# مقایسه روش های پیش پردازش داده‌های بیان ژن در ریزآرایه‌های افی‌متریکس

**هادی آتشی**

دانشیار بخش علوم دامی، دانشکده کشاورزی، دانشگاه شیراز، شیراز، ایران.

**شماره صفحات**
**۵–۱۶**

* نویسنده مسئول: **Atashi@Shirazu.ac.ir**

تاریخ دریافت: ۱۴۰۱/۰۷/۱۹ تاریخ پذیرش: ۱۴۰۱/۱۰/۲۱

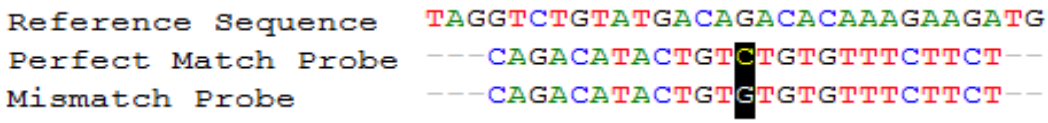
## چکیده

**امروزه، فنآوری ریزآرایه، روشی قدرتمند برای اندازه‌گیری هم‌زمان الگوهای بیان ژن، شمار زیادی از ژنها است. پیش از آنالیز، داده‌های ریزآرایه، باید پیش‌پردازش شوند تا، با حذف بسیاری از منابع تغییرات، نتایج آنالیز داده‌ها صحت لازم را داشته باشند. بدین منظور، فرآیند پیش‌پردازش چند گامه شامل: تصحیح پس‌زمینه، نرمال‌سازی، و چکیده‌سازی دارد که هر کدام به چندین روش انجام می‌شوند. هدف این پژوهش، مقایسه‌ی اثر روش‌های متفاوت پیش‌پردازش بر نتایج آنالیز داده‌های ریزآرایه است. در این راستا، داده‌های استفاده شده در این پژوهش، از وبگاه NCBI دانلود شدند. شماره‌ی دسترسی، شماره‌ی پلت‌فرم و نام داده‌ها به ترتیب GSE56589، GPL18534 و Affymetrix Bovine Genome Array است. دو روش تصحیح پس‌زمینه (MAS.5 و RMA.2)، دو روش نرمال‌سازی ( Scaling normalization و Quantile normalization) و دو روش چکیده‌سازی (Tukey biweight و Medianpolish) ارزیابی شد. در نهایت، نتایج حاصل از این مطالعه، نشان داد که تعداد و نوع ژن‌های با بیان متفاوت در روش‌های مختلف چکیده‌سازی تفاوت زیادی ندارند، اما، با تغییر در روش تصحیح پس‌زمینه یا روش نرمال‌سازی هم تعداد و هم نوع ژن‌های با بیان متفاوت تغییر زیادی می‌کند.**

**کلیدواژه‌ها**: بیان ژن، Affymetrix، ریزآرایه و پیش پردازش.

**مقدمه**

رایج‌ترین کاربرد ریزآرایه‌ها، اندازه‌گیری هم‌زمان الگوهای بیان شمار زیادی ژنها است. در فنآوری ریزآرایه‌ها از ویژگی "هیبرید شدن اسیدهای نوکلئیک" برای اندازه‌گیری رونوشت‌های یک اسید نوکلئیک ویژه در یک نمونه استفاده می‌شود. در هر ریزآرایه، دسته‌ای از اسیدهای نوکلئیک روی یک سطح جامد چسبیده‌اند[۱] که به آنها کاوشگر[۲] گویند و دسته‌ای دیگر از اسیدهای نوکلئیک که با کاوشگرها هیبرید می‌شوند، ملکول‌های هدف[۳] هستند که پیش از اضافه شدن به کاوشگرها نشاندار[۴] شده‌اند. در فنآوری ریزآرایه، از اسکنرهای ویژه‌ای برای اندازه‌گیری میزان ملکول‌های هدف هیبرید شده با هر کاوشگر استفاده می‌شود. در بین آرایه‌های الیگونوکلئوتیدی با تراکم بالا[۵]، استفاده از تراشه‌ی ژنی افی‌متریکس[۶] برای اندازه‌گیری بیان ژن‌ها رایج‌تر است. در تراشه‌های ژنی افی‌متریکس، الیگونوکلئوتیدهای کوتاه ۲۵ بازی به عنوان کاوشگر استفاده می‌شوند که به صورت جفت‌شده یا دوتایی هستند و به آنها جفت-کاوشگر گویند. هر جفت-کاوشگر، از یک کاوشگر جور[۷] (PM) و یک کاوشگر ناجور[۸] (MM) تشکیل شده است (Affymetrix, 2001). کاوشگر جور برای هیبرید شدن با رونوشت‌های ژن مورد نظر طراحی شده است؛ اگرچه هیبرید شدن این کاوشگرها با رونوشت‌های ژن‌های دیگر نیز اجتناب‌ناپذیر است. کاوشگر ناجور، برای اندازه‌گیری هیبریدهای نااختصاصی کاوشگر جور مربوطه، استفاده می‌شوند. در تراشه‌ی افی‌متریکس توالی همه‌ی بازها (۲۵ باز) به جز باز نوکلئوتیدی سیزدهم در کاوشگرهای جور و ناجور همانند است، باز نوکلئوتیدی سیزدهم در کاوشگر ناجور مکمل باز سیزدهم در کاوشگر جور است (شکل ۱). تعداد جفت-کاوشگرها در تراشه‌های مختلف افی‌متریکس ممکن است متفاوت باشد و به یک میلیون و سیصد هزار جفت-کاوشگر نیز می‌رسد.




