# به نام خدا

تمرین سری سوم
درس مقدمه ای بیوانفورماتیک
دکتر علی شریفی زارچی

فرزان رحمانی
۴۰۳۲۱۰۷۲۵

## سوال اول

الف) تفاوت بین رونویسی در پروکاریوت ها و یوکاریوت ها

۱. مکانیسم های اساسی:

- پروکاریوت ها: رونویسی و ترجمه به طور همزمان در سیتوپلاسم اتفاق می افتد. یک RNA پلیمراز منفرد رونویسی را انجام می دهد.
- یوکاریوت ها: رونویسی در هسته، و ترجمه در سیتوپلاسم اتفاق می افتد. سه RNA پلیمراز مجزا (I، II، III) برای انواع مختلف RNA وجود دارد.

۲. ساختار ژن:

- پروکاریوت‌ها: ژن‌ها اغلب به اپرون(operons) سازمان‌دهی می‌شوند و به ژن‌های متعدد اجازه می‌دهند تا به عنوان یک mRNA واحد (پلی سیسترونیک) رونویسی شوند.
- یوکاریوت ها: ژن ها به طور معمول مونوسیسترونیک (monocistronic) هستند که به صورت جداگانه با اینترون ها و اگزون ها رونویسی می شوند.

۳. پیچیدگی Regulation:

- پروکاریوت ها: Regulation عمدتاً در سطح رونویسی با استفاده از سرکوب کننده ها و فعال کننده های متصل به اپراتورها است.
- یوکاریوت ها: Regulation پیچیده تر است و شامل بازسازی کروماتین، تقویت کننده ها، خاموش کننده ها و فاکتورهای رونویسی می شود.

۴. تأثیر بر تنظیم ژن(Gene Regulation):

- پروکاریوت ها: تنظیم ساده تر اجازه می دهد تا به تغییرات محیطی واکنش سریع نشان دهد.
- یوکاریوت ها: تنظیم پیچیده، کنترل دقیق، بافت خاص و رشد بیان ژن را امکان پذیر می کند.

ب) اپی ژنتیک و تأثیر آن بر رونویسی ژن

۱. متیلاسیون DNA:

- متیلاسیون باقیمانده‌های سیتوزین (معمولاً در جزایر CpG) با مسدود کردن اتصال فاکتور رونویسی یا جذب پروتئین‌های سرکوب‌کننده، رونویسی را سرکوب می‌کند.
- مثال: Hypermethylation ژن های سرکوبگر تومور می تواند منجر به سرطان شود.

۲. اصلاحات هیستون (Histone Modifications):

- استیلاسیون (Acetylation): افزودن گروه های استیل به هیستون ها ساختار کروماتین را شل می کند و دسترسی به ژن و رونویسی را افزایش می دهد.
- متیلاسیون (Methylation): بسته به مکان و زمینه می تواند رونویسی را فعال یا سرکوب کند.
- فسفوریلاسیون(Phosphorylation) و یوبی کوئیتیناسیون (ubiquitination): سایر تغییرات موثر بر دینامیک کروماتین.

۳. نقش در Regulation:

- تغییرات اپی ژنتیک هویت سلول را ایجاد می کند و به سیگنال های محیطی و رشدی پاسخ های پویا می دهد.

پ) نقش فاکتور های رونویسی اختصاصی

۱. عملکرد در آغاز:

- فاکتورهای رونویسی خاص به توالی‌های تنظیم‌کننده DNA (پیش‌گیرنده، تقویت‌کننده یا همان promoters, enhancers) متصل می‌شوند و ماشین‌های رونویسی عمومی را به کار می‌گیرند.
- آنها به RNA پلیمراز در تشکیل کمپلکس شروع رونویسی در پروموتر کمک می کنند.

۲. تنظیم بیان ژن:

- عمل به عنوان فعال کننده یا سرکوب کننده:
  - فعال‌کننده‌ها (Activators): رونویسی را با به‌کارگیری فعال‌کننده‌های فعال یا تثبیت اتصال RNA پلیمراز تقویت می‌کنند.
  - سرکوبگرها (Repressors): رونویسی را با مسدود کردن اتصال فعال‌کننده یا به‌کارگیری corepressors مهار می‌کنند.

۳. مثال: فاکتور رونویسی p53 به عنوان یک سرکوب کننده تومور عمل می کند و ژن های دخیل در توقف چرخه سلولی و آپوپتوز (apoptosis) را تنظیم می کند.

ت) مهارکننده های رونویسی به عنوان داروهای ضد سرطان

۱. مکانیسم عمل:

- مهارکننده ها فرآیندهای رونویسی را برای سرکوب بیان oncogenes یا القای آپوپتوز (apoptosis) در سلول های سرطانی هدف قرار می دهند.

۲. مثال: Actinomycin D (Dactinomycin)

- در کمپلکس شروع رونویسی به DNA متصل می شود و از طویل شدن توسط RNA پلیمراز جلوگیری می کند.
- در درمان تومور ویلمز (Wilms' tumor)، rhabdomyosarcoma و سایر سرطان ها استفاده می شود.

۳. منطق و نقش آنها:

- سلول های سرطانی برای تکثیر سریع به شدت به برنامه های رونویسی خاصی متکی هستند و رونویسی را به یک هدف قابل دوام تبدیل می کنند.

ث) مفهوم Alternative Splicing

۱. تعریف:

- Alternative splicing یک فرآیند پس از رونویسی است که در آن یک pre-mRNA منفرد به روش های مختلف برای تولید ایزوفرم های مختلف mRNA بالغ متصل می شود.

۲. مکانیسم ها:

- شامل پرش اگزون، مکان‌های پیوند جایگزین '5 یا '3 (alternative 5' or 3' splice sites)، حفظ اینترون و اگزون‌های متقابل منحصر به فرد (mutually exclusive exons).

۳. تأثیر بر تنوع پروتئین:

- چندین پروتئین را از یک ژن تولید می کند و به تنوع عملکردی بدون افزایش اندازه ژنوم کمک می کند.
- مثال: ژن انسانی tropomyosin می تواند ایزوفرم های خاص بافت با عملکردهای متمایز تولید کند.

۴. اهمیت بیولوژیکی:

- برای توسعه، تمایز، و سازگاری ضروری است.
- خطا در alternative splicing می تواند منجر به بیماری هایی مانند سرطان یا اختلالات عصبی (neurodegenerative) شود.

## سوال دوم

الف) تجزیه و تحلیل داده های ریزآرایه و کاربرد آن

۱. تعریف:

تجزیه و تحلیل داده های ریزآرایه شامل پردازش و تفسیر داده های آزمایش های ریزآرایه است که سطح بیان هزاران ژن را به طور همزمان اندازه گیری می کند.

۲. کاربرد در بیوانفورماتیک:

برای مطالعه بیان ژن، شناسایی نشانگرهای زیستی (biomarkers)، درک مکانیسم های بیماری و طبقه بندی بیماری ها استفاده می شود.

۳. اندازه گیری بیان ژن:

ریزآرایه ها از نمونه های نشاندار cDNA یا RNA استفاده می کنند که با پروب های مکمل روی یک تراشه هیبرید می شوند. سیگنال‌های فلورسنت ساطع شده توسط پروب‌های متصل، میزان بیان ژن‌های مربوطه را کمیت بندی میکند.

ب) طراحی یک آزمایش ریزآرایه

۱. اصول اساسی:

- تعریف اهداف (Define objectives): ژن ها، شرایط یا مسیرهای مورد علاقه را شناسایی کنیم.
- کنترل ها و تکرارهای مناسب را برای اطمینان از قابلیت اطمینان آماری انتخاب کنیم.
- تنوع فنی و بیولوژیکی را در نظر بگیریم.

۲. اهمیت انتخاب نمونه (Sample Selection):

- relevance را تضمین می کند و عوامل مخدوش کننده را به حداقل می رساند.
- نمایش دقیق جمعیت یا شرایط مورد مطالعه، کیفیت و قابلیت تفسیر داده ها را افزایش می دهد.

پ) اصول و اهمیت پروب ها

۱. عملکرد پروب ها:

توالی های کوتاه DNA یا RNA روی تراشه ریزآرایه با توالی های هدف مکمل در نمونه hybridize می شوند.

۲. اهمیت انتخاب پروب:

- Specificity: از cross-hybridization برای تشخیص دقیق ژن اجتناب کنیم.
- Sensitivity: پروب ها باید transcript های کم فراوانی را تشخیص دهند.
- Coverage: همه ژن ها یا ایزوفرم های مورد نظر را نشان می دهد.

ت) مراحل تجزیه و تحلیل داده های ریزآرایه

۱. پیش پردازش (Preprocessing):

تصحیح پس زمینه، کاهش نویز، و افزایش سیگنال.

۲. نرمال سازی یا Normalization:

تغییرات فنی را برای اطمینان از مقایسه بین نمونه ها تنظیم کنیم.

۳. تجزیه و تحلیل آماری:

با استفاده از آزمون‌های آماری، ژن‌های با بیان متفاوت (DEGs) را شناسایی کنیم.

۴. تحلیل عملکردی:

DEG ها را به مسیرهای بیولوژیکی یا هستی شناسی ژن نگاشت کنیم.

۵. تفسیر داده ها (Data Interpretation):

ادغام یافته ها با دانش بیولوژیکی برای به دست آوردن نتایج معنادار.

ث) آرایه های یک رنگ در مقابل آرایه های دو رنگ

۱. Single-Color Array:

- از یک برچسب فلورسنت در هر نمونه استفاده می کند.
- مزیت: طراحی آزمایشی ساده تر.
- نقطه ضعف: به آرایه های بیشتری برای مقایسه نیاز دارد.

۲. Dual-Color Array:

- از دو برچسب فلورسنت برای مقایسه دو نمونه در یک آرایه استفاده می کند.
- مزیت: مقایسه مستقیم، تنوع فنی(technical variability) را کاهش می دهد.
- عیب: سوگیری بالقوه رنگ (Potential dye bias).

ج) مزایای ریزآرایه نسبت به روشهای سنتی

۱. توان عملیاتی بالا (High Throughput):

هزاران ژن را به طور همزمان تجزیه و تحلیل کنیم.

۲. کمی (Quantitative):

سطوح بیان نسبی (relative expression levels) را فراهم می کند.

۳. مقرون به صرفه (Cost-Effective):

ارزان تر از sequencing برای مطالعات در مقیاس بزرگ.

۴. همه کاره:

قابل استفاده برای موجودات و شرایط مختلف.

