تمرین: تحلیل مقدماتی داده‌های **RNA-seq** با استفاده از **limma-voom**

درس: مقدمه‌ای بر بیوانفورماتیک

هدف تمرین

در این تمرین با مراحل اولیه تحلیل داده‌های RNA-seq آشنا می‌شویم. هدف شامل دانلود داده از GEO، آماده‌سازی ماتریس بیان ژنی، انجام تحلیل‌های اکتشافی، و در نهایت اجرای تحلیل بیان تفاضلی است.

معرفی دیتاست

در این تمرین از دیتاست GSE270045 استفاده می‌کنیم که شامل داده‌های بیان ژنی از نمونه‌های بیماران Long COVID و گروه کنترل سالم است. این مجموعه شامل ۳۶ نمونه و حدود ۲۸ هزار ژن است.

۱ بخش اول: دانلود و آماده‌سازی داده

در این بخش هدف شما این است که ماتریس بیان ژنی مربوط به دیتاست GSE270045 را دریافت و در محیط R بارگذاری کنید. برای این کار لازم است فایل مکمل GSE270045_LC_counts.tsv.gz را دانلود کنید. این فایل ماتریس بیان ژن‌ها را شامل می‌شود که در آن:

- هر سطر یک ژن است،
- هر ستون یک نمونه‌ی بیولوژیک (از بیماران یا کنترل‌ها) است،
- و هر عدد مقدار بیان آن ژن در آن نمونه را نشان می‌دهد.

شما می‌توانید این فایل را با استفاده از تابع getGEOSuppFiles دانلود کرده و سپس با read.delim در R بخوانید.

پس از بارگذاری ماتریس بیان، موارد زیر را بررسی و تحلیل کنید:

۱. ابعاد ماتریس بیان را با تابع **dim** گزارش کنید. هر یک از این دو عدد (تعداد ژن‌ها و تعداد نمونه‌ها) چه معنایی دارند؟

۲. چند سطر و چند ستون اول ماتریس را بررسی کنید. به طور ویژه سطر مربوط به ژن 5S_rRNA را مشاهده کنید: مقدار بیان این ژن در تمام نمونه‌ها چقدر است و چرا این مقدار مشاهده می‌شود؟ (درباره‌ی ژن‌های بسیار پرخوانش یا بسیار کم‌خوانش توضیح دهید.)

۳. بررسی کیفیت اولیه داده:

- آیا در ماتریس بیان مقدار NA وجود دارد؟
- آیا مقدار منفی مشاهده می‌شود؟
- بازه‌ی مقادیر ماتریس (Minimum–Maximum) چقدر است؟

این موارد چه اطلاعاتی درباره‌ی نوع داده به شما می‌دهند؟

۴. آیا داده‌ها **normalized** هستند یا **log-transformed**؟ با توجه به شکل مقدارها (مثلاً وجود اعشار، نبودن مقدارهای منفی)، توضیح دهید که:

- آیا این داده‌ها raw counts (شمارش خام) هستند؟
- اگر نه، چه نوع پردازشی روی آن‌ها انجام شده است؟
- چرا این موضوع برای انتخاب روش تحلیل (مثلاً limma-voom) مهم است؟

۵. بررسی اندازه‌ی کتابخانه (**Library Size**): برای هر نمونه، مجموع مقادیر بیان ژن‌ها را محاسبه کنید. این مقدار را library size می‌نامیم. بررسی کنید:

- کدام نمونه‌ها library size بزرگ‌تر یا کوچک‌تر دارند؟
- آیا اختلاف library size بین نمونه‌ها زیاد است؟
- این موضوع چه تأثیری روی تحلیل تفاضلی دارد؟

(در صورت تمایل یک نمودار میله‌ای از library size رسم کنید.)

۶. هر نکته‌ی غیرعادی که مشاهده کردید گزارش کنید. برای مثال: وجود ژن‌هایی با مقدار بیان صفر در تمام نمونه‌ها، ستون‌هایی با مقدار مجموع غیرمعمول و غیره.