**شکل ۱: یک نمونه از توالی مرجع، کاوشگر جور و کاوشگر ناجور در تراشه‌ی ژنی Affymetrix**

**Fig 1: An example of reference sequence, perfect match probe and mismatch probe in Affymetrix gene Chip.**

[1] Immobilized
[2] Probe
[3] Target
[4] Labeled
5 High-density oligonucleotide arrays
[6] Affymetrix
[7] Perfect Match
[8] Mismatch

در هر پژوهش، افزون بر سازه‌های مورد آزمایش (دارو، بیماری، تیمار آزمایشی و ... )، پراکنش داده‌های خام[1] یا داده‌های در سطح کاوشگر تحت تأثیر تعداد زیادی عوامل ناشناخته دیگر است؛ از این رو، پیش از آنالیز داده‌های ریزآرایه و شناسایی ژن‌های با بیان متفاوت[2]، داده‌ها پیش‌پردازش می‌شوند (Bolstad *et al*., 2003). فرآیند پیش‌پردازش چند گامه شامل تصحیح پس زمینه[3]، نرمال‌سازی[4]، و چکیده‌سازی[5] دارد. در پیش‌پردازش، ابتداً، تصحیح پس‌زمینه، سپس، نرمال‌سازی و در آخر چکیده‌سازی انجام می‌شود (Freudenberg, 2005). اگرچه، برای هر کدام از گامه‌های پیش‌پردازش روش‌های زیادی ارایه شده است، هیچگونه، توافق نسبی در مورد اینکه کدام روش‌ها مناسب‌تر هستند، وجود ندارد (Ray *et al*., 2008; Wu, 2009).

**تصحیح پس‌زمینه:** برخی از هیبریدها در هر جفت-کاوشگر نااختصاصی[6] هستند، که باید پیش از تصحیح شوند تا صحت اندازه‌گیری‌ها بالا رود و همچنین، داده‌های بیان در یک دامنه‌ی قابل قبول قرار گیرند؛ این گونه تصحیح‌ها را تصحیح پس‌زمینه گویند که به روش‌های گوناگونی انجام می‌شود (Freudenberg, 2005; Tarca *et al*., 2006)). روش‌های MAS.5 و RMA از پرکاربردترین روش‌های تصحیح پس‌زمینه هستند. در روش تصحیح پس‌زمینه MAS.5، هم از کاوشگرهای جور و هم از کاوشگر های ناجور استفاده می‌کند. در این روش، ابتدا چیپ‌ها به k ناحیه‌ی مساوی تقسیم می‌شوند (به‌طور پیش‌فرض $k=16$). سپس، برای هر کاوشگر یک ضریب پس‌زمینه‌ی خاص $b(x,y)$ و یک ارزش خطا $n(x,y)$، بدست می‌آید. در اینجا، x و y ناحیه‌ی هندسی کاوشگرها روی چیپ را نشان می‌دهند. پس‌زمینه‌ی خاص ($b_k$) هر کاوشگر حاصل تفریق از شدت خام کاوشگر است مگر اینکه ارزشی کمتر از ارزش خطا داشته باشد که در این مورد شدت کاوشگر، برابر ارزش خطا در نظر گرفته می‌شود (Affymetrix, 2001).

روش تصحیح پس‌زمینه RMA توسط گروه Speed در دانشگاه Berkeley توسعه یافته است (۵). یکی از فرضیه‌های این روش استفاده از کاوشگرهای جور و نادیده گرفتن کاوشگرهای ناجور است. در این روش فرض می‌شود که شدت سیگنال هر کاوشگر جور از دو بخش تشکیل شده است؛ یک بخش به نام خطا که دارای توزیع نرمال است و بخش دیگر سیگنال واقعی مربوط به کاوشگر است که دارای توزیع نمایی است (Freudenberg, 2005).

**نرمال‌سازی:** داده‌های حاصل از آرایه‌های مختلف به دلیل تفاوت در کارایی واکنش‌های نسخه برداری معکوس[7]، نشان‌دار کردن، هیبریداسیون و همچنین، تفاوت در شرایط آزمایشگاهی آرایه‌ها، قابل مقایسه نیستند؛ و باید روی آنها نرمال‌سازی انجام

[1]Raw data
[2]Differentially expressed
[3]Background adjustment
[4]Normalization
[5]Summarization
[6]Non-specific hybridization
[7]Reverse transcription

