Βιοπληροφορική

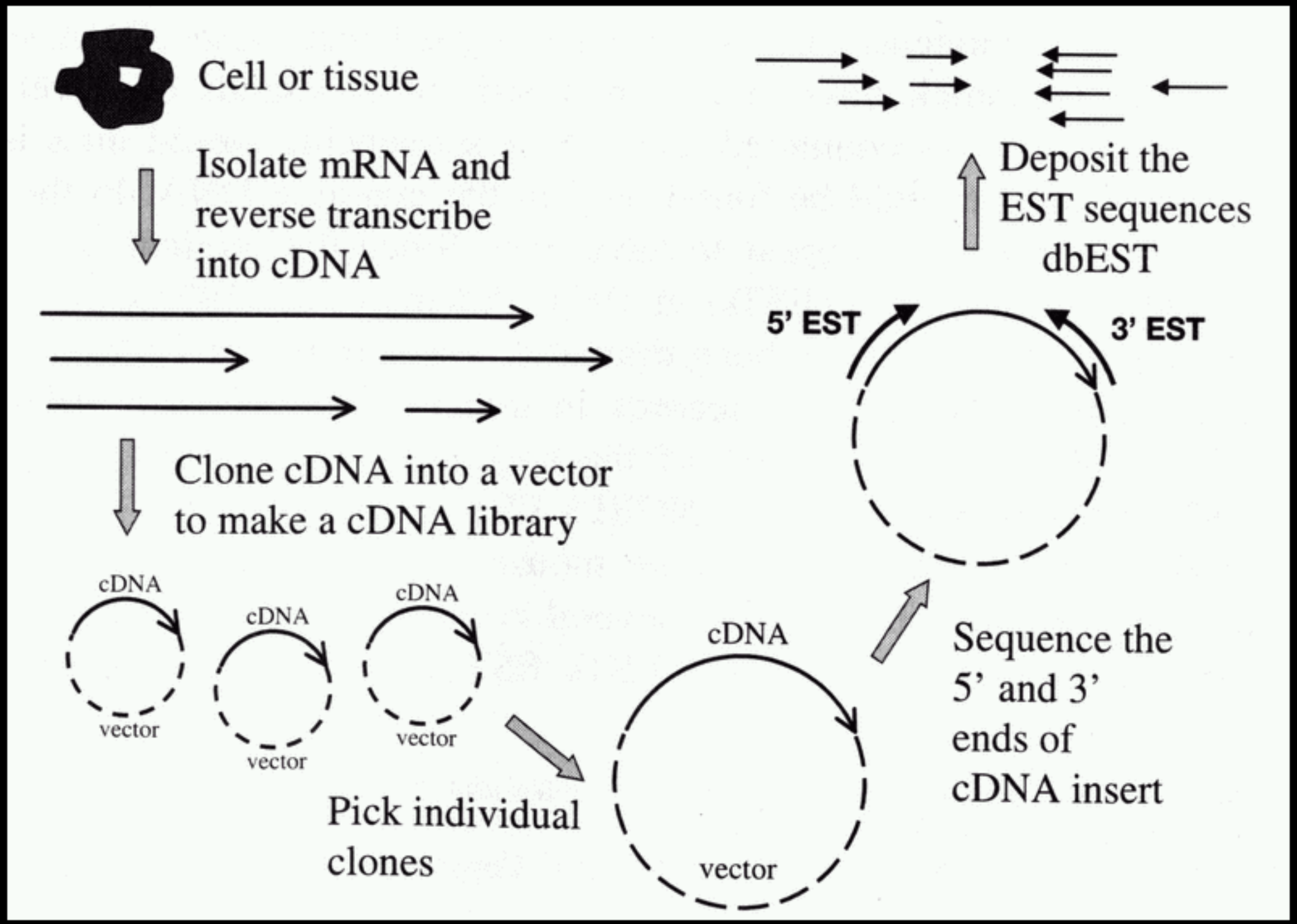
Διάλεξη 6η :

Expressed Sequence Tags, Microarrays.

Εισαγωγή

Τα ESTs είναι μια στρατηγική προσδιορισμού νουκλεοτιδικών αλληλουχιών η οποία δίνει μεγαλύτερη βαρύτητα στον αυτοματισμό και την ταχύτητα παραγωγής των αλληλουχιών παρά στην ποιότητα και την πληρότητα των προκυπτουσών αλληλουχιών. Λόγω της αυτοματοποίησης της διαδικασίας προσδιορισμού αυτών των αλληλουχιών, τα ESTs είναι πλέον η κυρίαρχη (από άποψη πλήθους) ομάδα αλληλουχιών στις βάσεις δεδομένων. Τα αποτελέσματα της χρήσης αυτών των δεδομένων επιβεβαίωσαν τις προβλέψεις των αρχικών υποστηρικτών της μεθόδου για το ότι τα ESTs θα επέτρεπαν τη γρήγορη εύρεση μίας πληθώρας νέων εκφραζόμενων γονιδίων.

Παραγωγή ESTs



Παραγωγή ESTs

Δηλαδή :

- Μια βιβλιοθήκη cDNA κατασκευάζεται από ένα όργανο, ιστό ή κυτταρική γραμμή.
- Για κάθε κλώνο, προσδιορίζεται η αλληλουχία από τα δύο άκρα του cDNA. Έτσι κάθε cDNA έχει ένα 5' και ένα 3' EST που του αντιστοιχεί. Σε μερικές περιπτώ-σεις προσδιορίζεται μόνο η 5' αλληλουχία.

Η διαδικασία αυτή είναι αυτοματοποιημένη σε τέτοιο βαθμό ώστε μερικά ερευνητικά κέντρα να μπορούν να παράγουν περισσότερα από 20000 ESTs ανά εβδομάδα

Προέλευση ESTs

Οι δημόσιες βάσεις δεδομένων περιέχουν ESTs από χιλιάδες cDNA βιβλιοθήκες από περισσότερους από 200 οργανισμούς. Υπάρχουν βιβλιοθήκες από ολόκληρα όργανα, ιστούς και κυτταρικές σειρές, καθώς και βιβλιοθήκες οι οποίες εστιάζουν σε διαφορετικά στάδια διαφοροποίησης ή στα διαφορικά μοτίβα γονιδιακής έκφρασης μεταξύ υγειών και μη κυττάρων. Ειδικά για το τελευταίο, υπάρχουν ειδικές τεχνικές για τη δημιουργία κανονικοποιημένων βιβλιοθηκών στις οποίες όλοι οι κλώνοι αντιπροσωπεύονται με παρόμοιες συχνότητες.

ESTs και βιβλιοθήκες

Ο αριθμός των κλώνων σε μια cDNA library είναι πολύ μεγαλύτερος από τον αριθμό των γονιδίων που εκφράζει το κύτταρο (ή που εκφράζονται σε ένα ιστό ή όργανο). Μια μεγάλη βιβλιοθήκη μπορεί να περιέχει περισσότερους από ένα εκατομμύριο κλώνους οι οποίοι αντιστοιχούν σε μερικές χιλιάδες εκφραζόμενα γονίδια. Τα ESTs προέρχονται από τον προσδιορισμό της αλληλουχίας σε ένα τυχαίο (μικρό) δείγμα αυτού του πληθυσμού.

Προβλήματα με τα ESTs

Ο προσδιορισμός της αλληλουχίας γίνεται σε ένα μόνο πέρασμα από τον αναλυτή. Για το λόγο αυτό οι αλληλουχίες είναι μικρές (μέχρι ~400 βάσεις) και η συχνότητα σφαλμάτων είναι σχετικά υψηλή (στο ~3%). Επίσης, η συχνότητα σφαλμάτων είναι μεγαλύτερη στην αρχή και το τέλος των αλληλουχιών, με το εύρος μεταξύ των θέσεων 100 και 300 να θεωρείται το πλέον αξιόπιστο. Άλλα προβλήματα είναι :

- Η μόλυνση των αλληλουχιών με βακτηριακές, μιτοχονδριακές ή πλασμιδιακές αλληλουχίες.
- Μη ειδική αλληλεπίδραση του πολυ(dT) με το mRNA κατά το στάδιο δημιουργίας της βιβλιοθήκης.

Προβλήματα με τα ESTs

- Τα ESTs μπορούν να πάσχουν από φαινόμενα τυχαίων εισαγωγών ή διαγραφών (με τις αναμενόμενες συνέπειες για τους αλγόριθμους ανάλυσης).
- Σε μερικές περιπτώσεις, αντί για πολυ(dT) χρησιμοποιούνται τυχαίοι εκκινητές για τη δημιουργία του cDNA. Σε αυτές τις περιπτώσεις η θέση του 3' EST είναι άγνωστη.
- Ένα άλλο συχνό πρόβλημα (~6%) είναι η ανεστραμμένη πολικότητα όπου το 5' EST είναι στην πραγματικότητα το 3' και αντίστροφα.
- Τέλος, υπάρχουν περιπτώσεις στις οποίες το 5' EST είναι ενός mRNA και το 3' EST ενός άλλου (χιμαιρικά ESTs).

