**Simulate RNAseq counts in relation to Copy Number dosage sensivity**

Simulating RNA-seq counts for genes with different types of dosage sensitivity — **dosage-sensitive genes**, **dosage-insensitive genes**, and **dosage-compensated genes** — involves adjusting gene expression based on changes in DNA copy number. Here's how these categories behave:

1. **Dosage-sensitive genes**: Gene expression is proportional to the copy number. For example, doubling the copy number leads to a doubling of expression.
2. **Dosage-insensitive genes**: Gene expression is independent of copy number. Even with changes in copy number, expression remains constant.
3. **Dosage-compensated genes**: These genes adjust their expression in response to changes in copy number to maintain constant levels of expression, so if the copy number increases, the gene downregulates expression, and vice versa.

Here’s a strategy for simulating RNA-seq counts for these genes:

1. **Simulating Copy Number Variation (CNV)**: First, let's assume different genes will have varying copy numbers in the genome, for instance:
   * Normal diploid copy number = 2 (baseline)
   * CN = 1 (haploid or loss of one copy)
   * CN = 3, 4, 5 (amplified)
2. **Simulating RNA-seq Counts**:
   * The baseline expression for each gene can be drawn from a distribution, like a **negative binomial** or **Poisson** distribution, which reflects typical RNA-seq count variability.
3. **Apply dosage sensitivity rules**:
   * For **dosage-sensitive** genes, expression increases/decreases proportionally to CN. Expression E=CN×EbaselineE = CN \times E\_{\text{baseline}}E=CN×Ebaseline​
   * For **dosage-insensitive** genes, expression remains constant regardless of CN. Expression E=EbaselineE = E\_{\text{baseline}}E=Ebaseline​
   * For **dosage-compensated** genes, expression adjusts inversely with CN changes to maintain a steady expression level. Expression E=2CN×EbaselineE = \frac{2}{CN} \times E\_{\text{baseline}}E=CN2​×Ebaseline​ (assuming normal CN=2 is the reference).

**Example Plan**

* **Number of Genes**: Simulate 1000 genes, split into:
  + 300 dosage-sensitive
  + 300 dosage-insensitive
  + 400 dosage-compensated
* **Copy Numbers**: Randomly assign CN values from {1, 2, 3, 4, 5} to each gene.
* **Baseline Expression**: Draw random counts from a negative binomial distribution to reflect typical RNA-seq variability.

For **dosage-insensitive genes**, the expression remains constant regardless of copy number, and for **dosage-compensated genes**, expression adjusts to compensate for the copy number change.

**Evaluate Methods Performance Using Simulated Data**

Create a dataset with known differential expression (true positives and true negatives).

Use several statistical methods to analyze the simulated dataset and identify DEGs.

**Determine DEGs**: For each method, define a significance threshold (e.g., adjusted p-value < 0.05) to identify DEGs.

**Evaluate Performance**:

* **True Positives (TP)**: Correctly identified DEGs.
* **False Positives (FP)**: Incorrectly identified as DEGs.
* **True Negatives (TN)**: Correctly identified non-DEGs.
* **False Negatives (FN)**: Missed DEGs.

**Calculate Performance Metrics**: Use the confusion matrix to calculate:

* **Sensitivity (Recall)**: TP/(TP+FN)\text{TP} / (\text{TP} + \text{FN})TP/(TP+FN)
* **Specificity**: TN/(TN+FP)\text{TN} / (\text{TN} + \text{FP})TN/(TN+FP)
* **Precision (Positive Predictive Value)**: TP/(TP+FP)\text{TP} / (\text{TP} + \text{FP})TP/(TP+FP)
* **F1 Score**: 2×(Precision×Recall)/(Precision+Recall)2 \times (\text{Precision} \times \text{Recall}) / (\text{Precision} + \text{Recall})2×(Precision×Recall)/(Precision+Recall)

**Compare Methods**: Summarize results across methods to assess their performance in identifying DEGs.

**DISCUSSION**

Incorporating CNV data as covariates in generalized linear models (GLMs) allows researchers to more accurately differentiate between genuine biological signals and variations stemming from changes in gene dosage. This integration is crucial for correctly identifying differentially expressed genes, particularly in cancer, where the expression levels of genes can fluctuate significantly in response to changes in copy number. By modeling CNV effects explicitly, researchers can mitigate the risk of statistical artifacts, ultimately leading to more reliable conclusions regarding gene expression profiles (references).

Understanding the interplay between CNVs and gene expression not only enhances the accuracy of analyses but also has profound implications for personalized medicine. By integrating CNV data into GLMs, researchers can uncover potential biomarkers for treatment response or resistance tailored to individual tumor profiles. This is especially relevant for targeted therapies, where the effectiveness of treatment may hinge on the expression levels of specific genes that are affected by CNVs (references).

Moreover, leveraging CN-aware approaches can facilitate a more nuanced understanding of tumor biology, enabling the identification of dosage-sensitive, dosage-insensitive and dosage-compensated genes. This knowledge is critical for elucidating the mechanisms of tumorigenesis and improving therapeutic strategies. Overall, the transition from CN-naive to CN-aware models represents a significant advancement in the quest for precision in cancer genomics.

Leveraging CN-aware analyses can refine the identification of driver genes and biomarkers. This has particular importance in designing precision therapies, as targeting CN-independent drivers may yield higher efficacy and specificity in aneuploid cancers.