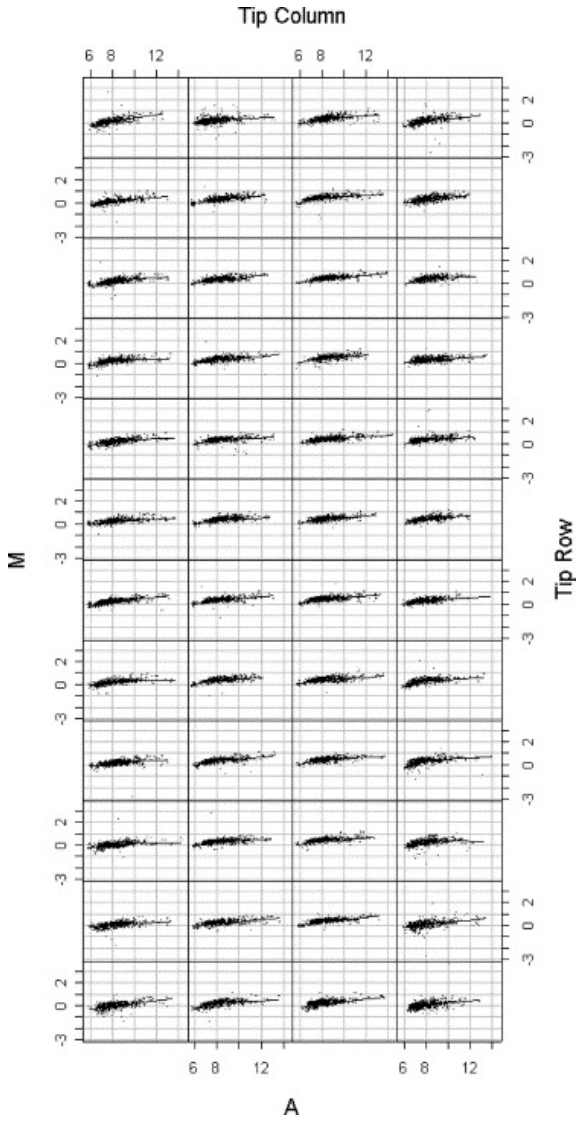


# Print-tip Loess



# Print-tip Loess

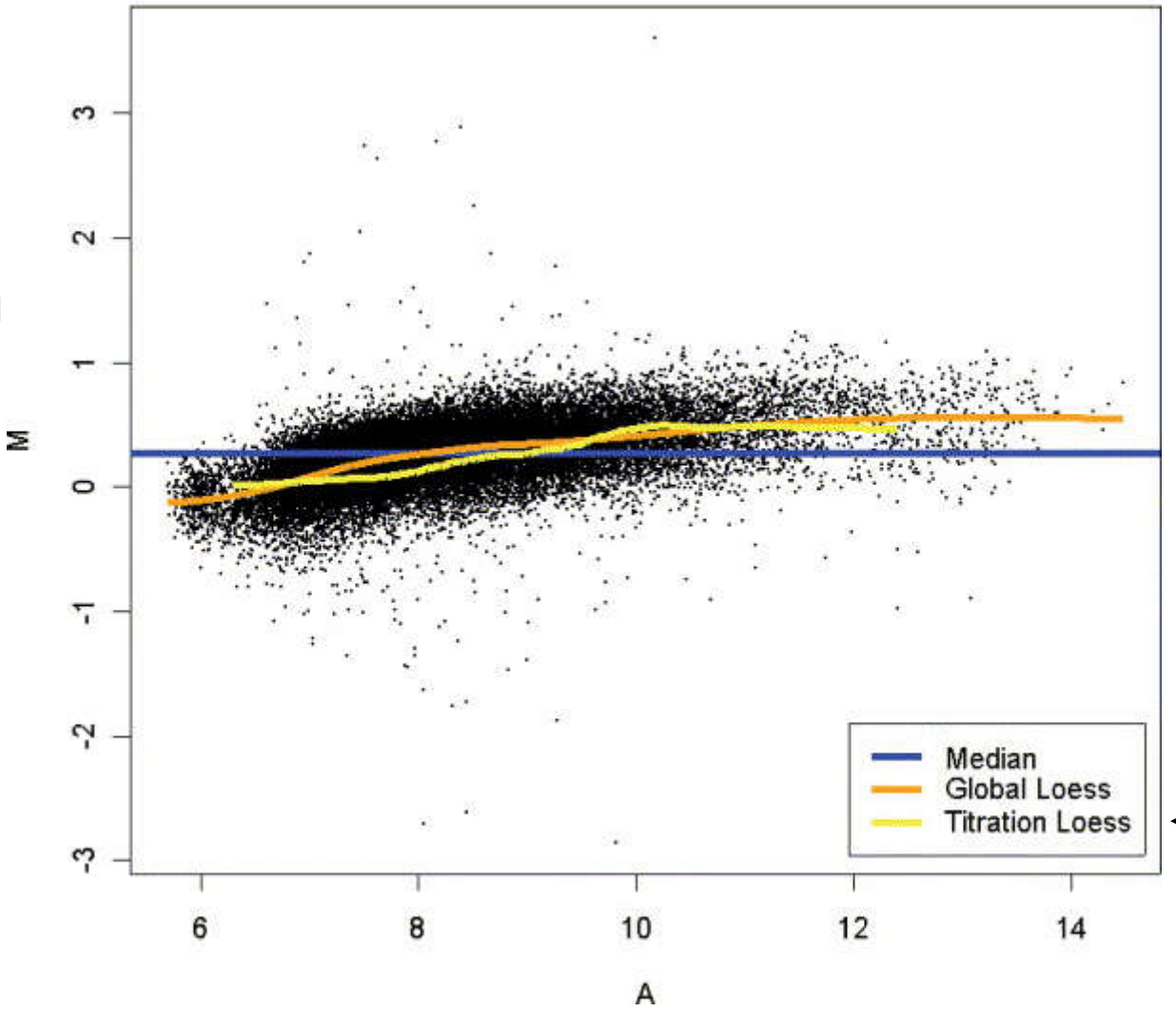
$$N = M - loess_i(A)$$



# Composite Loess

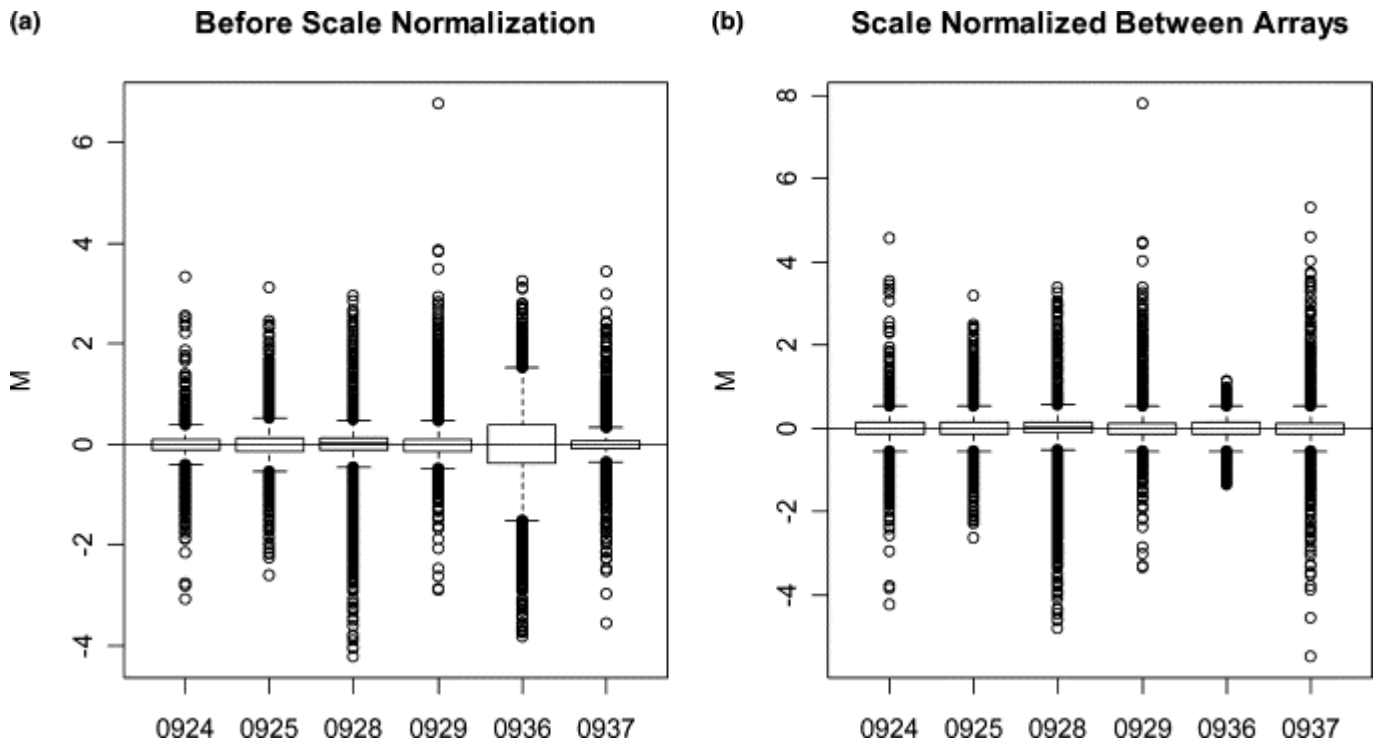
$$N = M - p(A) \cdot loess_{MSP}\,(A) - \{1 - p(A)\} \cdot loess_i(A)$$

όπου, $p(A)$ η αναλογία των ιχνηθετών με τιμές *A* μικρότερες του A και MSP είναι το ειδικά σχεδιασμένο Microarray Design Pool