شود (Bolstad *et al*., 2003; Chiogna *et al*., 2009). روش‌های Scaling normalization و Quantile normalization از مشهورترین روش‌های نرمال‌سازی داده‌های ریزآرایه هستند. روشScaling normalization به وسیله‌ی کمپانی افی‌متریکس پیشنهاد شده است (Affymetrix, 2001). در این روش، داده‌ها به گونه‌ای تبدیل می‌شوند که برخی پارامترهای آماری مانند میانه، میانگین، میانگین اصلاح‌شده[1] برای هر چیپ بعد از نرمال‌سازی باهم برابر باشند. روش Quantile normalization توسط بلستاد (Bolstad, 2001) پیشنهاد شده است. بیشتر پژوهش‌گران به‌دلیل سرعت بالا و فرض‌های کمتر از این روش برای نرمال‌سازی استفاده می‌کنند. در این روش ابتدا شدت‌های بیان برای هر نمونه در هر چیپ از کوچک‌ترین به بزرگ‌ترین مرتب می‌شود، سپس، میانگین مربوط به هر ردیف (شامل شدت بیان در همه‌ی چیپ‌ها) محاسبه می‌شود و از کوچک‌ترین به بزرگ‌ترین مرتب می‌شوند و در پایان، میانگین هر ردیف با شدت بیان کاوشگرهای آن ردیف جایگزین می‌شود.

**چکیده‌سازی**: در تراشه‌های Affymetrix، به ازای هر توالی مرجع[2]، بین ۱۱ تا ۲۵ جفت-کاوشگر وجود دارد، که با نام دسته-کاوشگر[3] شناخته می‌شوند. پیش از آنالیز داده‌ها، اطلاعات همه‌ی جفت-کاوشگرهای یک دسته-کاوشگر (۱۱ تا ۲۵ جفت-کاوشگر)، در یک ارزش خلاصه می‌شود که به آن چکیده‌سازی یا خلاصه‌سازی گویند (Bolstad *et al*., 2003; Freudenberg, 2005). روش‌های Tukey biweight و Medianpolish از روش‌های پرکاربرد چکیده‌سازی داده‌های ریزآرایه هستند. روش چکیده‌سازی Tukey biweight، از کاوشگرهای جور (PM) و ناجور (MM) برای چکیده‌سازی استفاده می‌کند (Affymetrix, 2001). در این روش ابتدا برای هر دسته-کاوشگر میزان IMها (ایده‌ال سازی MMها) با استفاده از توابع زیر محاسبه می‌شود:

رابطه ۱:
$$IM=\begin{cases} MM & if\ MM < PM \\ \frac{PM}{2} & if\ MM \geq PM \end{cases}$$

سپس، برای هر دسته-کاوشگر، با استفاده از تابع زیر میانگین تبدیل یافته محاسبه می‌شود:

رابطه ۲:
$$\sum_{i=1}^{11} log2\ (PMi - MMi)$$

روش چکیده‌سازی Medianpolish، شدت کاوشگرهای ناجور را نادیده می‌گیرد و برای بیان میانگین بیان در چند آرایه، از شدت‌های کاوشگرهای جور استفاده می‌کند (Irizarry & Jaffee, 2006; Li & Hung Wong, 2001; Ray *et al*.,2008). بنابراین، روش‌های زیادی برای هر کدام از گام‌های پیش‌پردازش وجود دارد و روش استفاده شده می‌تواند بر نتایج حاصل از آنالیز، تأثیر شایانی داشته باشد(Hoffmann & Dugas, 2002). هدف این مطالعه مقایسه‌ی اثر برخی روش‌های رایج پیش

[1]Trimmed mean
[2]Reference sequence
[3]Probeset

پردازش بر تعداد ژن‌های با بیان متفاوت در یک آزمایش ریزآرایه است.

## مواد و روش‌ها

### داده‌ها

در این مطالعه، از داده‌های خام آزمایش Morris *et al* و همکاران (2009) استفاد شد که هدف آن بررسی اثر شدت تعادل منفی انرژی بر بیان ژن‌ها در بافت طحال گاوهای شیری پرتولید بود (Morris *et al*., 2009). ۲۴ گاو هلشتاین دو هفته پیش از زایمان به گونه‌ی کاملا تصادفی در دو گروه تعادل منفی انرژی ملایم و شدید قرار گرفتند. گاوهای قرار گرفته در گروه تعادل منفی انرژی ملایم، روزانه ۸ کیلوگرم کنسانتره به همراه مصرف آزاد سیلاژ دریافت می‌کردند و هر روز یک‌بار دوشیده می‌شدند. گاوهای قرار گرفته در گروه تعادل منفی انرژی شدید، روزانه ۲۵ کیلوگرم سیلاژ و ۴ کیلوگرم کنسانتره دریافت کردند و هر روز سه بار دوشیده می‌شدند. پس از کشتار، از طحال گاوها نمونه‌گیری شد و برای بررسی بیان ژن استفاده شدند (Morris *et al*., 2009). داده‌ها از وبگاه NCBI (http://www.ncbi.nlm.nih.gov/geo) دانلود شدند. شماره‌ی دسترسی، شماره‌ی پلت‌فرم و نام داده‌ها به ترتیب GSE56589، GPL18534 و Affymetrix Bovine Genome Array است.