Βάσεις δεδομένων για ESTs

Οι πρωτοταγείς βάσεις είναι αυτές που έχουν ήδη αναφερθεί (EMBL, GenBank, DDBJ). Επιπλέον, το NCBI συντηρεί την dbEST, για 'database of Expressed Sequence Tags'. Ένα παράδειγμα μίας καταχώρησης EST μέσω της Entrez είναι :

Παράδειγμα καταχώρησης

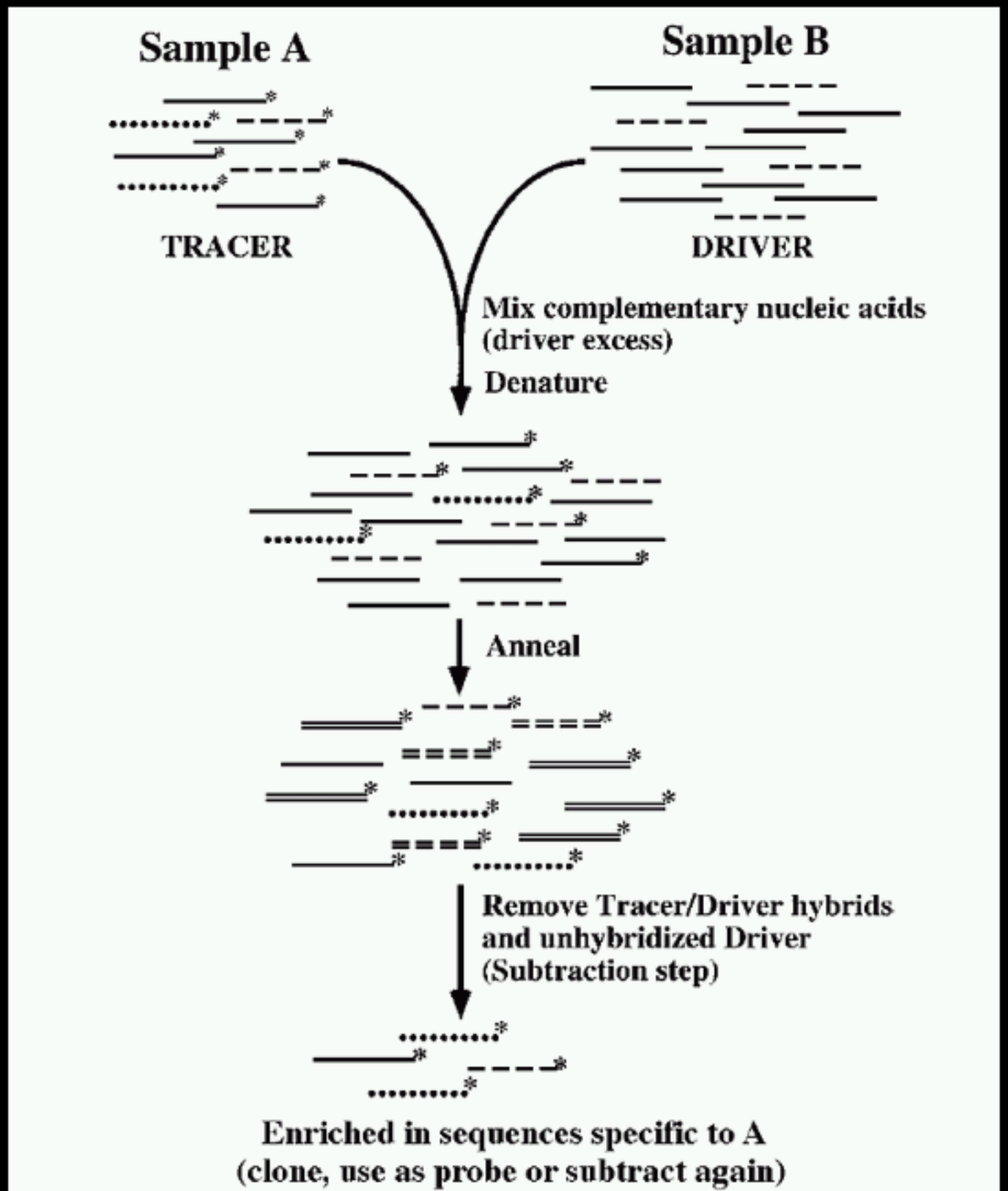
```
IDENTIFIERS
dbEST Id:          4025315
EST name:          hf43a02.x1
GenBank Acc:       AW592465
GenBank gi:        7279647
CLONE INFO
Clone Id:          IMAGE:2934602 (3')
Source:            NCI
DNA type:          cDNA
SEQUENCE
                   TTTTTTTTTAAATTGCCAAGTGATTTTACTTCAAGATGACATCAGAATTGCTAAAAGGTG
                   ATGTAACCGTCAGAGTGACTATTGATTATAACTCCCAGTAAGTGTCAACGTGATTTTCTC
                   ..................................................
                   CATTGTGTGGGCTTCCATTAGTATTTACTCATTAGGTTCAGTAGTTTTCATTATTTTCTC
                   TTTAAGACAGTAGCTGCCTGGGCCACAGGTTGCACCATCCACTGACCGCCCCATTTCTGG
                   CAAGTCTGGACCCTGGTGTGGCTAATAACCAAGGCATTTATT
Quality:           High quality sequence stops at base: 356
Entry Created:     Mar 22 2000
Last Updated:      Mar 22 2000
COMMENTS
                   This clone is available royalty-free through LLNL ; contact
                   the IMAGE Consortium (info@image.llnl.gov) for further
                   information.
```

Παράδειγμα καταχώρησης

```
PUTATIVE ID      Assigned by submitter
                 TR:Q60815 Q60815 ADAM 4 PROTEIN PRECURSOR ;

LIBRARY
Lib Name:        Soares_NFL_T_GBC_S1
Organism:        Homo sapiens
Organ:           pooled
Lab host:        DH10B
Vector:          pT7T3D-Pac (Pharmacia) with a modified polylinker
R. Site 1:       Not I
R. Site 2:       Eco RI
Description:     Equal amounts of plasmid DNA from three normalized libraries
                 (fetal lung NbHL19W, testis NHT, and B-cell NCI_CGAP_GCB1)
                 were mixed, and ss circles were made in vitro. Following HAP
                 purification, this DNA was used as tracer in a subtractive
                 hybridization reaction. The driver was PCR-amplified cDNAs
                 from pools of 5,000 clones made from the same 3 libraries.
                 The pools consisted of I.M.A.G.E. clones 297480-302087,
                 682632-687239, 726408-728711, and 729096-731399. Subtraction
                 by Bento Soares and M. Fatima Bonaldo.
```

Subtractive hybridization



Παράδειγμα καταχώρησης

```
SUBMITTER
Name:           Robert Strausberg, Ph.D.
E-mail:         cgapbs-r@mail.nih.gov

CITATIONS
Title:          National Cancer Institute, Cancer Genome Anatomy Project
                (CGAP), Tumor Gene Index
Authors:        NCI-CGAP http://www.ncbi.nlm.nih.gov/ncicgap
Year:           1997
Status:         Unpublished
```

Ομαδοποίηση των ESTs

Ο αριθμός των ESTs σε σχέση με τον αριθμό τον εκφραζόμενων γονιδίων είναι αστρονομικός (π.χ. για τον άνθρωπο υπάρχουν περισσότεροι από 4,5 εκατομμύρια ESTs για λιγότερα από 50 χιλιάδες γονίδια. Ακόμα και με τη χρήση κανονικοποιημένων βιβλιοθηκών, τα άφθονα mRNAs αντιπροσωπεύονται περισσότερο από τα πιο σπάνια (δεν είναι, για παράδειγμα, απίθανο να βρεθούν γονίδια στα οποία να αντιστοιχούν περισσότερα από 200 κατατεθειμένα ESTs). Για τους λόγους αυτούς, υπάρχουν μια σειρά από ερευνητικά προγράμματα στόχος των οποίων είναι να ενοποιήσουν το πλήθος των διαφόρων αλληλουχιών νουκλεϊκών οξέων με βάση τα γονίδια από τα οποία έχουν προέλθει.

