به نام خدا

مهدی کافی ۹۹۲۱۰۷۵۳
تحلیل داده‌های حجیم زیستی، تمرین چهارم

سوال ۱
ب)
روش CPM
یک روش بسیار ساده است که تنها در هر سمپل تعداد خوانش‌ها را به تعداد کل خوانش‌های آن سمپل در واحد میلیون خوانش تقسیم می‌کند.
مزیت این روش این است که بسیار ساده است و همینطور به هیچ داده‌ی اضافی‌ای مثل طول ژن‌ها نیاز ندارد.
این روش برای تحلیل بین replicateهای یک سمپل مناسب است و برای مقایسه درون سمپل و یا تحلیل differentially expressed genes مناسب نیست.
این روش نرمال کردن بر اساس طول ترنسکریپت را انجام نمی‌دهد و برای تحلیل‌هایی مناسب است که در آن‌ها خوانش‌ها مستقل از طول ژن تولید شده‌اند.

روش TPM
این روش در ابتدا تعداد خوانش‌ها را به طول ژن بر حسب kbp تقسیم می‌کند. سپس تمام این اعداد را در هر سمپل با یکدیگر جمع کرده و بر ۱ میلیون تقسیم می‌کند. برای هر سمپل مقادیر محاسبه شده در قسمت اول را به مقدار محاسبه شده در مرحله دوم تقسیم می‌کند.
این روش برای مقایسه تعداد ژن در یک سمپل و یا بین سمپل‌های یک گروه مناسب است ولی برای تحلیل differentially expressed ژن‌ها مناسب نیست.
این روش نیاز به طول ژن‌ها دارد و آن‌ها را در نرمال‌سازی لحاظ می‌کند.
این روش مناسب تحلیل‌هایی که است که در آن‌ها پروتکل تولید خوانش به طول ژن مرتبط است.
این روش به عنوان جایگزین برای RPKM ارائه شده‌است و به دلیل دقت کم روش RPKM؛ در TPM برخلاف RPKM، میانگین ثابت و متناسب با غلظت نسبی RNA است.

روش DESeq2
این روش برای مقایسه تعداد ژن‌ها بین سمپل‌ها و برای تحلیل DE مناسب است و برای مقایسه درون یک سمپل مناسب نیست.
این روش فرض میگیرد که تعداد کمی از ژن‌ها تفاوت بیان معنی‌داری دارند.
این روش طول ژن را در نظر نمی‌گیرد زیرا که فرض می‌کند طول ژن برای تمام نمونه‌ها یکسان است.
این روش برای تحلیل‌های بین سمپل‌ها بهتر عمل می‌کند.

منابع:

- https://hbctraining.github.io/DGE_workshop/lessons/02_DGE_count_normalization.html
- https://www.reneshbedre.com/blog/expression_units.html