### پیش‌پردازش داده‌ها

داده‌ها پس از دانلود از وبگاه NCBI، وارد نرم‌افزار R شدند و همه‌ی گامه‌های آنالیز داده‌ها در نرم‌افزار R (http://www.r-project.org) و Bioconductor (https://www.bioconductor.org) انجام شد. برای پیش‌پردازش داده‌ها دو روش تصحیح پس‌زمینه (MAS.5[1] و RMA.2[2])، دو روش نرمال‌سازی (Scaling normalization و Quantile normalization) و دو روش چکیده‌سازی (Tukey biweight و Medianpolish) استفاده شد (جدول ۱). سه گامه‌ی پیش‌پردازش با تابع threestep در بسته affyPLM از نرم‌افزار Bioconductor انجام شد (Bolstad *et al*., 2005).

**جدول ۱: هشت روش استفاده شده برای پیش‌پردازش داده‌ها (دو روش تصحیح پس‌زمینه × دو روش نرمال‌سازی × دو روش چکیده‌سازی)**

**Table 1: Eight methods used for data preprocessing (two background correction methods × two normalization methods × two abstracting methods)**

| چکیده‌سازی<br>**Abstracting** | نرمال‌سازی<br>**Normalization** | تصحیح پس‌زمینه<br>**Background correction** | شماره‌ی روش<br>**Number of method** |
|---|---|---|---|
| Medianpolish | Quantile normalization | RMA.2 | **1** |
| | | MAS.5 | **2** |
| | Scaling normalization | RMA.2 | **3** |
| | | MAS.5 | **4** |
| Tukey biweight | Quantile normalization | RMA.2 | **5** |
| | | MAS.5 | **6** |
| | Scaling normalization | RMA.2 | **7** |
| | | MAS.5 | **8** |

[1]Microarray Suite 5.0
[2]Robust Multi-array Average

**آنالیز داده‌ها**

تعداد کل جفت-کاوشگرها ۵۳۵۸۲۴ و تعداد کل دسته-کاوشگرها ۲۴۱۲۸ بود، که ۶۴۷۸تا از آنها اینترز آی‌دی[1] مشخص نداشتند و از آنالیز حذف شدند، و تعداد ۱۷۶۵۰ تا از آنها برای آنالیزهای بعدی باقی ماند. معمولا به ازای هر ژن یک یا چند توالی رفرنس، و یک یا چند دسته-کاوشگر وجود دارد؛ به عبارت دیگر برخی دسته-کاوشگرهای با آی‌دی های متفاوت، اینترز آی‌دی یکسان دارند و باید پس از پیش‌پردازش[2]، اطلاعات آن دسته-کاوشگرها در یک ارزش خلاصه شود، در این شرایط میانگین دسته-کاوشگرها، میانه‌ی دسته-کاوشگرها، ارزش دسته-کاوشگر دارای بیشترین مقدار بیان، یا ارزش دسته-کاوشگر دارای کمترین مقدار بیان انتخاب می‌شود (Irizarry, 2003). در این پژوهش، از ارزش دسته-کاوشگر دارای بیشترین مقدار بیان استفاده شد. در پایان این گامه، مقدار بیان ۱۱۸۵۰ ژن برای آنالیز باقی ماند. اثر تیمارها بر بیان ژن‌ها به وسیله‌ی آزمون eBayes[3] در بسته‌ی limma (Ritchie *et al*., 2015; Smyth *et al*., 2005) از نرم‌افزار R انجام شد. از آنجا که تعداد ژن‌ها در این‌گونه آنالیزها زیاد است (۱۱۸۵۰ ژن در این مطالعه)، بنابراین خطای نوع اول در آنها بالا است و برای کاهش آن، تصحیح آزمون‌های چندگانه[4] انجام می‌شود. در این مطالعه، برای تصحیح آزمون‌های چندگانه از روش BH[5] استفاده شد (Benjamini & Hochberg, 1995). در روش BH، ابتدا ژن‌ها بر اساس ارزش P محاسبه شده از کوچک‌ترین به بزرگ‌ترین ردیف و رتبه‌بندی می‌شوند (یعنی به ژن دارای کوچک‌ترین ارزش P، رتبه یک و ژن دارای بزرگ‌ترین ارزش P، رتبه‌ی N داده می‌شود و N نیز تعداد کل ژن‌ها است)؛ سپس ارزش P هر ژن در N ضرب و بر رتبه‌اش تقسیم می‌شود تا ارزش P تصحیح شده[6] به‌دست آید. تعداد ژن‌های با ارزش P تصحیح شده‌ی کمتر از یک درصد به عنوان ژن‌های معنی‌دار شمارش شدند. در هر روش پیش‌پردازش، همه‌ی ۱۱۸۵۰ ژن بر اساس ارزش P تصحیح شده، از کوچک‌ترین (ژن دارای کوچک‌ترین ارزش P تصحیح شده) به بزرگ‌ترین (ژن دارای بزرگ‌ترین ارزش P تصحیح شده) رتبه‌بندی شدند، و سپس همبستگی رتبه‌ای اسپیرمن[7] بین رتبه‌های ژن‌ها در روش‌های مختلف پیش‌پردازش، برآورد شد.

