Tehnologia ADN Microrray
- DNA chips-

I. Introducere

Abordările genomice din ultimul deceniu au determinat o transformare fundamentală a conceptului de cercetare în biologie și medicina. În schimbul evaluării unui singur biomarker, studiile de genomică funcțională sunt capabile sa furnizeze date despre întregul transcriptom al celulelor studiate.

Genomica funcțională, utilizând tehnica microarray, poate furniza informații asupra expresiei simultane a mii de gene sau chiar a întregului genom. Profilul de expresie genică al unui tip de celule îi determină funcționalitatea, fenotipul și răspunsul la terapie.

Tehnologia microarray a permis identificarea unor seturi de gene supra- și sub-exprimate în diferite patologii: cancerul de sân, cancerul de prostată, cancerul pulmonar, precum și în dereglarea anumitor procese fiziologice: inducerea apoptozei sau răspunsul la terapie. Analizele integrate ale mai multor studii, permise de numărul în creștere al publicațiilor în domeniu, au evidențiat generalități și particularități ale expresiei genice în anumite patologii.

Utilizarea microarray-urilor, în cercetarea biomedicală, nu se limitează doar la determinarea profilului de expresie genică, fiind de asemenea folosite pentru detecția SNP-urilor (single nucleotide polymorphism), aberațiilor de metilare, detecția splicing-ului alternativ, detecția de patogeni, amplificare genică, detecția genelor de fuziune.

În ultimul deceniu, array-urile de mare densitate, standardizarea protocoalelor de hibridizare, precizia tehnologiilor de scanare și dezvoltarea unor metode computaționale robuste, au impus tehnologia microarray ca un instrument genomic puternic și ușor de folosit. Tehnologiile microarray au devenit metodele preferate pentru evaluările la scara largă a expresiei genice datorită accesibilității, eficienței, costului și a protocoalelor standardizate.

Microarray-ul este o matrice solidă miniaturizată pe care sunt depozitate într-o ordine bine stabilită mii de fragmente de acizi nucleici. Principiul de bază al tehnicii microarray este similar principiului *Southern blot*-(hibridare ADN-ADN), respectiv *Northern blot* (hibridare ARN-ADN) și se bazează pe complementaritatea secvențelor genice de a se recunoaște între ele și de a detecta prezența sau absența ADN-ului sau al ARN-ului de interes, folosind o serie de detectori radiologici, fluorogenici sau chemoluminiscenți. ADNc-ul sau ARNc-ul, sintetizat prin reacția de revers transcripție de pe ARNm-ul extras din proba de interes, este marcat fluorescent și hibridizat pe suprafața array-ului. Cuantificarea nivelelor de expresie se bazează pe faptul ca intensitatea semnalului fluorescent al fiecărui spot este direct proporționala cu cantitatea de ARNm prezentă în proba analizată, care a hibridizat în spotul respectiv. Tehnica microarray compară nivelurile de expresie a unei gene în doua condiții diferite (ex. cancer vs. normal).

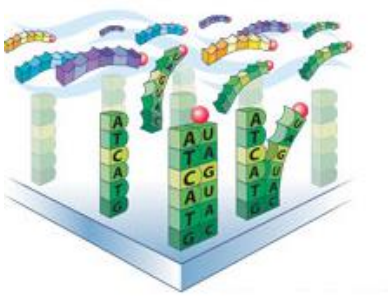


Fig. 1.1. Hibridizarea fragmentelor pe suprafața array-ului

La sfârșitul anilor ’90 au fost realizate sisteme robotizate sofisticate capabile sa poziționeze fragmente de ADN pe suporturi de sticlă la o densitate imposibil de realizat prin procese manuale. Pat Brown și colab. au fost primii care au dezvoltat tehnologia microarray de mare densitate. Din

acel moment a început o adevărată dezvoltare a acestor tehnologii atât la nivel academic cat și comercial.

Platformele actuale de microarray diferă între ele prin tipul de sondă utilizată (suportul de hibridare de pe suprafața array-ului), design (membrana de nylon, lama de sticla, array-ul în ansamblul său), tehnologia de fabricare a sistemelor de tip array (printare sau depunere), protocoalele de marcare și hibridare (monocolore sau bicolore).

În domeniul evaluării expresiei genice sunt utilizate mai multe platforme microarray care se bazează pe utilizarea suporturilor array ce conțin ADNc sau oligonucleotide[1].

Un sistem **ADNc microarray** cuprinde o colecție de mii de sonde (probe) care, în general, corespund produșilor PCR (Polymerase Chain Reaction) aparținând unor bănci de ADNc. Primul pas în producerea microarray-urilor cu ADNc este selectarea din bazele de date publice (UniGene, RefSeq, GenBank) a secvențelor ce urmează a fi poziționate pe array, urmată de amplificarea prin PCR a genelor de interes, folosind primeri specifici, și purificarea produșilor PCR. Fragmentele astfel obținute sunt depozitate, în poziții bine determinate, pe suprafața array-ului.

Dimensiunile zonelor de depozitare (spoturilor) au diametre de aproximativ 100-200 µm. Spoturile sunt separate de zone libere cu distanțe similare, ajungând la densități de 20.000 spoturi/cm^2. Array-urile pot fi din nylon, nitroceluloză sau sticlă.

Această tehnologie permite producerea la un cost relativ redus a array-urilor personalizate pentru fiecare studiu, cu un număr moderat de spoturi. Un avantaj al sistemelor microarray cu ADNc este posibilitatea producerii array-urilor cu structuri clonate care nu au fost încă secvențiate, ceea ce poate facilita descoperirea de noi gene. Dezavantajul sistemului ADNc constă în dificultatea menținerii și utilizării unor colecții voluminoase de clone de ADNc.

Oligonucleotidele microarray se referă la o tehnică specifică de producere a acestor array-uri și anume *sinteza in situ*. În loc de a fi sintetizate în prealabil, oligonucleotidele sunt în acest caz produse, bază cu bază, direct pe suprafața array-ului, fiecare oligonucleotidă fiind specifică pentru o anumită genă. Sistemele microarray bazate pe oligonucleotide sunt realizate în doua formate: structuri scurte (25-30 baze), respectiv structuri lungi (50-80 baze). Avantajul acestor sisteme comparativ cu microarray-ul ADNc constă în eliminarea menținerii unor mari colecții de clone ADNc sau produși PCR, respectiv a unei sensibilități de hibridare ridicate, ceea ce face ca variantele de transcriere genică să poată fi specific discriminate.