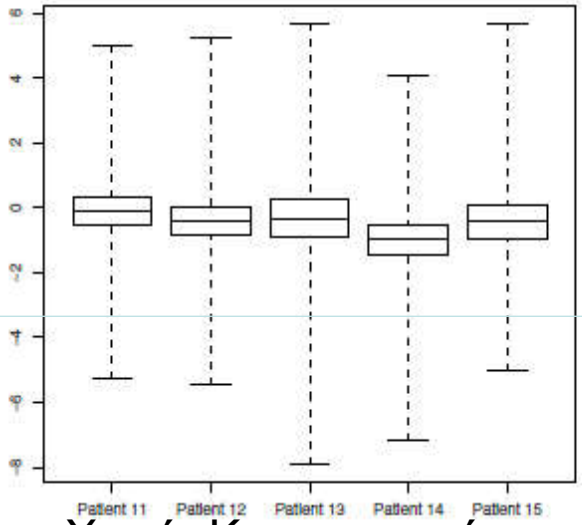



$loess_{MSP}(A)$

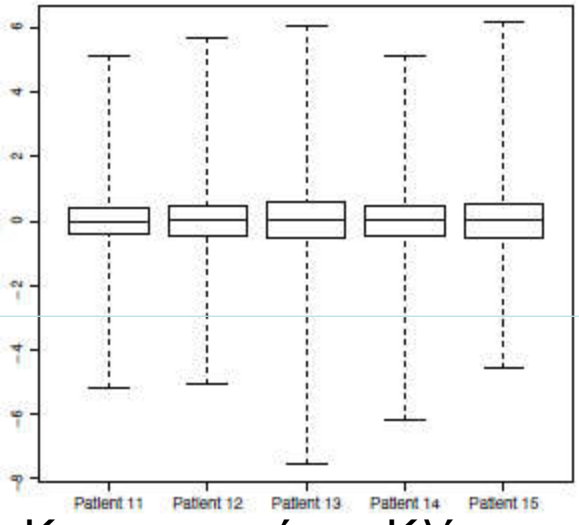
# Κανονικοποίηση μεταξύ συστοιχιών



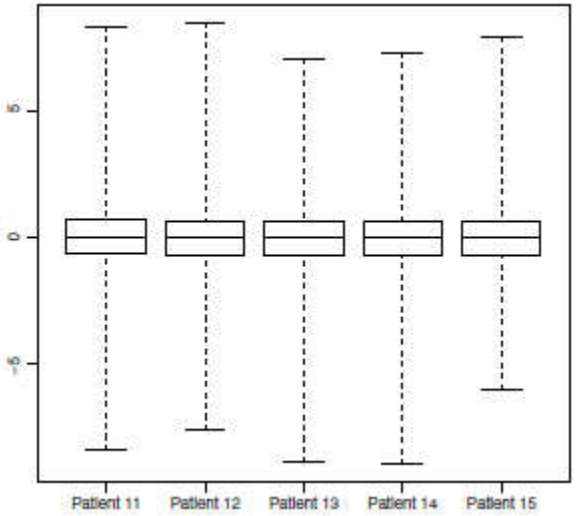
# Κανονικοποίηση μεταξύ συστοιχιών



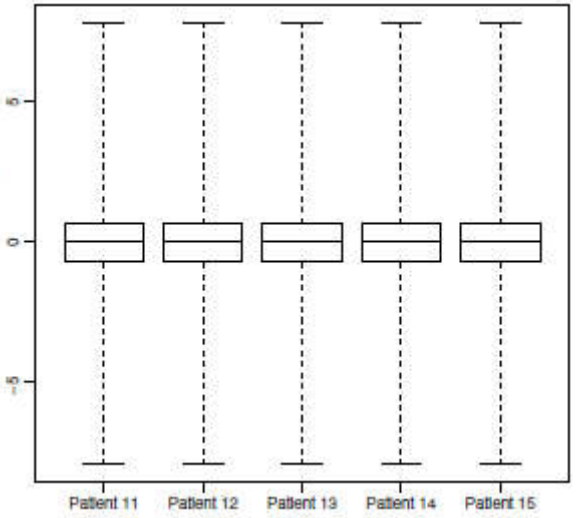

Χωρίς Κανονικοποίηση




Κανονικοποίηση Κλίμακας




Κανονικοποίηση Κέντρου




Κανονικοποίηση κατανομής

# Πόσα αντίγραφα;

- Βιολογικά:
  - Τουλάχιστον 3 αντίγραφα
  - Η επανάληψη των πειραμάτων δίνει μια αίσθηση της βιολογικής μεταβλητότητας
- Τεχνικά
  - Δεν είναι απαραίτητα εκτός αν μελετάται η απόδοση των μικροσυστοιχίων

# Πολλαπλές συγκρίσεις

Όταν γίνεται έλεγχος πολλαπλών ανεξάρτητων υποθέσεων, πρέπει να γίνει διόρθωση του p-value

# Bonferroni

| | wt | mut | p-value | **Bonf corr p-value** |
|---|---|---|---|---|
| A | 2.000±0.2887 | 5.667±0.8819 | 0.0168 | **0.084** |
| B | 1.167±0.3333 | 5.033±0.7424 | 0.0090 | **0.045** |
| C | 5.333±1.4530 | 3.167±0.4410 | 0.2268 | **1** |
| D | 1.167±0.4410 | 5.167±0.7265 | 0.0093 | **0.0465** |
| E | 4.833±0.6009 | 1.400±0.2082 | 0.0057 | 0.0285 |

Student's t-test

Mean±SEM

Standard Error of the Mean: $SEM = \frac{s}{\sqrt{n}}$

Πιθανότητα ύπαρξης ψευδοθετικών αποτελεσμάτων

•Οι διορθωμένες τιμές είναι το γινόμενο N·p-value, όπου N ο αριθμός των πολλαπλών ερωτημάτων (όταν το γινόμενο υπερβαίνει το 1, η τιμή γίνεται 1)

# False discovery rate (FDR) Benjamini-Hochberg

| | wt | mut | p-value |
|---|---|---|---|
| A | 2.000±0.2887 | 5.667±0.8819 | 0.0168 |
| B | 1.167±0.3333 | 5.033±0.7424 | 0.0090 |
| C | 5.333±1.4530 | 3.167±0.4410 | 0.2268 |
| D | 1.167±0.4410 | 5.167±0.7265 | 0.0093 |
| E | 4.833±0.6009 | 1.400±0.2082 | 0.0057 |

| | k | wt | mut | p-value | N·p-value/k | tmp | BH corr p-value |
|---|---|---|---|---|---|---|---|
| E | 1 | 4.833±0.6009 | 1.400±0.2082 | 0.0057 | 0.0285 | 0.0155 | 0.0155 |
| B | 2 | 1.167±0.3333 | 5.033±0.7424 | 0.0090 | 0.0225 | 0.0155 | 0.0155 |
| D | 3 | 1.167±0.4410 | 5.167±0.7265 | 0.0093 | 0.0155 | 0.0210 | 0.0155 |
| A | 4 | 2.000±0.2887 | 5.667±0.8819 | 0.0168 | 0.0210 | 0.2268 | 0.0210 |
| C | 5 | 5.333±1.4530 | 3.167±0.4410 | 0.2268 | 0.2268 | **1.0000** | 0.2268 |

Αναμενόμενο ποσοστό ψευδοθετικών αποτελεσμάτων

- Οι τιμές p-value ταξινομούνται κατά **αύξουσα** σειρά
- Υπολογίζεται το γινόμενο N·p-value/k και της προσωρινής τιμής, όπου N ο αριθμός των πολλαπλών ερωτημάτων και k η σειρά
- Ορίζεται το 1 ως προσωρινή τιμή
- Αρχίζοντας από το **μεγαλύτερο** p-value, κάθε διορθωμένη τιμή είναι η **μικρότερη** μεταξύ του γινομένου N·p-value/k και της προσωρινής τιμής (για την τήρηση της μονοτονίας, μετά το τέλος κάθε διόρθωσης προσωρινή τιμή γίνεται η διορθωμένη)

# Πολλαπλές συγκρίσεις

| | wt | mut | p-value | Bonf corr p-value | BH corr p-value |
|---|---|---|---|---|---|
| A | 2.000±0.2887 | 5.667±0.8819 | 0.0168 | 0.084 | 0.0210 |
| B | 1.167±0.3333 | 5.033±0.7424 | 0.0090 | 0.045 | 0.0155 |
| C | 5.333±1.4530 | 3.167±0.4410 | 0.2268 | 1 | 0.2268 |
| D | 1.167±0.4410 | 5.167±0.7265 | 0.0093 | 0.0465 | 0.0155 |
| E | 4.833±0.6009 | 1.400±0.2082 | 0.0057 | 0.0285 | 0.0155 |

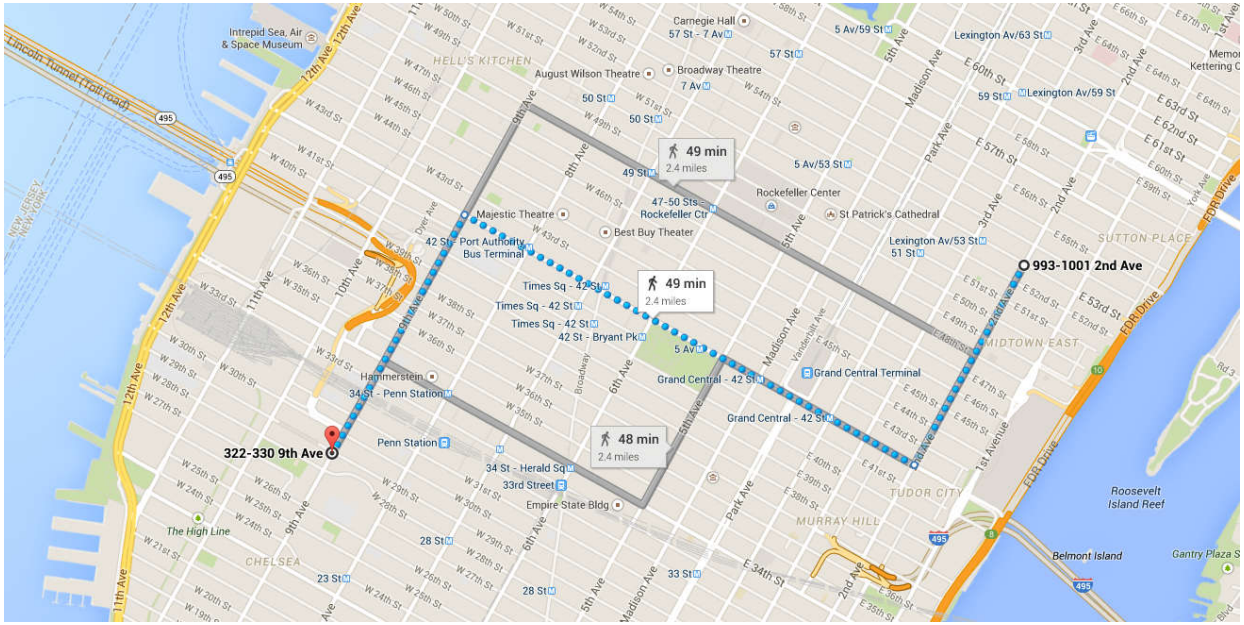
Η διόρθωση FDR είναι λιγότερο συντηρητική διαδικασία από τη διόρθωση Bonferroni

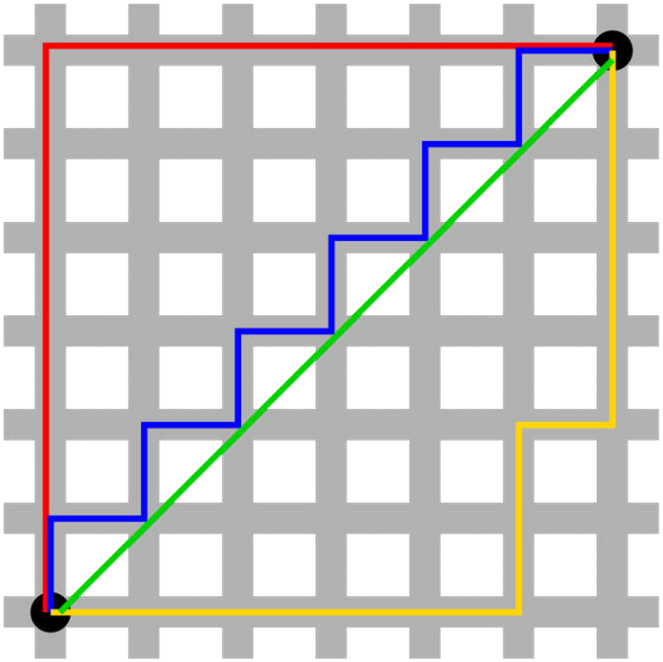
# Αποστάσεις

Minkowski $$d_{x,y} = \sqrt[m]{\sum_{i=1}^{n} |x_i - y_i|^m}$$

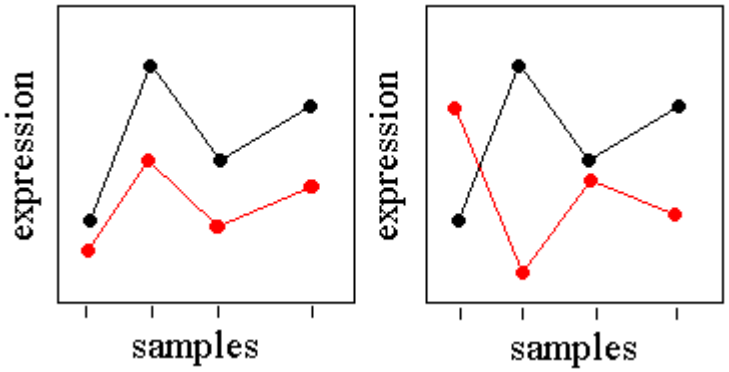
Manhattan $$d_{x,y} = \sum_{i=1}^{n} |x_i - y_i|$$ *m*=1

Ευκλείδεια $$d_{x,y} = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2}$$ *m*=2



# Συσχέτιση Pearson




$$r_{x,y} = \frac{1}{n-1}\sum_{i=1}^{n}\left(\frac{x_i - \bar{x}}{s_x}\right)\left(\frac{y_i - \bar{y}}{s_y}\right) = \frac{\sum_{i=1}^{n} z_{x_i} z_{y_i}}{n-1}$$

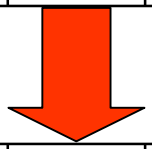
$$-1 \le r_{x,y} \le 1$$

$$d_{x,y} = 1 - r_{x,y}$$

# Μετατροπή *r* σε απόσταση *d*

Πίνακας τιμών *r*

| | A | B | C | D | E | F |
|---|---|---|---|---|---|---|
| A | 1 | 0.7 | 0.85 | -0.35 | -0.25 | 0.4 |
| B | 0.7 | 1 | 0.7 | -0.3 | -0.2 | 0.35 |
| C | 0.85 | 0.7 | 1 | -0.35 | -0.2 | 0.4 |
| D | -0.35 | -0.3 | -0.35 | 1 | 0.6 | -0.4 |
| E | -0.25 | -0.2 | -0.2 | 0.6 | 1 | -0.3 |
| F | 0.4 | 0.35 | 0.4 | -0.4 | -0.3 | 1 |