**مقایسه‌ی روش‌های پیش‌پردازش**

در این پژوهش، چهار معیار برای مقایسه‌ی روش‌های مختلف پیش‌پردازش داده‌های ریزآرایه بررسی شد: الف- همبستگی رتبه‌های داده شده به همه‌ی ژن‌ها (۱۱۸۵۰ ژن) بر اساس ارزش P تصحیح شده‌ی آنها در روش‌های مختلف پیش‌پردازش. به-

[1]Entrez ID
[2]Preprocessing
[3]Empirical Bayes Statistics
[4]Multiple testing correction
[5]Benjamini Hochberg
[6]Adjusted P value
[7]Spearman rank correlation

تعداد ژن‌های با بیان متفاوت[1] (ژن‌های معنی‌دار در سطح معنی‌داری یک درصد) در هر یک از روش‌های پیش‌پردازش. ج- همبستگی بین رتبه‌های داده شده به ۱۰۰ ژن نخست دارای کوچک‌ترین ارزش P تصحیح شده در روش‌های مختلف پیش-پردازش. د- تعداد ژن‌های مشترک بین ۱۰۰ ژن نخست دارای کوچک‌ترین ارزش P تصحیح شده در روش‌های مختلف پیش-پردازش.

## نتایج

### همبستگی رتبه‌های داده شده به همه‌ی ژن‌ها

همبستگی رتبه‌های داده شده به همه‌ی ژن‌ها (۱۱۸۵۰ ژن) بر اساس ارزش P تصحیح شده‌ی آنها و تعداد ژن‌های با بیان متفاوت[2] (ژن‌های معنی‌دار در سطح معنی‌داری یک درصد) در روش‌های مختلف پیش‌پردازش در جدول ۲ ارایه شده است. بالاترین همبستگی‌ها بین روش‌های ۱ و ۵ (۹۴ درصد)، ۲ و ۶ (۹۰ درصد)، ۳ و ۷ (۹۱ درصد) و ۴ و ۸ (۹۰ درصد) بود (جدول ۲). این روش‌ها فقط در چگونگی چکیده‌سازی با یکدیگر تفاوت دارند، به عنوان مثال، نرمال‌سازی و تصحیح پس‌زمینه در روش‌های ۱ و ۵ یکسان است، امّا چگونگی چکیده‌سازی در این دو روش متفاوت است (جدول ۱). بر اساس این نتایج می‌توان استنباط کرد که روش چکیده‌سازی به تنهایی تأثیر زیادی بر نتایج آنالیز ندارد. همبستگی بین روش‌های ۱و ۲ (۵۹ درصد)، ۳ و ۴ (۶۷ درصد)، ۵ و ۶ (۵۹ درصد) و ۷ و ۸ (۶۶ درصد)، نشان می‌دهد که تأثیر تغییر روش تصحصح پس‌زمینه بر نتایج قابل توجه است. زیرا این روش‌ها فقط در روش تصحیح پس‌زمینه با یکدیگر تفاوت دارند، به عنوان مثال، نرمال‌سازی و چکیده‌سازی در روش‌های ۱ و ۲ یکسان است، امّا چگونگی تصحیح پس‌زمینه در این دو روش متفاوت است (جدول ۱). همبستگی بین روش‌های ۱و ۳ (۴۰ درصد)، ۲ و ۴ (۴۷ درصد)، ۵ و ۷ (۴۵ درصد) و ۶ و ۸ (۵۰ درصد) نشان می‌دهد که تغییر روش نرمال‌سازی تأثیر بیشتری بر نتایج آنالیز در مقایسه با تغییر روش چکیده‌سازی و تصحیح پس‌زمینه دارد. به عبارت ساده‌تر، روش نرمال‌سازی بیشترین و روش چکیده‌سازی کمترین تأثیر را بر نتایج حاصل از آنالیز آماری دارد.

### تعداد ژن‌های با بیان متفاوت

نتایج بدست آمده از اثر روش‌های مختلف پیش‌پردازش بر تعداد ژن‌های معنی‌دار در جدول ۲ ارایه شده است. تعداد ژن‌های معنی‌دار در روش‌های مختلف پیش‌پردازش بین ۳۹ تا ۹۴ ژن متغیر است. روش ۸ بیشترین تعداد ژن‌های معنی‌دار (۹۴ ژن) و روش ۲ کمترین تعداد (۳۹ ژن) را شناسایی کردند.

### همبستگی بین رتبه‌های ۱۰۰ ژن دارای کوچک‌ترین ارزش P

همبستگی بین رتبه‌های داده شده به ۱۰۰ ژن نخست دارای کوچک‌ترین ارزش P تصحیح شده و تعداد ژن‌های مشترک

[1]Differentially expressed
[2]Differentially expressed

بین ۱۰۰ ژن نخست دارای کوچک‌ترین ارزش P تصحیح شده در روش‌های مختلف پیش‌پردازش در جدول ۳ ارایه شده است. تعداد ژن‌های مشترک بین ۱۰۰ ژن نخست در روش‌های ۱ و ۵ بیشترین بود (۹۰ ژن) و پس از آن تعداد ژن‌های مشترک بین روش‌های ۲ و ۶، ۳ و ۷، ۴ و ۸ به ترتیب ۸۰، ۸۴ و ۷۴ ژن بود که از دیگر روش‌ها بیشتر بود. این روش‌ها فقط در چگونگی چکیده‌سازی با یکدیگر تفاوت دارند، به عنوان مثال، نرمال‌سازی و تصحیح پس‌زمینه در روش‌های ۱ و ۵ یکسان است، امّا چگونگی چکیده‌سازی در این دو روش متفاوت است (جدول ۱). بنابراین می‌توان استنباط کرد که روش‌های چکیده‌سازی تفاوت زیادی با یکدیگر ندارند. تعداد ژن‌های مشترک در حالت ثابت بودن روش‌های چکیده‌سازی و نرمال‌سازی و متفاوت بودن روش تصحیح پس‌زمینه تغییر زیادی کرد (جدول ۳).