Tehnologiile two-color (bicromatice) permit hibridizarea pe același array a două probe (ex. țesut tumoral vs. țesut normal) marcate cu fluorocromi diferiți. Fluorocromii folosiți, cel mai adesea, sunt **Cy3** (fluorescență verde) și **Cy5** (fluorescență roșie). Probele marcate vor hibrida competitiv, intensitățile relative corespunzătoare celor două canale de fluorescență fiind exprimate ca rapoarte pentru a pune în evidență diferențele de expresie genică între cele două probe.

În marcajul one-color (monocromatic) toate probele sunt marcate cu același fluorocrom (**Cy3**) și hibridizate individual pe array-uri. Fiecare array furnizează intensitățile corespunzătoare unei probe. Compararea nivelelor de expresie a unei gene între două probe implică folosirea a două astfel de array-uri.

I.1. Etapele reacției microarray

Reacția microarray implică următoarele etape: prepararea și marcarea fluorescenta a probelor, hibridizarea pe lamă și spălarea lamelor și achiziția de imagini.

I.1.1. Prepararea și marcarea fluorescentă a probelor

Probele biologice utilizate trebuie colectate proaspete și procesate într-un timp cât mai scurt, deoarece ARN-ul are o stabilitate redusă, fiind ținta enzimelor de degradare. Extracția ARN-ului din

aceste probe se poate face fie prin metoda clasică (fenol-cloroform), fie cu ajutorul kiturilor de extracție specializate pentru diferitele tipuri de probe biologice (țesut, sânge, urină). ARN-ul extras este evaluat cantitativ (cu un spectrofotometru Nanodrop) și calitativ în vederea utilizării lui ulterioare. Studiile de genomică funcțională necesită utilizarea unor cantități cât mai precise de acizi nucleici (ARN) pentru a putea estima cat mai corect modificările de expresie genică.

De pe ARN-ul extras este sintetizat ADNc-ul de pe care va fi din nou sintetizat ARNc-ul marcat fluorescent. Marcarea fluorescentă a țintelor de ARNc se realizează prin incorporarea unor nucleotide analoage conjugate cu un fluorocrom în timpul procesului de revers transcripție. Fluorocromii Cy3 și Cy5 sunt cei mai folosiți pentru sinteza de sonde microarray datorită emisiei distincte la lungimi de undă diferite. Ca și în cazul ARN-ului de origine, și pentru ARNc-ul țintă marcat fluorescent se vor face evaluări de calitate și cantitate premergătoare fazei de hibridare pe lamele microarray. Măsurarea cu precizie a concentrației celor două sonde microarray este extrem de importantă deoarece reacția microarray cu marcaj two-color se bazează pe o reacție de hibridare competitivă între două sonde marcate diferit și dispuse în cantitate egală pe aceeași lama microarray.

I.1.2. Hibridizarea pe lamă și spălarea lamelor

Hibridizarea este etapa în care sondele imobilizate pe suprafața slide-ului și țintele de ARNc sau ADNc marcate fluorescent vor forma hetero-duplicați pe baza complementarității dintre secvențele acestora. Pe fiecare lamă microarray se va depune aceeași cantitate de ținte marcate. Procesul de hibridizare durează între 12 și 24 de ore și se realizează la temperaturi între 45^0 - 65^0C, în funcție de tipul de array utilizat. Apoi, lamele sunt supuse unor spălări succesive pentru a se elimina excesul de ținte de ADNc sau ARNc care nu au hibridizat. Astfel, pe lamă va rămâne doar ADNc-ul sau ARNc-ul de interes care urmează a fi cuantificat.

I.1.3. Achiziția de imagini

Sondele de pe array în care au hibridizat țintele marcate vor emite fluorescent sub excitația unui laser. Fiecare pixel din imaginea digitală generată de scanner-ul cu laser reprezintă intensitatea fluorescenței indusă prin focusarea laserului în acel punct al array-ului.

Scanerele pentru array-urile two-color conțin, de regulă, doua lasere. În cazul array-urilor two-color, scanerul generează două imagini monocrome, separat una pentru fiecare din cele doua lasere corespunzătoare tehnologiei two-color. În final, acestea sunt combinate pentru a crea imaginea roșu-verde. Atât imaginile monocrome (corespunzătoare tehnologiei one-color) cât și imaginile corespunzătoare tehnologiei two-color sunt stocate, de obicei, în fișiere TIFF (Tagged Image File Format).

I.2. Procesarea imaginilor

Imaginile obținute prin scanare cu laser reprezintă **datele brute** ale experimentului. Folosind anumiți algoritmi, imaginile sunt convertite în variantă digitală (informație numerică), informație care va cuantifica nivelul de expresie genică. Procesarea imaginilor are un impact major în calitatea datelor obținute și în concluziile biologice ce se desprind din aceste date. Procesarea imaginilor implică următoarele etape:

- Identificarea poziției spoturilor pe array (localizarea spoturilor).
- Identificarea pixelilor care vor intra în calculul intensității date de spot, respectiv a pixelilor din apropierea spotului care vor fi folosiți pentru calculul background-ului local.
- Controlul de calitate.

Majoritatea array-urilor au spoturile dispuse într-o grila dreptunghiulară, fiecare spot având o poziție bine definită pe această grilă. Numărul grilelor, locația acestora pe lamă, numărul de linii și coloane a fiecărei grile și poziția exactă a spoturilor pe aceasta grilă este specificată de fiecare

producător în funcție de tipul de array-ului. Aceste informații sunt folosite pentru poziționarea grilei de citire peste imaginea lamei.

I.2.1. Identificarea poziției spoturilor pe array (localizarea spoturilor).