Πίνακας αποστάσεων

| | A | B | C | D | E | F |
|---|---|---|---|---|---|---|
| A | 0 | 0.3 | 0.15 | 1.35 | 1.25 | 0.6 |
| B | 0.3 | 0 | 0.3 | 1.3 | 1.2 | 0.65 |
| C | 0.15 | 0.3 | 0 | 1.35 | 1.2 | 0.6 |
| D | 1.35 | 1.3 | 1.35 | 0 | 0.4 | 1.4 |
| E | 1.25 | 1.2 | 1.2 | 0.4 | 0 | 1.3 |
| F | 0.6 | 0.65 | 0.6 | 1.4 | 1.3 | 0 |

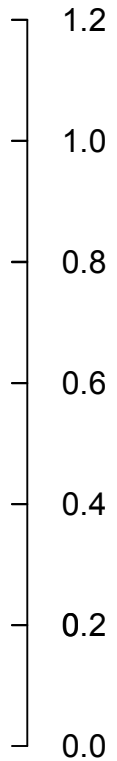
$$d_{x,y} = 1 - r_{x,y}$$

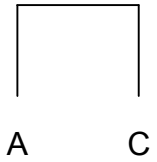
# Ιεραρχική Ομαδοποίηση

| | A | B | C | D | E | F |
|---|---|---|---|---|---|---|
| A | 0 | 0.3 | **0.15** | 1.35 | 1.25 | 0.6 |
| B | 0.3 | 0 | 0.3 | 1.3 | 1.2 | 0.65 |
| C | **0.15** | 0.3 | 0 | 1.35 | 1.2 | 0.6 |
| D | 1.35 | 1.3 | 1.35 | 0 | 0.4 | 1.4 |
| E | 1.25 | 1.2 | 1.2 | 0.4 | 0 | 1.3 |
| F | 0.6 | 0.65 | 0.6 | 1.4 | 1.3 | 0 |

| | B | D | E | F | AC |
|---|---|---|---|---|---|
| B | 0 | 1.3 | 1.2 | 0.65 | 0.3 |
| D | 1.3 | 0 | 0.4 | 1.4 | 1.35 |
| E | 1.2 | 0.4 | 0 | 1.3 | 1.2 |
| F | 0.65 | 1.4 | 1.3 | 0 | 0.6 |
| AC | 0.3 | 1.35 | 1.2 | 0.6 | 0 |

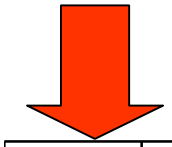
Complete: maximum (1.25)
Average: mean (1.225)
Single: minimum (1.2)



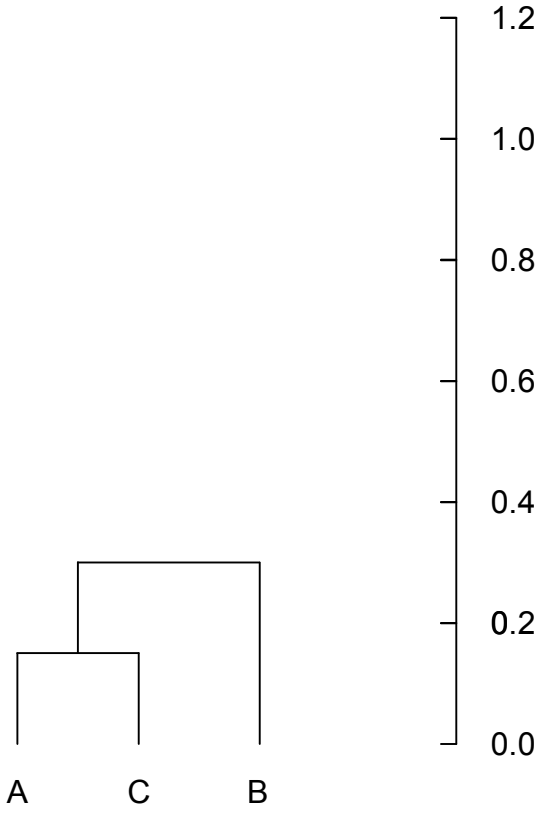



# Ιεραρχική Ομαδοποίηση

| | B | D | E | F | AC |
|---|---|---|---|---|---|
| B | 0 | 1.3 | 1.2 | 0.65 | **0.3** |
| D | 1.3 | 0 | 0.4 | 1.4 | 1.35 |
| E | 1.2 | 0.4 | 0 | 1.3 | 1.2 |
| F | 0.65 | 1.4 | 1.3 | 0 | 0.6 |
| AC | **0.3** | 1.35 | 1.2 | 0.6 | 0 |



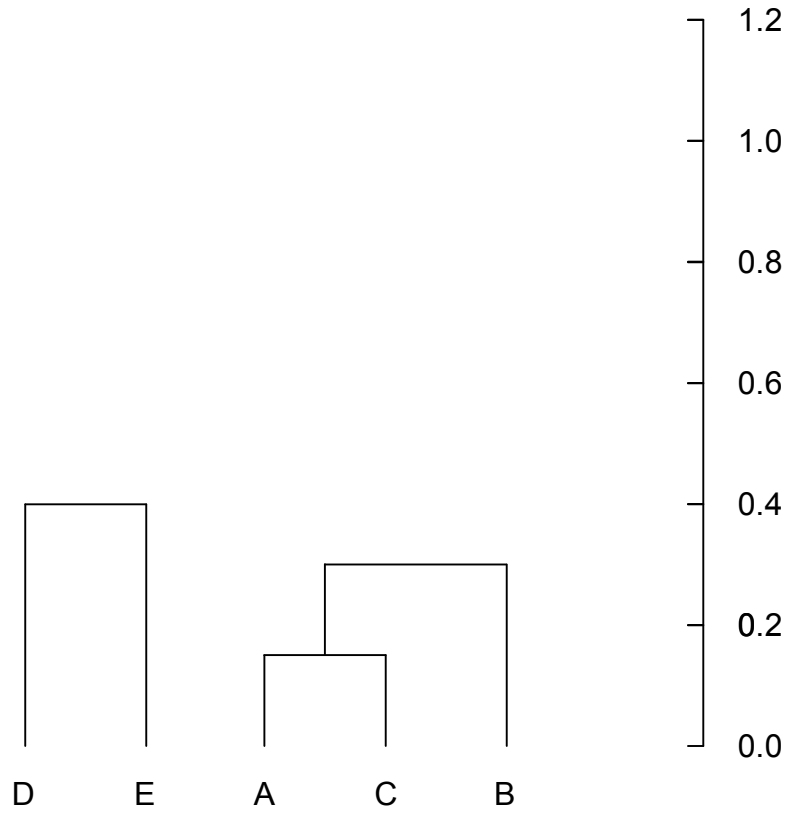
| | D | E | F | ABC |
|---|---|---|---|---|
| D | 0 | 0.4 | 1.4 | 1.3 |
| E | 0.4 | 0 | 1.3 | 1.2 |
| F | 1.4 | 1.3 | 0 | 0.6 |
| ABC | 1.3 | 1.2 | 0.6 | 0 |



# Ιεραρχική Ομαδοποίηση

| | D | E | F | ABC |
|---|---|---|---|---|
| D | 0 | **0.4** | 1.4 | 1.3 |
| E | **0.4** | 0 | 1.3 | 1.2 |
| F | 1.4 | 1.3 | 0 | 0.6 |
| ABC | 1.3 | 1.2 | 0.6 | 0 |

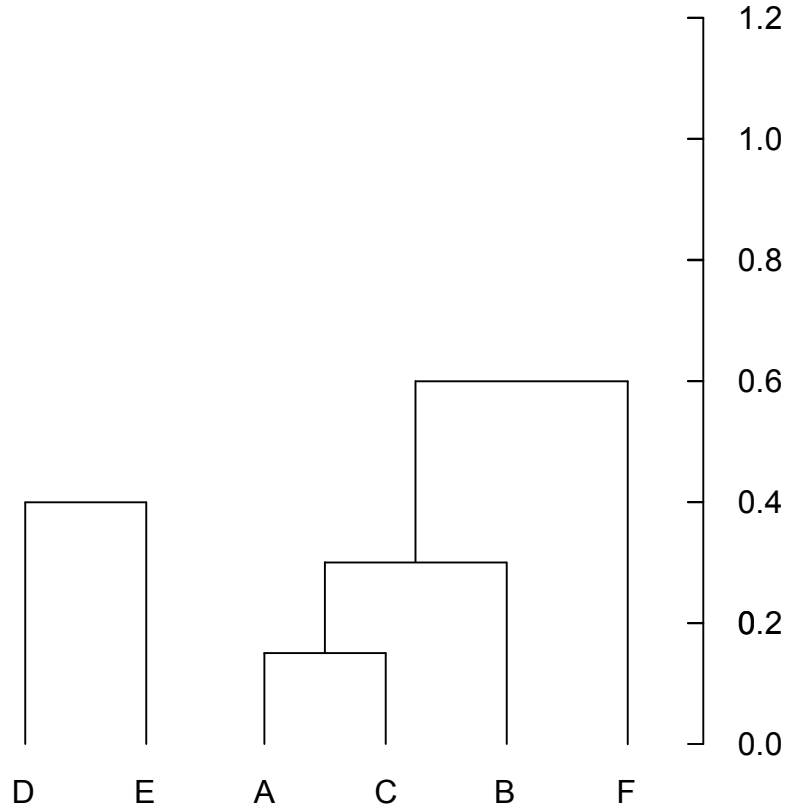
| | F | ABC | DE |
|---|---|---|---|
| F | 0 | 0.6 | 1.3 |
| ABC | 0.6 | 0 | 1.2 |
| DE | 1.3 | 1.2 | 0 |



# Ιεραρχική Ομαδοποίηση

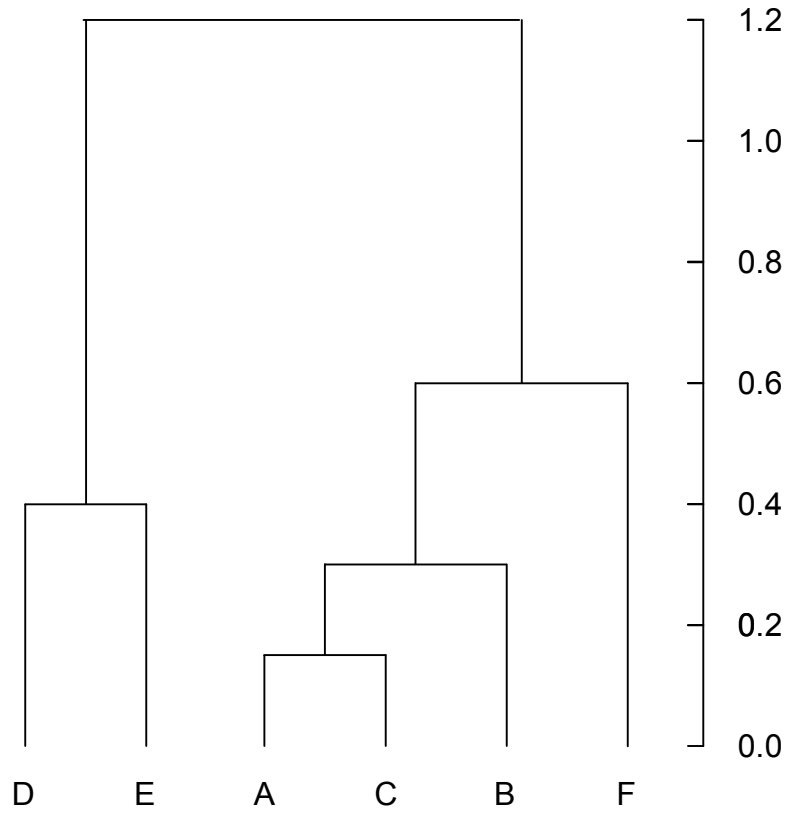
| | F | ABC | DE |
|---|---|---|---|
| F | 0 | **0.6** | 1.3 |
| ABC | **0.6** | 0 | 1.2 |
| DE | 1.3 | 1.2 | 0 |