**جدول ۲: همبستگی اسپیرمن بین رتبه‌های ژن‌ها (اعداد بالای قطر) و تعداد ژن‌های با بیان متفاوت (اعداد قطری) در روش‌های مختلف پیش‌پردازش (۱۱۸۵۰ ژن).**

**Table 2: The Spearman correlation between the ranks of genes (above the diameter) and the numbers of the genes with different expression (on the diameter) in different pre-processing methods (11850 genes).**

| 8 | 7 | 6 | 5 | 4 | 3 | 2 | 1 | روش پیش‌پردازش **pre-processing method** |
|---|---|---|---|---|---|---|---|---|
| 34 | 43 | 59 | 94 | 32 | 40 | 59 | **58** | 1 |
| 48 | 34 | ۹۰ | 58 | 47 | 32 | **39** | | 2 |
| 66 | 91 | 33 | 41 | 67 | **57** | | | 3 |
| 90 | 65 | 55 | 32 | **51** | | | | 4 |
| 34 | 45 | 59 | **54** | | | | | 5 |
| 50 | 34 | **73** | | | | | | 6 |
| 66 | **94** | | | | | | | 7 |
| **92** | | | | | | | | 8 |

**جدول ۳: همبستگی اسپیرمن (بالای قطر)، تعداد ژن‌های معنی‌دار (روی قطر)، و تعداد ژن‌های مشترک (زیر قطر) بین رتبه‌های ۱۰۰ ژن نخست دارای کوچک‌ترین ارزش‌های P تصحیح شده در روش‌های مختلف پیش‌پردازش.**

**Table 3: The Spearman's correlation (above the diameter), the number of significant genes (on the diameter), and the number of common genes (below the diameter) between the ranks of the first 100 genes with the smallest P values corrected in the method various pre-processing methods.**

| 8 | 7 | 6 | 5 | 4 | 3 | 2 | 1 | |
|---|---|---|---|---|---|---|---|---|
| 67 | 34 | 57 | 94 | 59 | 43 | 54 | 58 | 1 |
| 71 | 44 | 84 | 51 | 72 | 40 | 39 | 43 | 2 |
| 59 | 83 | 20 | 37 | 38 | 57 | 39 | 60 | 3 |
| 77 | 57 | 60 | 50 | 51 | 46 | 57 | 36 | 4 |
| 54 | 33 | 61 | 54 | 37 | 57 | 44 | 90 | 5 |
| 71 | 32 | 73 | 41 | 51 | 37 | 80 | 40 | 6 |
| 61 | 94 | 40 | 61 | 46 | 84 | 41 | 61 | 7 |
| 92 | 48 | 64 | 39 | 74 | 46 | 58 | 39 | 8 |

**تعداد ژن‌های مشترک بین ۱۰۰ ژن دارای کوچک‌ترین ارزش P**

تعداد ژن‌های مشترک بین ۱۰۰ ژن نخست در روش‌های ۱و۲، ۳ و ۴، ۵ و ۶، ۷ و ۸ به ترتیب ۴۳، ۴۶، ۴۱ و ۴۸ ژن بود. این روش‌ها فقط در روش تصحیح پس‌زمینه با یکدیگر تفاوت دارند، بنابراین استنباط می‌شود که روش‌های تصحیح پس‌زمینه

تفاوت شایانی با یکدیگر دارند. تعداد ژن‌های مشترک بین ۱۰۰ ژن نخست در روش‌های ۱و۳، ۲ و ۴، ۵ و ۷، ۶ و ۸ به ترتیب ۶۰، ۵۷، ۶۱ و ۶۴ ژن بود. این روش‌ها تنها در روش نرمال‌سازی با یکدیگر تفاوت دارند، بنابراین تغییر روش نرمال‌سازی اثر زیادی بر نتایج و نوع ژن‌های با بیان متفاوت دارد.

