Transfer Learning for Large-Scale Genomic AI in Cancer Genomics

J. Yu, M.Baudis

Department of Molecular Life Sciences and Swiss Institute of Bioinformatics, University of Zurich, Switzerland

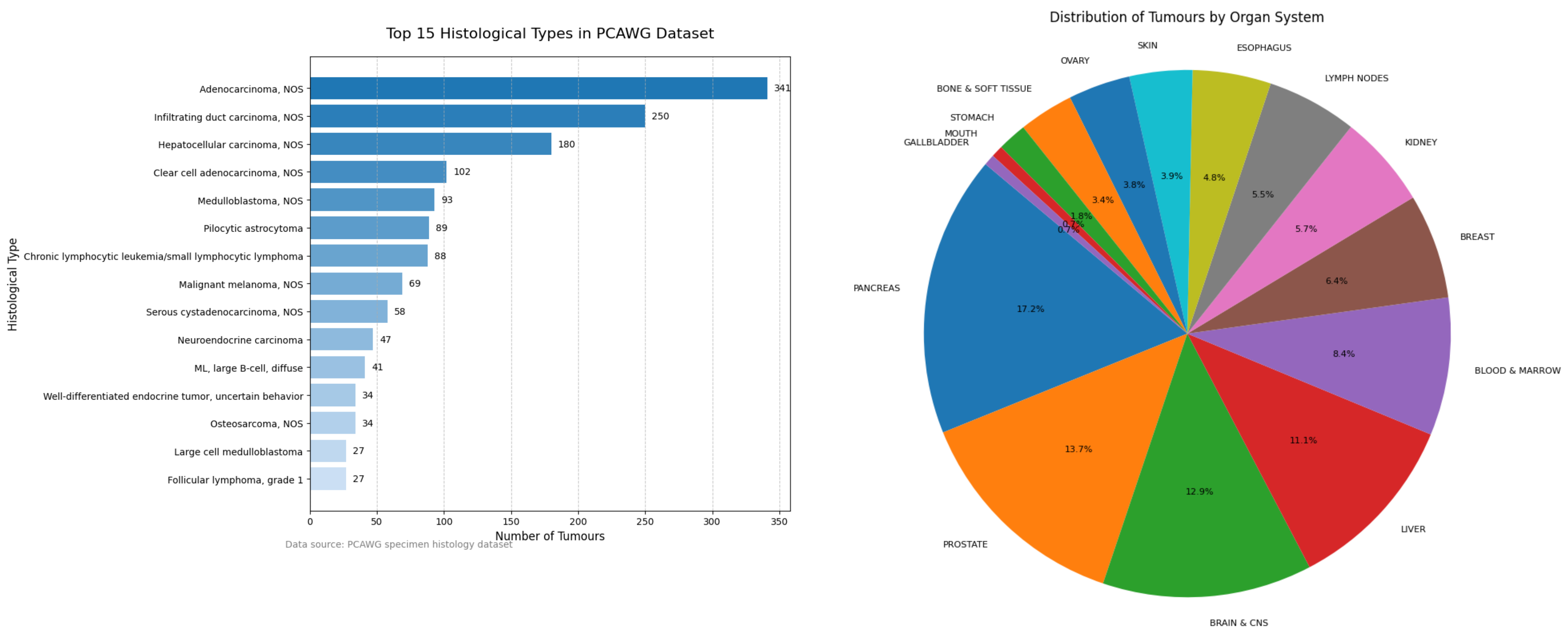
Genomic Language Models and pipeline

Genomic foundation models are deep neural networks pre-trained on billions of DNA sequences through self-supervised learning. By predicting masked bases across diverse genomic regions, they develop contextual embeddings that capture functional features from protein-binding motifs to genetic variants. These adaptable representations could enable transfer learning for specialized tasks with limited labeled data. We focus on applying this approach to cancer genomics.

Cancer Genomics Analysis Pipeline

Fig.1 shows our proposed pipeline for adapting the Nucleotide Transformer—a genomic foundation model—to cancer genomics. The pipeline begins by processing tumor/normal miniBAM files and variant calls, extracting focused 6kb genomic windows centered on mutations with padding tailored to each variant type. These DNA sequences are converted into tokens and used to fine-tune the model through two training strategies: masked language modeling, where 15% of bases are randomly hidden for the model to predict, and contrastive learning that directly compares matched tumor-normal sequence pairs to capture biologically relevant differences. The fine-tuned model is expected to generate genomic embeddings that capture general sequence context and cancer-specific patterns, enabling downstream applications without requiring extensive retraining and labelling. Possible downstream applications include tumour classification, variant pathogenicity scoring, structure variants breakpoint analysis, etc.