سؤالات بخش اول (۲۰ نمره)

- سؤال ۱.۱ (۴ نمره): چرا برای تحلیل نیاز به فایل GSE270045_LC_counts.tsv.gz داریم و نه فایل Series Matrix؟
- سؤال ۲.۱ (۴ نمره): ابعاد ماتریس بیان چه اطلاعاتی از ساختار داده به شما می‌دهد؟

- سؤال ۳.۱ (۴ نمره): چرا ژن 5S_rRNA مقدار بیان بسیار کم یا صفر دارد؟ چه عواملی می‌توانند آن را توضیح دهند؟
- سؤال ۴.۱ (۴ نمره): آیا شکل توزیع داده‌ها نشان می‌دهد که داده‌ها raw counts هستند یا normalized؟ استدلال کنید.
- سؤال ۵.۱ (۴ نمره): library size چه نقشی در تحلیل RNA-seq دارد و چرا بررسی آن پیش از تحلیل مهم است؟

۲ بخش دوم: ساخت جدول اطلاعات نمونه‌ها (**Metadata**)

برای انجام تحلیل بیان تفاضلی، تنها داشتن ماتریس بیان ژن‌ها کافی نیست؛ باید بدانیم هر ستون مربوط به کدام نوع نمونه است (بیمار یا کنترل). به همین دلیل لازم است جدولی به نام metadata بسازیم که در آن برای هر نمونه، اطلاعاتی مانند گروه آزمایشی (مثلاً Control یا ME_CFS) ثبت شده باشد.
در این بخش هدف شما ساختن یک جدول متادیتاست که:

- سطرهای آن نمونه‌ها (sampleها) باشند،
- یک ستون شامل شناسه‌ی نمونه (مثلاً CCI031، S013_11 و ...)،
- و یک ستون شامل گروه هر نمونه (مثلاً Control یا ME_CFS).

برای این کار مراحل زیر را انجام دهید:

۱. با استفاده از تابع getGEO اطلاعات سری GSE270045 را دریافت کنید و با کمک تابع pData جدول اطلاعات نمونه‌ها (phenotype data) را استخراج کنید. چند ستون اول این جدول را بررسی کنید (برای مثال ستون‌هایی مانند title و characteristics_ch1).

۲. ستونی را پیدا کنید که از روی آن بتوانید وضعیت سلامت نمونه‌ها را تشخیص دهید؛ برای مثال در این دیتاست، در ستون title عبارات Healthy Control برای نمونه‌های کنترل و عبارات دیگر برای بیماران دیده می‌شود. بر اساس این ستون، یک متغیر جدید (مثلاً group) تعریف کنید که دو سطح داشته باشد: Control و ME_CFS (یا نام مناسب دیگر برای بیماران).

۳. با استفاده از یک جدول فراوانی (مثلاً table(group)) بررسی کنید چند نمونه در گروه Control و چند نمونه در گروه ME_CFS قرار می‌گیرند. این اعداد را گزارش کنید.

۴. در جدول اطلاعات نمونه‌ها، ستونی را پیدا کنید که حاوی شناسه‌ی واقعی نمونه‌ها باشد؛ برای مثال ممکن است عباراتی مانند "sample id: CCI031" در یکی از ستون‌های characteristics_ch1 ذخیره شده باشند. شناسه‌ی خام (مانند CCI031، HP1193، S013_11) را از این ستون استخراج کنید و آن را در ستونی به نام sample_id ذخیره کنید.

۵. بررسی کنید که آیا مقادیر sample_id دقیقاً با نام ستون‌های ماتریس بیان (counts) همخوانی دارند یا خیر. اگر لازم بود، ترتیب ستون‌های ماتریس counts را طوری مرتب کنید که با ترتیب سطرهای جدول متادیتا یکسان شود (یعنی هر ستون بیان، با سطر متناظر در metadata مربوط به همان نمونه باشد).

۶. در نهایت یک جدول متادیتا (data frame) بسازید که:

- هر سطر یک نمونه،
- یک ستون شامل sample_id،
- و یک ستون شامل group باشد،

و نام سطرها (rownames) را برابر sample_id قرار دهید.