Plasarea corecta a grilei este extrem de importantă deoarece coordonatele acesteia sunt folosite pentru identificarea și atribuirea identității fiecărui spot. Deplasarea sau alinierea greșita a grilei poate determina obținerea unor rezultate eronate prin atribuirea incorectă a nivelelor de expresie unei alte gene. Pentru evitarea acestui inconvenient sunt folosite pe array spoturi de control în poziții predeterminate, în general la începutul și la sfârșitul fiecărei grile. Aceste spoturi furnizează semnale de intensitate cunoscută care permit plasarea corectă a grilei pe suprafața array-ului. Designul spoturilor de control este astfel conceput încât în ele să hibridizeze *controale spike-in exogene*. Controalele spike-in exogene sunt de obicei fragmente de ARN adăugate în cantități și concentrații cunoscute, într-o probă, într-un anumit stadiu de preparare și marcare fluorescentă.

Plasarea grilei permite determinarea poziției fiecărui spot și calcularea ariei acestuia. Aria va determinarea intensitatea semnalului fluorescent al fiecărui spot respectiv al background-ului. Exista mai multe metode pentru determinarea acestei arii.

Metoda cercului fix plasează un cerc de dimensiune prestabilită peste suprafața spotului luând în considerare la calculul intensității toți pixelii cuprinși în interiorul acestei zone. În cazul în care spoturile au dimensiuni variabile, situație frecvent întâlnită în cazul array-urilor, metoda furnizează rezultate inexacte.

Metoda cercului variabil poziționează un cerc a cărui dimensiune poate fi ajustată pentru fiecare spot. Această metodă este capabilă sa rezolve problema spoturilor cu dimensiuni variabile dar are rezultate mai puțin satisfăcătoare în cazul spoturilor cu forme neregulate.

Metoda histogramelor plasează cercuri pe suprafața spotului și a background-ului iar pe baza histogramei intensităților pixelilor, elimina din calcul pixelii cu intensități extrem de mari sau extrem de scăzute. Această metodă este fiabilă și pentru spoturile neregulate dar poate prezenta probleme pentru spoturile cu dimensiuni mici, dacă diametrul cercului este prea mare.

Metoda CookieCutter este o metodă eficientă în analizarea microarray-urilor de tip Agilent, ea combină metoda cercului variabil și a histogramelor și consideră ca zonă de interes o regiune circulară cu centrul în mijlocul spotului a cărei rază este inferioară razei nominale a spotului (fig. 1.2.). Procentul din raza nominală care definește raza zonei de interes poate fi fixat de către utilizator. Astfel sunt delimitate următoarele zone:

- *Zona centrala a spotului* (zona de interes) care include toți pixelii localizați în interiorul elipsei sau a zonei circulare, pe baza cărora se calculează parametrii statistici.
- *Zona de excludere* care reprezintă aria circulară exterioară zonei centrale a spotului. Partea exterioară a acestei zone delimitează conturul spotului. Informația care se găsește în aceasta zonă (sub forma de pixeli) nu este inclusă în analizele statistice folosite pentru zona centrală sau background.
- *Background-ul local* (fond local) care definește aria exterioară zonei de excludere și se stabilește funcție de raza nominală.

Această metodă este în general folosită cu rezultate bune în analiza array-urilor de tip Agilent datorită uniformității spoturilor.

I.2.2. Identificarea pixelilor care vor intra în calculul intensității generate de spot, respectiv a pixelilor din apropierea spotului care vor fi folosiți pentru calculul background-ului local.

După delimitarea zonei centrale și a background-ului local, sunt evaluați parametrii statistici ai pixelilor din aceste zone: media, mediana și deviația standard. Din calculul intensității spotului,

respectiv din calculul intensității background-ului sunt eliminați pixelii care prezintă valori extreme ale intensității (*pixeli outliers*), aceștia fiind în afara cutoff-ului stabilit. Astfel, sunt luați în considerare doar pixelii cu valori cuprinse în acest interval (*pixeli inliers*) (fig. 1.3.) Acest proces este aplicat pentru zona centrală și background-ul local al fiecărui spot și al fiecărui canal de fluorescență în parte. În cazul marcajului two-color, dacă un pixel dintr-un canal de fluorescență este eliminat, se elimină automat și pixelul corespunzător celuilalt canal de fluorescență. Parametrii statistici ai intensității de hibridizare sunt apoi reevaluați doar pentru valorile rămase. În general, este de preferat ca pentru intensitatea semnalului de hibridizare să fie folosită valoarea medianei intensității pixelilor deoarece aceasta nu este influențată de eventualele valori extreme ale pixelilor. Media aritmetică este sensibilă la aceste valori.

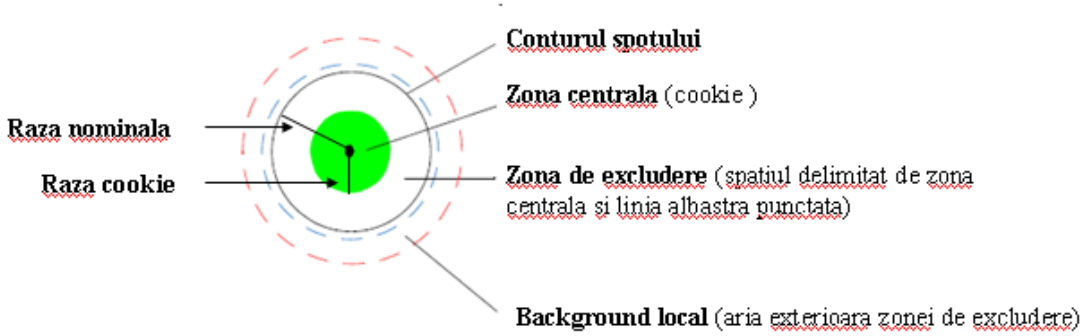


Fig. 1.2. Definirea background-ului local și a zonei centrale prin metoda *CookieCutter*