چ) اطلاعات قابل استخراج و ارتباط آنها با عملکرد ژن ها

۱. اطلاعات قابل استخراج:

- پروفایل های بیان ژن (Gene expression profiles)
- ژن های بیان شده متفاوت (Differentially expressed genes)
- شبکه های هم بیان ژن. (Gene co-expression networks)

۲. Relevance:

- نشانگرهای زیستی را شناسایی کنیم، عملکرد ژن را درک کنیم و مسیرهای بیولوژیکی(biological pathways) را مشخص کنیم.

ح) کاهش نویز پس زمینه

روش ها:

- پردازش تصویر: برای فلورسانس غیر اختصاصی (non-specific fluorescence) تنظیم کنیم.
- فیلتر کردن پروب: پروب های بی کیفیت یا نویزی را برداریم.
- تکنیک های عادی سازی: technical variability را کاهش دهیم.
- از positive control و negative control استفاده کنیم.

خ) نرمال سازی در تجزیه و تحلیل داده های ریزآرایه

۱. ضرورت:

تغییرات در آماده سازی نمونه (variations in sample preparation)، hybridization و scanning را تصحیح می کند.

۲. روش ها:

- Global Normalization: سیگنال ها به یک میانگین یا میانه یکنواخت مقیاس میکند.
- Quantile Normalization: توزیع مقادیر expression ها را در میان آرایه ها تراز(align) می کند.

د) شناسایی ژن های بیان شده متفاوت (Differentially Expressed Genes)

۱. مراحل:

- پیش پردازش و نرمال سازی داده ها.
- انجام تست های آماری (به عنوان مثال، آزمون t، ANOVA).
- اصلاحات آزمایشی متعدد (multiple testing corrections) اعمال کنیم (به عنوان مثال، FDR).
- رتبه بندی ژن ها بر اساس fold change و اهمیت آنها و تحلیل p - value.

۲. خروجی:

- فهرست ژن هایی که در شرایط خاص، تغییر بیان یافته اند.

ذ) محدودیت های ریزآرایه در مقابل RNA-Seq

۱. محدودیت ها:

- طراحی پروب استاتیک (Static Probe Design): فقط توالی های از پیش تعریف شده را تشخیص می دهد.
- Limited Sensitivity: با transcriptsهای دارای فراوانی کم دست و پنجه نرم می کند.
- محدوده دینامیکی پایین (Lower Dynamic Range): دقت کمتر در تعیین کمیت سطوح بیان ( Less precise quantification of expression levels).

۲. مزایای RNA-Seq:

- تشخیص Unbiased رونوشت ها (transcripts).
- شناسایی رونوشت های novel و رویدادهای alternative splicing.
- sensitivity و محدوده دینامیکی بیشتر

## سوال سوم

الف) آزمون آماری مورد استفاده

برای بررسی تفاوت بین دو گروه از آزمون t مستقل (independent t-test) استفاده می شود. independent t-test مناسب است زیرا:

- داده ها از دو گروه مستقل (افراد سالم و بیماران) تشکیل شده است.
- هدف مقایسه میانگین سطوح بیان GLUT4 بین این گروه ها است.

ب) فرضیه ها

- فرض صفر ($H_0$: Null Hypothesis): تفاوت معنی داری در سطح بیان ژن GLUT4 بین افراد سالم و بیماران مبتلا به دیابت نوع ۲ وجود ندارد.
- فرض مقابل ($H_1$: Alternative Hypothesis): تفاوت معنی داری در سطح بیان ژن GLUT4 بین افراد سالم و بیماران مبتلا به دیابت نوع ۲ وجود دارد.

ج) نتایج آزمون تی (t-test)

- آماره t (T-statistic): $t = 12.61$
- p-value: $p = 9.998 \times 10^{-10}$

در ادامه راه حل و توضیح گام به گام نحوه محاسبه آماره t و p-value برای مقایسه سطح بیان ژن GLUT4 بین افراد سالم و بیماران آورده شده است:

مرحله ۱: داده ها را شناسایی کنیم.

سطوح بیان برای دو گروه عبارتند از:

- نمونه های سالم: [12.3, 12.2, 11.7, 12.0, 12.4, 11.9, 12.1, 12.5, 11.8, 12.3]
- نمونه های بیمار: [11.0, 10.7, 10.8, 11.1, 10.9, 10.6, 10.7, 11.0, 10.8, 10.5]

مرحله ۲: فرمول T-Test را تعریف میکنیم.

فرمول آماره t (t-statistic) در آزمون t مستقل (independent t-test) به صورت زیر است:

$$t = \frac{\overline{X_1} - \overline{X_2}}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

که:

- $\overline{X_1}, \overline{X_2}$ : میانگین دو گروه اند.
- $s_1^2, s_2^2$ : واریانس نمونه دو گروه اند.
- $n_1, n_2$ : اندازه نمونه از دو گروه اند.

مرحله ۳: محاسبه میانگین گروه

$$\overline{X_1} = \frac{\text{Sum of healthy samples}}{n_1} = \frac{12.3 + 12.2 + \cdots + 12.3}{10} = 12.12$$

$$\overline{X_2} = \frac{\text{Sum of patient samples}}{n_2} = \frac{11.0 + 10.7 + \cdots + 10.5}{10} = 10.81$$

مرحله ۴: واریانس ها را محاسبه می کنیم.

واریانس به صورت زیر محاسبه می شود:

$$s^2 = \frac{\sum(X - \bar{X})^2}{n - 1}$$

برای نمونه های سالم ($s_1^2$):

$$s_1^2 = \frac{(12.3 - 12.12)^2 + (12.2 - 12.12)^2 + \cdots + (12.3 - 12.12)^2}{10 - 1} = \frac{0.636}{9} = 0.071$$

برای نمونه های بیمار ($s_2^2$):

$$s_2^2 = \frac{(11.0 - 10.81)^2 + (10.7 - 10.81)^2 + \cdots + (10.5 - 10.81)^2}{10 - 1} = \frac{0.329}{9} = 0.037$$

مرحله ۵: مقادیر را در فرمول قرار میدهیم.

$$t = \frac{12.12 - 10.81}{\sqrt{\frac{0.071}{10} + \frac{0.037}{10}}}$$

$$t = \frac{1.31}{\sqrt{0.0071 + 0.0037}} = \frac{1.31}{\sqrt{0.0108}} = \frac{1.31}{0.1039} = 12.61$$

مرحله ۶: محاسبه درجه آزادی

با استفاده از فرمول آزمون Welch t برای درجات آزادی (Welch's t-test formula for degrees of freedom):

$$df = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{\left(\frac{s_1^2}{n_1}\right)^2}{n_1 - 1} + \frac{\left(\frac{s_2^2}{n_2}\right)^2}{n_2 - 1}}$$

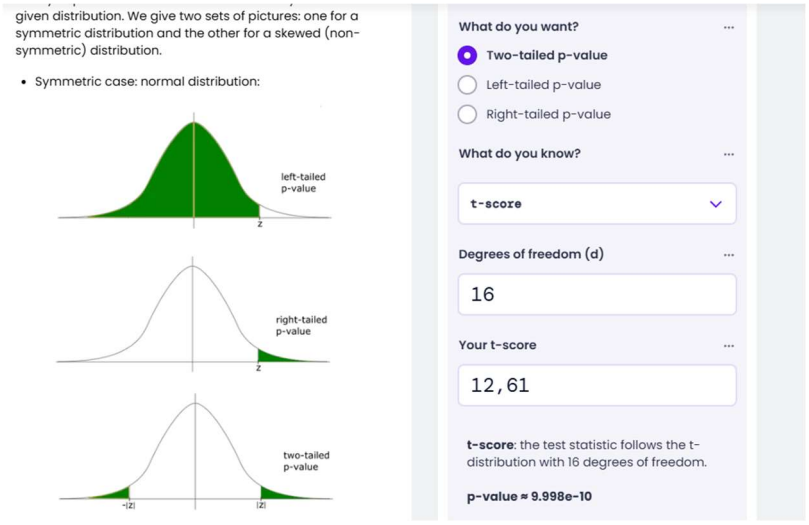
$$df = \frac{\left(\frac{0.071}{10} + \frac{0.037}{10}\right)^2}{\frac{\left(\frac{0.071}{10}\right)^2}{9} + \frac{\left(\frac{0.037}{10}\right)^2}{9}} = 16.377 \text{ (approximately 16)}$$

مرحله ۷: P-Value را پیدا کنیم.

با استفاده از یک جدول توزیع t یا نرم افزار آماری یا کد پایتون یا یک وب سایت مانند https://www.omnicalculator.com/statistics/p-value، مقدار p برای موارد زیر محاسبه می شود (توجه: مقدار p-value به صورت two-tailed محاسبه شده است که دو برابر one-tailed یا همان right-tailed است):

- $t = 12.61$
- $Degrees\ of\ freedom = 16$

همان طور که میبینیم مقدار p-value بسیار کوچک است ($p = 9.998 \times 10^{-10}$).




```python
from scipy.stats import t

# Given values
t_statistic = 12.61
df = 16  # Degrees of freedom

# Calculate the two-tailed p-value
p_value = 2 * t.sf(abs(t_statistic), df)
# p_value_right_tailed = t.sf(abs(t_statistic), df)

print(f"P-value: {p_value:.10e}")
```

```
P-value: 9.9984873396e-10
```

نتایج نهایی

- آماره t (T-statistic): $t = 12.61$
- p-value: $p = 9.998 \times 10^{-1}$

د) تصمیم گیری

مقدار p-value ( $9.998 \times 10^{-1}$) بسیار کوچکتر از سطح معنی‌داری مانند $\alpha = 0.05$ است. بنابراین:

- فرض صفر ($H_0$: Null Hypothesis) رد می شود.
- فرض مقابل ($H_1$: Alternative Hypothesis) درست است و در نتیجه تفاوت معنی داری در سطح بیان ژن GLUT4 بین افراد سالم و بیماران مبتلا به دیابت نوع ۲ وجود دارد.

$$9.998 \times 10^{-10} = p < \alpha = 0.05$$

هـ) نتیجه گیری

بر اساس نتایج، تفاوت آماری معنی داری در سطح بیان ژن GLUT4 بین افراد سالم و بیماران مبتلا به دیابت نوع ۲ وجود دارد چرا که مقدار p-value حاصل از t-test بسیار کوچکتر از سطح معنی‌داری مانند $\alpha = 0.05$ است. این امر نشان می دهد که دیابت نوع ۲ تأثیر قابل توجهی بر کاهش سطح بیان ژن GLUT4 دارد. به بیان دیگر چون مقدار p-value بسیار کوچکتر از سطح معنی‌داری مانند $\alpha = 0.05$ است، پس این اتفاق نمی تواند شانسی رویداده باشد و فرض مقابل ($H_1$: Alternative Hypothesis) درست است یعنی تفاوت معنی داری در سطح بیان ژن GLUT4 بین افراد سالم و بیماران مبتلا به دیابت نوع ۲ وجود دارد.

