Utilizing Transformer Models for Copy Number Calling

Objective
The primary objective of this project is to develop an advanced computational tool based on transformer models to accurately classify genomic sites for CNV analysis. This tool aims to address current limitations in CNV detection by leveraging the powerful capabilities of transformer models, known for their effectiveness in handling complex patterns in large-scale data.

Background and Significance
CNVs contribute significantly to genetic diversity and have been implicated in various diseases, including cancer, neurodevelopmental disorders, and autoimmune diseases. Traditional methods for CNV detection, such as array comparative genomic hybridization (aCGH) and quantitative PCR (qPCR), have limitations in terms of resolution, scalability, and the ability to handle complex genomic data. The proposed transformer-based tool aims to overcome these limitations, providing a more robust and comprehensive approach to CNV analysis.

Data Acquisition and Preprocessing:

Utilize publicly available genomic datasets, focusing on coverage and allele frequency (AAF) data.
Implement preprocessing steps, including normalization and scaling, to prepare the data for analysis by the transformer model.

Transformer Model Development

Develop a transformer model tailored for genomic data, focusing on the encoder mechanism to capture the sequential dependencies in genomic sequences.
Optimize the model architecture, including the number of layers, heads, and model dimensions, to suit the specific requirements of CNV detection.

Training and Validation

Train the model on labeled datasets from HPRC (human pangenome project) genomes run with hmcnc, an hmm based caller as reference.
Implement a rigorous validation process, using separate datasets to evaluate the model's accuracy and reliability.

Hyperparameter Tuning
Employ techniques such as grid search or Bayesian optimization to fine-tune the model's hyperparameters for optimal performance.