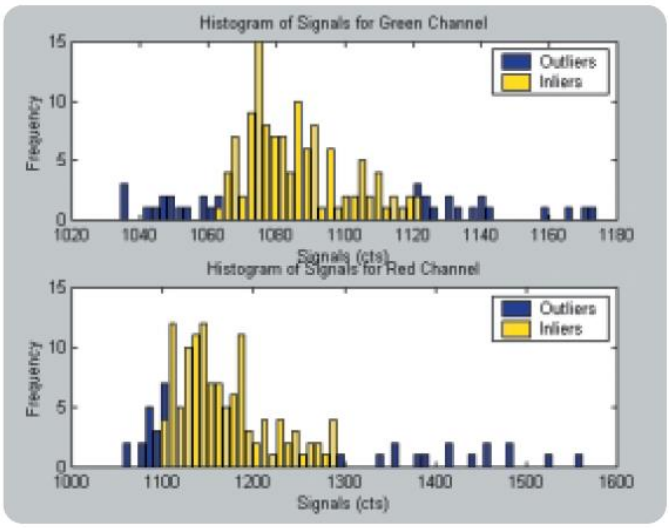


Fig. 1.3. Analiza pixelilor outliers. Histograma reprezintă pixelii din zona de interes ai unui spot, exemplificarea este valabilă pentru marcajul bicolor (two-color): primul grafic reprezintă marcajul verde, al doilea grafic reprezintă marcajul roșu. Pixelii outliers sunt reprezentați cu albastru, iar pixelii inliers sunt reprezentați cu galben.

I.2.3. Controlul de calitate

Controlul de calitate presupune, de cele mai multe ori, **eliminarea pixelilor saturați**. Saturația este definită ca reprezentând valoarea maximă permisă a intensității pixelilor (65535 pentru marea majoritate a scanerelor). Începând cu anul 2006, Agilent folosește sistemul *XDR (eXtended Dynamic Range),* care se bazează pe combinarea datelor obținute prin scanare la două rezoluții, o rezoluție înaltă și una joasă. Aceasta permite creșterea limitei de saturație de 10-20 de ori

(20*65535), obținându-se astfel intensități mult peste semnalele biologice care ar putea fi obținute în mod normal.

Spoturile saturate au pixeli de aceeași intensitate (intensitatea maximă generată de scanner) și în consecință au o deviație standard egala cu zero. Astfel, valoarea deviației standard poate fi folosită pentru identificarea acestora. Deoarece intensitatea saturată nu reprezintă intensitatea reală dată de spot, ea nu poate fi folosită pentru cuantificarea nivelurilor de expresie genică. Din acest motiv spoturile saturate sunt considerate nesatisfăcătoare și eliminate din analizele ulterioare.

I.3. Transformarea și normalizarea datelor

Normalizarea datelor este un termen general pentru un set de metode aplicate cu scopul de a corecta erorile sistematice introduse de experiment și de a genera valori care să poată fi comparate între ele, păstrând nealterată informația biologică. Sursele de erori în microarray sunt multiple: variații ale cantităților inițiale de ARNc folosite, diferențe în marcarea și detecția fluorescentă[2,3], probleme de spațialitate a array-ului manifestate printr-o puternică dependență a intensității de locația pe array[4]. Din aceste considerente, nivelele brute de expresie nu pot fi direct comparate între ele deoarece aceste erori ar putea determina diferențe nerealiste între nivelurile de expresie a diferitor gene. Procesul de normalizare implică o etapă premergătoare de filtrare și transformare a datelor, urmată de normalizarea propriu-zisă care depinde de designul experimental:

- ***normalizarea intra array:*** cuprinde metode care permit compararea celor două canale de fluorescență corespunzătoare Cy3 respectiv Cy5. Aceste metode au relevanță doar pentru tehnologiile two-color.
- ***normalizarea inter array:*** cuprinde metode care permit compararea diferitelor array-uri. Această normalizare poate fi aplicată atât tehnologiilor one-color cât și tehnologiilor two-color.

Datele obținute prin procesarea imaginilor sunt furnizate sub forma unor fișiere text. Extragerea informațiilor biologice corecte implică filtrarea acestora pentru a obține date cu acuratețe crescută și transformarea lor pentru a putea fi mai ușor utilizate în analizele următoare. În procesul de filtrare și transformare exista 3 etape importante:

- eliminarea spoturilor marcate.
- eliminarea influenței background-ului.
- transformarea datelor.

I.3.1. Eliminarea spoturilor marcate

În această etapă pot fi eliminate din analiză spoturile marcate la controlul de calitate al array-ului.

I.3.2. Eliminarea influenței backgroundului

Intensitatea semnalului microarray nu este datorată doar hibridizării între sondele și țintele de ARN marcate fluorescent, ci și a unor surse sistematice de zgomot introduse de o anumită fluorescență naturală a array-ului sau de hibridizările nespecifice între ARN-urile țintă marcate fluorescent și suprafața array-ului. Acest zgomot, numit semnal de background, trebuie extras din semnalul dat de spot. În cazul în care această corecție nu este făcută corect, semnalul raportat poate fi semnificativ mai mare sau mai mic decât valoarea reală. Procedura de eliminare a background-ului funcționează bine dacă intensitatea spotului este mai mare decât a background-ului. Cu toate acestea, când intensitatea background-ului depășește valoarea intensității spotului, rezultatul va fi un număr negativ lipsit de semnificație biologică. Câteva metode de extragere a semnalului de background sunt detaliate în cele ce urmează.

Metoda locală de estimare a backgroundului extrage media/mediana semnalului dat de pixelii background-ului local ai fiecărui spot, identificați prin metodele amintite, din media/mediana semnalului spotului corespunzător (ex. media/mediana semnalului background-ului local al spotului X este scăzută din media/mediana semnalului brut al spotului X, ș.a.m.d.). Metoda de corecție locală se bazează pe presupunerea că semnalul background-ului local este aditiv, adică este prezent și în semnalul dat de spot. Această metodă este indicată atunci când semnalul background-ului de pe suprafața array-ului este neomogen. În cazul în care semnalul background-ului local nu este aditiv, se impune folosirea unor metode globale de estimare a acestuia**.** Aceste metode impun eliminarea tuturor spoturilor și zonelor de background care nu au trecut controlul de calitate și care au fost marcate ca neuniforme.

Metoda globală de estimare a background-ului evaluează intensitatea background-ului la nivelul întregului array. Media/mediana pixelilor background-ului este extrasă din intensitatea brută a fiecărui spot. (ex. media/mediana semnalului background-ului este scăzută din media/mediana semnalului brut al spotului X, Y, Z, ș.a.m.d.).

