**Methods Summary: Neuroblastoma Gene Ranking Analysis Using Copy Number and Expression Data**

**1. Data Preparation and Genome Coordinate Alignment**

We used the Neuroblastoma Research Consortium (NRC) dataset (GSE85047), which includes gene expression and copy number variation (CNV) data mapped to the hg18 genome build. To ensure compatibility with modern genomic annotations, we performed a two-step liftOver of CNV regions:

* **Step 1:** hg18 to hg19 using the hg18ToHg19 chain file
* **Step 2:** hg19 to hg38 using the hg19ToHg38 chain file

Both steps were performed using the liftOver function from the rtracklayer R package. Unmapped regions, including chromosomes 23 and 24, were discarded. Successfully mapped regions were annotated with gene information (chromosome, start/end positions, gene symbols) from Ensembl using the biomaRt package (getBM() function, autosomal chromosomes 1–22).

CNV data in Circular Binary Segmentation (CBS) format were classified using thresholds from Depuydt et al. (2018):

* **Gain:** Mean CNV ratio ≥ 0.15
* **Loss:** Mean CNV ratio ≤ –0.25
* **Normal:** Ratios between these thresholds

These classifications were saved in a compiled file: gene\_level\_cnvS\_SBK\_hg18\_hg19\_hg38.csv.

**2. Expression and Metadata Processing**

Gene expression data were obtained from mRNA\_expression\_data.txt, and clinical metadata from 20111216\_NRC\_samples.xlsx. Sample IDs were standardized (lowercased, whitespace removed) to ensure consistency. Probe-level expression data were mapped to gene symbols using mRNA info.txt, with unmapped probes discarded and duplicate mappings resolved by retaining the first occurrence.

Expression values were normalized across genes using L2 normalization (sklearn.preprocessing.normalize). CNV and expression matrices were aligned, reshaped, and stored in an xarray.Dataset to support efficient multidimensional computation.

Metadata categories were defined based on **INSS stage** and **MYCN amplification**:

* **highstage\_amp**: Stage 3/4 with MYCN amplification
* **highstage\_sc**: Stage 3/4 without MYCN amplification
* **lowstage\_amp / lowstage\_sc**: Stage 1/2/4S with or without MYCN amplification

A total of 219 samples with complete expression, CNV, and metadata information were retained for modeling.

**3. Linear Modeling of Gene Expression**

For each gene, a linear model was fit using the stats package in R via the rpy2 interface in Python. The model used:

* **Dependent variable**: Normalized gene expression
* **Independent variables**: CNV status (gain/loss/normal) and metadata category

CNV status was treated as a categorical variable (reference: "normal"). Genes with >44 missing CNV values (out of 219) were excluded. Remaining missing values were handled by removing samples with NaNs for either CNV or expression.

Model outputs (coefficients, p-values, etc.) were stored as an xarray.DataArray, then flattened and saved as gene\_results\_lm\_SLB\_V4\_hg38.csv. This analysis is implemented in the Expression\_LinearModel\_Final\_SLB.py script.

**4. Chromosomal Arm and Gene Ranking**

Chromosomal arm ranking was conducted using the NB\_Ranking\_SLB.py script, based on CNV spread and model-based expression sensitivity.

**4.1. Frequency of CNVs (analyzeCNAspread)**  
We used CNV data from 556 high-risk NB patients (Depuydt et al., 2018; PDP dataset). Segmented CNV regions were mapped to genes (hg38), and penetrance scores were calculated as the proportion of patients with gains or losses per region.

**4.2. Group-specific CNA Analysis (rankNBaccordingToMYCNstatus)**  
Patients were grouped by MYCN amplification status. CNV frequency was calculated separately for each group, and only chromosomal arms altered in ≥25% of patients were retained. Additional filters included:

* **Gene density** (genes per region)
* **Gain/loss ratio**
* **Binomial test** for significant imbalance between gains and losses

Gene-level penetrance was computed per region.

**4.3. Multi-Criteria Gene Ranking**  
Genes within significantly altered chromosomal arms were ranked using three independent criteria:

1. **CNrank**: Frequency of CNV per gene
2. **Dosagerank**: Expression sensitivity to CNV, from linear model
3. **Riskrank**: Association with clinical risk or survival

Genes missing data in ≥2 of these categories were excluded. A final **composite ranking score** was calculated as:

combined\_rank = percentile(CNrank) × percentile(dosagerank) × percentile(riskrank)

This approach prioritizes genes that are frequently altered, dosage-sensitive, and associated with disease severity.

**Summary**

This comprehensive pipeline combines genome remapping, CNV and expression integration, metadata-based stratification, and multi-layered modeling to identify and rank potential neuroblastoma driver genes across chromosomal arms. The approach leverages both data-driven statistical modeling and biological filtering, providing a strong foundation for further functional validation.