سؤالات بخش دوم (۲۰ نمره)

- سؤال ۱.۲ (۴ نمره): کدام ستون یا ستون‌ها در جدول اطلاعات نمونه‌ها (pData) برای تشخیص این‌که یک نمونه Healthy Control است یا بیمار (ME_CFS) بیشترین کمک را به شما کردند؟ دلیل خود را توضیح دهید.

- سؤال ۲.۲ (۴ نمره): چند نمونه در هر گروه (Control و ME_CFS) دارید؟ آیا تعادل بین تعداد نمونه‌های دو گروه را مناسب می‌دانید؟

- سؤال ۳.۲ (۴ نمره): چرا مهم است که نام ستون‌های ماتریس بیان ژن‌ها (counts) دقیقاً با شناسه‌ی نمونه‌ها در جدول متادیتا (metadata) منطبق باشد؟ اگر این همخوانی به‌هم بخورد چه مشکلاتی ممکن است در تحلیل تفاضلی رخ دهد؟

- سؤال ۴.۲ (۴ نمره): اگر به‌اشتباه برخی نمونه‌های بیمار را در گروه Control (یا برعکس) قرار دهید، انتظار دارید چه اثری بر نتایج تحلیل بیان تفاضلی (DE) داشته باشد؟

- سؤال ۵.۲ (۴ نمره): چرا معمولاً بهتر است سطح مرجع (reference level) متغیر group را گروه Control قرار دهیم؟ این انتخاب چه تأثیری بر تفسیر مقدار logFC دارد؟

۳ بخش سوم: بررسی اکتشافی داده‌های بیان ژنی

در این بخش هدف شما این است که با استفاده از چند نمودار و شاخص ساده، درک بهتری از توزیع داده‌های بیان ژنی و تفاوت‌های کلی بین نمونه‌ها به‌دست آورید. قبل از انجام هر تحلیل آماری (مثل تحلیل بیان تفاضلی)، این بررسی‌های اولیه برای شناسایی داده‌های غیرعادی و درک نوع داده بسیار مهم هستند.

فرض کنید که در این مرحله ماتریس بیان ژن‌ها (counts) و جدول متادیتا (coldata) را مانند بخش‌های قبل آماده کرده‌اید و نام ستون‌های counts با نام سطرهای coldata منطبق است.

۱. بررسی عددی کلی روی ماتریس بیان با استفاده از توابعی مانند summary، range، any و is.na توصیف عددی زیر را برای ماتریس بیان به‌دست آورید:

- حداقل، چارک‌ها، میانه، میانگین و حداکثر مقادیر،
- وجود یا عدم وجود مقادیر NA،
- وجود یا عدم وجود مقادیر منفی.

بر اساس این مقادیر، توضیح دهید که این داده‌ها بیشتر شبیه شمارش خام (raw counts) هستند یا داده‌های نرمال‌شده.

۲. محاسبه و بررسی اندازه‌ی کتابخانه (**Library Size**) برای هر نمونه، مجموع مقادیر بیان ژن‌ها را محاسبه کنید (جمع هر ستون ماتریس counts). این مقدار تقریباً نشان‌دهنده‌ی library size یا عمق توالی‌یابی آن نمونه است.

- یک نمودار میله‌ای از مقادیر library size برای همه‌ی نمونه‌ها رسم کنید.
- در نمودار خود نمونه‌ها را به دو گروه Control و ME_CFS تفکیک کنید (مثلاً با رنگ یا برچسب مناسب).

۳. بررسی توزیع بیان در مقیاس **log2** ماتریس بیان را به صورت log2(expression + 1) تبدیل کنید و برای هر نمونه (هر ستون)، یک boxplot رسم کنید تا توزیع مقادیر بیان در مقیاس لگاریتمی را مقایسه کنید. دقت کنید که تمام نمونه‌ها در یک نمودار واحد رسم شوند تا قابل مقایسه باشند.