## سوال چهارم

الف) محاسبه نسبت شانس (Odds Ratio)

تعریف نسبت شانس (OR)

نسبت شانس، شانس وقوع یک رویداد در یک گروه را نسبت به گروه دیگر مقایسه می کند. برای این مطالعه:

$$\text{Odds Ratio} = \frac{\text{Odds in Aspirin group}}{\text{Odds in Ibuprofen group}}$$

محاسبه گام به گام

- شانس در گروه آسپرین:

$$\text{Odds}_{\text{Aspirin}} = \frac{\text{Number of Heart Attacks}}{\text{Number of No Heart Attacks}} = \frac{80}{220} = 0.3636$$

- شانس در گروه ایبوپروفن:

$$\text{Odds}_{\text{Ibuprofen}} = \frac{\text{Number of Heart Attacks}}{\text{Number of No Heart Attacks}} = \frac{120}{180} = 0.6667$$

- نسبت شانس (OR):

$$\text{OR} = \frac{\text{Odds}_{\text{Aspirin}}}{\text{Odds}_{\text{Ibuprofen}}} = \frac{0.3636}{0.6667} = 0.5455$$

شانس متقابل(reciprocal) یا معکوس آن هم برابر مقدار زیر است:

$$\text{OR}_{\text{Ibuprofen vs Aspirin}} = \frac{1}{0.5455} = 1.83$$

بنابراین، افرادی که ایبوپروفن مصرف می کنند، ۱/۸۳ برابر بیشتر از افرادی که آسپرین مصرف می کنند، دچار حمله قلبی می شوند. به بیان دیگر، افرادی که آسپرین مصرف می کنند، ۰/۵۴۵۵ برابر کمتر از افرادی که ایبوپروفن مصرف می کنند، دچار حمله قلبی می شوند.

تفسیر مفصل تر نسبت شانس ها:

شانس در هر گروه

- شانس در گروه آسپرین: ۰/۳۶۳۶
  - به این معنی که به ازای هر ۱ عدم حمله قلبی، تقریباً ۰/۳۶۳۶ حمله قلبی در گروه آسپرین وجود دارد.
  - متناوباً، می توانید آن را تقریباً ۱ حمله قلبی در هر ۲/۷۵ عدم حمله قلبی تفسیر کنید.
- شانس در گروه ایبوپروفن: ۰/۶۶۶۷
  - به این معنی که به ازای هر ۱ عدم حمله قلبی، تقریباً ۰/۶۶۶۷ حمله قلبی در گروه ایبوپروفن وجود دارد.
  - متناوباً، نشان دهنده ۱ حمله قلبی در هر ۱/۵ عدم حمله قلبی است.

نسبت شانس (OR)

- نسبت شانس برای آسپرین در مقابل ایبوپروفن: ۰/۵۴۵۵
  - احتمال حمله قلبی در گروه آسپرین ۰/۵۴۵۵ برابر کمتر از گروه ایبوپروفن است.
  - این نشان می دهد که آسپرین با کاهش خطر حمله قلبی در مقایسه با ایبوپروفن همراه است.

نسبت شانس متقابل (Reciprocal Odds Ratio)

- نسبت شانس برای ایبوپروفن در مقابل آسپرین: ۱/۸۳
  - احتمال حمله قلبی در گروه ایبوپروفن ۱/۸۳ برابر بیشتر از گروه آسپرین است.

این نسبت ها به چه معناست؟

۱. اثربخشی داروها:

- به نظر می رسد آسپرین در کاهش خطر حملات قلبی در مقایسه با ایبوپروفن موثرتر باشد.

۲. پیامدهای بالینی:

- نسبت شانس کمتر (OR < 1) برای آسپرین نشان دهنده اثر محافظتی بالقوه آن است.
- نسبت معکوس (OR > 1) برای ایبوپروفن نشان می دهد که ممکن است خطر بیشتری نسبت به آسپرین داشته باشد.

۳. تصمیم گیری:

- این یافته‌ها می‌توانند بر تصمیم‌گیری‌های درمانی تأثیر بگذارند و از آسپرین به جای ایبوپروفن برای پیشگیری از حمله قلبی، با اعتبارسنجی آماری بیشتر (مثلاً p-value) تأثیر بگذارند.

ب) فرضیه ها و نتیجه محاسبات

- فرضیه صفر (Null Hypothesis: $H_0$): هیچ تفاوتی در خطر حمله قلبی بین دو دارو وجود ندارد.

- فرضیه جایگزین (Alternative Hypothesis: $H_1$): آسپرین در مقایسه با ایبوپروفن خطر حمله قلبی را کاهش می دهد.

از آنجایی که نسبت شانس مخالف صفر و کوچکتر از ۱ است ($OR = 0.5455, OR_{Ibuprofen\ vs\ Aspirin} = 1.83$)، خطر کمتر حمله قلبی را برای کسانی که آسپرین مصرف می کنند نشان می دهد که از فرضیه جایگزین (H1) پشتیبانی می کند. به بیان دیگر فرضیه صفر رد می شود و فرضیه جایگزین تایید میشود البته باید مقدار p-value نیز برای تایید چک شود.

توضیحات مفصل تر در ادامه آمده است:

وقتی فرضیه صفر (H0) درست باشد

- فرضیه صفر (H0) بیان می کند که:

احتمال حمله قلبی برای مصرف کنندگان آسپرین و ایبوپروفن یکسان است

- از نظر ریاضی این یعنی:

$$\text{Odds Ratio (OR)} = 1$$

- اگر $\text{OR} = 1$ باشد، هیچ تفاوتی بین دو گروه از نظر خطر حمله قلبی وجود ندارد.

وقتی $\text{OR} \neq 1$

- $\text{OR} > 1$:
  - احتمال حمله قلبی در گروه آسپرین بیشتر از گروه ایبوپروفن است.
  - این نشان می دهد که آسپرین ممکن است خطر را نسبت به ایبوپروفن افزایش دهد.
- $\text{OR} < 1$ :
  - احتمال حمله قلبی در گروه آسپرین کمتر از گروه ایبوپروفن است.
  - این نشان می دهد که آسپرین خطر حمله قلبی را در مقایسه با ایبوپروفن کاهش می دهد، همانطور که در این مطالعه مشاهده شد ($\text{OR} = 0.5455$).

نکات کلیدی در مورد نسبت شانس (OR)

۱. چگونه OR خطر را منعکس می کند:

- OR = 1: تفاوتی وجود ندارد (H0 درست است).
- OR < 1: گروه اول (آسپرین) شانس کمتری برای نتیجه (حمله قلبی) دارند.
- OR > 1: گروه اول (آسپرین) شانس بیشتری برای نتیجه دارد.

۲. اهمیت آماری:

- حتی اگر $\text{OR} < 1$ یا $\text{OR} > 1$ باشد، باید از نظر آماری معنی‌دار باشد (p-value <0.05) تا $H_0$ با اطمینان رد شود.

توجه به عدم تفسیر نادرست از نسبت شانس

- OR همیشه مثبت است زیرا شانس به صورت مقادیر مثبت بیان می شود.
- $\text{OR} = 0$ نامعتبر است زیرا به این معنی است که شانس یک گروه برای نتیجه صفر است که واقع بینانه نیست.
- یک OR کوچک (به عنوان مثال، نزدیک به ۰/۱) یک اثر محافظتی بسیار قوی از حمله قلبی را نشان می دهد.

در این مطالعه OR = ۰/۵۴۵۵:

- نشان دهنده اثر محافظتی آسپرین در مقایسه با ایبوپروفن است.
- خطر حمله قلبی برای کسانی که آسپرین مصرف می کنند در مقایسه با ایبوپروفن تقریباً ۴۵ درصد کمتر است.

ارتباط با فرضیه ها:

- اگر p-value از آزمایش دقیق فیشر کمتر از ۰/۰۵ باشد، $H_0$ را رد می کنیم و می پذیریم که آسپرین به طور قابل توجهی خطر حمله قلبی را کاهش می دهد.

ج) آزمون دقیق فیشر: آزمون دقیق فیشر یک روش آماری است که برای تعیین اینکه آیا ارتباط معنی‌داری بین دو متغیر طبقه‌بندی شده (categorical variable) در جدول احتمالی(contingency table) وجود دارد یا خیر.

- P-value محاسبه شده: $p = 0.00071$

**ابتدا راجع به خود آزمون، نحوه کارکرد آن و اینکه چرا در این مطالعه استفاده میشود توضیح میدهیم و سپس وارد محاسبات می شویم:**

هدف این آزمون: این آزمون تعیین می کند که آیا ارتباط معنی داری بین دو متغیر طبقه بندی شده در جدول احتمالی وجود دارد یا خیر. احتمال دقیق مشاهده داده های داده شده (یا شدیدتر: or more extreme) را تحت فرضیه صفر محاسبه می کند.

- آزمون دقیق فیشر برای بررسی اهمیت ارتباط بین دو متغیر طبقه‌بندی شده(categorical variables) در جدول احتمالی ۲×۲ استفاده می‌شود.
- برای اندازه های نمونه کوچک یا زمانی که فرکانس های مورد انتظار در جدول پایین است ایده آل است چرا که که باعث می شود تست های دیگری مانند تست Chi-square کمتر قابل اعتماد باشند.
- احتمال دقیق مشاهده داده ها را تحت فرضیه صفر محاسبه می کند.

چگونه کار می کند؟

۱. فرضیه های صفر و جایگزین:

- H0: هیچ ارتباطی بین این دو متغیر وجود ندارد (به عنوان مثال، داروی مصرف شده و احتمال حمله قلبی).
- H1: بین این دو متغیر ارتباط وجود دارد (به عنوان مثال، داروی مصرف شده بر احتمال حمله قلبی تأثیر می گذارد).

۲. محاسبه دقیق احتمال:

- آزمون فیشر احتمال دقیق (p-value) مشاهده داده ها را تحت فرضیه صفر محاسبه می کند.
- از توزیع فوق هندسی(hypergeometric distribution) برای محاسبه احتمال تمام نتایج ممکن به صورت افراطی یا شدیدتر از آنچه مشاهده شده است (extreme or more extreme than the observed one)، با توجه به مجموع سطرها و ستون ها استفاده می کند.