**بحث**

در این مطالعه، اثر دو روش تصحیح پس‌زمینه (MAS.5 و RMA.2)، دو روش نرمال‌سازی (Scaling normalization و Quantile normalization) و دو روش چکیده‌سازی (Tukey biweight و Medianpolish) بر تعداد ژن‌های با بیان متفاوت در یک آزمایش ریزآرایه بررسی شد. نتایج نشان داد که روش چکیده‌سازی تأثیر زیادی بر نتایج آنالیز ندارد، اما روش‌های تصحیح پس‌زمینه و نرمال‌سازی تأثیر قابل توجهی بر نتایج آنالیز دارند. به بیان دیگر، تعداد و نوع ژن‌های با بیان متفاوت در روش‌های مختلف چکیده‌سازی تفاوت زیادی ندارند؛ اما با تغییر در روش تصحیح پس‌زمینه یا روش نرمال‌سازی هم تعداد و هم نوع ژن‌های با بیان متفاوت تغییر زیادی می‌کند. در یک پژوهش نشان داده شد که روش‌های پیش‌پردازش به ویژه نرمال‌سازی و تصحیح پس‌زمینه بر تعداد و نوع ژن‌های با بیان متفاوت تأثیر قابل توجهی دارند (Ray *et al*., 2008). چکیده‌سازی فقط از مقادیر بیان کاوشگرهای مربوط به هر ژن برای محاسبه میزان بیان همان ژن استفاده می‌کند و از مقادیر بیان دیگر ژن‌ها استفاده نمی‌کند، بنابراین تأثیر کمتری بر پیش‌پردازش داده‌ها دارد. روش‌های نرمال‌سازی از بیان مربوط به همه‌ی ژن‌ها برای نرمال کردن همه‌ی داده‌ها استفاده می‌کند و می‌تواند تأثیر بیشتری بر ساختار داده‌ها و در نهایت بر نتایج حاصل از آنالیزها داشته باشد. این یافته‌ها توسط پژوهش‌گران دیگری نیز گزارش شده است (Irizarry & Jaffee, 2006; Li & Hung Wong, 2001; Ray *et al*.,2008). به عنوان مثال، Irizarry *et al* (2006) نشان دادند که تصحیح پس‌زمینه بیشترین تأثیر را بر نتایج آنالیز داده‌های ریزآرایه دارد (Li & Hung Wong, 2001). در یک پژوهش دیگر نشان داده شد که تنها هنگامی نتایج آنالیز داده‌های ریزآرایه، همبستگی قابل قبولی با یکدیگر دارند که روش نرمال‌سازی در آنها یکسان باشد; به عبارت دیگر روش نرمال‌سازی تأثیر بسیار زیادی بر نتایج آنالیز دارد (Irizarry & Jaffee, 2006). این موضوع به ویژه در هنگام مقایسه نتایج پژوهش‌های یکسان انجام شده توسط افراد مختلف بسیار با اهمیت است. روش‌های پیش‌پردازش از نظر معیارهای دیگری مانند حجم و سرعت محاسبات نیز با یکدیگر تفاوت دارند (Bolstad *et al*., 2003; Schwender & Belousov, 2013; Shakya *et al*., 2010). داده‌های ریزآرایه پس از پیش‌پردازش باید آنالیز شوند تا ژن‌های با بیان متفاوت شناسایی شوند. روش‌های آمار پارامتری و ناپارامتری متفاوتی برای آنالیز داده‌ها ریزآرایه وجود دارد که هر کدام نتایج متفاوتی در پی دارند (Khondoker *et al*., 2007; Miklos, & Maleszka, 2004). بنابراین، تعداد و نوع ژن‌های با بیان متفاوت علاوه بر پیش‌پردازش، می‌توانند تحت تأثیر روش آنالیز نیز قرار گیرند (Draghici *et al*., 2006; Ioannidis, 2005a). از سازه‌های دیگری که می‌تواند بر نتایج آنالیز داده‌های

ریزآرایه تأثیر زیادی داشته باشد، اندازه نمونه است; برای اینکه نتایج داده‌های ریزآرایه اعتبار کافی داشته باشند باید تعداد نمونه‌ها زیاد باشد (Ioannidis, 2005a). گزارش شده است که برای شناسایی ژن‌های مهم مؤثر بر یک بیماری با وراثت پیچیده[1] باید تعداد نمونه‌ها چند صد برابر تعدادی باشد که اکنون استفاده می‌شود (Ioannidis, 2003; Ein-Dor *et al*., 2006). همه‌ی این عوامل باعث شده است تا نتایج حاصل از آنالیز ریزآرایه‌ها با چالش تکرارپذیر نبودن روبرو باشند. اکنون چالش تکرارپذیر نبودن نتایج ریزآرایه‌ها به یک موضوع پرمجادله بین دانشمندان رشته‌ی بیوانفورماتیک تبدیل شده و مقاله‌های زیادی درباره‌ی چالش تکرارپذیری نتایج حاصل از آنالیز ریزآرایه‌ها منتشر شده است (Draghici *et al*., 2006; Ioannidis, 2005a; Ioannidis, 2003; Ioannidis, 2005)

## نتیجه‌گیری

فنآوری ریزآرایه روشی قدرتمند برای اندازه‌گیری هم‌زمان بیان شمار زیادی ژن است. داده‌های ریزآرایه، باید پیش از آنالیز پیش‌پردازش شوند تا تغییرات ناخواسته حذف شوند و نتایج آنالیز صحت لازم را داشته باشند. پیش‌پردازش داده‌های ریزآرایه سه گامه تصحیح پس‌زمینه، نرمال‌سازی و چکیده‌سازی دارد، که هر کدام می‌توانند به چندین روش انجام شوند و روش استفاده شده تأثیر قابل توجهی بر نتایج آنالیز آماری دارد. بر اساس نتایج این آزمایش می‌توان استنباط کرد که روش چکیده‌سازی تأثیر زیادی بر نتایج آنالیز ندارد; با این حال، روش‌های تصحیح پس‌زمینه و نرمال‌سازی تأثیر قابل توجهی بر نتایج آنالیز دارند و در انتخاب آنها باید دقت زیادی شود تا یک روش مناسب با هدف پژوهش انتخاب شود.