| | ABCF | DE |
|---|---|---|
| ABCF | 0 | 1.2 |
| DE | 1.2 | 0 |

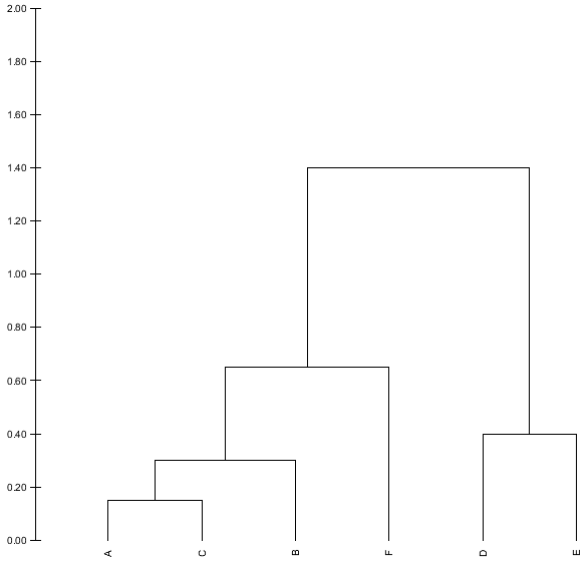


# Ιεραρχική Ομαδοποίηση

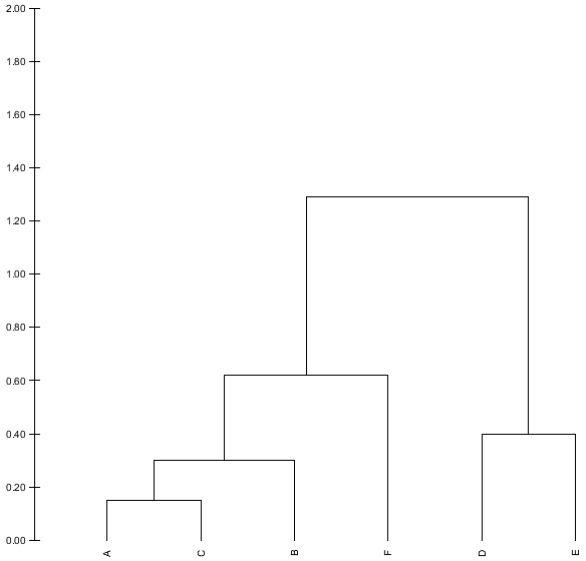
| | ABCF | DE |
|---|---|---|
| ABCF | 0 | **1.2** |
| DE | **1.2** | 0 |



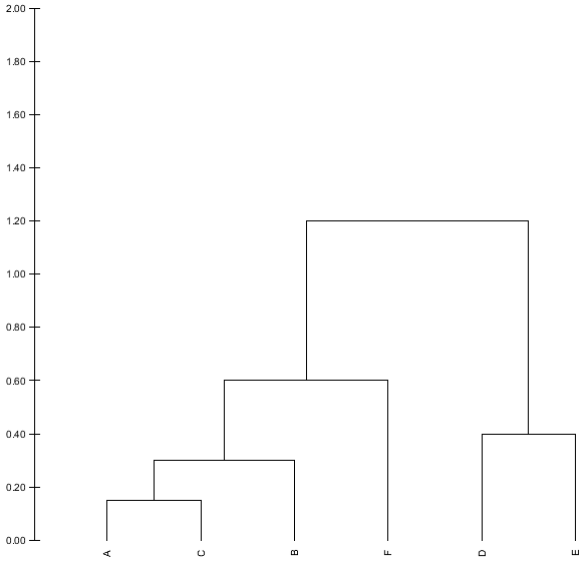
## Complete




## Average

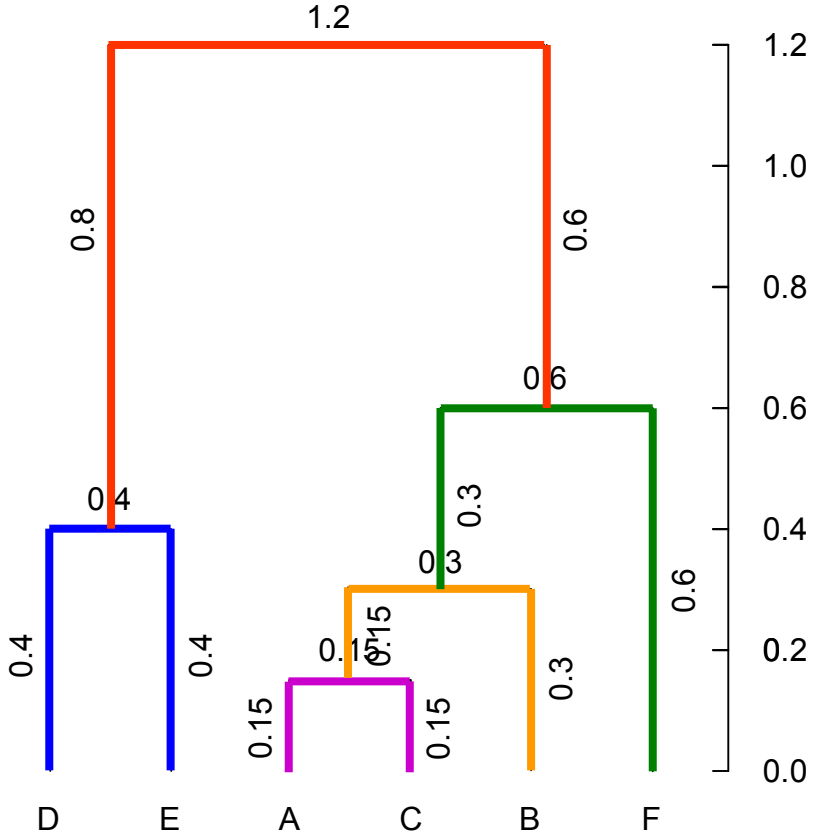



## Single

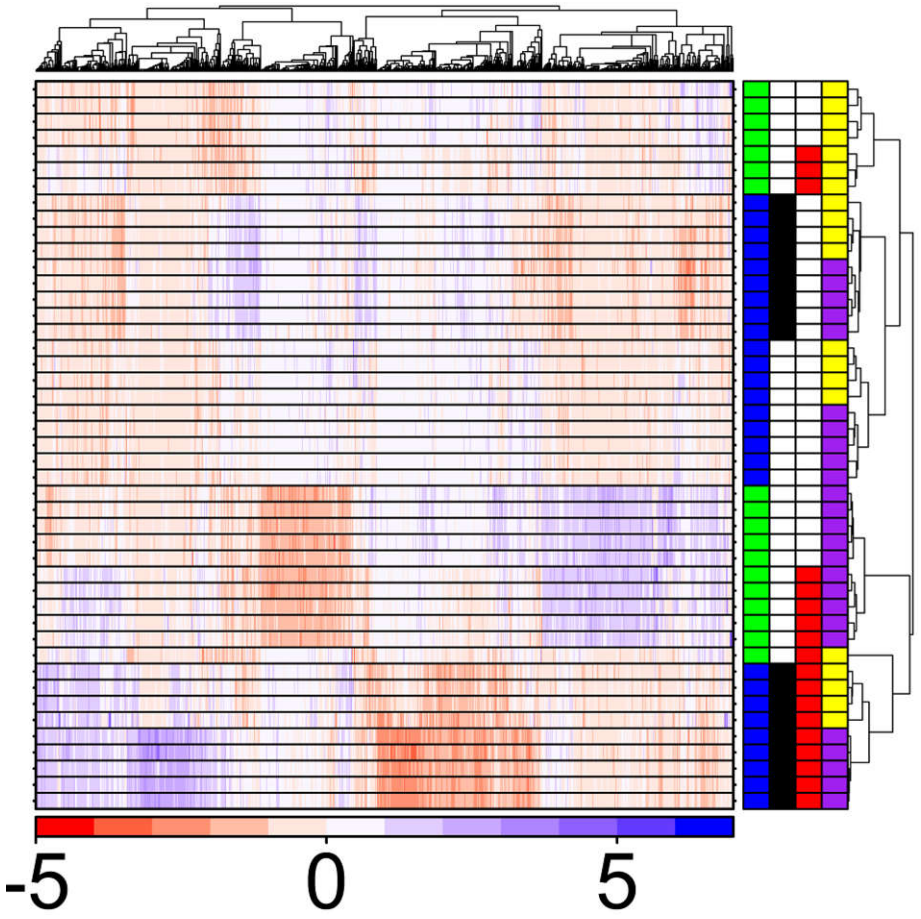


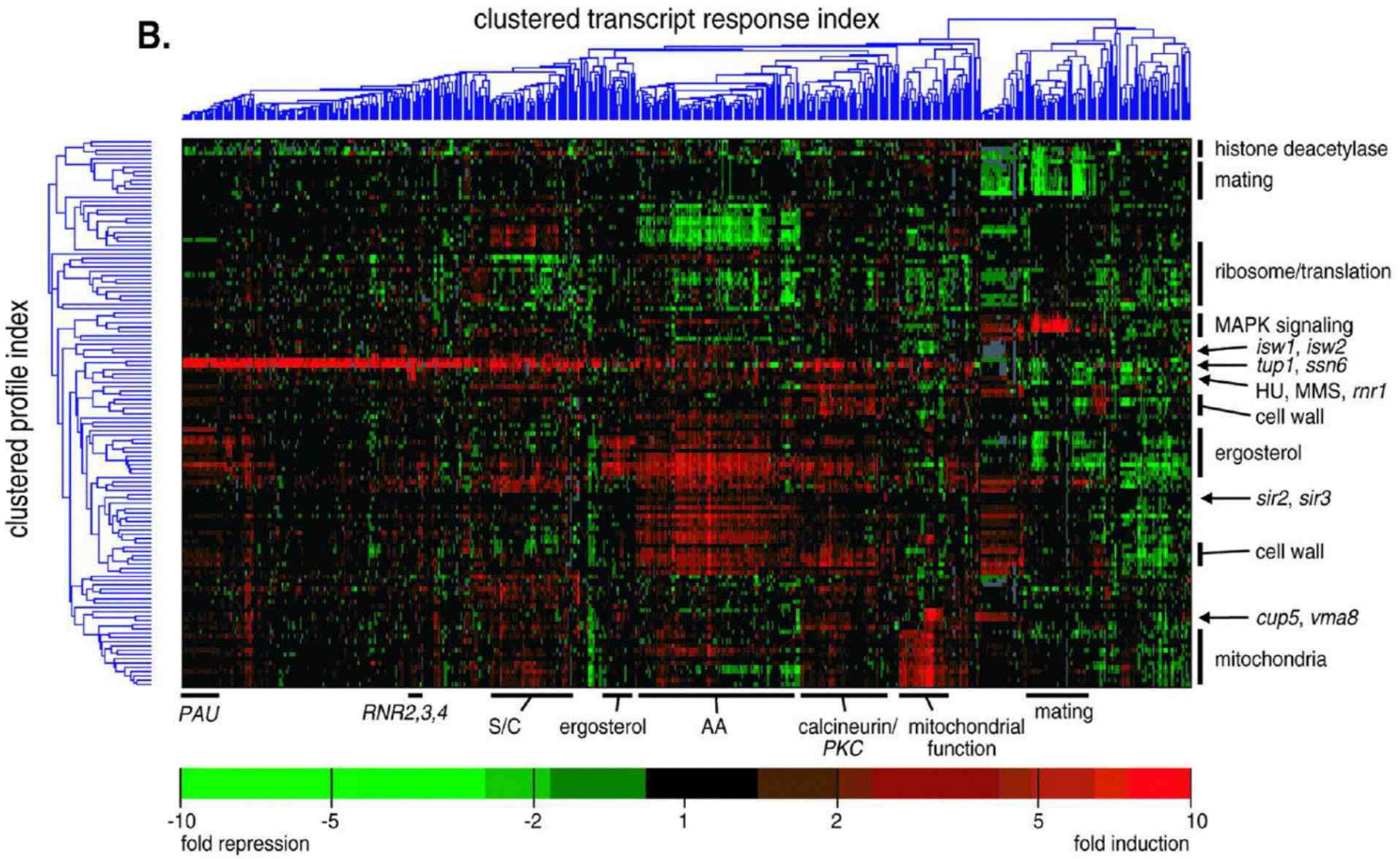
# Newick Format

```
(
  (
    D:0.4
    ,
    E:0.4
  )
  ,
  (
    A:0.15
    
  )
);
```