۳. داده های ورودی:

این آزمون از یک جدول احتمالی ۲×۲ استفاده می کند.

۴. فرمول کلیدی:

$$P = \frac{\binom{a+b}{a}\binom{c+d}{c}}{\binom{n}{a+c}}$$

که:

- a، b، c، d: تعداد سلول ها در جدول اند.
- n: تعداد کل مشاهدات است.

به هر جدول ممکن یک احتمال اختصاص داده می شود و آزمون فیشر احتمالات همه جداول را به صورت افراطی یا شدیدتر از جدول مشاهده شده جمع می کند.

چرا از آزمون دقیق فیشر در این مطالعه استفاده می شود؟

۱. حجم نمونه کوچک: حجم کل نمونه (۶۰۰ شرکت کننده) متوسط است، اما برای برخی از پیامدها (به عنوان مثال، ۸۰ حمله قلبی در گروه آسپرین)، فرکانس سلولی نسبتاً کوچک است.

۲. نتایج دقیق: بر خلاف روش‌های تقریبی مانند آزمون Chi-square، آزمون فیشر مقادیر p دقیقی را ارائه می‌کند که وقتی اندازه نمونه یا فرکانس‌های مورد انتظار پایین باشد قابل اعتمادتر هستند.

۳. نتیجه باینری: داده ها شامل متغیرهای باینری هستند (حمله قلبی: بله/خیر) و ب =ه خوبی در جدول ۲×۲ قرار می گیرند و آزمایش فیشر را مناسب می کند.

مزایای تست دقیق فیشر

- دقیق برای نمونه های کوچک: نتایج دقیق را بدون توجه به حجم نمونه ارائه می دهد.
- غیر پارامتریک: هیچ فرضی در مورد توزیع داده ها ندارد.
- مقاوم: ایده آل برای مواردی که فرکانس های مورد انتظار پایین است.

محدودیت ها

- از نظر محاسباتی برای جداول یا مجموعه داده های بزرگ فشرده است (اما قابل مدیریت برای جداول ۲*۲ است).

در این مطالعه، آزمایش فیشر بسیار مفید است زیرا به طور دقیق رابطه بین مصرف آسپرین و ایبوپروفن و خطر حمله قلبی را ارزیابی می‌کند و حتی با توزیع داده‌های مشاهده‌شده نتایج قابل اعتمادی را تضمین می‌کند.

**محاسبات به شرح زیر است:**

۱. فرضیه صفر (H0): هیچ ارتباطی بین نوع دارو (آسپرین یا ایبوپروفن) و خطر حمله قلبی وجود ندارد.

۲. درک معنای "Extreme or More Extreme"

در یک جدول 2*2، extreme به جدول هایی اشاره دارد که در آنها:

- انحراف از توزیع مورد انتظار تحت فرضیه صفر به اندازه یا بیشتر از آنچه مشاهده شد است.

به عنوان مثال:

- داده های مشاهده شده: آسپرین در مقایسه با ایبوپروفن حملات قلبی را کاهش می دهد.
- داده های شدیدتر بیشتر به نفع آسپرین است یا تفاوت آشکارتری بین گروه ها خواهد داشت.

۳. فرمول آزمون دقیق فیشر

احتمال هر پیکربندی جدول ۲*۲ خاص با استفاده از توزیع ابر هندسی محاسبه می شود:

$$P = \frac{\binom{a+b}{a}\binom{c+d}{c}}{\binom{n}{a+c}}$$

- a، b، c، d: مقادیر در جدول اند.
- n: تعداد کل مشاهدات است (n = a+b+c+d).

در صفحه بعد حل گام به گام آمده است.

۴. حل گام به گام

داده های مشاهده شده عبارتند از:

| | Aspirin | Ibuprofen | Total |
|---|---|---|---|
| **No Heart Attack** | a=220 | c=180 | a+c=400 |
| **Heart Attack** | b=80 | d=120 | b+d =200 |
| **Total** | a+b=300 | c+d=300 | n=a+b+c+d=600 |

احتمال مشاهده این جدول خاص:

$$P = \frac{\binom{a+b}{a}\binom{c+d}{c}}{\binom{n}{a+c}}$$

$$P = \frac{\binom{220+80}{220}\binom{180+120}{180}}{\binom{600}{400}}$$

محاسبات فاکتوریل ها:

$$\binom{a+b}{a} = \frac{(300)!}{(220)!\,(80)!}, \quad \binom{c+d}{c} = \frac{(300)!}{(180)!\,(120)!}, \quad \binom{n}{a+b} = \frac{(600)!}{(400)!\,(200)!}$$

جداول Extreme تر:

- آزمون دقیق فیشر احتمالات (P) را برای همه جداول جمع می کند که اختلاف نسبت ها به اندازه جدول مشاهده شده بزرگتر و بزرگتر است.
- جداول افراطی (Extreme tables) مواردی را شامل میشود که $a > 220$ باشد.

فرآیند محاسبه:

- تمام جداول extreme ممکن را با مجموع ردیف ها و ستون های یکسان (marginal sums) برمی شماریم.
- احتمال هر جدول را با استفاده از فرمول هایپرهندسی محاسبه میکنیم.
- احتمالات همه جداول را به صورت extreme یا شدیدتر از جدول مشاهده شده جمع میکنیم.

۴. محاسبه P-Value

مقدار P-Value در این مطالعه (P = 0.00071) با جمع کردن احتمالات جدول مشاهده شده و همه جداول افراطی تر ( more extreme tables) به دست می آید. این محاسبات به دلیل پیچیدگی شمارش تمام جداول ممکن، معمولاً به صورت برنامه نویسی شده ای انجام می شوند. برای محاسبه میتوانیم از کد پایتون یا نرم افزار های آماری یا وب سایت های مختلف استفاده کنیم. همان طور که در زیر دیده میشود:

**Easy Fisher Exact Test Calculator**

Success! The Fisher exact test statistic and statement of significance appear beneath the table. Blue means you're dealing with dependent variables; red, independent.

| Results | | | |
|---|---|---|---|
| | No Heart Attack | Heart Attack | ***Marginal Row Totals*** |
| Aspirin | 220 | 80 | 300 |
| Ibuprofen | 180 | 120 | 300 |
| ***Marginal Column Totals*** | 400 | 200 | 600 (Grand Total) |

The Fisher exact test statistic value is 0.0007. The result is significant at p < .05.

Start Again

```python
from scipy.stats import fisher_exact
import numpy as np

contingency_table = np.array([[80, 220], [120, 180]])

odds_ratio, p_value_fisher = fisher_exact(contingency_table)

odds_ratio, p_value_fisher
```

```
(0.5454545454545454, 0.0007120141417653778)
```

۵. چرا از تست دقیق فیشر استفاده می شود؟

- احتمالات دقیق: بر خلاف تست های تقریبی (مثلاً Chi-square)، تست فیشر مقادیر p دقیق را محاسبه می کند و آن را برای نمونه های کوچک یا تعداد نمونه های کم قابل اعتماد می کند.
- تناسب سناریو: این مطالعه شامل یک جدول ۲×۲ با داده‌های طبقه‌بندی شده و تعداد نسبتاً کوچک در برخی سلول‌ها است (به عنوان مثال، b = 80، d = 120)، که آزمایش دقیق فیشر را مناسب می‌کند.

خلاصه

مقدار p به صورت زیر محاسبه می شود:

۱. محاسبه احتمال جدول مشاهده شده.
۲. جمع کردن احتمالات برای همه جداول به صورت افراطی یا افراطی تر (extreme or more extreme) تحت $H_0$.
۳. استفاده از احتمال مجموع برای ارزیابی اینکه آیا فرضیه صفر باید رد شود یا خیر.

در این مطالعه، مقدار p-value بسیار کوچک (۰/۰۰۰۷۱) و کوچکتر از سطح معنی دار $\alpha = 0.05$ قویاً نشان می دهد که آسپرین در مقایسه با ایبوپروفن خطر حمله قلبی را کاهش می دهد.

د) تجزیه و تحلیل نتایج و فرضیه

- سطح معنی داری مانند $\alpha = 0.05$ در نظر میگیریم.
- مقدار p-value (۰/۰۰۰۷۱) بسیار کوچکتر از $\alpha$ است که منجر به رد فرضیه صفر ($H_0$) می شود. این نشان می دهد که آسپرین به طور قابل توجهی خطر حمله قلبی را در مقایسه با ایبوپروفن کاهش می دهد.
- در واقع فرضیه جایگزین (Alternative Hypothesis: $H_1$) تایید میشود. یعنی این اتفاق شانسی نبوده و افرادی که ایبوپروفن مصرف می کنند، بیشتر از افرادی که آسپرین مصرف می کنند، دچار حمله قلبی می شوند. به بیان دیگر، داروی آسپرین تاثیرگذار بوده و افرادی که آسپرین مصرف می کنند، کمتر از افرادی که ایبوپروفن مصرف می کنند، دچار حمله قلبی می شوند.
- همچنین در بخش های قبل هم دیدیم که با توجه به OR=0.5455 مصرف کنندگان آسپرین در مقایسه با مصرف کنندگان ایبوپروفن شانس کمتری برای حمله قلبی دارند.

نتیجه گیری:

آسپرین در مقایسه با ایبوپروفن از نظر آماری تاثیر معناداری در کاهش خطر حمله قلبی دارد.

## سوال پنجم

الف) تکنیک RNA-Seq و مزایای آن

۱. RNA-Seq چیست؟

- RNA-Seq (توالی یابی RNA یا RNA sequencing) یک تکنیک توالی یابی با توان عملیاتی بالا است که از توالی یابی نسل بعدی (NGS: next-generation sequencing) برای تجزیه و تحلیل transcriptome یا رونوشت (همه مولکول های RNA، از جمله mRNA، ncRNA و غیره) از نمونه استفاده می کند.

۲. چگونه به مطالعه بیان ژن کمک می کند:

- بیان ژن را با اندازه گیری فراوانی رونوشت های RNA تعیین می کند.
- رونوشت‌های جدید، ایزوفرم‌ها و انواع مختلف اسپلایس را شناسایی می‌کند.
- تغییرات پویا در بیان ژن در شرایط یا نقاط زمانی را ثبت می کند.

۳. مزایای نسبت به Microarray:

- تشخیص بی طرفانه (Unbiased Detection): نیازی به دانش قبلی از توالی ندارد.
- Higher Sensitivity and Specificity: رونوشت ها و ایزوفرم های کم فراوانی را تشخیص می دهد.
- Wide Dynamic Range: ژن های با بیان بالا و کم را با دقت بیشتری تعیین می کند.
- کشف رونوشت های جدید (Discovery of Novel Transcripts): ژن های unannotated و رویدادهای splicing را شناسایی می کند.