În cazul în care există hibridizări nespecifice semnificative pe suprafața array-ului, ambele metode pot supra- sau subestima semnalul de background. Multe studii au raportat prezența „spoturilor negative" datorită blocării incomplete a suprafeței array-ului în procesul de fabricație. Acest fenomen se manifestă prin legarea cu o eficacitate mai mare a ARN-ului țintă marcat fluorescent de suprafața array-ului decât de sondele plasate în spoturi. Pentru rezolvarea acestui inconvenient sunt folosite pe suprafața array-ului seturi de ***controale negative,*** care conțin sonde astfel selectate încât să prezinte o omologie cât mai redusă cu ARN-ul țintă care practic nu va hibridiza în aceste spoturi. Astfel, tot ceea ce este măsurat va reprezenta doar fluorescența suprafeței slide-ului și eventualele hibridizări nespecifice.

Metoda de estimare a background-ului prin controale negative este una dintre metodele globale cele mai recomandate. La estimarea background-ului sunt folosite doar controalele negative pentru care intensitatea semnalului este minimă, între care nu există diferențe semnificative și care au trecut atât controlul de calitate cât și marcajul de neuniformitate. Media/mediana valorilor acestor controale este scăzută din media/mediana fiecărui spot.

Metoda semnalului minim al spoturilor de pe array este tot o metodă globală, recomandată în cazul în care nu sunt validate controalele negative. Din intensitatea fiecărui spot este extrasă valoarea minimă a intensității spoturilor de pe array.

I.3.3. Transformarea datelor

După corecția background-ului datele sunt, în general, transformate prin logaritmare. Această transformare îmbunătățește caracteristicile distribuției datelor, aducându-le la distribuții normale ce permit folosirea parametrilor statistici clasici pentru analiza lor.

În cazul marcajului two-color, intensitățile corespunzătoare celor două canale de fluorescență sunt exprimate ca rapoarte (rații) pentru a pune în evidență diferențele de expresie genică între cele două probe (proba de interes vs. proba de referință). Raportul este dat de relația:

$$T_k = R_k / G_k \text{ pentru fiecare gena k de pe array.}$$

unde R_k reprezintă intensitatea probei de interes marcată cu Cy5, iar G_k reprezintă intensitatea probei de referință, marcată cu Cy3.

Aceste rapoarte (rații) tratează însă genele supra- și sub-exprimate într-un mod diferit. Genele cu un nivel de expresie de 2 ori mai mare în proba de interes față de referință vor avea un raport T_k=2, în timp ce genele cu un nivel de expresie de 2 ori mai mic în proba de interes față de referință vor avea un raport T_k= 0,5 (½). Ca urmare, nivelele de expresie ale genelor sub-exprimate

vor fi comprimate în intervalul 0 și 1, în timp ce nivelurile de expresie ale genelor supra-exprimate vor fi mai mari decât 1.

Această problemă este rectificată aplicând transformarea logaritmică, în general, folosind logaritmul în baza 2. Logaritmarea permite transformarea raportului intensităților în diferența acestora, după relația matematică:

$$log\ A/B = log\ A - log\ B$$

producând astfel un spectru continuu al valorilor. Logaritmarea rațiilor intensităților tratează simetric genele supra- și sub-exprimate. Astfel, în exemplul de mai sus, o gena supra-exprimată de 2 ori în proba de interes vs. referință va avea un logaritm al raportului $\log_2 T_k = 1$ ($\log_2 2 = 1$) în timp ce o genă sub-exprimată de 2 ori în proba de interes vs. referință va avea un logaritm al raportului $\log_2 T_k = -1$ ($\log_2 (1/2) = -1$). Pentru genele între care nu există o diferență de expresie $\log_2 T_k = 0$.

În cazul marcajului monocolor este logaritmat nivelul „absolut" de expresie genică deoarece în această situație doar o singură probă este hibridizată pe array[5].

În figura 1.4 sunt ilustrate datele unui experiment two-color în două tipuri de reprezentări: intensitățile brute și valorile logaritmate ale acestora. Fiecare punct de pe grafic reprezintă intensitatea unui spot de pe array.

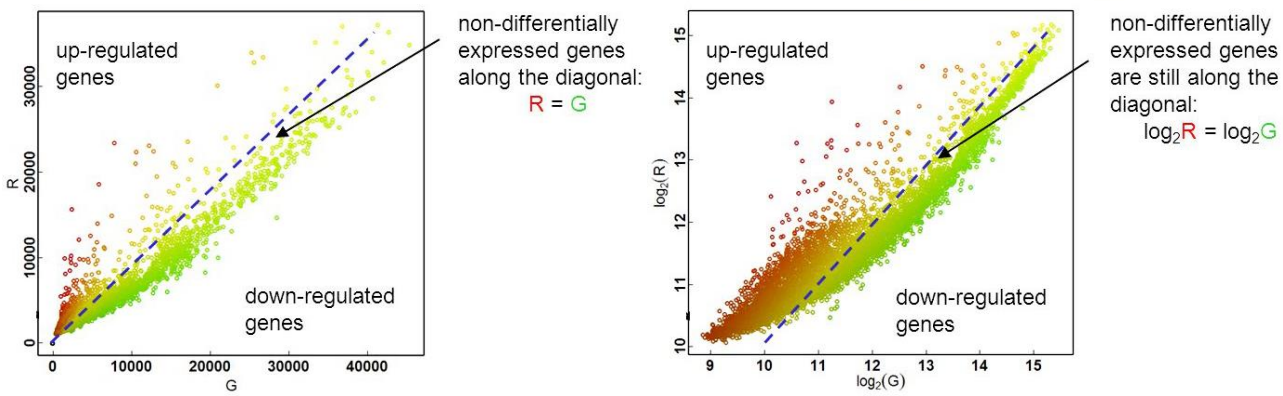


Fig. 1.4. Intensitățile brute (stânga) și valorile logaritmate ale acestora (dreapta).

În cazul intensităților brute se constată că majoritatea valorilor sunt concentrate în colțul stânga jos al graficului, variabilitatea datelor crescând cu intensitatea. O astfel de distribuție a datelor nu satisface condițiile pentru aplicarea ulterioară a parametrilor statistici clasici. Datele logaritmate sunt împrăștiate uniform, variabilitatea fiind aproximativ constantă la toate intensitățile. Aceste date tind spre o distribuție normală care satisface condițiile cerute de analizele ulterioare.