# Heat Map




B.
clustered transcript response index
clustered profile index
histone deacetylase
mating
ribosome/translation
MAPK signaling
isw1, isw2
tup1, ssn6
HU, MMS, rnr1
cell wall
ergosterol
sir2, sir3
cell wall
cup5, vma8
mitochondria
PAU
RNR2,3,4
S/C
ergosterol
AA
calcineurin/ PKC
mitochondrial function
mating
-10
-5
-2
1
2
5
10
fold repression
fold induction

# Neighbour Joining

| | a | b | c | d | e |
|---|---|---|---|---|---|
| a | 0.0 | 1.0 | 1.8 | 1.8 | 1.6 |
| b | 1.0 | 0.0 | 2.0 | 2.0 | 1.8 |
| c | 1.8 | 2.0 | 0.0 | 1.6 | 1.4 |
| d | 1.8 | 2.0 | 1.6 | 0.0 | 0.6 |
| e | 1.6 | 1.8 | 1.4 | 0.6 | 0.0 |
| $\sum_{k=1}^{r} d(i,k)$ | 6.2 | 6.8 | 6.8 | 6.0 | 5.4 |

$$Q(i,j) = (n-2)d(i,j) - \sum_{k=1}^{n} d(i,k) - \sum_{k=1}^{n} d(j,k)$$

Q-matrix

| | a | b | c | e | d |
|---|---|---|---|---|---|
| a | | -10.0 | -7.6 | -6.8 | -6.8 |
| b | -10.0 | | -7.6 | -6.8 | -6.8 |
| c | -7.6 | -7.6 | | -8.0 | -8.0 |
| d | -6.8 | -6.8 | -8.0 | | -9.6 |
| e | -6.8 | -6.8 | -8.0 | -9.6 | |

| | a | b | c | d | e |
|---|---|---|---|---|---|
| a | 0.0 | 1.0 | 1.8 | 1.8 | 1.6 |
| b | 1.0 | 0.0 | 2.0 | 2.0 | 1.8 |
| c | 1.8 | 2.0 | 0.0 | 1.6 | 1.4 |
| d | 1.8 | 2.0 | 1.6 | 0.0 | 0.6 |
| e | 1.6 | 1.8 | 1.4 | 0.6 | 0.0 |

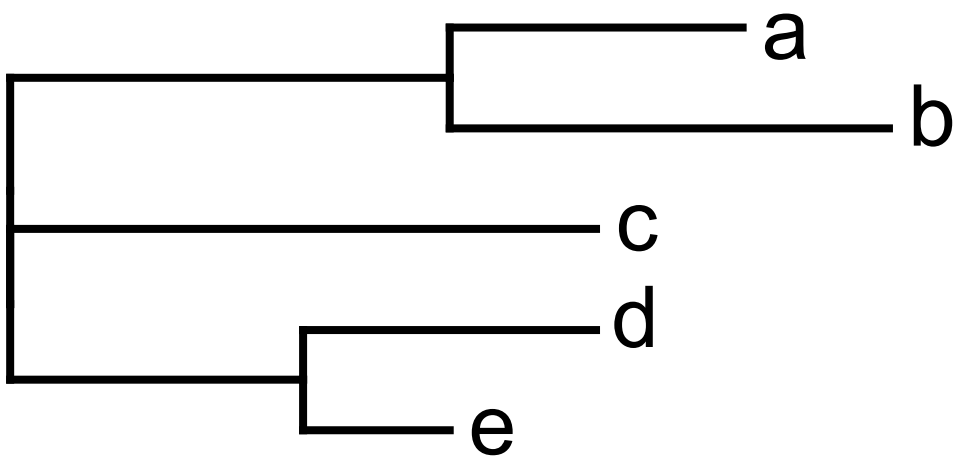
$$d(f,u)=\frac{1}{2}\left(d(f,g)+\frac{\left[\sum_{k=1}^{n}d(f,k)-\sum_{k=1}^{n}d(g,k)\right]}{(n-2)}\right)$$

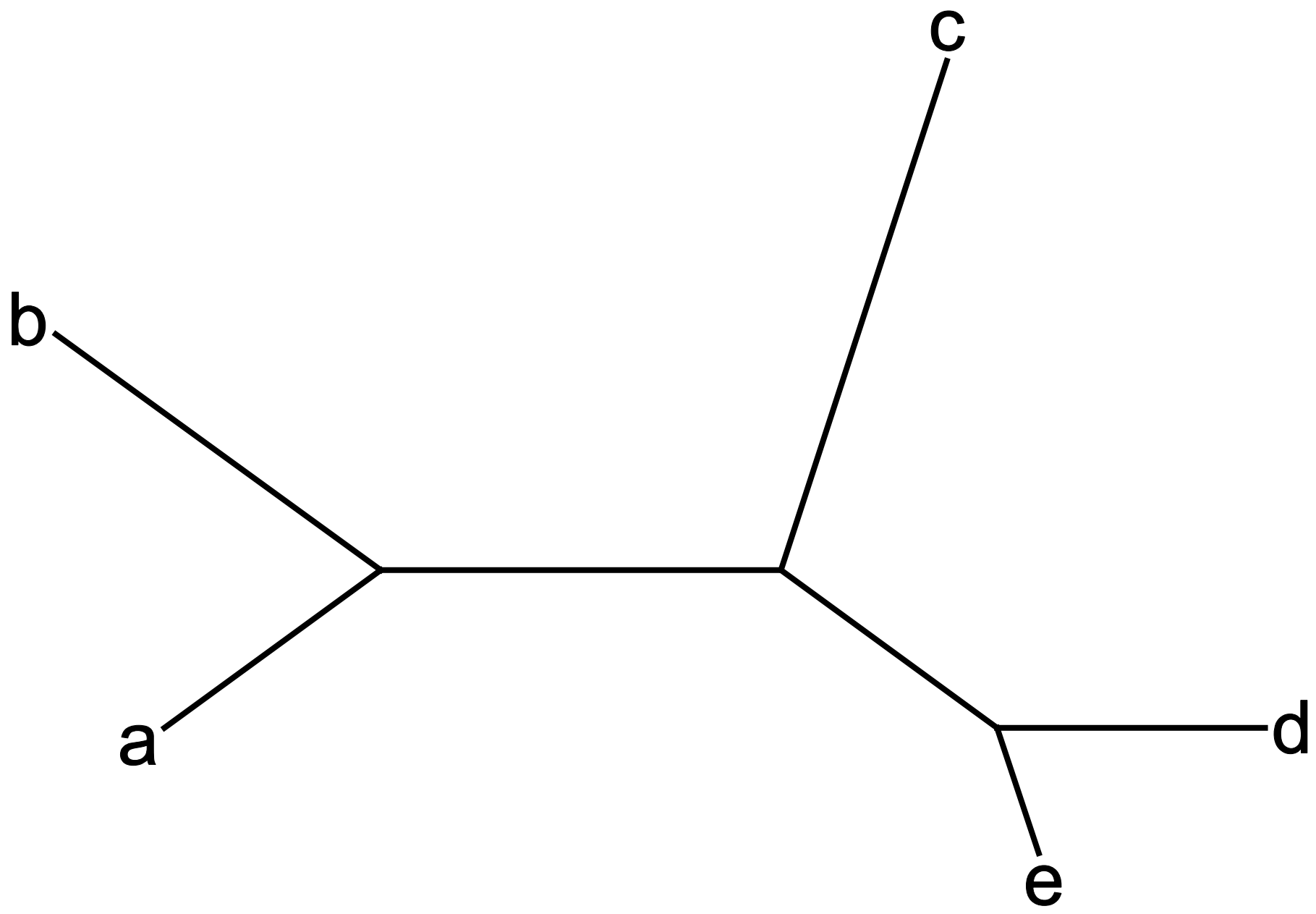
| | ab |
|---|---|
| a | 0.4 |
| b | 0.6 |

$$d(u,k)=\frac{d(f,k)+d(g,k)-d(f,g)}{2}$$

| | ab | c | d | e |
|---|---|---|---|---|
| ab | 0.0 | 1.4 | 1.4 | 1.2 |
| c | 1.4 | 0.0 | 1.6 | 1.4 |
| d | 1.4 | 1.6 | 0.0 | 0.6 |
| e | 1.2 | 1.4 | 0.6 | 0.0 |

((a:0.4,b:0.6):0.6,(c:0.8,(d:0.4,e:0.2):0.4):0);




c
b
a
d
e

# Γενικά

•Όταν έχουμε πολλές παρατηρήσεις (δείγματα ή γονίδια), καθεμία από τις μεταβλητές τους (γονίδια ή δείγματα, αντίστοιχα), θα μπορούσε να θεωρηθεί ως μια διαφορετική διάσταση, σε ένα d-διάστατο χώρο

# Ανάλυση Κυρίων Συνιστωσών

- Ένας πολυδιάστατος χώρος είναι συχνά δύσκολο να απεικονιστεί
  - Κύριος στόχος των μεθόδων μάθησης χωρίς επίβλεψη είναι η μείωση του αριθμού των διαστάσεων, βαθμολογώντας όλες τις παρατηρήσεις, ώστε να ομαδοποιήσουν παρόμοιες παρατηρήσεις μαζί, με βάση πολλαπλές μεταβλητές
- Η μείωση των πολλαπλών μεταβλητών σε μία, δύο ή τρεις, που μπορεί να παρασταθεί γραφικά με ελάχιστη απώλεια πληροφορίας, είναι χρήσιμη στην ανακάλυψη γνώσης
- Η ανάλυση κυρίων συνιστωσών (PCA), μια δημοφιλής τεχνική πολλών μεταβλητών, χρησιμοποιείται κυρίως για την μείωση των διαστάσεων των d μεταβλητών σε δύο ή τρεις διαστάσεις

# Ανάλυση Κυρίων Συνιστωσών

•Η ανάλυση κυρίων συνιστωσών συνίσταται στον ορθογώνιο μετασχηματισμό που μετατρέπει ένα σύνολο παρατηρήσεων -πιθανώς σχετιζόμενω-ν μεταβλητών σε ένα σύνολο τιμών γραμμικώς μη συσχετιζόμενων μεταβλητών που ονομάζονται κύριες συνιστώσες

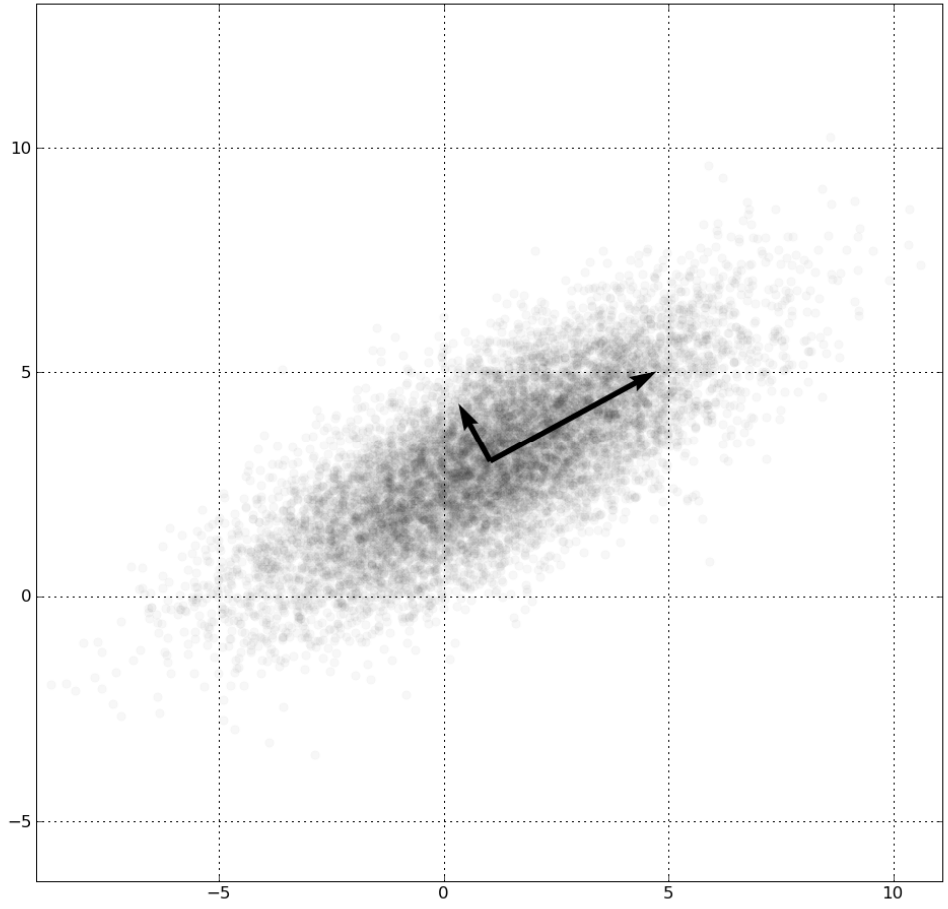
•Ο αριθμός των κυρίων συνιστωσών είναι μικρότερος ή ίσος με τον αριθμό των μεταβλητών