ب) داده های خام در مقابل داده های پردازش شده در RNA-Seq

۱. داده های خام (Raw Data):

- حاوی توالی خام read ها در قالب FASTQ.
- شامل base call quality scores است.
- مفید برای تجزیه و تحلیل مجدد یا خطوط لوله پردازش جایگزین (alternative processing pipelines).

۲. داده های پردازش شده (Processed Data):

- از داده های خام پس از کنترل کیفیت، هم ترازی و کمی سازی (quality control, alignment, and quantification) به دست می آید.
- شامل تعداد عبارات، داده های نرمال شده، یا رونوشت های حاشیه نویسی شده ( expression counts, normalized data, or annotated transcripts) است.
- آماده برای تفسیر آماری یا بیولوژیکی پایین دستی (downstream statistical or biological interpretation).

۳. اهمیت:

- داده های خام: تکرارپذیری(reproducibility) و انعطاف پذیری (flexibility) در تجزیه و تحلیل را تضمین می کند.
- داده های پردازش شده: بینش و مقایسه سریع را تسهیل می کند.

ج) پایگاه داده SRA

۱. تعریف:

- The Sequence Read Archive (SRA) یک مخزن عمومی است که توسط NCBI مدیریت می شود و داده های توالی یابی خام از مطالعات مختلف از جمله RNA-Seq را ذخیره می کند.

۲. داده های ذخیره شده:

- توالی خام read ها (Raw sequencing reads) در قالب FASTQ.
- Metadata در مورد نمونه ها، روش های توالی یابی و شرایط.

۳. دسترسی به داده های RNA-Seq:

- پایگاه داده SRA را از طریق وب سایت NCBI با کلمات کلیدی، ارگانیسم، نوع مطالعه یا شماره های الحاق ( keywords, organism, study type, or accession numbers) قابل جستجو است.
- داده ها را با استفاده از NCBI SRA Toolkit میتوان دانلود کرد.

د) یافتن یک مطالعه مرتبط با سرطان در پایگاه داده SRA

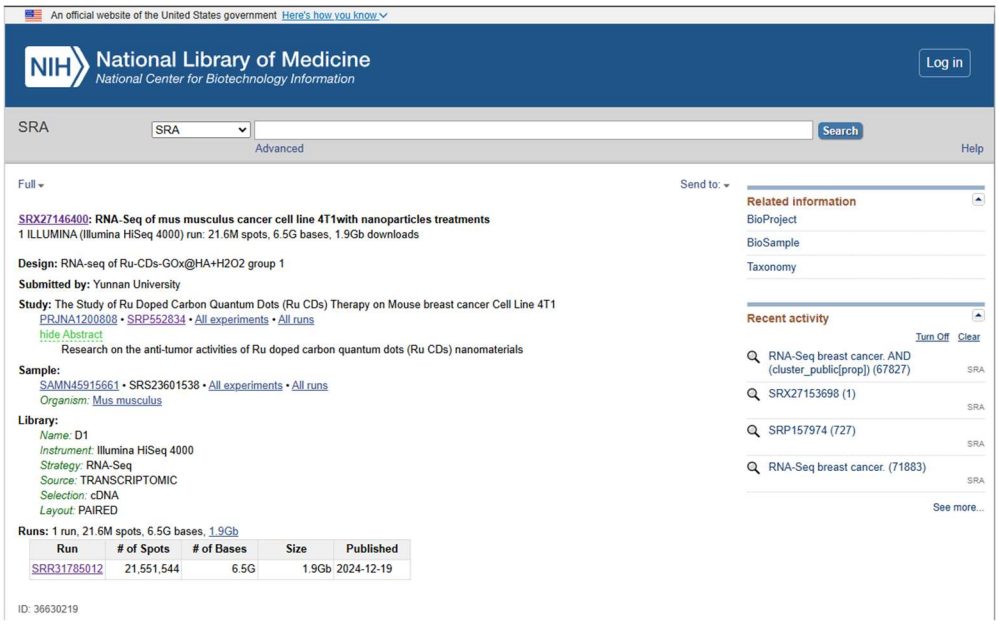
۱. مراحل جستجو:

- از پایگاه داده NCBI SRA ( https://www.ncbi.nlm.nih.gov/sra) دیدن میکنیم.
- جستجوی عباراتی مانند «RNA-Seq breast cancer.» انجام میدهیم.
- نتایج را بر اساس نوع مطالعه، ارگانیسم یا ارتباط (study type, organism, or relevance) فیلتر میکنیم.
- یک مطالعه مناسب را به دلخواه شناسایی میکنیم و شماره الحاق (accession number) آن را یادداشت میکنیم.

۲. مطالعه انتخابی:

- لینک: https://www.ncbi.nlm.nih.gov/sra/SRX27146400[accn]
- Accession Number: SRX27146400
- SRA: SRP552834
- Run Accession: SRR31785012
- توضیحات مطالعه:

The Study of Ru Doped Carbon Quantum Dots (Ru CDs) Therapy on Mouse breast cancer Cell Line 4T1
Abstract: Research on the anti-tumor activities of Ru doped carbon quantum dots (Ru CDs) nanomaterials



هـ) دانلود داده های خام با استفاده از NCBI SRA Toolkit

۱. دستورات:

- NCBI SRA Toolkit را نصب کنیم:

```
sudo apt-get install sra-toolkit
```

- دانلود داده های خاص:

```
prefetch Run_Accession_ID # Replace with specific Run Accession ID
```

- تبدیل SRA به FASTQ:

```
fastq-dump --split-files Run_Accession_ID
```

۲. مثال برای Accession انتخابی:

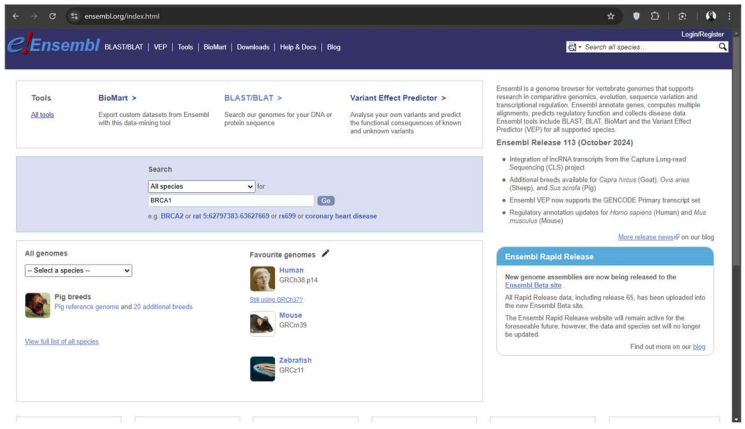
```
prefetch SRR31785012
fastq-dump --split-files SRR31785012
```

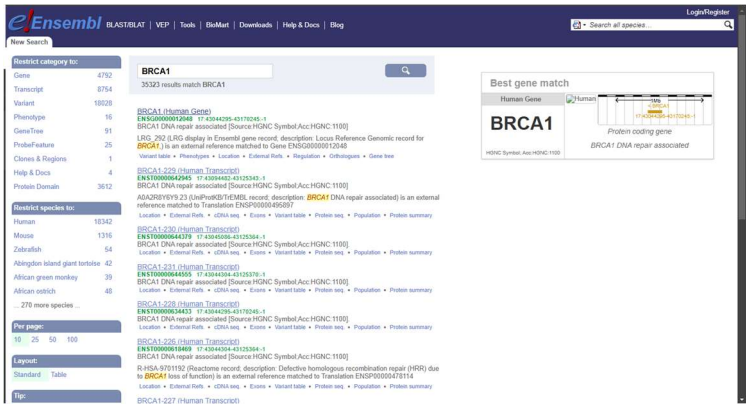
```
1m  1  !prefetch SRR31785012     # Replace with specific Run Accession ID

2024-12-26T16:36:56 prefetch.2.11.3: Current preference is set to retrieve SRA Normalized Format files with full base quality scores.
2024-12-26T16:36:56 prefetch.2.11.3: 1) Downloading 'SRR31785012'...
2024-12-26T16:36:56 prefetch.2.11.3: SRA Normalized Format file is being retrieved, if this is different from your preference, it may be due to current fi
2024-12-26T16:36:56 prefetch.2.11.3:  Downloading via HTTPS...
2024-12-26T16:37:51 prefetch.2.11.3:  HTTPS download succeed
2024-12-26T16:38:04 prefetch.2.11.3:  'SRR31785012' is valid
2024-12-26T16:38:04 prefetch.2.11.3: 1) 'SRR31785012' was downloaded successfully
2024-12-26T16:38:04 prefetch.2.11.3: 'SRR31785012' has 0 unresolved dependencies

8m  [7]  1  !fastq-dump --split-files SRR31785012

Read 21551544 spots for SRR31785012
Written 21551544 spots for SRR31785012
```

و) استفاده از پایگاه داده Ensembl

۱. یافتن یک ژن:

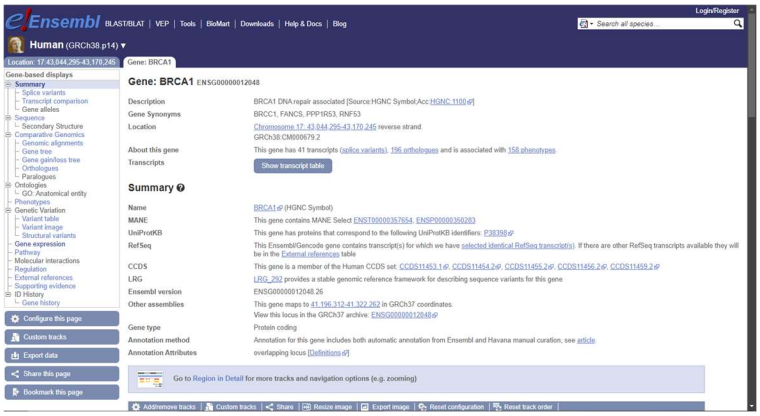
- از ( https://www.ensembl.org) دیدن میکنیم.
- "BRCA1" را با استفاده از نوار جستجو، جستجو میکنیم.