## منابع

**Affymetrix (2001).** Statistical algorithms reference guide, Technical report, Affymetrix.

**Benjamini, Y. & Hochberg, Y. (1995).** Controlling the false discovery rate: A practical and powerful approach to multiple testing. Journal of the Royal Statistical Society: Series B 57: 289-300.

**Bolstad, B.M. (2001).** Probe level quantile normalization of high density oligonucleotide array data. Division of Biostatistics, University of California, Berkeley.

**Bolstad, B.M., Collin, F., Brettschneider, J., Simpson, K., Cope, L., Irizarry, R.A, & Speed, T.P. (2005).** Quality Assessment of Affymetrix GeneChip Data. In: R, Gentleman, V, Carey, W, Huber, R, Irizarry & S, Dudoit (Eds.). Bioinformatics and Computational Biology Solutions Using R and Bioconductor (pp. 33-48). Springer, New York.

**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 19: 185-193.

[1] Complex genetic disease

**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 10: 61.

**Draghici, S., Khatri, P., Eklund, A.C. & Szallasi, Z. (2006).** Reliability and reproducibility issues in DNA microarray measurements. Trends in Genetics 22:101–109.

**Ein-Dor, L., Zuk, O. & Domany, E. (2006).** Thousands of samples are needed to generate a robust gene list for predicting outcome in cancer. In Proc. National Academy of Sciences of the United States of America 103: 5923–5928.

**Freudenberg, J.M. (2005).** Comparison of background correction and normalization procedures for high-density oligonucleotide microarrays. Technical Report 3, Leipzig Bioinformatics Working Paper.

**Hoffmann, R., Seidl, T. & Dugas, M. (2002).** Profound effect of normalization on detection of differentially expressed genes in oligonucleotide microarray data analysis. Genome Biology 22:789–794.

**Ioannidis, J.P.A. (2003)**. Genetic associations: False or true? Trends in Molecular Medicine 9: 135–138.

**Ioannidis, J.P.A. (2005b).** Why most published research findings are false. PLoS Medicine, 2: e124.

**Ioannidis, J.P.A. (2005a):** Microarrays and molecular research: Noise discovery? The Lancet 365: 454–455.

**Irizarry, R.A. (2003).** Summaries of Affymetrix GeneChip probe level data. Nucleic Acids Research. 31: 15e

**Irizarry, R.A., Wu, Z. & Jaffee, H.A. (2006).** Comparison of Affymetrix GeneChip expression measures. Bioinformatics 22:789–794.

**Khondoker, M.R., Glasbey, C.A. & Worton, B.J. (2007).** A comparison of parametric and nonparametric methods for normalising cDNA microarray data. Biometrical Journal 49: 815– 823.

**Li, C. & Hung Wong, W. (2001).** Model-based analysis of oligonucleotide arrays: model validation, design issues and standard error application. Genome Biology 2: 8.

**Miklos, G.L.G. & Maleszka, R. (2004).** Microarray reality checks in the context of a complex disease. Nature Biotechnology 22: 615–621.

**Morris, D.G., Waters S.M, McCarthy, S.D., Patton, J., Earley, B., Fitzpatrick, R., Murphy, J.J., Diskin, M.G., Kenny, D.A., Brass, A. & Wathes, D.C. (2009).** Pleiotropic effects of negative energy balance in the postpartum dairy cow on splenic gene expression: repercussions for innate and adaptive immunity. Physiological Genomics 39:28-37.

**Ray, M., Freudenberg, J. & Zhang, W. (2008).** A comprehensive analysis of the effect of microarray data preprocessing methods on differentially expressed transcript selection. In: P, Stafford (Ed.), Methods in Microarray Normalization (pp.1-18). CRC Press, Boca Raton.

**Ritchie, M.E., Phipson, B., Wu, D., Hu, Y., Law, C.W., Shi, W. & Smyth, G.K. (2015).** *Limma* powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Research 43:e47.

**Schwender, H. & Belousov, A. (2013).** Comparison of Preprocessing Methods for Affymetrix Microarrays. CHANCE, 19: 15-20.

**Shakya, K., Ruskin, H.J., Kerr, G., Crane, M. & Becker, J. (2010).** Comparison of microarray preprocessing methods. Advances in Computational Biology: 680: 139-147

**Smyth, G.K., Gentleman, R., Carey, V., Dudoit, S., Irizarry, R. & Huber, W. (2005).** Limma: linear models for microarray data. In: R, Gentleman, V, Carey, W, Huber, R, Irizarry & S, Dudoit (Eds.). Bioinformatics and Computational Biology Solutions Using R and Bioconductor (pp. 397-420). Springer, New York.

**Tarca, A.L., Romero, R. & Draghici, S. (2006).** Analysis of microarray experiments of gene expression profiling. American Journal of Obstetrics Gynecology 195: 373–388.

**Wu, Z. (2009).** A review of statistical methods for preprocessing oligonucleotide microarrays. Statistical Methods in Medical Research 18: 533-541.