Ομαδοποίηση : UniGene

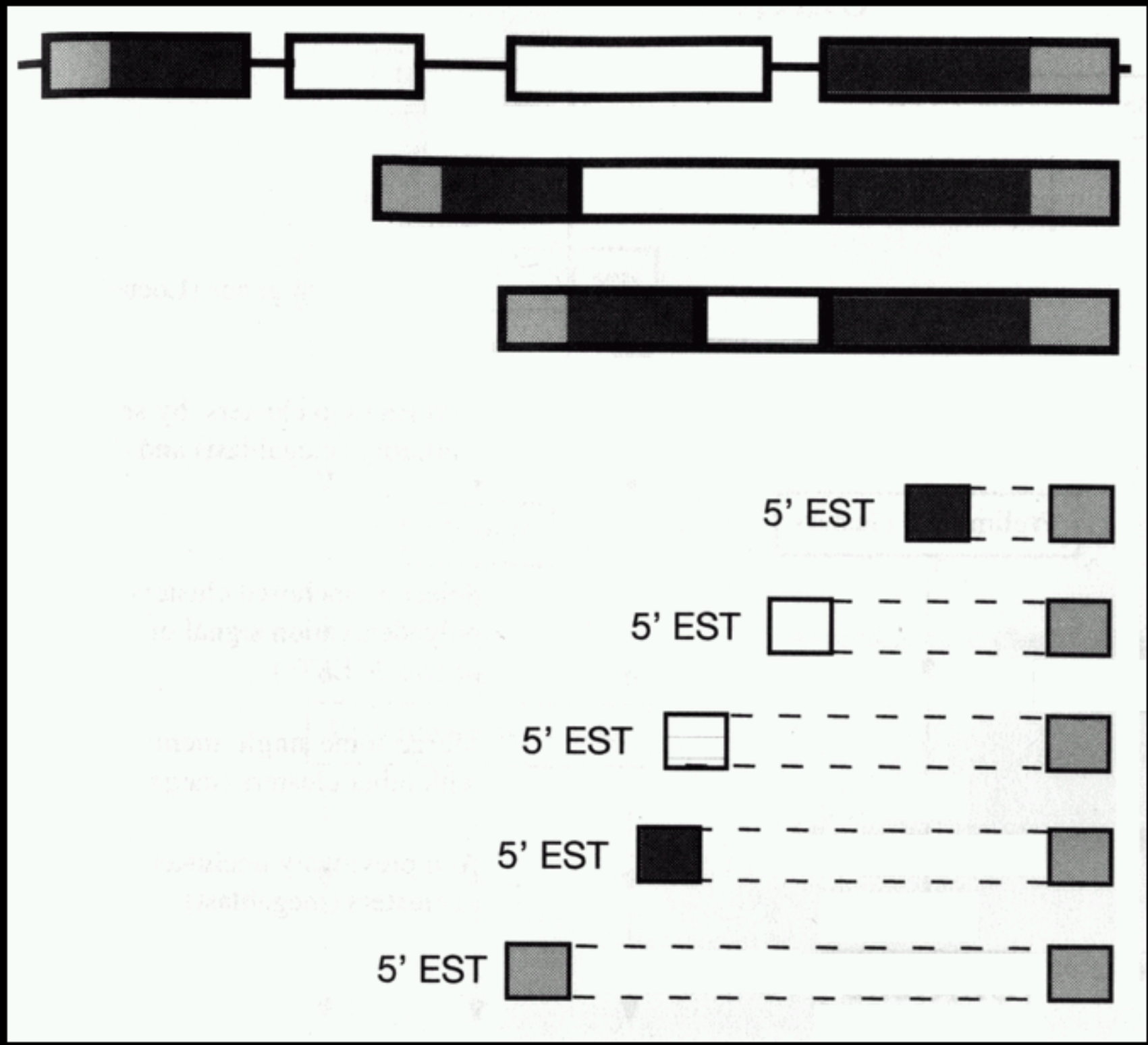
Το UniGene (στο NCBI) ομαδοποιεί ESTs, άλλες mRNA αλληλουχίες και CDS γενωμικού DNA σε ομάδες συσχετιζόμενων αλληλουχιών. Στις περισσότερες περιπτώσεις, κάθε ομάδα αποτελείται από αλληλουχίες που προέρχονται από ένα γονίδιο, συμπεριλαμβανο-μένης της πιθανότητας του alternative splicing. Αυτές οι ομάδες είναι ειδικές για τους οργανισμούς με τους οποίους ασχολείται το UniGene. Η δημιουργία μίας τέτοιας ομάδας γίνεται σε στάδια ως εξής :

- Οι αλληλουχίες ελέγχονται για τυχόν ύπαρξη μολύνσεων από μιτοχονδριακές, ριβοσωμικές ή βακτηριακές αλληλουχίες.

Ομαδοποίηση : UniGene

- Οι αλληλουχίες ελέγχονται για επαναληπτικά στοιχεία και LCRs (low complexity regions). Επίσης, το μήκος τους δεν θα πρέπει να είναι μικρότερο από 100 βάσεις.
- Η πρώτη ομαδοποίηση (χρησιμοποιώντας αλγόριθμους στοίχισης) γίνεται μεταξύ αλληλουχιών γενωμικού DNA και mRNAs.
- Η δεύτερη ομαδοποίηση αφορά εύρεση ESTs που είναι ομόλογα τόσο μεταξύ τους όσο και με τις ομάδες DNA-mRNA από το προηγούμενο βήμα.
- Όσα ESTs ή άλλες αλληλουχίες δεν έχουν ήδη ενταχθεί σε κάποια ομάδα, επανασυγκρίνονται με τις υπάρχουσες ομάδες χρησιμοποιώντας λιγότερο αυστηρά κριτήρια σύγκρισης.

Ομαδοποίηση : UniGene



Ομαδοποίηση : UniGene

```
> gnl|UG|Mm#S10112544 BY091318 RIKEN full-length enriched,
  16 days embryo whole body Mus musculus cDNA clone K630085I11 5',
  mRNA sequence /clone=K630085I11 /clone_end=5' /gb=BY091318
  /gi=26200457 /ug=Mm.39968 /len=353
GAGAGGGTCATCCAAGACCTGAGGAAGATAGAGAGGCAGAGAGTGGGAGCTATACCACGA
.....................................................
GGGTCCTCAGAGGACCTATCAACTGAGTTTGGTCACCACATCCACCGGGGATA
> gnl|UG|Mm#S8041744 Mus musculus histidine rich calcium binding
  protein (Hrc), mRNA /cds=(29,2209) /gb=NM_010473 /gi=6754241
  /ug=Mm.39968 /len=2270
CCCAGACGCTCAGCTGCTAAACGTCCCCATGGGCTTCCAGGGGCCATGGTTGCACACTTG
.....................................................
GGCAAGAGCTGCATCTATTTCTTTGAATAAATGTGCTCCTAGAAAAAAAA
> gnl|UG|Mm#S8071022 L0201C11-3 NIA Mouse Newborn Ovary cDNA
  Library Mus musculus cDNA clone L0201C11 3', mRNA sequence
  /clone=L0201C11 /clone_end=3' /gb=AW551690 /gi=31567559
  /ug=Mm.39968 /len=586
CTAGGAGCACATTTATTCAAAGAAATAGATGCAGCTCTTGCCAGGCGGCCAGGTCAGGGC
.....................................................
TCTTCCTCTTCATCATCTTCCTGGCTCATGGGGAGGCTATGGCCCC
> gnl|UG|Mm#S8113658 Mus musculus histidine-rich calcium-binding
  protein mRNA, complete cds /cds=(138,2354) /gb=AF132218
  /gi=5326837 /ug=Mm.39968 /len=2407
GGAAGATAGAGAGGCAGAGAGTGGGAGCTATACCACGACAAAAGGGACAATCTGAAAGTC
```

Ομαδοποίηση : UniGene

Η ομαδοποίηση που προσφέρει η UniGene φαίνεται να αποδίδει ικανοποιητικά : περισσότερες από 4,5 εκατομμύρια ανθρωπινές αλληλουχίες ομαδοποιούνται σε λιγότερες από 130 χιλιάδες ομάδες (με το 98% των αλληλουχιών να είναι ESTs). Από τις ~130000 ομάδες, μόνο μία στις έξι περιλαμβάνει και αλληλουχίες γενωμικού DNA. Άρα, εάν οι περισσότερες από τις ομάδες αντιστοιχούν σε εκφραζόμενα γονίδια, θα πρέπει να συναχθεί ότι η πλειοψηφία των γονιδίων αντιπροσωπεύεται μόνο από ESTs.

28/11/2010 :

6,954,505 seq - 123,448 clusters - 33,302 με mRNA

Ομαδοποίηση : UniGene

Η ομαδοποίηση που προσφέρει η UniGene φαίνεται να αποδίδει ικανοποιητικά : περισσότερες από 4,5 εκατομμύρια ανθρωπινές αλληλουχίες ομαδοποιούνται σε λιγότερες από 130 χιλιάδες ομάδες (με το 98% των αλληλουχιών να είναι ESTs). Από τις ~130000 ομάδες, μόνο μία στις έξι περιλαμβάνει και αλληλουχίες γενωμικού DNA. Άρα, εάν οι περισσότερες από τις ομάδες αντιστοιχούν σε εκφραζόμενα γονίδια, θα πρέπει να συναχθεί ότι η πλειοψηφία των γονιδίων αντιπροσωπεύεται μόνο από ESTs.