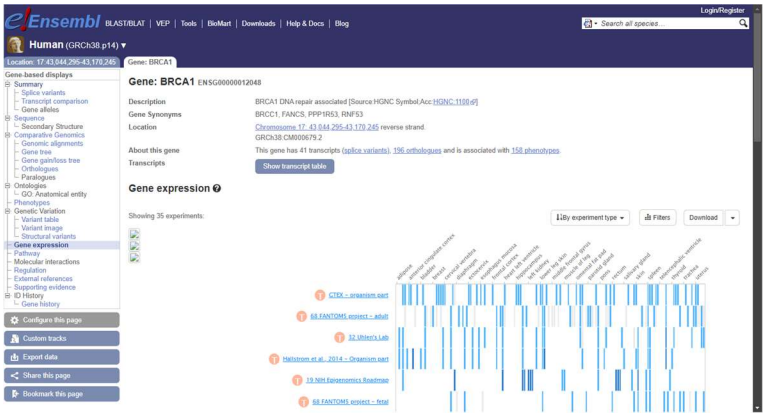



۲. داده های بیان ژن:

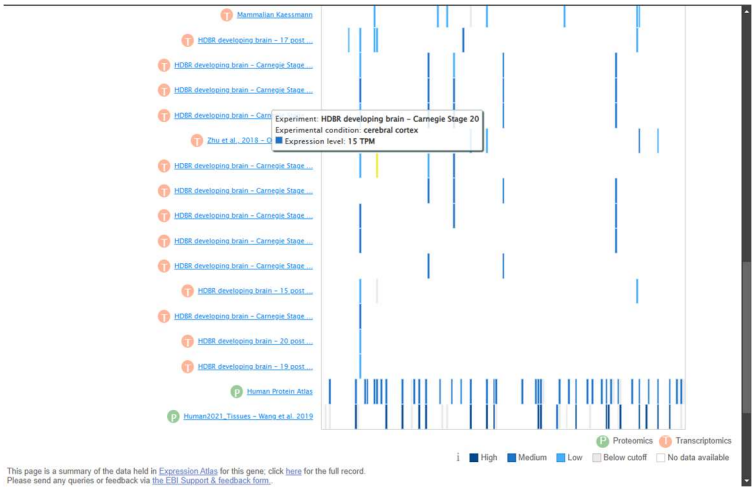
- ابتدا صفحه مربوط به ژن را باز میکنیم.




- حال به بخش "Gene expression" صفحه ژن می رویم.




- اطلاعاتی مختلفی در این بخش ارائه می شود مانند:
  - Transcript table
  - سطوح بیان (Expression Levels): در بافت ها و شرایط مختلف.
  - Disease Associations: پیوند بین بیان ژن و بیماری های خاص.
  - داده های مقایسه ای (Comparative Data): پروفایل های بیان بین گونه ها.



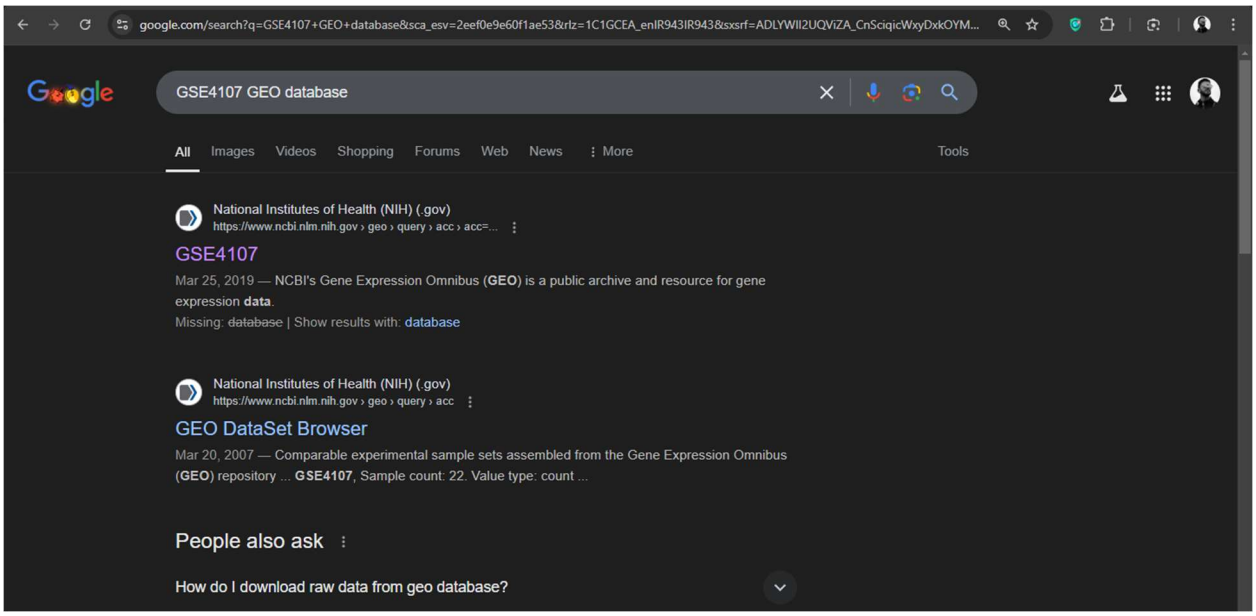

۳. به عنوان مثال برای BRCA1 اطلاعات زیر را میتوان مشاهده کرد:

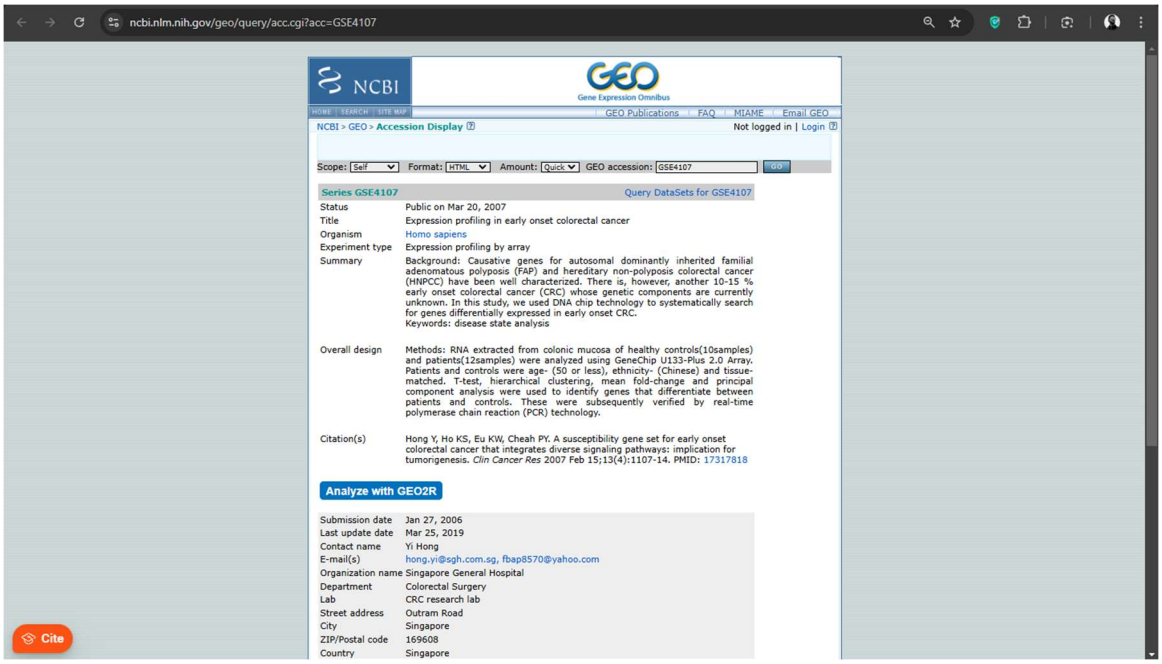
- در بافت سینه و تخمدان به شدت بیان می شود.
- بی نظمی مرتبط با سرطان سینه و تخمدان.
- مقایسه سطح بیان بین بافت های سالم و سرطانی.

## سوال اول عملی

الف) هر داده ای را به دلخواه (با سرچ در GEO DataSets) انتخاب میکنیم و شماره دسترسی (Accession Number) آن را مشخص میکنیم. (به عنوان مثال: GSE4107). سپس آن را با استفاده از GEO2R تجزیه و تحلیل میکنیم.

ما در این سوال از مجموعه داده با شماره دسترسی GSE4107 که در این لینک موجود است ( https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE4107) استفاده خواهیم کرد، که شامل پروفایل بیان در سرطان کولورکتال در شروع اولیه (expression profiling in early-onset colorectal cancer) است.






خلاصه مطالعه به شرح زیر است:

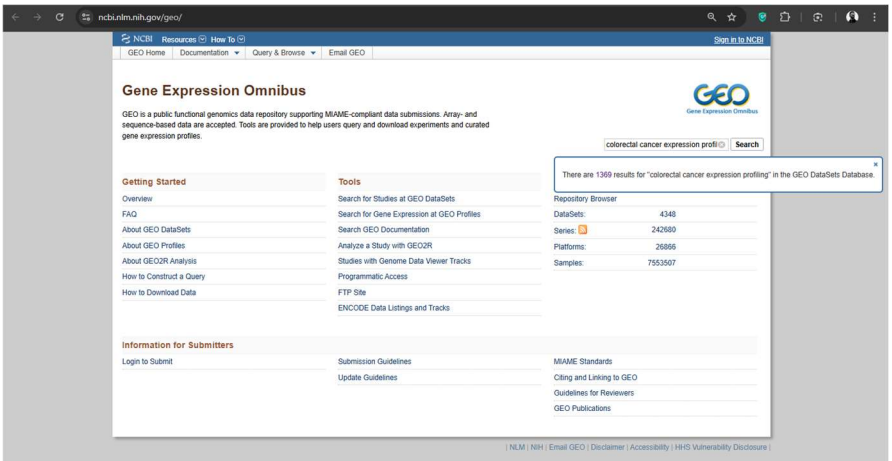
Background: Causative genes for autosomal dominantly inherited familial adenomatous polyposis (FAP) and hereditary non-polyposis colorectal cancer (HNPCC) have been well characterized. There is, however, another 10-15 % early onset colorectal cancer (CRC) whose genetic components are currently unknown. In this study, we used DNA chip technology to systematically search for genes differentially expressed in early onset CRC.
Keywords: disease state analysis