•Ο μετασχηματισμός γίνεται έτσι ώστε η πρώτη κύρια συνιστώσα να περιέχει τη μέγιστη δυνατή διακύμανση (δηλαδή τη μεγαλύτερη δυνατή διασπορά των δεδομένων) και κάθε επόμενη συνιστώσα να περιέχει τη μεγαλύτερη διακύμανση από τις υπόλοιπες, με τον περιορισμό να είναι ορθογώνια (μη συσχετιζόμενες), ως προς τις προηγούμενες

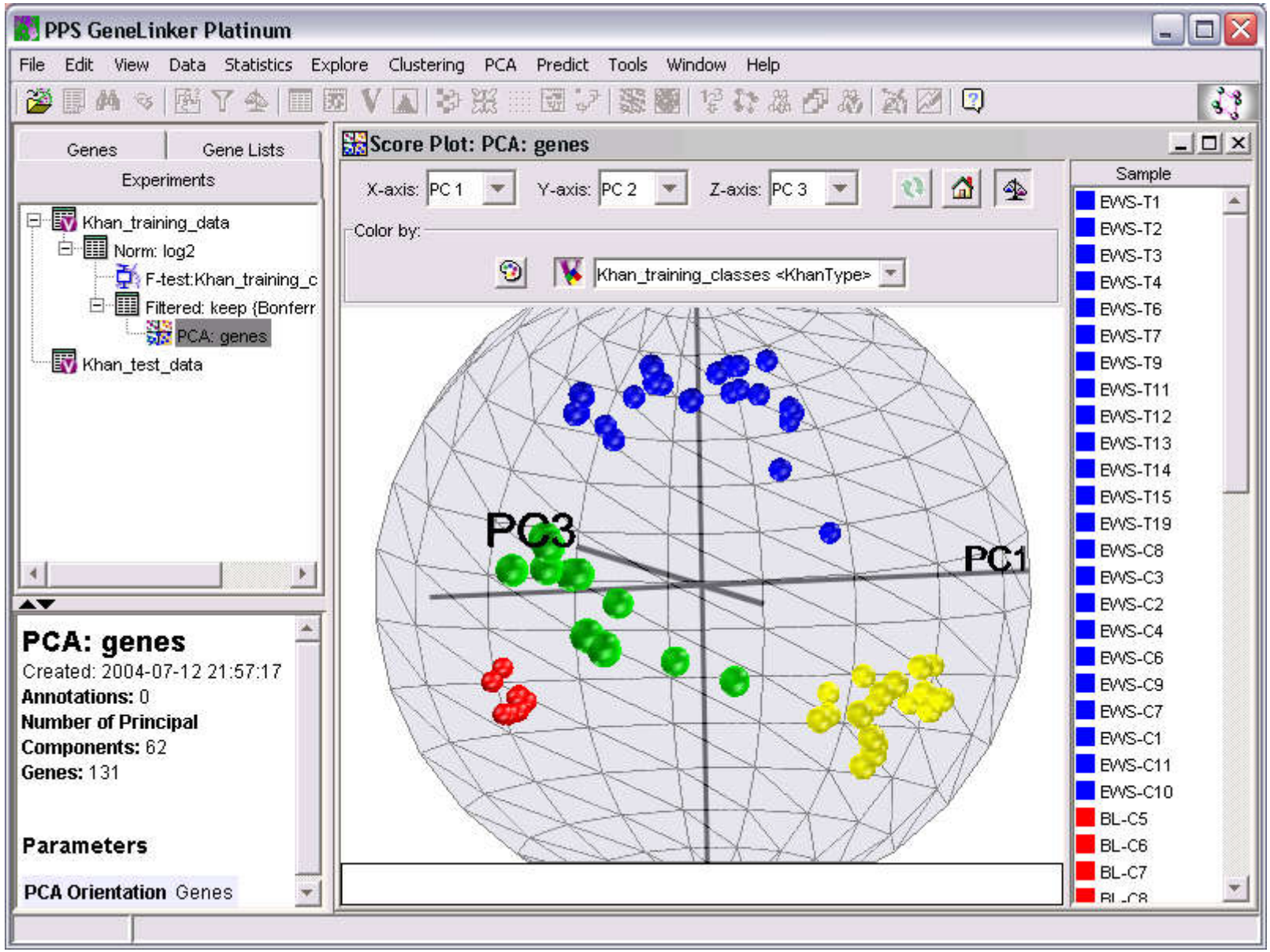
# Ανάλυση Κυρίων Συνιστωσών

- Κάθε συνιστώσα έχει ένα eigenvector και μια eigenvalue
- Οι eigenvalues μετρούν το ποσό της διακύμανσης που εξηγείται από κάθε κύρια συνιστώσα (είναι μέγιστη στην πρώτη βασική συνιστώσα και μειώνεται σταδιακά στις υπόλοιπες)
- Μια eigenvalue άνω του 1 σημαίνει ότι η συνιστώσα παρέχει περισσότερη διακύμανση από μια από τις αρχικές μεταβλητές των τυποποιημένων δεδομένων.
    - Το άθροισμα των eigenvalues είναι ίσο με τον αριθμό των αρχικών μεταβλητών
    - Έτσι, διαιρώντας το άθροισμα των eigenvalues των πρώτων κυρίων μεταβλητών με τον αριθμό των μεταβλητών, μπορούμε να γνωρίζουμε το ποσοστό της διακύμανσης που περιέχουν.
        - Στις μικροσυστοιχίες, οι τρεις πρώτες κύριες συνιστώσες είναι περιέχουν πάνω από το 97% της συνολικής διακύμανσης
        - Επιτρέπεται η απεικόνιση σε 2 ή 3 διαστάσεις των σχέσεων των δειγμάτων ή γονιδίων

# Ανάλυση Κυρίων Συνιστωσών



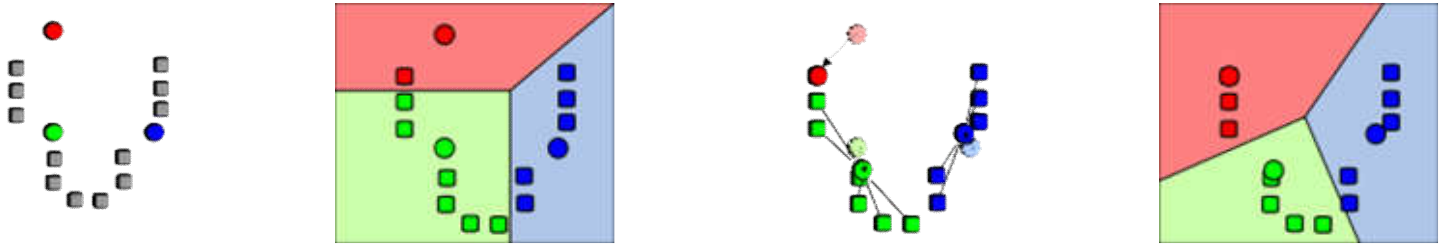
# Ανάλυση Κυρίων Συνιστωσών



# Ομαδοποίηση K-means

- Ορίζεται ο αριθμός των ομάδων (K)
- Αρχικός τυχαίος διαχωρισμός σε K ομάδες
- Βαθμιαία βελτίωση της ομαδοποίησης, μέχρι σύγκλιση (μη περαιτέρω βελτίωση)

# Ομαδοποίηση K-means



1. Τυχαία δημιουργία k αρχικών «μέσων τιμών» (σε αυτήν την περίπτωση k=3) εντός της περιοχής των δεδομένων (δείχνονται με χρώμα)
2. Δημιουργούνται *k* ομάδες μέσω της σύνδεσης κάθε παρατήρησης με την εγγύτερη μέση τιμή. Οι διαχωρισμοί εδώ αναπαρίστανται με διαγράμματα Voronoi που παράγονται από τις μέσες τιμές
3. Το κεντροειδές καθενός από τις k ομάδες γίνεται η μέση τιμή του
4. Τα βήματα 2 και 3 επαναλαμβάνονται ώσπου να υπάρχει σύγκλιση

# Ομαδοποίηση K-means

Με δεδομένων ένα σύνολο παρατηρήσεων ($\mathbf{x}_1$, $\mathbf{x}_2$, …, $\mathbf{x}_n$), όπου κάθε παρατήρηση είναι ένα d-διάστατο πραγματικό διάνυσμα, η ομαδοποίηση k-means στοχεύει στο διαχωρισμό των $n$ παρατηρήσεων σε $k$ (≤ $n$) ομάδες $\mathbf{S} = \{S_1, S_2, \ldots, S_k\}$, ώστε να ελαχιστοποιήσει το εντός ομάδων άθροισμα των τετραγώνων, δηλαδή να ανακαλύψει το:

$$\arg\min_{S} \sum_{i=1}^{k} \sum_{x \in S_i} \left\| x - \mu_i \right\|^2$$

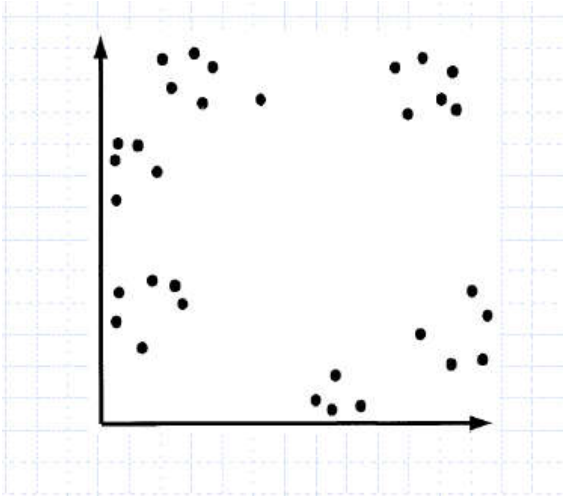
όπου $\boldsymbol{\mu}_i$ είναι η μέση τιμή των σημείων του $S_i$.

