Anna Poetsch
Biomedical Genomics, Technische Universität Dresden

DNA sequence preferences of DNA damage and repair are dependent on the damage, the underlying epigenome and sequence content. Also transcription and replication determine sequence susceptibility and repair preferences. The blueprint for these other processes are themselves largely encoded, which creates a tissue-specific and individual balance between functionality and stability. Consequently, there is a sequence-inherent risk for somatic genome evolution, which leads to phenotypes of ageing and cancer.

We use DNA language models to investigate how stability is encoded in the DNA.

GROVER (Genome Rules Obtained via Extracted Representations) is a foundation model that we have built and fine-tune to learn probabilities of DNA double-strand breaks. Their location is enriched in DNA that encodes for early DNA replication. Interestingly, data for actual replication timing adds little additional information. H3K36me3 is a histone mark for active transcription and represents for the model a mark for tissue identity. Unlike replication timing, it is required for the most accurate DNA double strand break predictions.

EAGLE-MUT (Efficient Analysis with a Genome-wide LSTM to Evaluate per-nucleotide MUTation susceptibility) learns sequence context of single base substitutions in individual tumor samples. The distribution of mutation probability over individual genomes is heterogenous and individually different. The sequence patterns we derive from the model reveal motifs of up to 200 bp and define mutagenesis cold- and hot-spots. They suggest mechanisms of susceptibity to and protection from specific mutagenic processes that result from DNA damage and repair preferences. For example, the deficiency to resolve mismatches with gastric acid-associated oxidative DNA damage leads to very distinct mutation hotspots in gastric and esophageal adenocarcinoma.

Models like GROVER and EAGLE-MUT overcome the sparse nature of genome instability data, can be used in an interpretable way, and thus lead to novel mechanistic insight into the DNA sequence relationships with DNA damage, repair, and mutagenesis.

Host: Martin Svensson