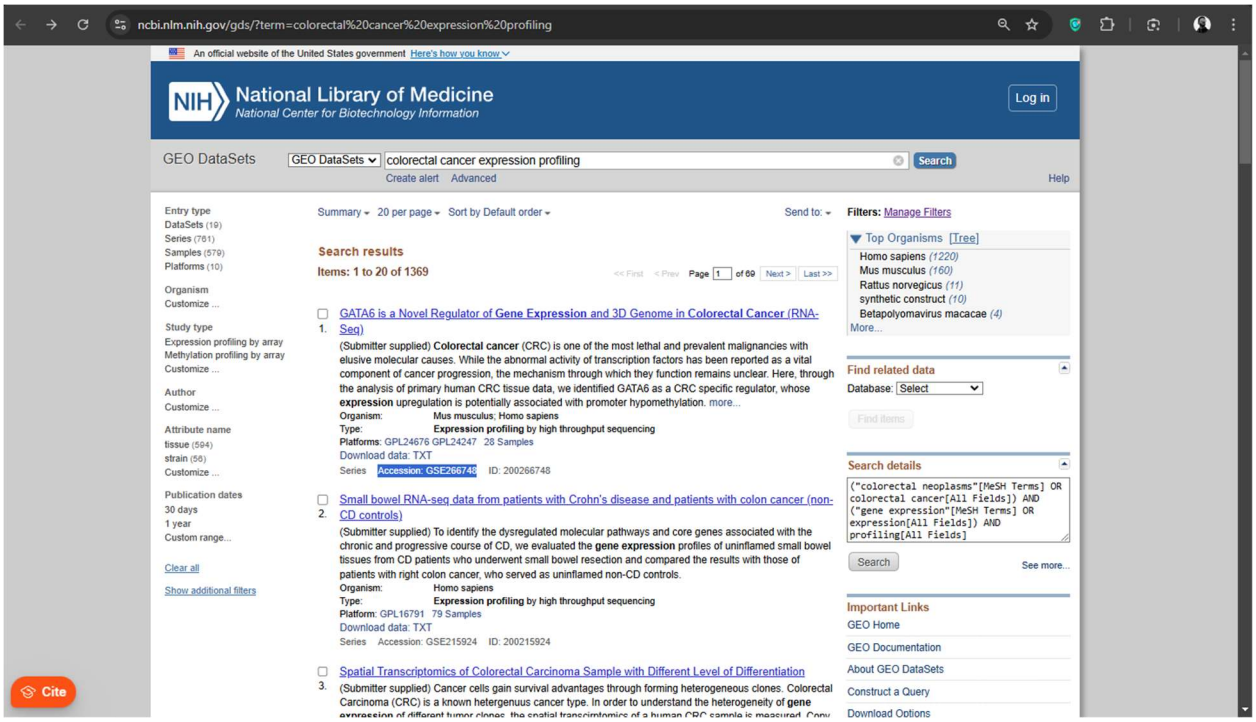
ب) توضیح مراحل جستجو و انتخاب داده ها.

۱. دسترسی به پایگاه داده GEO: به صفحه اصلی GEO ( https://www.ncbi.nlm.nih.gov/geo/ ) میرویم.

۲. جستجو برای مجموعه داده ها:

- از نوار جستجو برای وارد کردن کلمات کلیدی مرتبط با علاقه تحقیقاتی خود مانند " colorectal cancer expression profiling." استفاده میکنیم.
- فهرست مجموعه داده هایی که با معیارهای جستجوی شما مطابقت دارند را مرور میکنیم و با فیلتر های گوناگون به مجموعه داده دلخواه می رسیم.






۳. یک مجموعه داده را انتخاب میکنیم به عنوان مثال GSE4107:

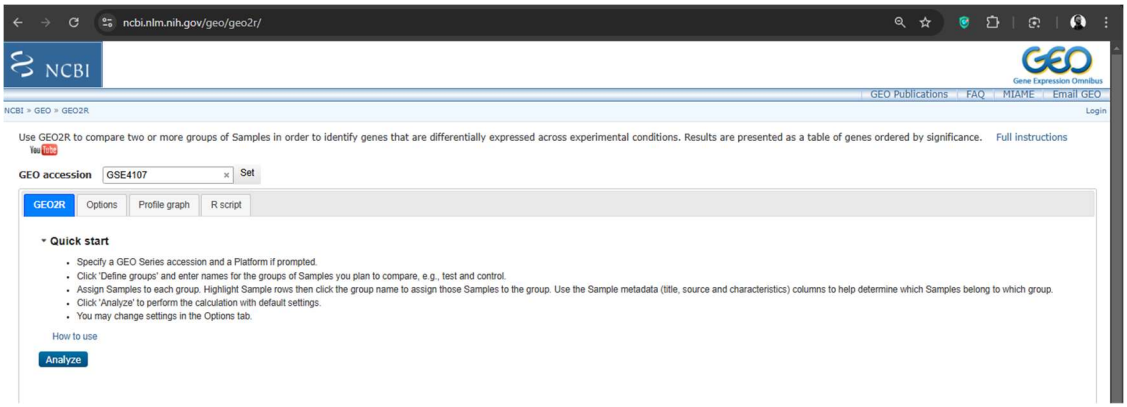
- روی مجموعه داده مورد علاقه کلیک میکنیم تا اطلاعات دقیق آن شامل Summary، Overall design و Sample های موجود را مشاهده کنیم.
- مطمئن می شویم که مجموعه داده شامل نمونه‌های کنترل و بیمار(control and patient samples) مرتبط با مطالعه ما باشد.

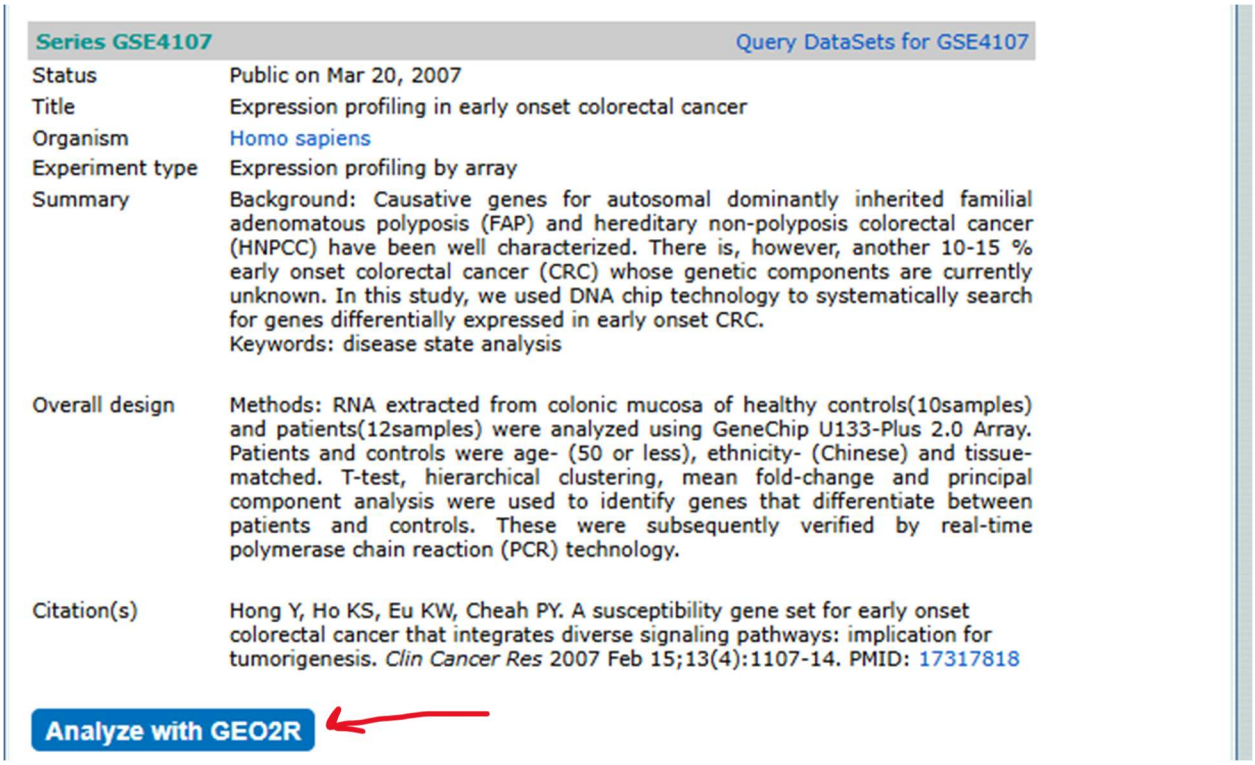
۴. به Accession Number توجه داریم: هر مجموعه داده دارای یک شماره دسترسی (Accession Number) منحصر به فرد (به عنوان مثال، GSE4107) است که برای ارجاع و تجزیه و تحلیل استفاده می شود.

ج) ژن هایی را با تفاوت بیان معنی دار بین گروه کنترل و بیمار با استفاده از GEO2R شناسایی میکنیم.

۱. دسترسی به GEO2R: به GEO2R tool موجود در این لینک ( https://www.ncbi.nlm.nih.gov/geo/geo2r/) میریم.

۲. شماره دسترسی(Accession Number) را وارد میکنیم: "GSE4107" را در کادر GEO accession وارد کرده و روی "Set" و "Analyze" کلیک میکنیم. یا اینکه در صفحه مطالعه مربوطه روی Analyze with GEO2R کلیک میکنیم.

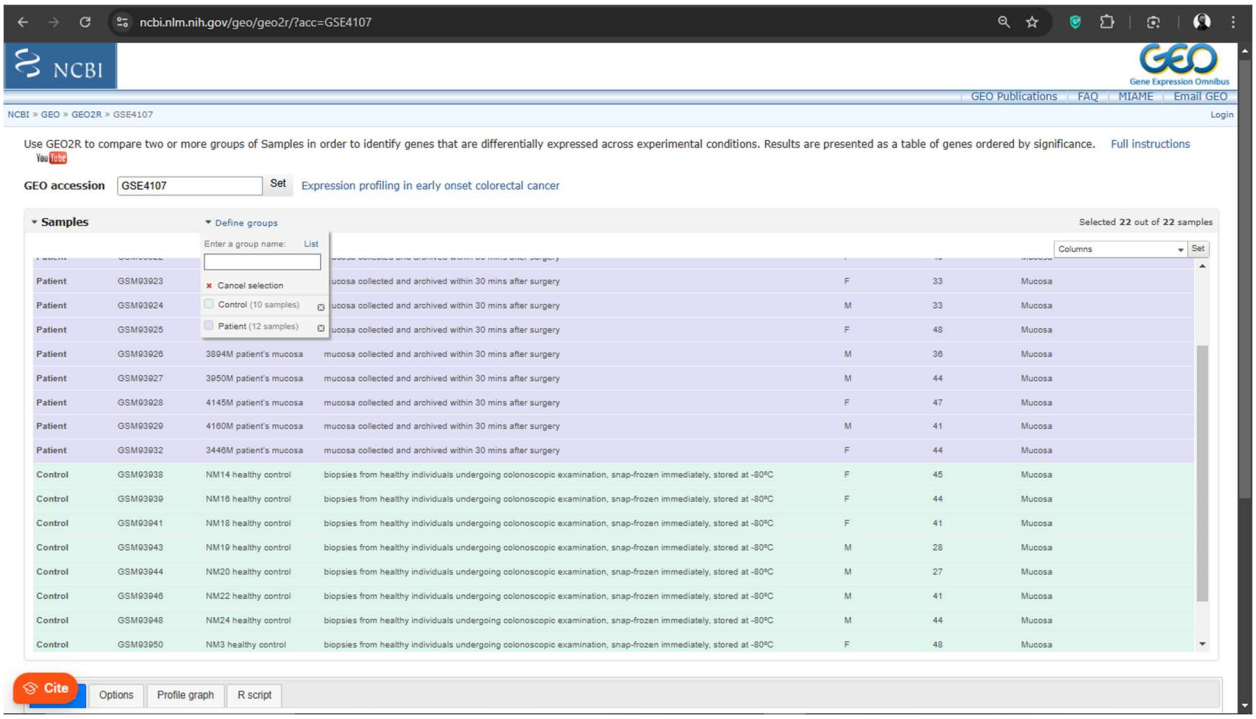





۳. گروه های نمونه را تعریف میکنیم:

- روی «Define group» کلیک میکنیم و دو گروه ایجاد میکنیم، به عنوان مثال، «Control» و «Patient».