۴. رسم نمودار **voom** برای روند میانگین-واریانس با استفاده از بسته‌های edgeR و limma، از ماتریس counts یک شیء DGEList ساخته، یک design matrix ساده با متغیر group تعریف کنید و تابع voom را اجرا کنید تا نمودار mean–variance trend رسم شود.

سؤالات بخش سوم (۲۰ نمره)

- سؤال ۱.۳ (۴ نمره): طبق خروجی تابع summary و بازه‌ی مقادیر، حداقل، میانه و حداکثر مقدار بیان ژن‌ها در این داده تقریباً چقدر است؟ نبودن مقادیر منفی و وجود اعداد اعشاری چه چیزی درباره‌ی نوع داده (خام یا نرمال‌شده) به شما می‌گوید؟
- سؤال ۲.۳ (۴ نمره): در نمودار library size:
  - کدام نمونه‌ها کوچک‌ترین و بزرگ‌ترین library size را دارند؟
  - آیا به‌طور کلی بین دو گروه Control و ME_CFS تفاوت چشمگیری در library size مشاهده می‌کنید؟
  - چرا اختلاف زیاد در library size می‌تواند روی تحلیل‌های بعدی (مثلاً DE) اثر بگذارد؟
- سؤال ۳.۳ (۴ نمره): در نمودارهای boxplot از مقادیر log2(expression + 1):
  - آیا شکل کلی توزیع نمونه‌ها شبیه هم است یا برخی نمونه‌ها از نظر میانه یا گستره‌ی مقادیر به وضوح متفاوت‌اند؟
  - اگر نمونه‌ای وجود دارد که توزیع بسیار متفاوتی دارد، چه توضیح‌های احتمالی برای این رفتار می‌توانید ارائه دهید؟
- سؤال ۴.۳ (۴ نمره): نمودار voom چه رابطه‌ای بین میانگین بیان (محور افقی) و واریانس (محور عمودی) نشان می‌دهد؟ چرا مشاهده‌ی این رابطه برای استفاده از روش limma-voom مهم است؟
- سؤال ۵.۳ (۴ نمره): با توجه به اینکه مقادیر ماتریس بیان اعشاری و غیرصحیح هستند، توضیح دهید چرا استفاده از روش‌هایی مانند DESeq2 (که فرض شمارش‌های صحیح را دارند) مناسب نیست و در عوض روش limma-voom انتخاب می‌شود.

۴ بخش چهارم: تحلیل **PCA** و بررسی خوشه‌بندی نمونه‌ها

در این بخش هدف شما این است که با استفاده از تحلیل مولفه‌های اصلی (Principal Component Analysis) ساختار کلی داده را بررسی کنید و ببینید نمونه‌ها بر اساس گروه آزمایشی (Control در برابر ME_CFS) چگونه در فضاهای کم‌بعد قرار می‌گیرند. این تحلیل یکی از مهم‌ترین مراحل اکتشافی در داده‌های RNA-seq است و می‌تواند به شناسایی الگوهای کلی، تفاوت‌های گروهی و نمونه‌های دورافتاده (outliers) کمک کند.

نکته‌ی آموزشی: پس از تبدیل داده‌های بیان ژنی به مقیاس log2، تمام ژن‌ها تقریباً روی یک دامنه‌ی مشابه (معمولاً بین ۰ تا ۱۵) قرار می‌گیرند. در این شرایط، استفاده از scale.=TRUE در تابع prcomp

باعث می‌شود ژن‌های کم‌واریانس (که اطلاعات کمی دارند و اغلب نویز هستند) همان اندازه‌ی ژن‌های پُرواریانس بر نتیجه‌ی PCA تأثیر بگذارند، که این موضوع می‌تواند تفسیر PCA را مخدوش کند. به همین دلیل در تحلیل‌های RNA-seq معمولاً از scale.=FALSE استفاده می‌شود.

مراحل پیشنهادی برای این بخش:

۱. ماتریس بیان ژنی را به صورت $\log2(\text{expression} + 1)$ تبدیل کنید تا مقادیر قابل مقایسه‌تر شوند و تأثیر ژن‌های بسیار پربیان کاهش یابد.