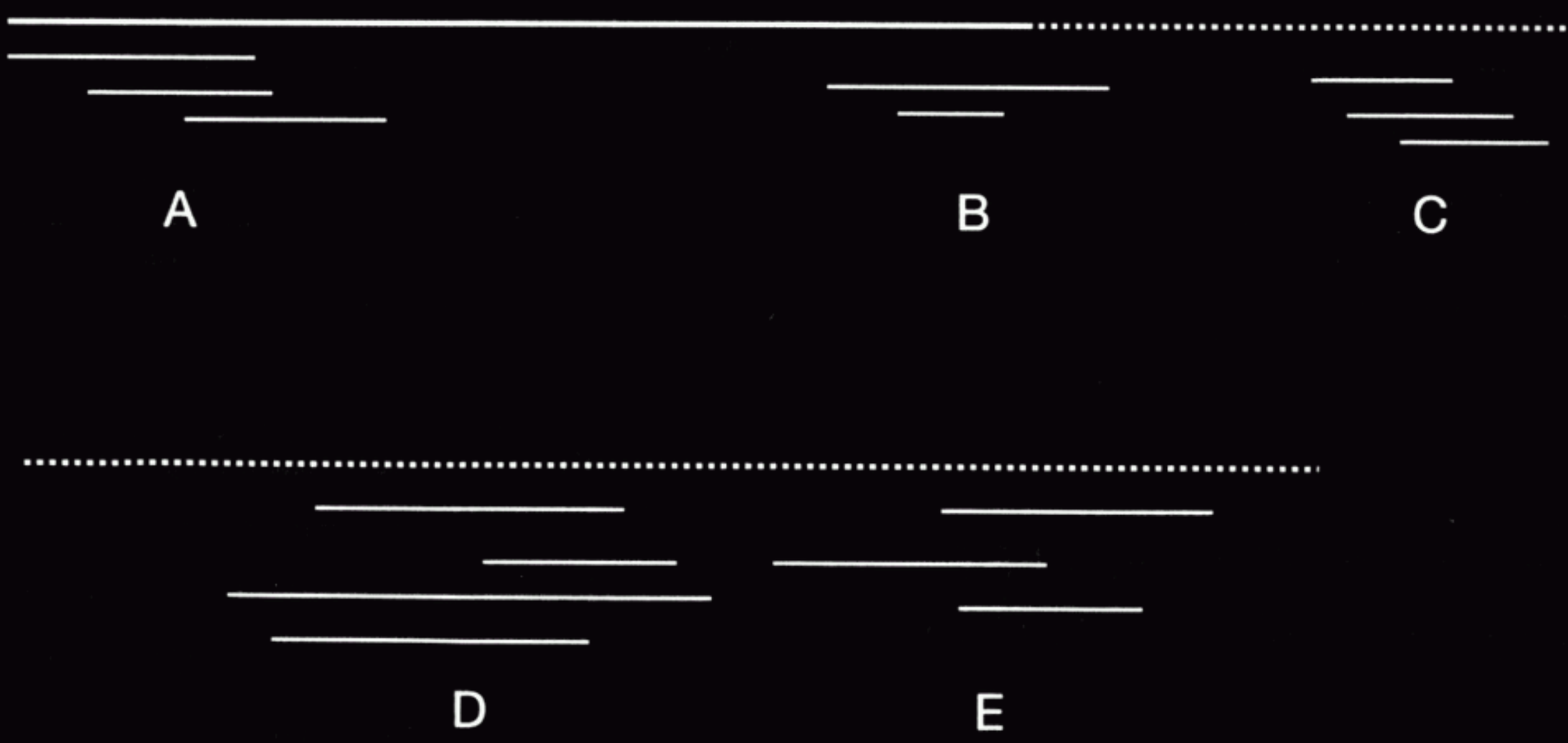
23/11/2012 :

6,998,600 seq - 130,029 clusters - 34,554 με mRNA

Εύρεση ορθόλογων αλληλουχιών

Μία από τις χρήσεις της UniGene είναι η εύρεση (προαιρετικά) ορθόλογων πρωτεϊνών. Αυτό επιτυγχάνεται μέσω της στοίχισης ομάδων από διαφορετικούς οργανισμούς. Όταν δύο ομάδες (από δύο οργανισμούς) είναι η μία η πλέον όμοια της άλλης και το αντίστροφο (στο σύνολο των ομάδων των δυο πληθυσμών) τότε και μόνο τότε οι εν λόγο ομάδες θεωρούνται προαιρετικά ορθόλογες. Το "προαιρετικά" προκύπτει από το ότι η μοναδική ένδειξη ομολογίας (και 'ορθολογίας') έχει προκύψει από καθαρά υπολογιστικές μεθόδους.

Ομαδοποίηση : προβλήματα



Εργαλεία ανάλυσης των ESTs

Τα πρωταρχικά εργαλεία ανάλυσης είναι οι αλγόριθμοι στοίχισης αλληλουχιών που έχουν ήδη αναφερθεί. Ιδιαίτερης σημασία για την ανάλυση EST αλληλουχιών έχουν τα προγράμματα που μπορούν να κάνουν έρευνα των βάσεων δεδομένων μέσω υπολογιστικά προσδιοριζόμενων μεταφράσεων αλληλουχιών (π.χ. BLASTX, TBLASTN, κλπ). Όταν μια έρευνα των βάσεων δεδομένων αποκαλύπτει πολλά ESTs τα οποία είναι ταιριάζουν (ομοιάζουν) με την αλληλουχία-στόχος, τότε αυτές οι EST αλληλουχίες στοιχίζονται μεταξύ τους (στοίχιση πολλών αλληλουχιών) για να αποκαλυφθεί η κοινή (consensus) αλληλουχία.

Εφαρμογές : Gene hunting

Η αναζήτηση ομολογιών στα ESTs είναι μια από τις πλέον 'προσοδοφόρες' μεθόδους για την ανακάλυψη νέων γονιδίων. Ο λόγος, βέβαια, είναι ότι οι πλήρους μήκους αλληλουχίες χαρακτηρίζονται (από αυτούς που τις προσδιόρισαν) πολύ καλύτερα απ'ότι τα ESTs. Έτσι το EST τμήμα της GenBank είναι ταυτόχρονα το λιγό-τερο ακριβές, αλλά και το πλέον πολλά υποσχόμενο. Τα παραδείγματα τέτοιων χρήσεων ξεκινούν από έρευνα για παράλογες αλληλουχίες (ίδιος οργανισμός, παρόμοια λειτουργία) και ορθόλογες αλληλουχίες, μέχρι την εύρεση alternative spliced μορφών γνωστών γονιδίων.

Εφαρμογές : πρόβλεψη γονιδίων

Μια άλλη χρήση των ESTs είναι στην πρόβλεψη ή βελτίωση της πρόβλεψης γονιδίων σε γενωμικό DNA. Ο λόγος βρίσκεται στο ότι περίπου το 90% των ταυτοποιημένων γονιδίων του ανθρώπινου γενωμικού DNA μπορούν να ανιχνευθούν μέσω της ομοιότητας τους με κάποιο ή κάποια ESTs (όπως προκύπτει από τις μεταξύ τους στοιχίσεις). Η εύρεση ομολογιών με ESTs μπορεί λοιπόν να συμπληρώσει τις προβλέψεις των αλγορίθμων πρόβλεψης γονιδίων, ιδιαίτερα για το χαρακτηρισμό alternatively spiced μορφών.

Εφαρμογές : SNPs

Τα SNPs (για Single Nucleotide Polymorphisms) έχουν προσελκύσει τόσο ενδιαφέρον ώστε να υπάρχει μια δημόσια βάση δεδομένων για αυτούς τους πολυμορφισμούς, η dbSNP. Ο λόγος είναι ότι τα SNPs μπορούν να επιτρέψουν τη σύνδεση ανάμεσα στην ποικιλότητα σε επίπεδο αλληλουχίας και κληρονομούμενα φαινοτυπικά χαρακτηριστικά, και είναι χρήσιμο υλικό για μελέτες πληθυσμιακής και εξελικτικής βιολογίας. Επειδή τα ESTs παράγονται από βιβλιοθήκες που προέρχονται από διαφορετικά άτομα, θα μπορούσαν να γίνουν μια πηγή για το χαρακτηρισμό νέων SNPs.

Εφαρμογές : SNPs

Το πρόβλημα με τη χρήση των ESTs για το χαρακτηρισμό πολυμορφισμών βρίσκεται στη δυσκολία διάκρισης μεταξύ πραγματικών πολυμορφισμών και λαθών στον προσδιορισμό των EST αλληλουχιών. Το πρόβλημα μεγεθύνεται από το ότι τα σφάλματα στις EST αλληλουχίες δεν είναι πάντα τυχαία αλλά εξαρτώνται από το ποιόν των αλληλουχιών (GC-rich, επαναλαμβανόμενες αλληλουχίες, ...). Το αποτέλεσμα είναι ότι η αβεβαιότητα στο χαρακτηρισμό SNPs μέσω της στοίχισης ESTs είναι τόσο μεγάλη ώστε να απαιτούνται ισχυρά στατιστικά κριτήρια για να αποφασιστεί τι θα πρέπει να θεωρείται σημαντικό.

Εφαρμογές : Γονιδιακή έκφραση

Η χρήση των ESTs για τη μελέτη των επιπέδων γονιδιακής έκφρασης είναι προβληματική λόγω του ότι οι περισσότερες cDNA βιβλιοθήκες είναι κανονικοποιημένες (και συνεπώς η συχνότητα εμφάνισης των κλώνων δεν είναι ανάλογη του επιπέδου έκφρασης των αντίστοιχων μεταγράφων). Εξαίρεση αποτελούν ESTs τα οποία έχουν προσδιοριστεί με στόχο τη μελέτη της γονιδιακής έκφρασης και προέρχονται από μη κανονικοποιημένες βιβλιοθήκες. Γνωστά παραδείγματα τέτοιων βιβλιοθηκών είναι οι βιβλιοθήκες του Cancer Genome Anatomy Project (CGAP) για φυσιολογικά, προ-καρκινικά και καρκινικά κύτταρα.

Εφαρμογές : microarrays

Οι εφαρμογή των ESTs για την παραγωγή microarrays περιορίζεται σε γονιδιώματα τα οποία δεν είναι πλήρως ταυτοποιημένα και σχολιασμένα. Σε αυτές τις περιπτώσεις αυτό που συνήθως γίνεται είναι να χρησιμοποιούνται ομαδοποιημένα ESTs από π.χ. την UniGene, για τον σχεδιασμό ολιγονουκλεοτιδίων για την κατασκευή των arrays. Τα ολιγονουκλεοτίδια αυτά είναι είτε consensus αλληλουχίες που προκύπτουν από τη στοίχιση πολλών ESTs ή και μεμονωμένα αντιπροσωπευτικά (για μια ομάδα) ESTs. Τα ολιγονουκλεοτίδια αυτά χρησιμοποιούνται για τη δημιουργία ενός cDNA το οποίο ενσωματώνεται στο microarray.