- بر اساس اطلاعات نمونه مجموعه داده، نمونه های مناسب را به هر گروه اختصاص می دهیم.

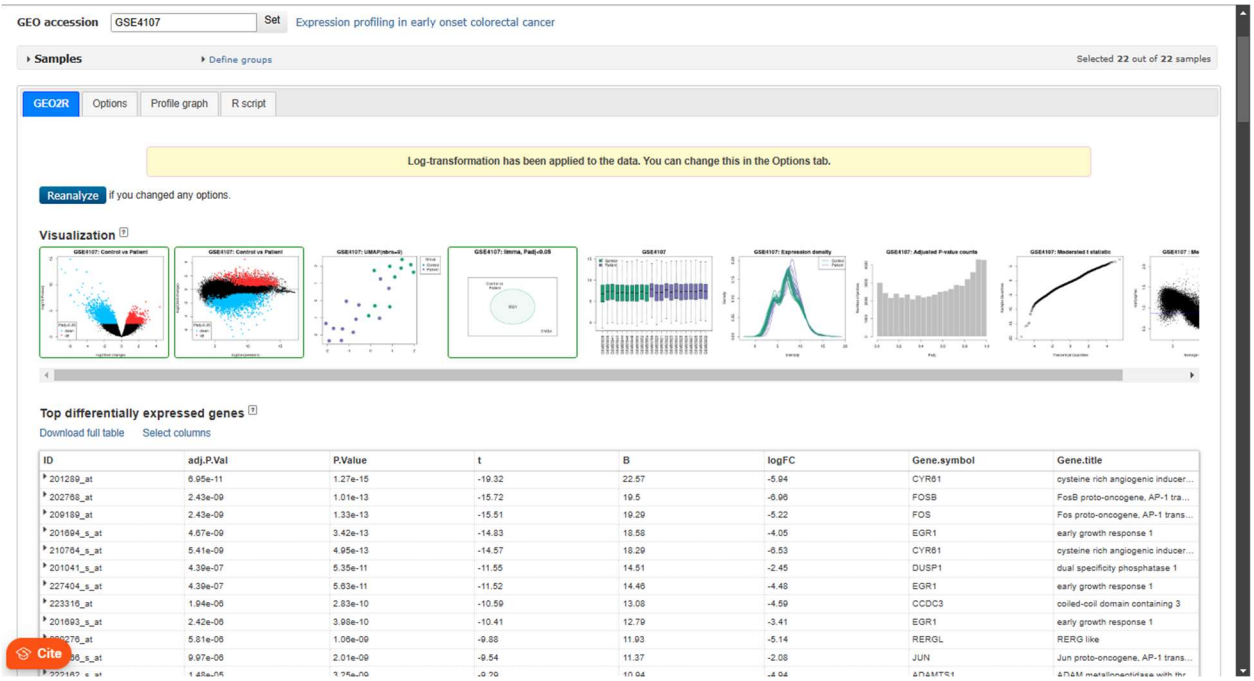


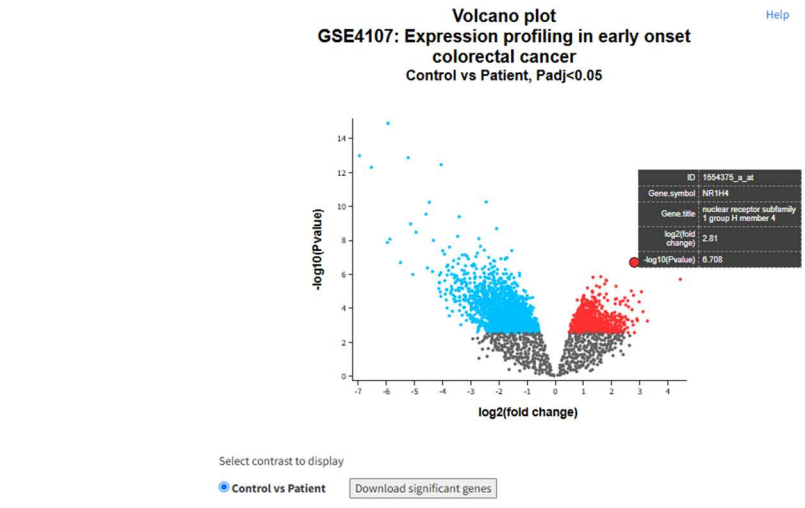
۴. تجزیه و تحلیل را انجام میدهیم:

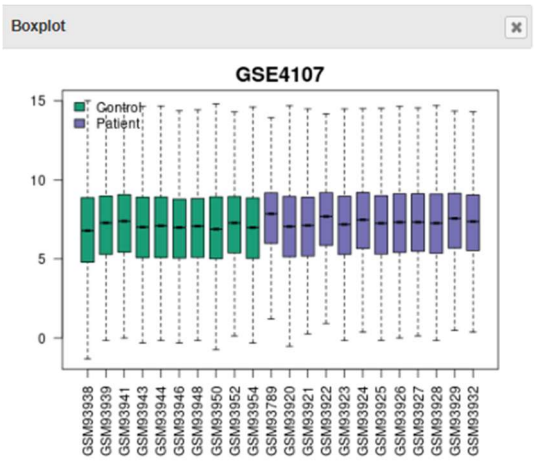
- پس از تخصیص نمونه ها به گروه ها، برای شناسایی ژن های بیان شده متفاوت بین دو گروه، روی "Analyze" کلیک میکنیم.

۵. بررسی نتایج:

- GEO2R جدولی از ژن ها را با آماره هایی مانند adjusted p-values و log fold changes نمایش می دهد که نشان دهنده اهمیت و بزرگی تفاوت های بیان است. همچنین نمودار های متنوعی برای visualization ارائه می دهد. در زیر نتیجه آن آمده است.
- ژن های با تفاوت بیان معنی دار بین گروه کنترل و بیمار: به عنوان مثال در volcano plot ژن های قرمز و آبی ژن های با تفاوت بیان معنی دار بین گروه کنترل و بیمار هستند که هر چه بالاتر باشند و از نقطه ۰ فاصله بیشتری داشته باشند تفاوت بیان معنی دار تری دارند. همچنین در جدول داده شده این ژن ها به ترتیب اهمیت مرتب شده اند که همان طور که میبینید ژن های اولیه در جدول دارای adjusted p-values کوچک و log fold changes بزرگ هستند که به معنای تفاوت بیان معنی دار بین گروه کنترل و بیمار است.







د) نتایج را در جدولی شامل ژن های مهم و p-value ارائه میدهیم.

در زیر جدولی از ژن های انتخاب شده از مجموعه داده GSE4107 وجود دارد که تفاوت های بیانی قابل توجهی را بین نمونه های کنترل و بیمار نشان می دهد. همان طور که میبینید تمام ژن ها را مشاهده نمیکنید بلکه فقط ۱۰ ژن با کمترین adj.P.Val آمده اند (۱۰ ژن ابتدایی در جدول خروجی ابزار GEO2R):

| ID | adj.P.Val | P.Value | t | logFC | Gene.symbol |
|---|---|---|---|---|---|
| **201289_at** | 6.95E-11 | 1.27E-15 | -19.316086 | -5.94112291 | CYR61 |
| **202768_at** | 2.43E-09 | 1.01E-13 | -15.7194503 | -6.95512411 | FOSB |
| **209189_at** | 2.43E-09 | 1.33E-13 | -15.514348 | -5.22397202 | FOS |
| **201694_s_at** | 4.67E-09 | 3.42E-13 | -14.8278778 | -4.0543598 | EGR1 |
| **210764_s_at** | 5.41E-09 | 4.95E-13 | -14.5653311 | -6.53075681 | CYR61 |
| **201041_s_at** | 4.39E-07 | 5.35E-11 | -11.5465574 | -2.45413476 | DUSP1 |
| **227404_s_at** | 4.39E-07 | 5.63E-11 | -11.516511 | -4.47587421 | EGR1 |
| **223316_at** | 1.94E-06 | 2.83E-10 | -10.5931505 | -4.59230998 | CCDC3 |
| **201693_s_at** | 2.42E-06 | 3.98E-10 | -10.4060636 | -3.41023566 | EGR1 |
| **220276_at** | 5.81E-06 | 1.06E-09 | -9.8761837 | -5.13632928 | RERGL |

همچنین جدول کامل در فایل GSE4107.top.table.tsv در پیوست موجود است. در این جدول که از پراهمیت به کم اهمیت مرتب شده است در سطر های اولیه میتوانید مهم ترین ژن ها را مشاهده کنید. در این جدول تمام ژن های با adjusted p-value کوچک تر از ۰/۰۵ از کوچک به بزرگ آمده اند.

توجه: adjusted p-value اصلاحات آزمایشی متعدد (multiple testing corrections) را به حساب می‌آورد، و log fold change نشان‌دهنده جهت و میزان تغییر بیان است (مقادیر منفی نشان‌دهنده کاهش در بیماران در مقایسه با گروه کنترل است).

برای نمایش بصری استفاده از GEO2R، آموزش زیر مفید بود که از آن استفاده کردم:

https://www.youtube.com/watch?v=9RyWjzSnaE0

# پایان