۲. ژن‌هایی را که واریانس آن‌ها بسیار کم است حذف کنید؛ این ژن‌ها اطلاعات چندانی برای خوشه‌بندی ندارند و تنها باعث افزایش نویز می‌شوند. برای مثال می‌توانید ژن‌هایی که واریانس آن‌ها در پایین‌ترین چند درصد قرار دارد حذف کنید.

۳. تحلیل PCA را با استفاده از تابع prcomp اجرا کنید. توجه کنید که prcomp انتظار دارد سطرها نمونه و ستون‌ها متغیر باشند، بنابراین ماتریس را باید transpose کنید.

۴. درصد واریانس توضیح‌داده‌شده توسط مؤلفه‌های اصلی را محاسبه و بررسی کنید:

- چه مقدار از واریانس کل توسط PC1 توضیح داده می‌شود؟
- درصدهای مربوط به PC2 و PC3 چقدر است؟

۵. یک نمودار دوبُعدی از PC1 و PC2 رسم کنید و نمونه‌ها را بر اساس گروه (Control یا ME_CFS) با رنگ‌های متفاوت مشخص کنید. می‌توانید برچسب نمونه‌ها را نیز روی نمودار اضافه کنید.

۶. نمودار میله‌ای درصد واریانس توضیح‌داده‌شده توسط اولین چند مؤلفه‌ی اصلی را رسم کنید.

سؤالات بخش چهارم (۱۸ نمره)

- سؤال ۱.۴ (۳ نمره): PC1 چه درصدی از واریانس کل داده را توضیح می‌دهد؟ آیا این مقدار نشان می‌دهد که یک روند غالب در داده وجود دارد؟
- سؤال ۲.۴ (۳ نمره): آیا نمونه‌های دو گروه Control و ME_CFS در نمودار PC1–PC2 از یکدیگر جدا می‌شوند؟ این موضوع چه معنایی از نظر وجود تفاوت بین دو گروه دارد؟
- سؤال ۳.۴ (۳ نمره): کدام نمونه‌ها از نظر مکانی در PCA «دورتر از بقیه» (outlier) به نظر می‌رسند؟ چند دلیل احتمالی برای دورافتادگی یک نمونه پیشنهاد کنید (مثلاً کیفیت پایین نمونه، خطای تکنیکی، تفاوت بیولوژیک واقعی و غیره).

- سؤال ۴.۴ (۳ نمره): چرا قبل از اجرای PCA باید ژن‌های با واریانس بسیار کم را حذف کرد؟ این کار چه تأثیری روی کیفیت خوشه‌بندی دارد؟

- سؤال ۵.۴ (۳ نمره): در اجرای PCA از پارامتر scale.=FALSE استفاده کردیم. توضیح دهید چرا معمولاً در داده‌های RNA-seq (پس از تبدیل log2) نباید داده‌ها را به صورت استاندارد شده (واریانس ۱ و میانگین ۰) مقیاس‌بندی کنیم. اگر به اشتباه از scale.=TRUE استفاده شود، انتظار دارید چه تغییری در نتایج PCA رخ دهد؟

- سؤال ۶.۴ (۳ نمره): اگر به‌جای log2(expression + 1) از داده‌ی خام استفاده می‌کردید، فکر می‌کنید چه اتفاقی برای نمودار PCA می‌افتاد؟ چرا؟

۵ بخش پنجم: تحلیل بیان تفاضلی با **limma-voom**

در این بخش هدف شما بررسی تفاوت بیان ژن‌ها بین دو گروه Control و ME_CFS است. برای این کار از روش limma-voom استفاده می‌کنیم که با تبدیل voom وابستگی میانگین-واریانس را مدل می‌کند و سپس یک مدل خطی برای هر ژن برازش می‌دهد.

مراحل پیشنهادی برای انجام تحلیل:

۱. یک شیء DGEList از ماتریس counts بسازید و در صورت نیاز عوامل نرمال‌سازی را با تابع calcNormFactors محاسبه کنید.

