## Can Deep Learning Precisely Predict BRCA1 Gene Expression in Ovary Tissue? A Case Study with Enformer

To use on my capstone project, I selected the paper "Benchmarking of deep neural networks for predicting personal gene expression from DNA sequence highlights shortcomings." I was intrigued by this paper because it calls into question the degree to which deep learning models can actually predict gene expression from DNA sequences, a gigantic promise in the field, but one that does not always end up playing out as intended. Instead of analyzing the entire genome as they did, I chose to focus on one example and important one: BRCA1 expression in the ovary tissue.

BRCA1 is a gene with an essential role in breast and ovarian cancer. My main question was simple: can deep learning model Enformer, which was trained on the human genome, predict the level of expression of BRCA1 in the ovary tissue correctly? I was interested to know how its predictions compared with actual experimental data from the GTEx gene expression data in tissues.

### What I Did

To verify, I carefully prepared the DNA sequence surrounding BRCA1, about 393,216 base pairs long, such that Enformer would only act on it. I inputted this sequence into the model and reaped the prediction for the ovary tissue.

Enformer gives arbitrary units, as opposed to the TPM (transcripts per million) units that GTEx uses, but I was still able to get a rough idea of whether the model was predicting low, mid, or high expression.

For BRCA1 in ovary tissue:

- Enformer produced a value of around 0.23 (arbitrary units).
- GTEx has a median TPM of about 2.0.

Both suggest BRCA1 is expressed moderately in the ovary, so at least qualitatively, the model and the data agree. But since the units are totally different, I can't compare the numbers quantitatively. It's more of a "does this feel close?" type of test.

**What This Means (and What's Next)**

So far, things seem to make sense: Enformer does capture the general trend of BRCA1 expression in ovary tissue. But there are limits. Since there's no way to convert Enformer's output into real TPM values, I can't tell exactly how close the prediction is — just that it's in the right range.

For next steps, I'd like to:

- Test how well Enformer predicts BRCA1 expression in other tissues to make sure its tissue-specific patterns match what's expected.
- Try this same process with other genes linked to ovarian cancer.
- Find a way to make Enformer's outputs easier to compare with real data like GTEx — maybe by applying some kind of scaling or normalization.

**Final Thoughts**

I learned that while models like Enformer are exciting and can pick up patterns in DNA that line up with real gene expression, they aren't ready yet for making accurate disease-related predictions without more work. Still, this small project gave me a good sense of both the potential and the limits of deep learning in this area — and raised ideas for how to make these tools more useful in the future.