Εφαρμογές : microarrays

Τα προβλήματα με τη χρήση των ESTs για την κατασκευή microarrays βρίσκονται (1) στο ευμετάβλητο των ομαδοποιήσεων των ESTs και στα προβλήματα που έχει αυτή καθ'αυτή η διαδικασία ομαδοποίησης, (2) Την μικρή αντιπροσώπευση σπανίων μεταγράφων στις EST βιβλιοθήκες, και, (3) στο ότι για να είναι χρήσιμο ένα microarray θα πρέπει για κάθε χρησιμοποιούμενο EST να είναι γνωστά όσο το δυνατό περισσότερα για τη βιολογία του γονιδίου απ'το οποίο προέρχεται. Το πρόβλημα περιπλέκεται από το ονοματολογικό χάος της σύγχρονης μοριακής βιολογίας.

Microarrays

Τα microarrays είναι τεχνικές που χρησιμοποιούνται για εύρεση και χαρακτηρισμό του μοτίβου γονιδιακής έκφρασης σε επίπεδο ολόκληρων γονιδιωμάτων.
Η αρχή της μεθόδου είναι η ίδια με τις αναλύσεις κατά Southern (υβριδοποίηση με σημασμένες αλληλουχίες) μόνο που η κλίμακα του πειράματος και της ανάλυσης είναι μερικές τάξεις μεγέθους μεγαλύτερη.

Microarrays

Το ζητούμενο είναι να προσδιοριστεί το επίπεδο έκφρασης (σε 'αντίγραφα mRNA ανά κύτταρο') για κάθε γονίδιο. Μια τέτοια μέτρηση δεν είναι υλοποιήσιμη από καμία από τις διαθέσιμες μεθόδους. Οι πλέον αξιόπιστες μετρήσεις φαίνεται να προέρχονται μέσω της χρήσης τυπωμένων cDNA microarrays με δυο κανάλια μέτρησης (ένα για το προς προσδιορισμό δείγμα και ένα για το δείγμα αναφοράς). Τα κανάλια αυτά αντιστοιχούν σε διαφορετικές (φθορίζουσες) χρωστικές σήμανσης του DNA οι ποσότητες των οποίων ανιχνεύονται μέσω της της χρήσης δύο διαφορετικών μήκους κύματος LASER.