۲. ژن‌های کم‌بیان را با تابع filterByExpr حذف کنید تا تنها ژن‌های دارای سیگنال کافی وارد تحلیل شوند.

۳. یک ماتریس طراحی (design matrix) بر اساس متغیر group بسازید. توجه کنید که گروه Control باید سطح مرجع باشد تا مقدار logFC تفاوت بیان ME_CFS نسبت به Control را نشان دهد.

۴. تبدیل voom را اجرا کنید تا روند میانگین-واریانس مدل شود و وزن‌ها برای تحلیل خطی محاسبه شود. نمودار mean–variance trend را مشاهده و تفسیر کنید.

۵. مدل خطی را با lmFit برازش دهید و با eBayes آماره‌های تعدیل‌شده را محاسبه کنید.

۶. جدول نتایج را با topTable استخراج کنید و ستون‌های مهم زیر را بررسی کنید:

- logFC - جهت و اندازه‌ی تغییر بیان،
- AveExpr - میانگین بیان ژن،

- P.Value و adj.P.Val - مقادیر معنی‌داری،
- B - آمار لیما برای احتمال تفاضلی بودن ژن.

۷. یک نمودار volcano رسم کنید و ژن‌های معنی‌دار را برجسته کنید (مثلاً adj.P.Val < 0.05 و |logFC| > 1).

۸. از ۳۰ تا ۵۰ ژن برتر (بر اساس adj.P.Val) برای رسم یک heatmap استفاده کنید و الگوهای خوشه‌بندی نمونه‌ها را بررسی کنید.

سؤالات بخش پنجم (۲۱ نمره)

- سؤال ۱.۵ (۳ نمره): تابع filterByExpr چه نوع ژن‌هایی را حذف می‌کند و چرا این کار برای تحلیل DE ضروری است؟
- سؤال ۲.۵ (۳ نمره): نمودار voom چه اطلاعاتی درباره‌ی رابطه‌ی میانگین و واریانس در داده‌ها نشان می‌دهد؟ چرا مدل‌سازی این روند برای limma ضروری است؟
- سؤال ۳.۵ (۳ نمره): در جدول نتایج، سه ژن برتر بر اساس adj.P.Val را نام ببرید و مقدار logFC آن‌ها را تفسیر کنید. هرکدام بیان‌شان در کدام گروه بیشتر است؟
- سؤال ۴.۵ (۳ نمره): بر اساس نمودار volcano، الگوی کلی تفاوت بیان چگونه است؟ آیا به‌طور کلی ژن‌های بیشتری افزایش بیان دارند یا کاهش بیان؟
- سؤال ۵.۵ (۳ نمره): چند ژن دارای adj.P.Val < 0.05 و |logFC| > 1 یافتید؟ این عدد چه چیزی درباره‌ی اختلاف بین دو گروه بیان می‌کند؟
- سؤال ۶.۵ (۳ نمره): در نقشه‌ی گرمایی (heatmap)، آیا نمونه‌های دو گروه Control و ME_CFS به‌طور طبیعی در دو خوشه قرار گرفته‌اند؟ اگر نمونه‌ای در خوشه‌ی «غیرمنتظره» قرار گرفته، چه توضیح‌هایی می‌تواند داشته باشد؟
- سؤال ۷.۵ (۳ نمره): جهت مثبت یا منفی بودن logFC چه معنایی دارد؟ در این مدل دقیقاً عبارت ME_CFS vs Control چگونه تفسیر می‌شود؟

خروجی نهایی مورد نیاز و نمره‌ی کلی کدنویسی (۱ نمره)

دانشجویان باید موارد زیر را تحویل دهند:

- یک یا چند فایل R Script شامل تمام کدهای اجرا شده

- یک گزارش PDF شامل نمودارها و پاسخ سؤالات هر بخش
- سؤال ۱.۶ (۱ نمره): کیفیت کلی کدنویسی، مستندسازی و مرتب‌بودن گزارش (PDF)؛ شامل نام‌گذاری مناسب متغیرها، وجود توضیح (comment) در بخش‌های اصلی، و خوانایی شکل‌ها.