# Αυτό-οργανωμένοι χάρτες

- Ορισμός αριθμού ομάδων k σε ένα πλέγμα k=p*g κεντροειδών
- Μετακίνηση του πλέγματος προς τα σημεία

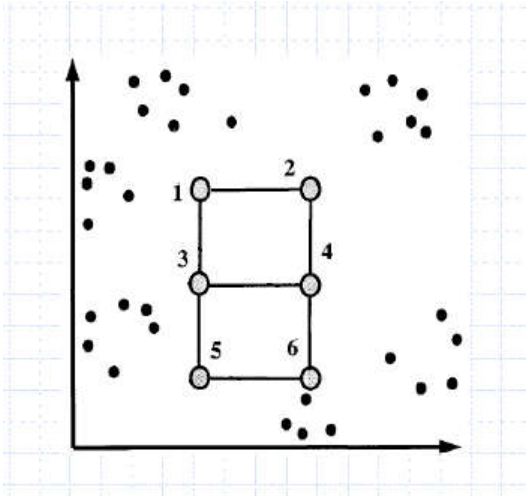
# Αυτό-οργανωμένοι χάρτες

Επιλογή k=6 ομάδων



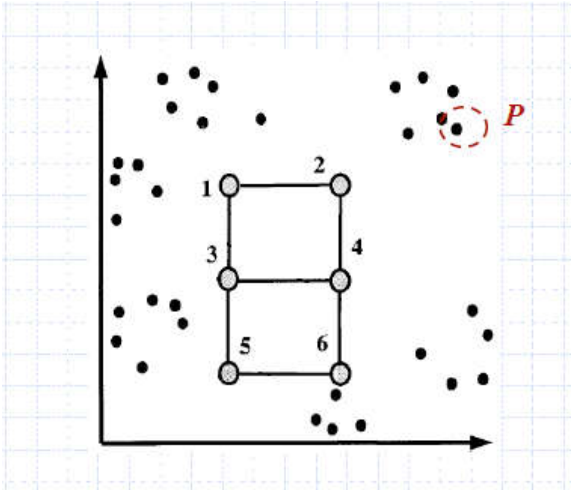
# Αυτό-οργανωμένοι χάρτες

Επιλογή τυχαίας θέσης
πλέγματος 6 κεντροειδών



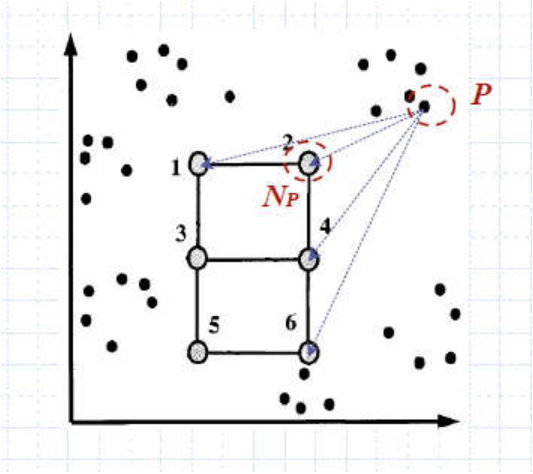
# Αυτό-οργανωμένοι χάρτες

Επιλογή τυχαίου σημείου



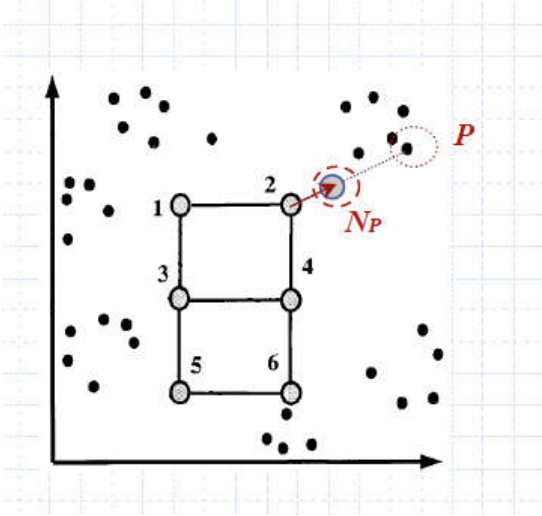
# Αυτό-οργανωμένοι χάρτες

Εύρεση πλησιέστερου κεντροειδούς



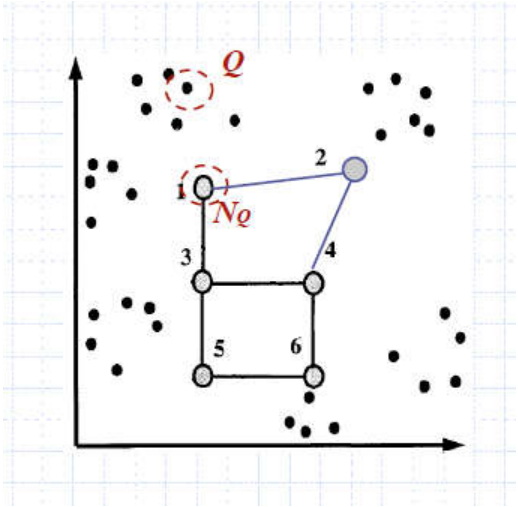
# Αυτό-οργανωμένοι χάρτες

Μετακίνηση κεντροειδούς
προς το σημείο



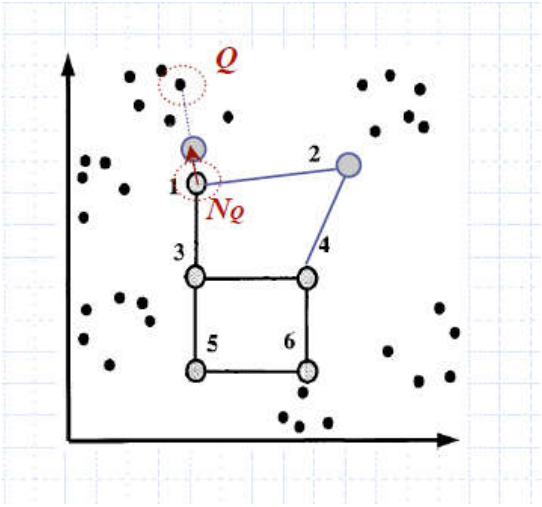
# Αυτό-οργανωμένοι χάρτες

Επανάληψη για νέο τυχαίο σημείο



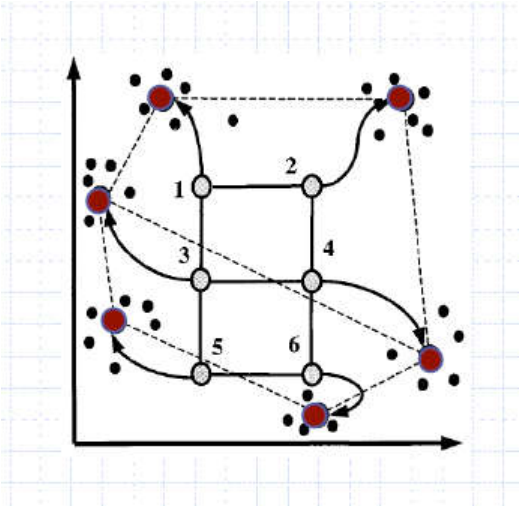
# Αυτό-οργανωμένοι χάρτες

Επανάληψη



# Αυτό-οργανωμένοι χάρτες

Επανάληψη μέχρι να υπάρξει σύγκλιση



# Υπεργεωμετρική κατανομή

| | drawn | not drawn | total |
|---|---|---|---|
| **white** | $k$ | $m - k$ | $m$ |
| **black** | $n - k$ | $N + k - n - m$ | $N - m$ |
| **total** | $n$ | $N - n$ | $N$ |

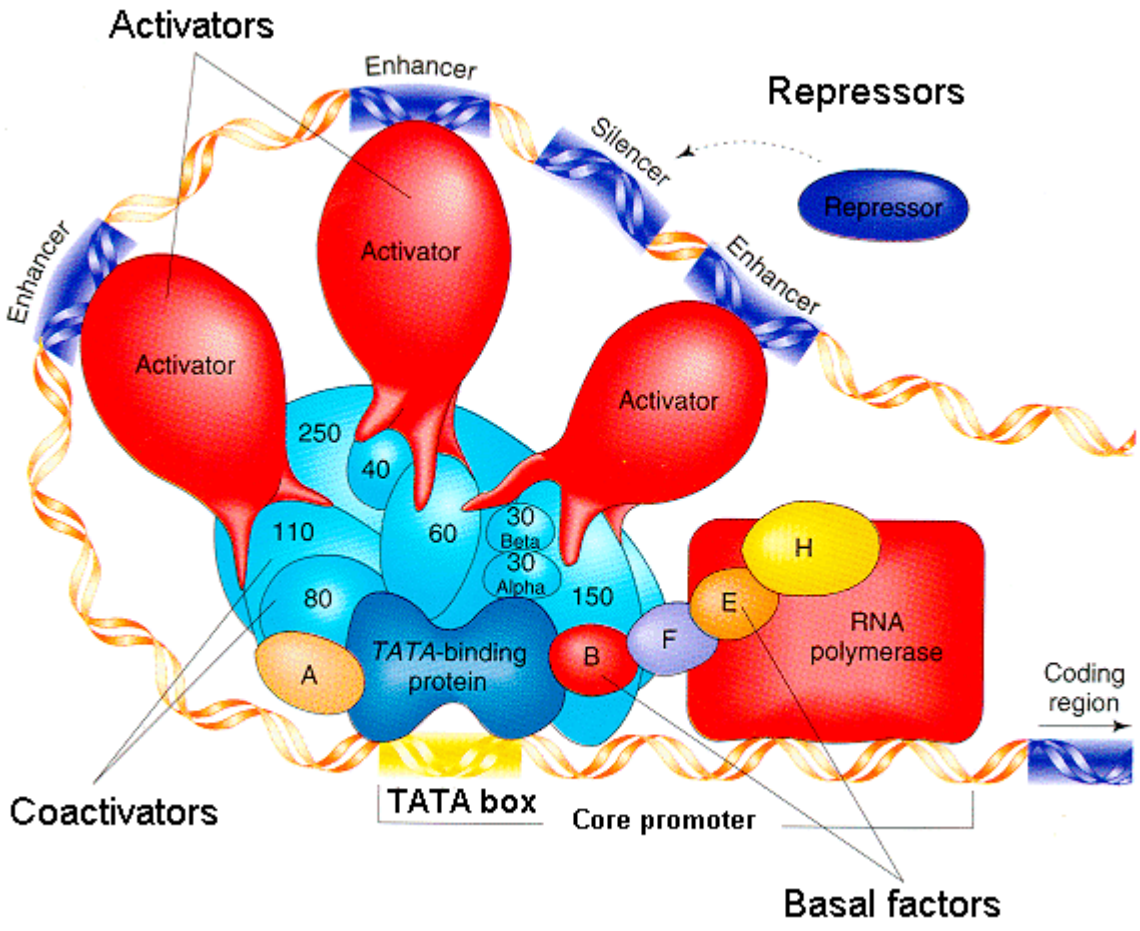
$$P(X = k) = \frac{\binom{m}{k}\binom{N-m}{n-k}}{\binom{N}{n}}$$

$$\binom{n}{k} = \frac{n!}{k!(n-k)!}$$

$$P(X \leq K) = \sum_{k=0}^{K} P(X = k)$$

$$P(X \geq K) = \sum_{k=K}^{n} P(X = k)$$

# Μεταγραφή

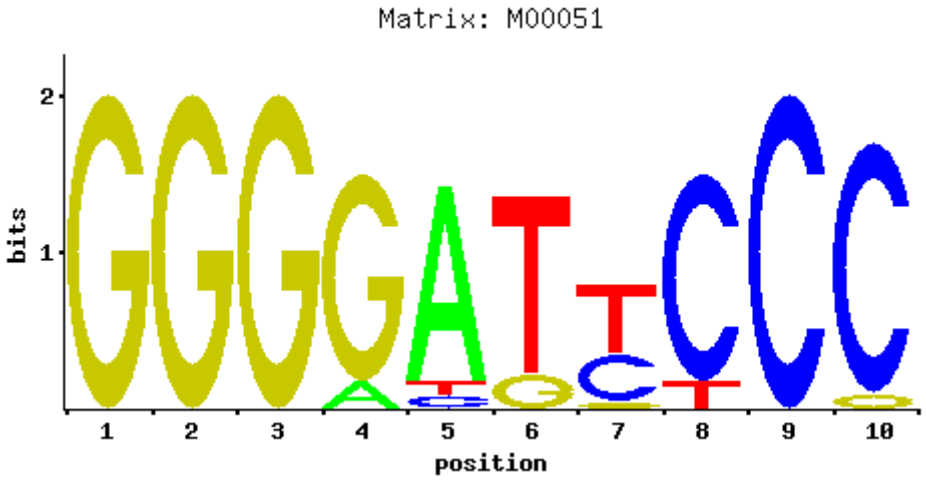

Activators
Enhancer
Repressors
Silencer
Repressor
Enhancer
Enhancer
Activator
Activator
Activator
250
40
110
60
30
Beta
30
Alpha
80
150
H
E
RNA
polymerase
F
B
A
TATA-binding
protein
Coding
region
Coactivators
TATA box
Core promoter
Basal factors

# TransFac

Βάση δεδομένων ευκαρυωτικών cis-δρώντων ρυθμιστικών στοιχείων του DNA και trans-δρώντων παραγόντων. Καλύπτει όλο το φάσμα από τη ζύμη ως τον άνθρωπο.