Microarrays

Prepare cDNA Probe

Test

Reference

Label with Fluorescent Dyes

Combine cDNAs

Hybridize probe to microarray

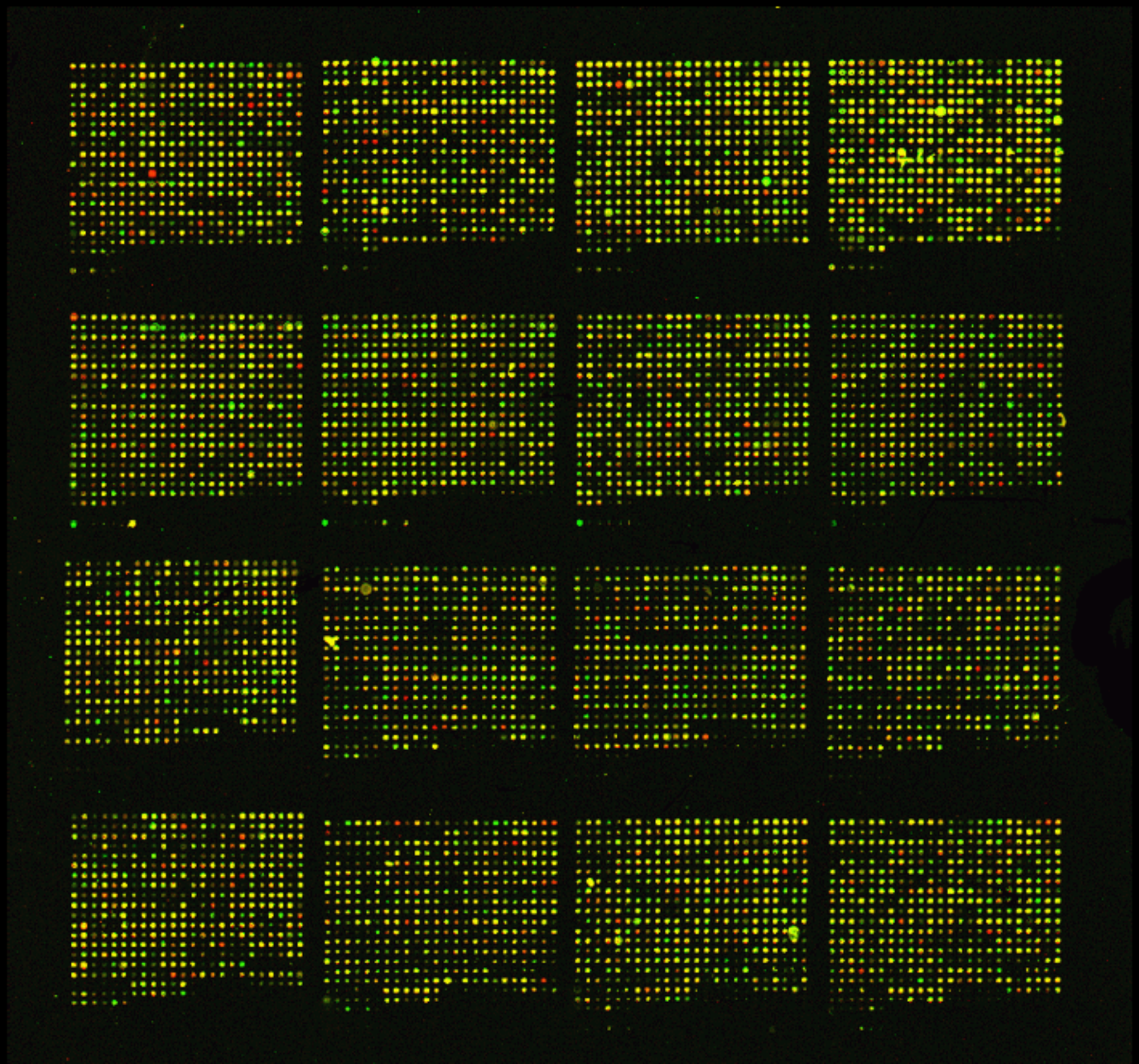
Scan

Print Microarray

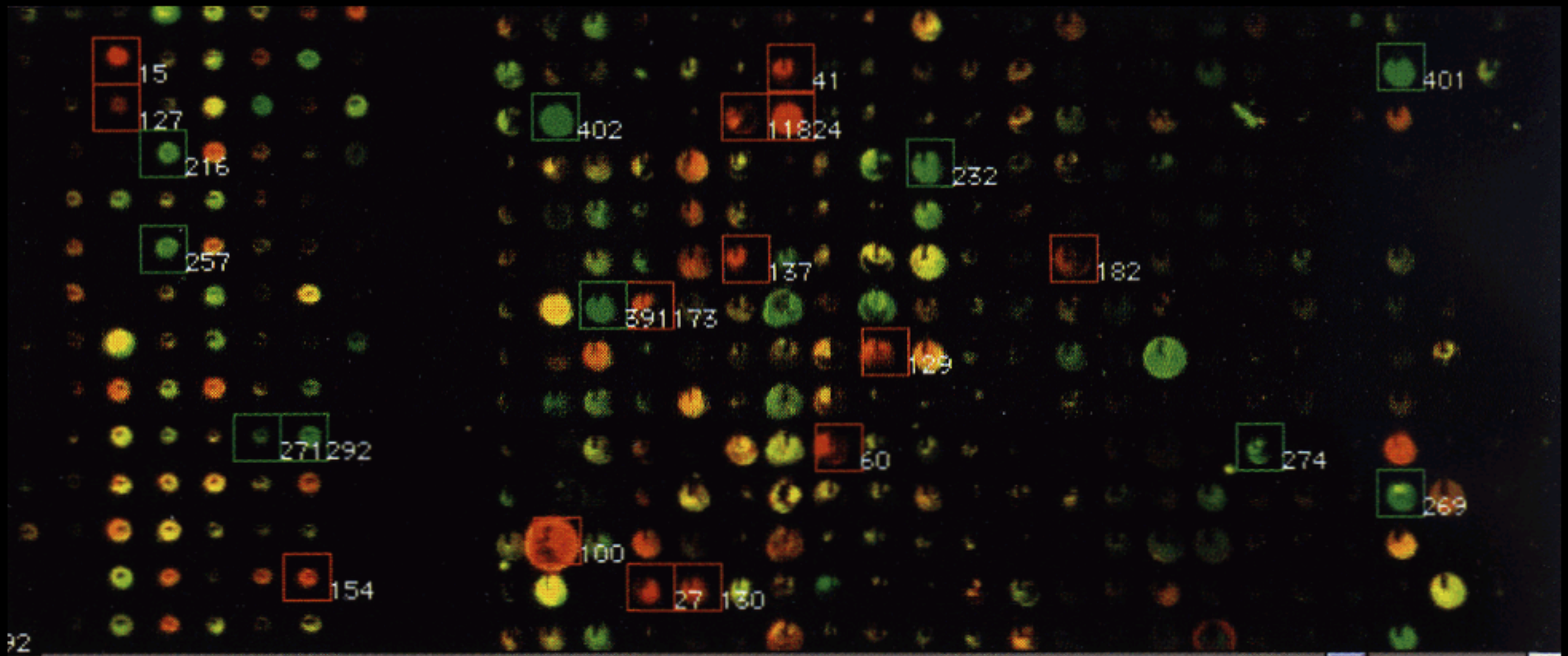
cDNA Library

Microarray Slides

Microarrays



Microarrays



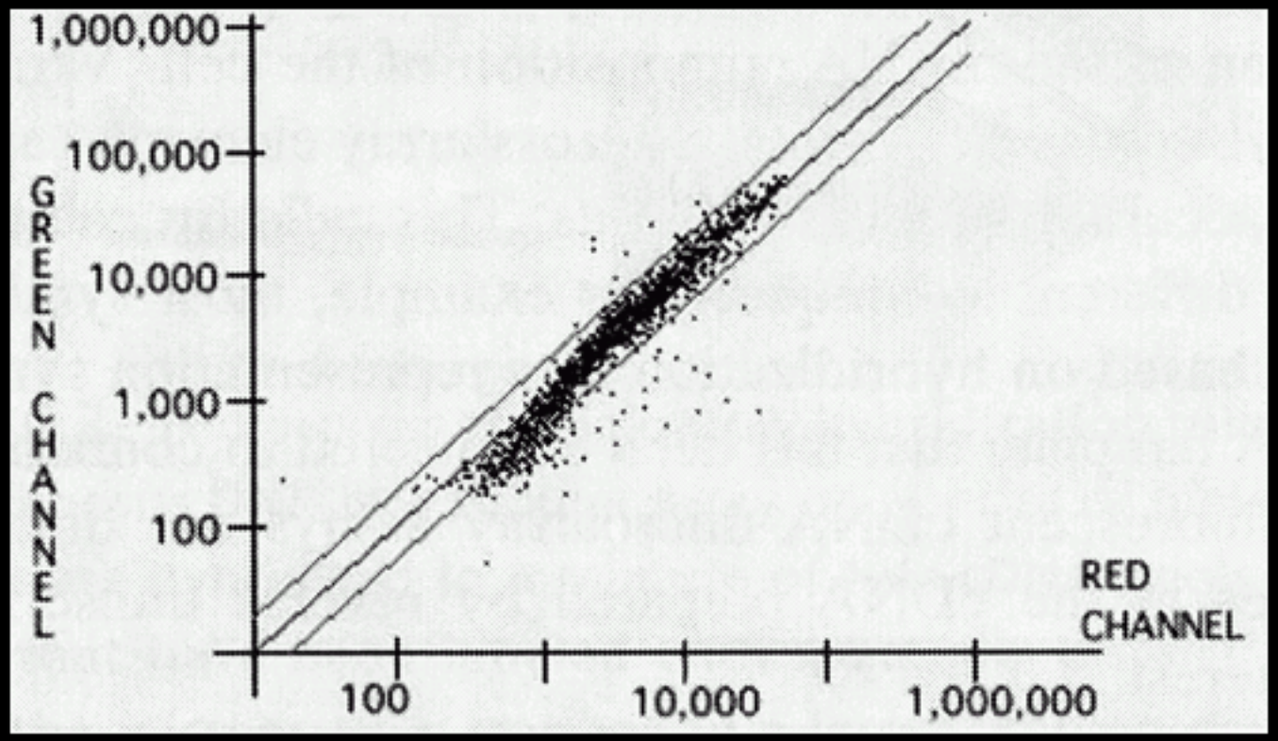
#	Clone Id	Ratio	Red Inten.	Green Inten.	R Size	G Size	Title
1	897910	36.6331	33860.4	1009.10	75	61	osteoblast specific factor 2 (fasciclin I-like)
2	755663	23.2327	29693.9	1395.35	94	82	retinoic acid receptor, beta
3	815542	21.3878	12398	632.851	68	32	myxovirus (influenza) resistance 1, homolog of murine
4	49164	18.0264	22464.3	1360.51	96	83	vascular cell adhesion molecule 1
5	108837	16.6373	41163.5	2701.14	91	67	small inducible cytokine A2 (monocyte chemotactic pro

Microarrays

Το βασικό πλεονέκτημα της μεθόδου αυτής (με τα δύο κανάλια μέτρησης) είναι ότι η μέτρηση από το cDNA αναφοράς επιτρέπει την απαλοιφή προβλημάτων από μη ειδική υβριδοποίηση, διαφορική ικανότητα υβριδοποίησης για διαφορετικά τμήματα του array, κοκ. Η απαλοιφή γίνεται με το χρησιμοποιούνται όχι οι καθ'αυτό μετρήσεις στα δύο κανάλια, αλλά ο λόγος των μετρήσεων. Εάν το επίπεδο έκφρασης ενός γονιδίου είναι περίπου το ίδιο στο δείγμα και το μάρτυρα, το πηλίκο θα είναι περίπου ίσο με ένα. Εάν υπάρχει σημαντική διαφορά στο επίπεδο έκφρασης ο λόγος θα διαφέρει σημαντικά από τη μονάδα.

Microarrays

Επιπλέον, επειδή για τη μεγάλη πλειοψηφία των γονιδίων τα επίπεδα έκφρασης θα είναι περίπου τα ίδια, είναι εφικτό να προσδιοριστούν τα επίπεδα του θορύβου για τις μετρήσεις από τα δυο κανάλια.



Microarrays : βάσεις

Παράδειγμα : Stanford Microarray Database

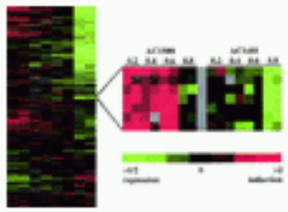
Stanford Microarray Database

SMD stores raw and normalized data from microarray experiments, as well as their corresponding image files. In addition, SMD provides interfaces for data retrieval, analysis and visualization. Data is released to the public at the researcher's discretion or upon publication

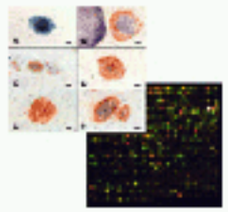
SMD Home
Public Search
Software & Tools
Staff
SMD Code
About SMD
SMD Specifications
Help Index

Researcher Login:
User Name:

Items of Interest



MgrA, an Orthologue of Mga, Acts as a Transcriptional Repressor of the Genes within the rlrA Pathogenicity Islet in Streptococcus pneumoniae. Hemsley C, et al. (2003) J Bacteriol 185 (22):6640-6647



Gene expression patterns in human embryonic stem cells and human pluripotent germ cell tumors Sperger et al. PNAS, 10.1073/pnas.2235735100

Growth Phase-Dependent Response of Helicobacter pylori to Iron Starvation Merrell DS, et al. (2003)Infect Immun

Transcriptome analysis of Arabidopsis colonized by a plant-growth promoting rhizobacterium reveals a general effect on disease