Studiile de genomică funcțională realizate cu ajutorul tehnologiei microarray permit, astfel, cuantificarea simultană a nivelurilor de expresie a zeci de mii de gene. Aceste studii generează cantități enorme de date care furnizează informații cu rol crucial în cercetarea biomedicală privind înțelegerea proceselor biologice, identificarea de noi biomarkeri, identificarea mecanismelor de toxicitate, predicția răspunsului la terapie. Alături de folosirea unor probe de calitate adaptate condițiilor de studiu, strategiile în stabilirea designului, alegerea celor mai potrivite metode și validările repetate ale acestora, sunt esențiale pentru a garanta obținerea unor informații de calitate.

I.3.4. Normalizarea intra array

Acest tip de normalizare se aplică tehnologiilor two-color. Compararea genelor diferit exprimate se reduce, în acest caz, la compararea fluorescențelor corespunzătoare Cy3 și Cy5. Principalele surse de erori, în acest caz pot fi:

A. erori de incorporare și detecție fluorescentă datorate:

- incorporării diferite a Cy3 și Cy5 în ARNc-ul țintă.
- emisiei fluorescente diferite ale Cy3 și Cy5 cu grad diferit de incorporare.
- detecției diferențiate în fotomultiplicator a emisiilor fluorescente ale Cy3 și Cy5 de diferite intensități.

B. erori spațiale:

- variația gradientului intensității celor doi fluorocromi în diferite zone ale array-ului datorită diferențelor de focusare a laserului pe suprafața array-ului.

Metodele de normalizare sunt numeroase și depind de tehnologia utilizată, existând un număr mare de publicații pe această tema[6,7,8]. Pentru datele Agilent, se aplică mai întâi o normalizare liniară (regresie liniară) pentru corectarea erorilor apărute la incorporarea și detecția fluorocromilor, urmată de o normalizare neliniară de tip LOWESS (LOcally WEighted Scatterplot Smoothing) pentru corectarea erorilor spațiale.

Toate aceste metode pornesc de la premisa că, în general, majoritatea genelor dintr-un experiment microarray nu sunt diferit exprimate. Dacă această presupunere nu este adevărată, folosirea unor probe de referință este mult mai potrivită. Probele de referință pot fi *gene housekeeping* sau *controale spike-in exogene*.

I.3.4.1. Regresia liniară

Conceptul de regresie liniară permite estimarea valorii unei variabile pe baza alteia dacă între cele două variabile se presupune că există o relație liniară:

$$Y = bX + a$$

unde: a reprezintă interceptul (locul pe ordonată unde dreapta de regresie se intersectează cu OY, sau valoarea lui Y pentru $X = 0$), b reprezintă gradientul (indică cu cât se modifică Y atunci când X crește sau scade cu o unitate). Constantele a și b țin cont de abaterea standard și de media variabilelor după următoarele formule:

$$b = r \cdot S_y / S_x$$

unde r reprezintă valoarea coeficientului de corelație dintre X și Y, S_y reprezintă abaterea standard a variabilei Y, S_x este abaterea standard a variabilei X.

$$a = M_y - b \cdot M_x$$

unde M_y reprezintă media variabilei Y, M_x reprezintă media variabilei X.

Folosind aceste formule, valorile estimate a lui Y vor fi calculate pe baza valorilor date ale lui X. În cazul experimentelor microarray reprezentarea grafică a celor două fluorescențe permite calcularea, pe baza ecuației de regresie liniară, a factorilor de normalizare care permit rescalarea unei intensități pe seama celeilalte pentru fiecare genă de pe array.

Pentru normalizarea intensităților celor două fluorescențe se folosește reprezentarea grafică a mediei logaritmilor intensităților vs. logaritmul rațiilor pentru fiecare spot (fig. 1.5.). Această reprezentare se numește grafic de tip MA.

Acest grafic are pe OX: (media) A = ½ (log R + log G) și pe OY: (raportul) M= log R- log G. Fiecare punct de pe grafic reprezintă un spot/genă.

Această reprezentare este foarte sugestivă pentru comportamentul celor două canale de fluorescență. În cazul în care ele răspund similar, punctele se vor așeza simetric de-a lungul unei linii orizontale care trece prin 0, orice variație de la această linie reprezentând răspunsuri diferite ale celor două canale de fluorescență. Linia de regresie trasată prin norul de puncte arată tendința

celor două fluorescențe. Exemplul din figura 1.5 (stânga) sugerează ca Cy5 răspunde mai puternic la intensități mici decât Cy3, în timp ce la intensități mari Cy3 răspunde mai puternic decât Cy5.