# Πίνακες TransFac

```
AC   M00051
XX
ID   V$NFKAPPAB50_01
XX
DT   13.04.1995 (created); hiwi.
DT   04.12.2003 (updated); vma.
CO   Copyright (C), Biobase GmbH.
XX
NA   NF-kappaB (p50)
XX
DE   NF-kappaB (p50)
XX
BF   T00593 p50; Species: human, Homo sapiens.
XX
PO       A       C       G       T
01       0       0      18       0       G
02       0       0      18       0       G
03       0       0      18       0       G
04       2       0      16       0       G
05      16       1       0       1       A
06       0       0       3      15       T
07       0       7       1      10       Y
08       0      16       0       2       C
09       0      18       0       0       C
10       0      17       1       0       C
XX
BA   18 selected binding sequences
XX
CC   oligonucleotides binding to bacterially expressed NF-kappaB (p50) were selected (gel shift) and
amplified (PCR) in 3 cycles
XX
RN   [1]; RE0002922.
RX   PUBMED: 1406630.
RA   Kunsch C., Ruben S. M., Rosen C. A.
RT   Selection of optimal kappaB/Rel DNA-binding motifs: interaction of both subunits of NF-kappaB with
DNA is required for transcriptional activation
RL   Mol. Cell. Biol. 12:4412-4421 (1992).
XX
//
```




Πίνακας Βεβαρημένων Θέσεων

# Match

Αναζητά πιθανές θέσεις πρόσδεσης μεταγραφικών παραγόντων συγκρίνοντας μια αλληλουχία DNA με πίνακες βεβαρημένων θέσεων της TransFac

- Ο αλγόριθμος στηρίζεται στο πληροφοριακό περιεχόμενο του πίνακα βεβαρημένων θέσεων της TransFac
- Χρησιμοποιεί δύο τιμές αξιολόγησης (από 0 έως 1):
  - Βαθμολογία ομοιότητας πίνακα
  - Βαθμολογία ομοιότητας πυρήνα (5 πιο συντηρημένες συνεχόμενες θέσεις)
- Θετική αντιστοίχιση υπάρχει όταν και οι δύο τιμές υπερβαίνουν την ουδό που έχει οριστεί

# Match

- Υπολογίζουμε την τιμή Αβεβαιότητας κατά Shannon *H(i)* στη θέση *i* του πίνακα συχνο-τήτων, όπου *b* η κάθε βάση και *f(b,i)* η συχνότητα εμφάνισης της βάσης *b* στη θέση *i*

$$H(i) = -\sum_{b=a}^{t} f(b,i)\log_2 f(b,i)$$

- Υπολογίζουμε το ποσό της πληροφορίας I(i) στη θέση *i* του πίνακα:

$$I(i) = 2 - H(i)$$

- Υπολογίζουμε το ποσό της πληροφορίας *I(b,i)* κάθε βάσης *b* στη θέση *i* του πίνακα:

$$I(b,i) = f(b,i)I(i)$$

- Υπολογίζουμε το σκορ ομοιότητας πίνακα με κάθε παράθυρο της αλληλουχίας μας:

$$mSS = \frac{Current - Min}{Max - Min}$$

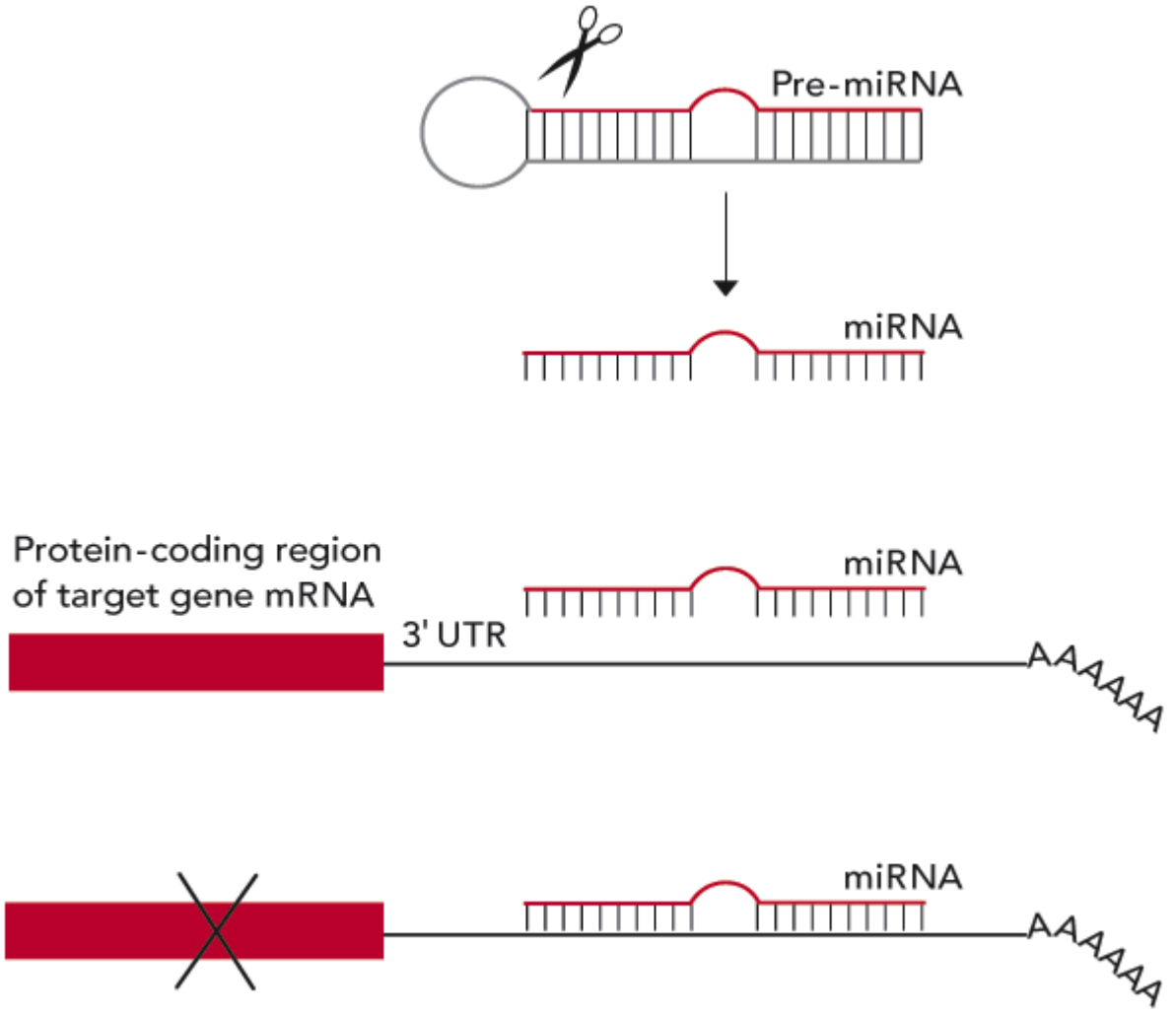
όπου:

$$Current = \sum_{i=1}^{L} I(b,i)$$

$$Min = \sum_{i=1}^{L} \min\left(I(b,i)\right)$$

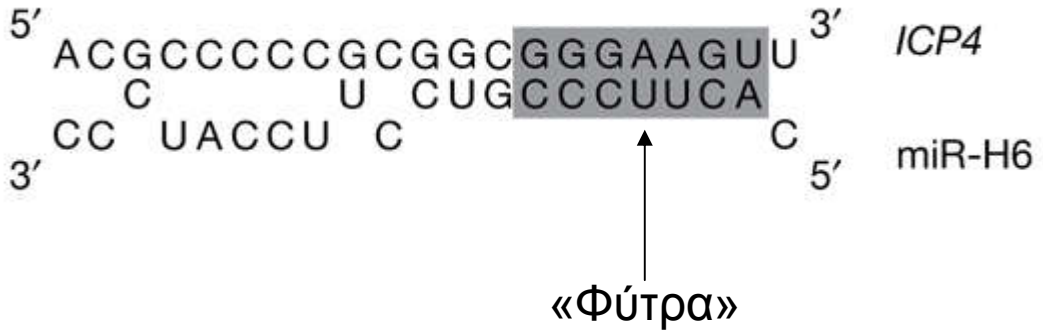
$$Max = \sum_{i=1}^{L} \max\left(I(b,i)\right)$$

# RNA Interference



# TargetScan

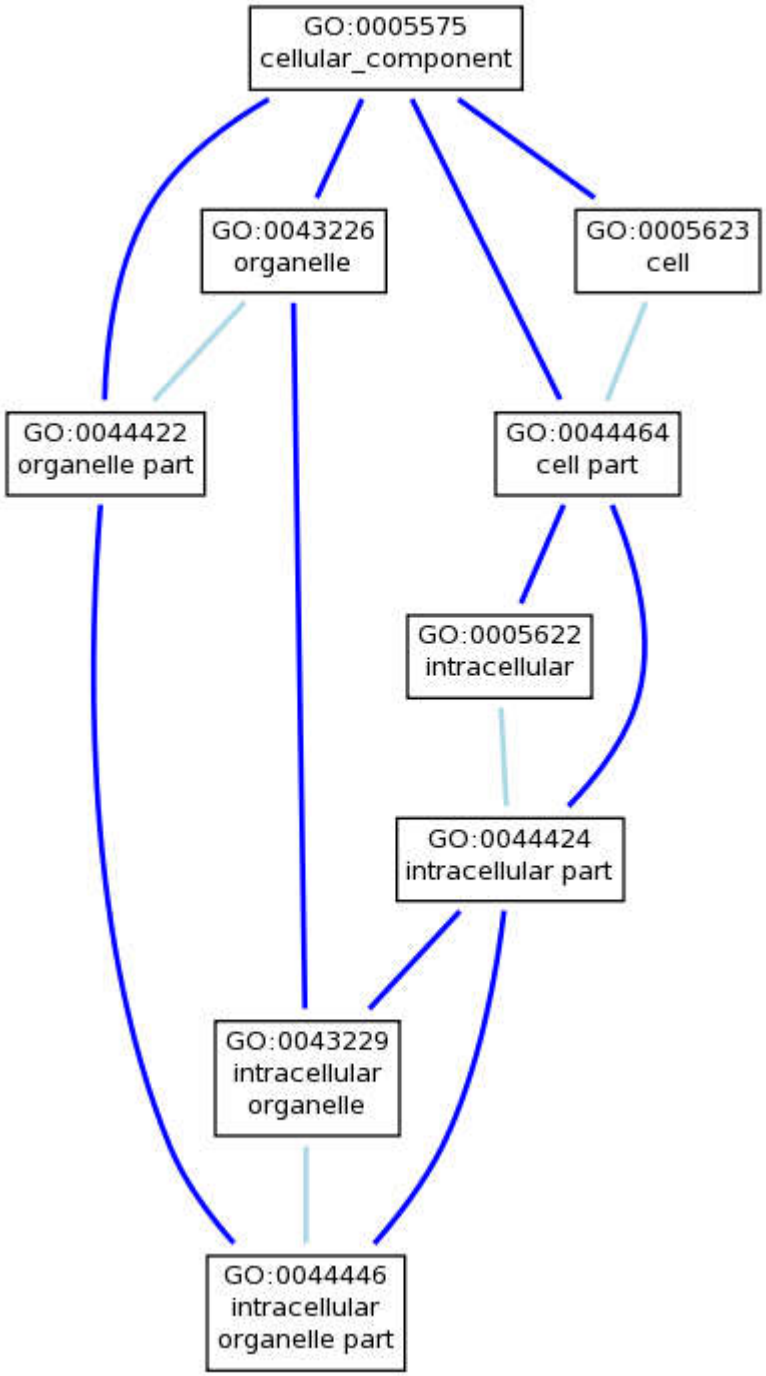
Προβλέπει βιολογικούς στόχους των miRNAs ψάχνοντας για την παρουσία συντηρημένων (ή μη) 8μερών και 7μερών θέσεων που αντιστοιχούν με τη «φύτρα» κάθε miRNA



# Gene Ontology

Σημαντική βιοπληροφορική πρωτοβουλία με στόχο την τυποποίηση της εκπροσώπησης των γονιδίων και των γνωρισμάτων των γονιδιακών προϊόντων μεταξύ ειδών και βάσεων δεδομένων. Παρέχει ένα ελεγχόμενο λεξιλόγιο όρων για την περιγραφή των γνωρισμάτων και των δεδομένων σχολιασμών των γονιδιακών προϊόντων από τα μέλη της κοινοπραξίας GO, καθώς και εργαλεία για την πρόσβαση και την επεξεργασία αυτών των δεδομένων.

Δομή GO: Μη κυκλικός κατευθυνόμενος γράφος πολλαπλής πατρότητας



# KEGG PATHWAY

Συλλογή ενημερωμένων χαρτών οδών που αναπαριστούν τις γνώσεις μας σχετικά με τις μοριακές αλληλεπιδράσεις, δίκτυα αντιδράσεων και δομικές σχέσεις