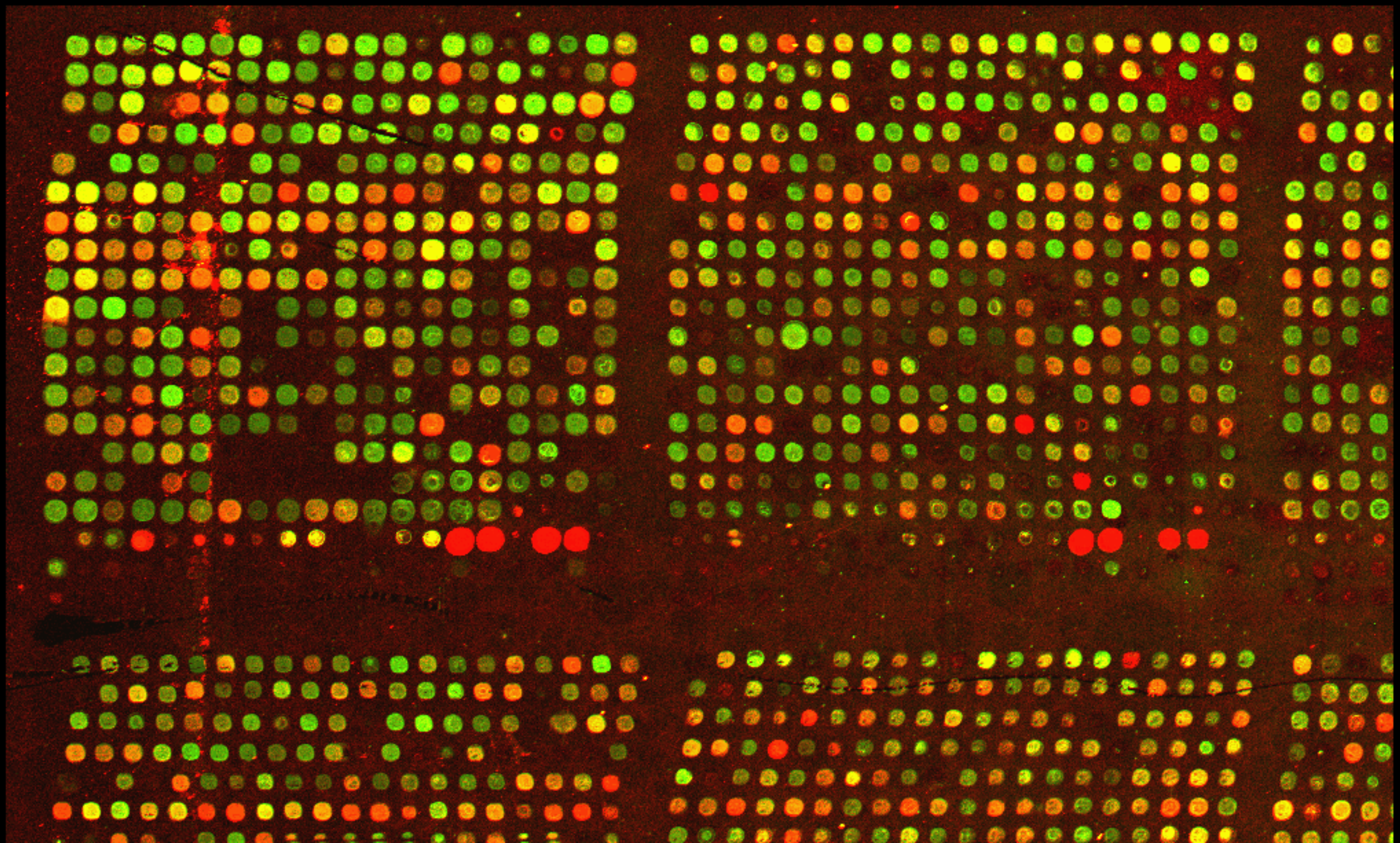
Microarrays

Υπολογιστικά προβλήματα : θόρυβος

Τα πρωταρχικά δεδομένα είναι τιμές εντάσεως για κάθε κανάλι και για κάθε σημείο του array (με πολλά σημεία για κάθε κηλίδα). Οι τιμές αυτές είναι ψηφιακές με περιορισμένο δυναμικό εύρος (π.χ. 0-65535). Το πρώτο πρόβλημα είναι να μετατραπούν αυτά τα δεδομένα (raw data) σε τιμές έντασης ανά κηλίδα/cDNA. Η διαδικασία περιπλέκεται από την ύπαρξη τυχαίου θορύβου, πειραματικών ατελειών, και συστηματικών σφαλμάτων : διαφορετικά μεγέθη κηλίδων, κηλίδες χρωστικής, σκόνη, γρατσουνιές του array, μη γραμμική απόκριση των καναλιών, κοκ.

Microarrays

Υπολογιστικά προβλήματα : θόρυβος



Microarrays

Υπολογιστικά προβλήματα : θόρυβος

Συνοπτικά, η αρχική επεξεργασία των δεδομένων περιλαμβάνει : εύρεση του κέντρου και της διαμέτρου των κηλίδων, υπολογισμός του επιπέδου θορύβου στην περιοχή της κηλίδας (local background), αφαίρεση ή διόρθωση μετρήσεων από προβληματικές κηλίδες, υπολογισμός της καθαρής έντασης της κηλίδας (δηλ. μετά από διόρθωση για το θόρυβο υποβάθρου).
Το αποτέλεσμα από την αρχική επεξεργασία των δεδομένων είναι δύο μετρήσεις (Ri,Gi) για κάθε cDNA.
Εάν το πείραμα περιλαμβάνει περισσότερες της μίας arrays, τότε τα δεδομένα θα είναι της μορφής (Rij,Gij) [για κάθε cDNA και κάθε array].

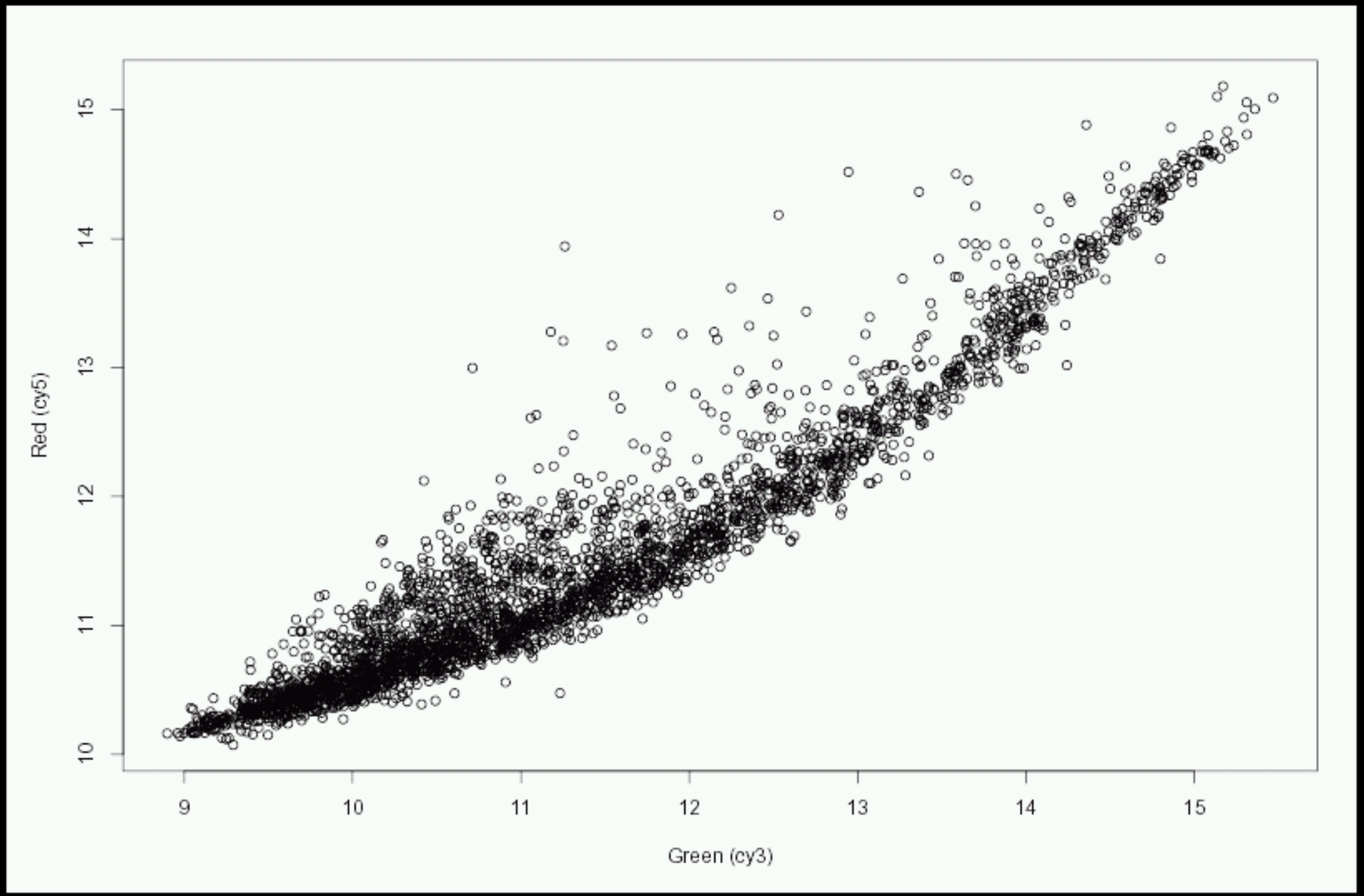
Microarrays

Υπολογιστικά προβλήματα : κανονικοποίηση

Τα δεδομένα (Rij,Gij) πρέπει επίσης να διορθωθούν για τυχόν συστηματικές (αλλά μη βιολογικού περιεχομένου) διαφορές μεταξύ δειγμάτων του ίδιου array και μεταξύ των δειγμάτων διαφορετικών arrays. Τέτοιες συστηματικές (και διορθώσιμες) διαφορές προκύπτουν από διαφορική πρόσδεση των χρωστικών, διαφορετικές συνθήκες μέτρησης, διαφορετικές ποσότητες σημασμένων cDNAs, κοκ. Η διάγνωση και διόρθωση τέτοιων προβλημάτων γίνεται μέσω διαγραμμάτων που συσχετίζουν τις μετρήσεις των δύο καναλιών π.χ. log(Ri) vs. log(Gi) ή
[log(Ri) - log(Gi)] vs [log(Ri) + log(Gi)] / 2