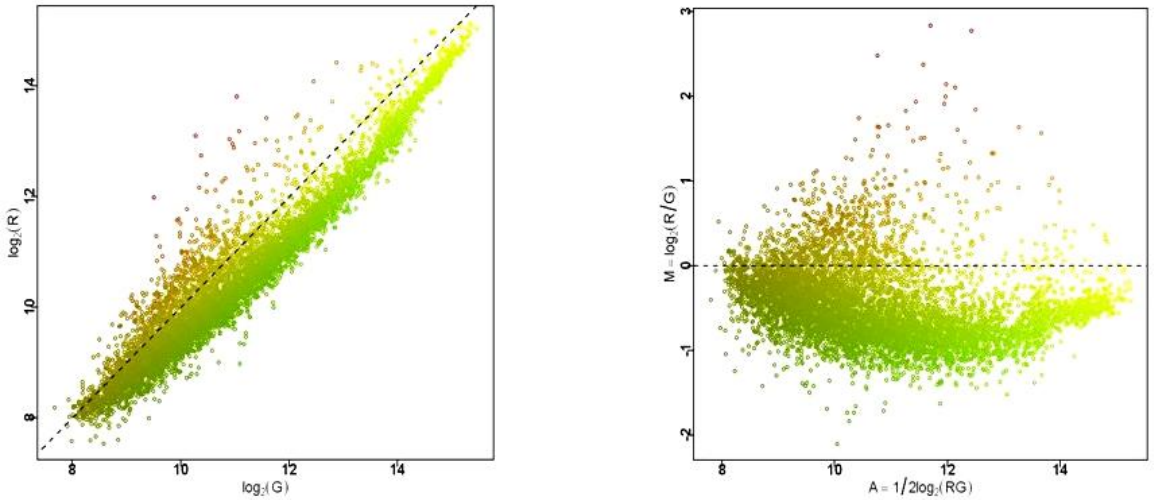


Fig. 1.5. Reprezentare a datelor de tip MA (stânga), reprezentare a datelor de tip MA normalizate prin regresie (dreapta). Datele brute $\{(R,G)\}_{n = 1..5184}$ unde R = semnalul canalului roșu și G = semnalul canalului verde. Date normalizate $\{(M,A)\}_{n = 1..5184}$ unde $M = \log_2(R/G)$ (raport) și $A = 1/2 \cdot \log_2(R \cdot G)$ (intensitatea semnalului) deci, $R = (2^{2A+M})^{1/2}$ și $G = (2^{2A-M})^{1/2}$.

Deoarece, așa cum s-a menționat mai sus, se presupune ca majoritatea genelor nu sunt diferit exprimate, acest efect este atribuit artefactelor experimentale, care vor fi înlăturate prin normalizarea datelor pe baza ecuației de regresie. Sunt identificate gradientul și interceptul liniei de regresie a norului de puncte, după care este calculată valoarea estimată a logaritmului rației pe baza ecuației de regresie din valorile mediei logaritmilor intensităților. Pentru fiecare spot valoarea normalizata a logaritmului rației se obține prin extragere a valorii estimate prin regresie a logaritmului rației din valoarea brută a acestuia, obținându-se un set de date liniarizate (fig. 1.5. dreapta).

I.3.4.2. Regresia LOWESS (LOcally WEighted Scatterplot Smoothing)

Regresia LOWESS în forma ei generalizată LOESS (LOcal regrESSion) efectuează un număr mare de regresii locale în ferestre suprapuse de-a lungul norului de date unind aceste linii de regresie pentru a genera o suprafață polinomială bidimensională pe graficul MA. La fel ca și în cazul regresiei liniare, valoarea estimată prin regresia față de suprafața LOESS a logaritmului rației este extrasă din valoarea brută a acestuia pentru fiecare spot, obținându-se astfel intensități normalizate fără erori spațiale.

I.3.5. Normalizarea inter array

Metodele de normalizare inter array permit compararea probelor hibridizate pe diferite array-uri, indiferent de metoda folosită: one-color sau two-color (fig. 1.6.a). Deoarece hibridizările au loc în experimente diferite, fiecare reacție poate introduce mici erori. Astfel, pot să apară diferențe între intensitățile globale ale array-urilor care nu sunt datorate diferențelor biologice. Pentru o comparare corectă a probelor este necesară corectarea variabilității introduse prin utilizarea mai multor array-uri. Pentru aceasta, datele trebuie transformate astfel încât toate experimentele să aibă aceeași distribuție a valorilor. Pentru tehnologia two-color această normalizare este opțională și este aplicată numai după normalizarea intra array. Pentru tehnologia one-color normalizarea se face folosind metode pentru egalizarea distribuțiilor după transformarea logaritmică a datelor. Metodele de normalizare între array-uri sunt: scalarea, centrarea sau normalizarea distribuției.

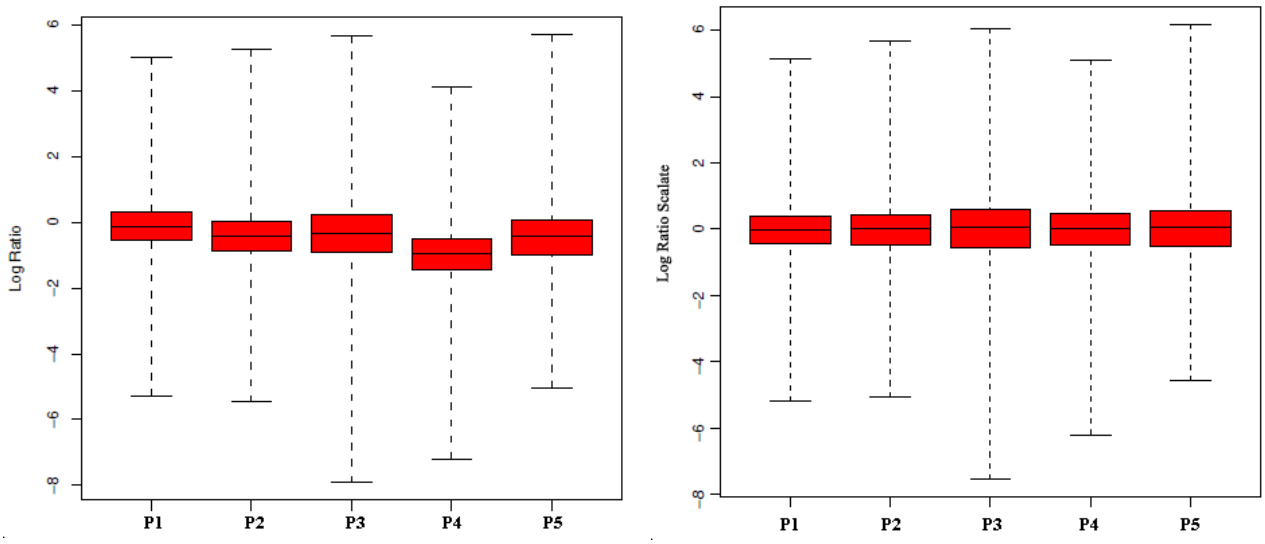
Scalarea. Datele sunt scalate pentru ca mediile tuturor distribuțiilor să fie egale. Media intensităților logaritmate sau a rațiilor logaritmate a tuturor spoturilor de pe array este scăzută din valoarea intensității logaritmate respectiv a rației logaritmate a fiecărui spot de pe array. După scalare, media tuturor log intensităților sau a log rațiilor de pe array va fi 0 (fig. 1.6.b). O alternativă ar fi folosirea medianei în locul mediei, aceasta furnizând rezultate mai robuste deoarece nu este sensibila la valorile extreme ale distribuției valorilor.