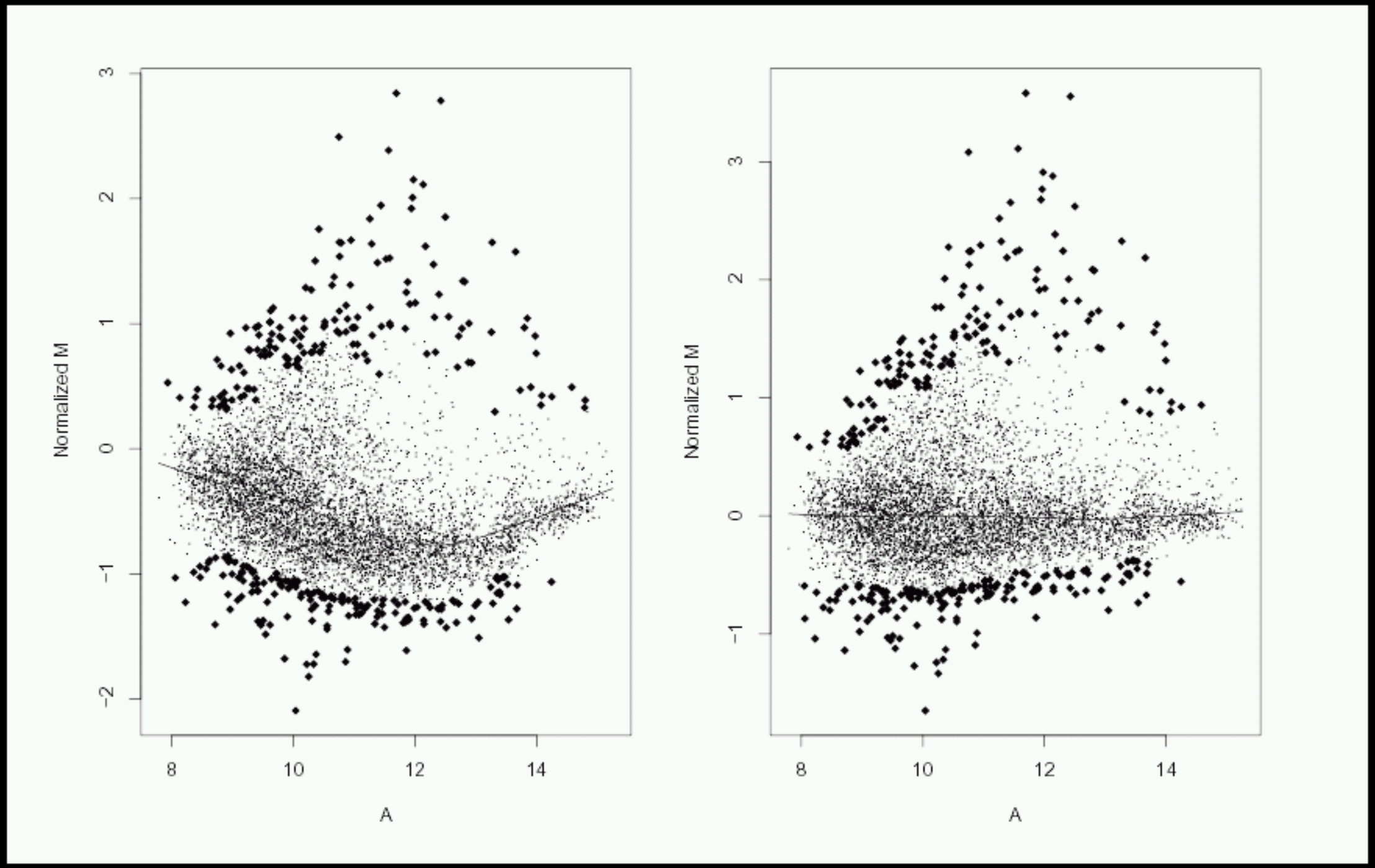
Microarrays

Υπολογιστικά προβλήματα : κανονικοποίηση



Microarrays

Υπολογιστικά προβλήματα : κανονικοποίηση



Microarrays

Υπολογιστικά προβλήματα : ανάλυση δεδομένων

Για πολλά προβλήματα, απλή εξέταση των δεδομένων μπορεί να είναι αρκετή. Για παράδειγμα, εάν αυτό που εξετάζεται είναι το ποιά γονίδια επάγονται από τη χορήγηση ενός φαρμάκου, ταξινόμηση των δεδομένων σε φθίνουσα σειρά αρκεί για την απάντηση του ερωτήματος. Καθώς το μέγεθος του προβλήματος αυξάνει, αυξάνει και η δυσκολία εύρεσης των σχέσεων μεταξύ των μοτίβων έκφρασης των γονιδίων. Για αυτές τις περιπτώσεις αυτό που απαιτείται είναι μια υπολογιστική μέθοδος η οποία να ταυτοποιεί σχέσεις στην έκφραση των γονιδίων και να τις παρουσιάζει με μια φιλική προς τον χρήστη μέθοδο απεικόνισης.

Microarrays

Υπολογιστικά προβλήματα : ομαδοποίηση

Τα δεδομένα έχουν τη μορφή :

```
gene1   1.1   0.4   0.7 -0.9 ...   1.0
gene2   0.9 -0.1   1.1   1.2 ... -0.4
.....................................
geneN   1.7 -1.1 -0.3   2.2 ...   0.0
```

Η ομαδοποίηση τους γίνεται με τη μορφή δενδρογράμ-
ματος μέσω του υπολογισμού ενός πίνακα αποστάσεων
(αναλογία με την κατασκευή του δενδρογράμματος-
οδηγού για τη στοίχιση πολλών αλληλουχιών). Οι
αποστάσεις μπορεί να είναι Ευκλείδιες αποστάσεις
αν και συνηθέστερα είναι συντελεστές συσχέτισης.

Microarrays

Παρένθεση : γραμμικός συντελεστής συσχέτισης

Ο γραμμικός συντελεστής συσχέτισης δύο ομάδων παρατηρήσεων (Xi,Yi) είναι :

$$\mathcal{C}(x_i, y_i) = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_i (x_i - \bar{x})^2} \sqrt{\sum_i (y_i - \bar{y})^2}}$$

και παίρνει τιμές από +1.0 (πλήρως συσχετιζόμενες τιμές), μέσω του μηδενός (μη συσχετιζόμενες τιμές), μέχρι το -1.0 (πλήρως αντι-συσχετιζόμενες τιμές).
Η αναλογία με την περίπτωση που έχουμε Ν παρατηρήσεις για τα επίπεδα έκφρασης του γονιδίου Χ, και Ν παρατηρήσεις για τα επίπεδα έκφρασης του γονιδίου Υ, είναι προφανής.

Microarrays

Υπολογιστικά προβλήματα : ομαδοποίηση

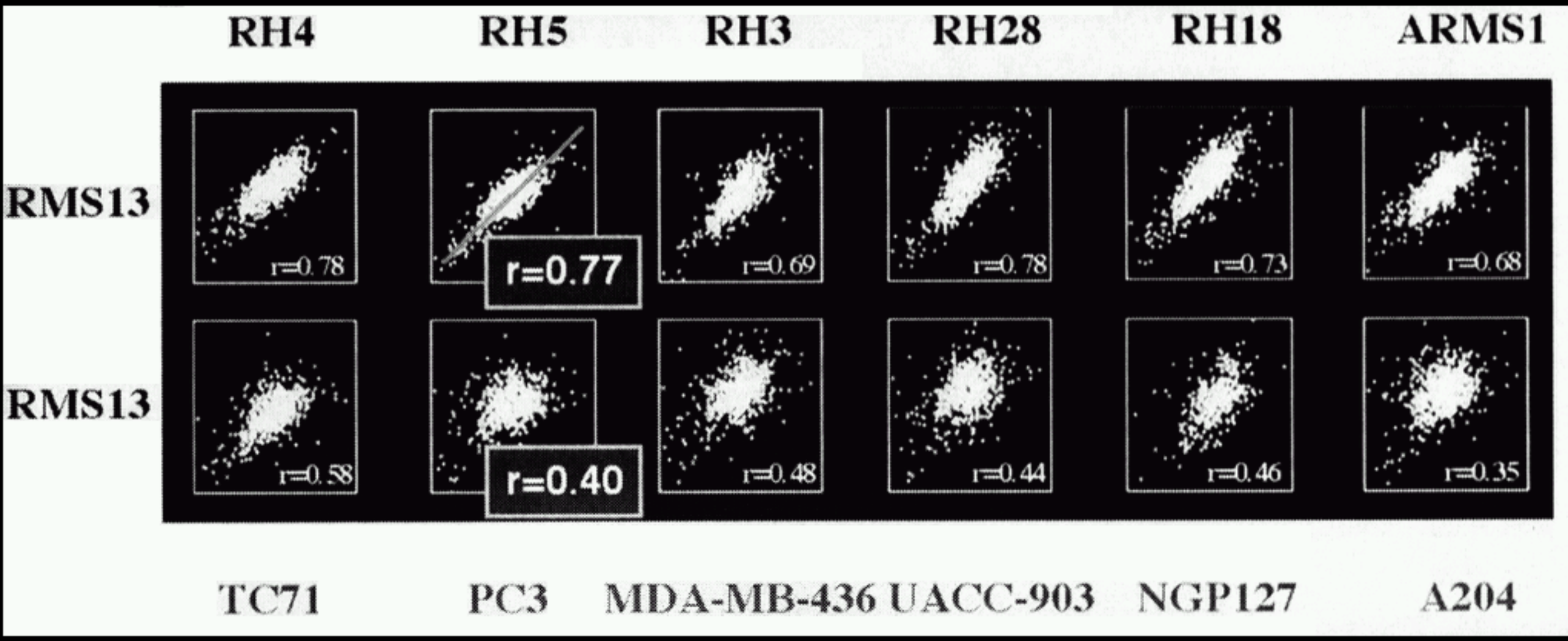
Με βάση έναν πίνακα αποστάσεων μεταξύ των μοτίβων έκφρασης των γονιδίων

	g1	g2	.....	gn
g1	1.0	0.4	0.7 -0.9 ...	1.0
g2	0.9	1.0	0.9 0.7 ...	-0.4
.....				
gN	0.7	-1.0	-0.3 0.2 ...	1.0

κατασκευάζεται ένα δενδρόγραμμα που αναπαριστά τις μεταξύ τους σχέσεις συνηθέστατα με τους αλγόριθμους UPGMA και Neighbor Joining. Αυτοί θα αναφερθούν στην επόμενη διάλεξη.

Microarrays

Υπολογιστικά προβλήματα : ομαδοποίηση



Microarrays