Centrarea. Datele sunt centrate pentru ca mediile și deviațiile standard ale tuturor distribuțiilor să fie egale. Metoda este similară scalării. Diferența dintre valoarea intensității logaritmate sau a rației logaritmate a fiecărui spot de pe array și media intensităților logaritmate respectiv a rațiilor logaritmate a tuturor spoturilor de pe array este împărțită la deviația standard. După centrare media tuturor intensităților logaritmate sau a rațiilor logaritmate de pe array va fi 0, iar deviația standard va fi 1 (fig. 1.6.c). Metoda centrării este frecvent folosită pentru compararea array-urilor.

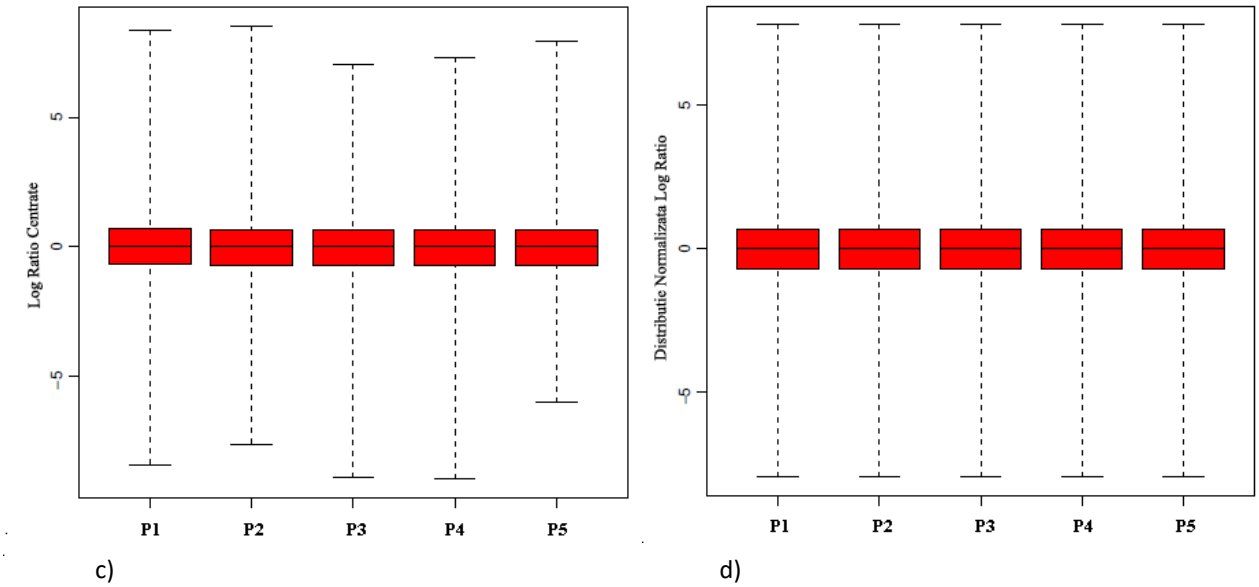
Distribuția normalizată urmărește obținerea unor distribuții identice a datelor pe fiecare array. Pentru aceasta:

- datele sunt centrate.
- pentru fiecare array datele centrate sunt ordonate ascendent.
- se calculează o nouă distribuție pentru care:
    - *valoarea minimă* = media celor mai mici valori de pe fiecare array.
    - *următoarea valoare* = media valorilor de pe array-uri care corespund aceleiași poziții.
    - și tot așa până la *valoare maximă* = media celor mai mari valori de pe fiecare array.
- se înlocuiește fiecare valoare de pe fiecare array cu media corespunzătoare în noua distribuție.

După această normalizare datele de pe fiecare array vor avea o medie de 0, o deviație standard de 1 și distribuții identice pe fiecare array (fig. 1.6.d). Acest tip de normalizare este alternativa la metoda de centrare, însă, totuși, metoda de centrare este mult mai simplă și mai frecvent utilizată pentru normalizarea microarray-urilor.



a) b)



c) d)

Fig. 1.6. Distribuția valorilor rațiilor logaritmate pentru 5 pacienți. Linia din centrul fiecărei distribuții reprezintă media (mediana) valorilor distribuției, iar dimensiunea casetei reprezintă deviația standard. a) rații logaritmate brute, b) rații logaritmate normalizate prin scalare, c) rații logaritmate normalizate prin centrare, d) distribuție normalizată a rațiilor logaritmate.

I.4. Referințe bibliografice

[1] **Gu W., Wang Y.,**Ed., 2011, *Gene discovery for disease models*, John Wiley & Sons, 11-30.

[2] **Quackenbush J.**, 2001, *Computational analysis of microarray data*. Nat Rev Genet 2001; 2:418-27.

[3] **Irizarry R.A., Hobbs B., Collin F., Beazer-Barclay Y.D., Antonellis K.J., Scherf U., et al**., 2003, *Exploration, normalization, and summaries of high density oligonucleotide array probe level data*. Biostatistics 2003; 4:249-64.

[4] **Tarca A.L., Cooke J.E., Mackay J**., 2005, *A robust neural networks approach for spatial and intensity-dependent normalization of cDNA microarray data*. Bioinformatics 2005; 21:2674-83.

[5] **Cui X., Kerr M.K., Churchill G.A**., 2003, *Transformations for cDNA microarray data*, Stat Appl Genet Mol Biol 2003; 2: Article4.

[6] **Bolstad B.M., Irizarry R.A., Astrand M., Speed T.P**., 2003, *A comparison of normalization methods for high density oligonucleotide array data based on variance and bias*. Bioinformatics 2003; 19:185-93.

[7] **Chiogna M., Massa M.S., Risso D., Romualdi C**., 2009, *A comparison on effects of normalisations in the detection of differentially expressed genes*. BMC Bioinformatics. 2009 Feb 13;10:61.

[8] **Khondoker M.R., Glasbey C.A., Worton B.J**., 2007, *A comparison of parametric and nonparametric methods for normalising cDNA microarray data*. Biom J. 2007 Dec; 49(6):815-23.