Fig 2. Top 15 Histological Types and Organ system distribution across 1788 PCAWG samples



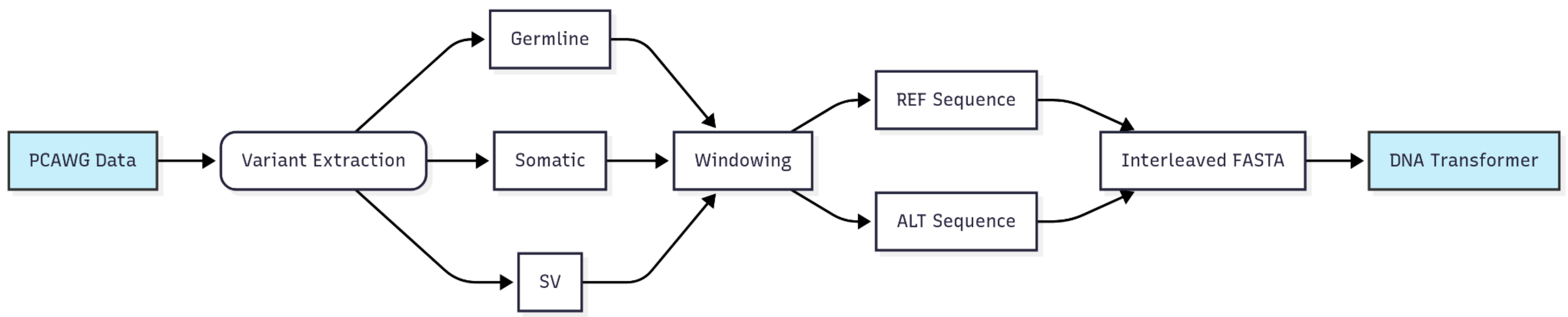


Fig 3. Data preprocessing pipeline: from BAM to paired REF/ALT windows

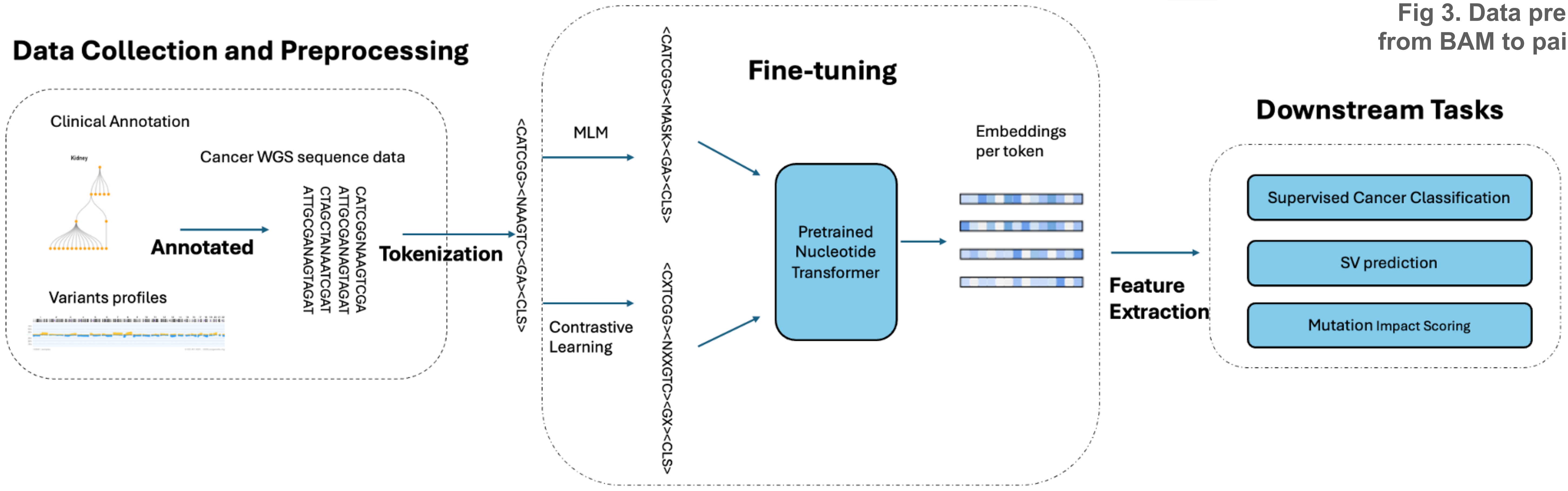


Fig 1. Proposed pipeline overview: from PCAWG variant-centric windows, through transformer fine-tuning, to downstream cancer-genome analyses.

PCAWG WGS Dataset

The Pan-Cancer Analysis of Whole Genomes (PCAWG) dataset is used as our primary data source. PCAWG is an international cohort of whole-genome sequences from many different cancers – it provides both the tumor genomes and their matched normal genomes for comparison. In total, we're using 1,788 tumor–normal pairs from PCAWG. We use the PCAWG "miniBAM" files. These are essentially reduced versions of the genome data that keep only the reads surrounding each called variant (reads within ±10 bp of each single-nucleotide variant, ±200 bp of small indels, and ±500 bp around structural variant breakpoints).

The PCAWG cohort covers a wide range of cancer types (as illustrated in Fig.2) – about 25 different organ systems (the top contributors include pancreas, prostate, brain/central nervous system, liver, and blood-related cancers) and 47 histological subtypes.

Preprocessing and Example Data

Fig.3 shows the data preprocessing pipeline. The preprocessing pipeline converts tumor/normal minibams and variant VCFs into compressed FASTA files with REF/ALT sequence pairs using variant-centered windows. These outputs provide tokenizable inputs for transformer fine-tuning. Figure 4 displays a preprocessed sample example with its window size distribution and the most frequent nucleotide patterns.

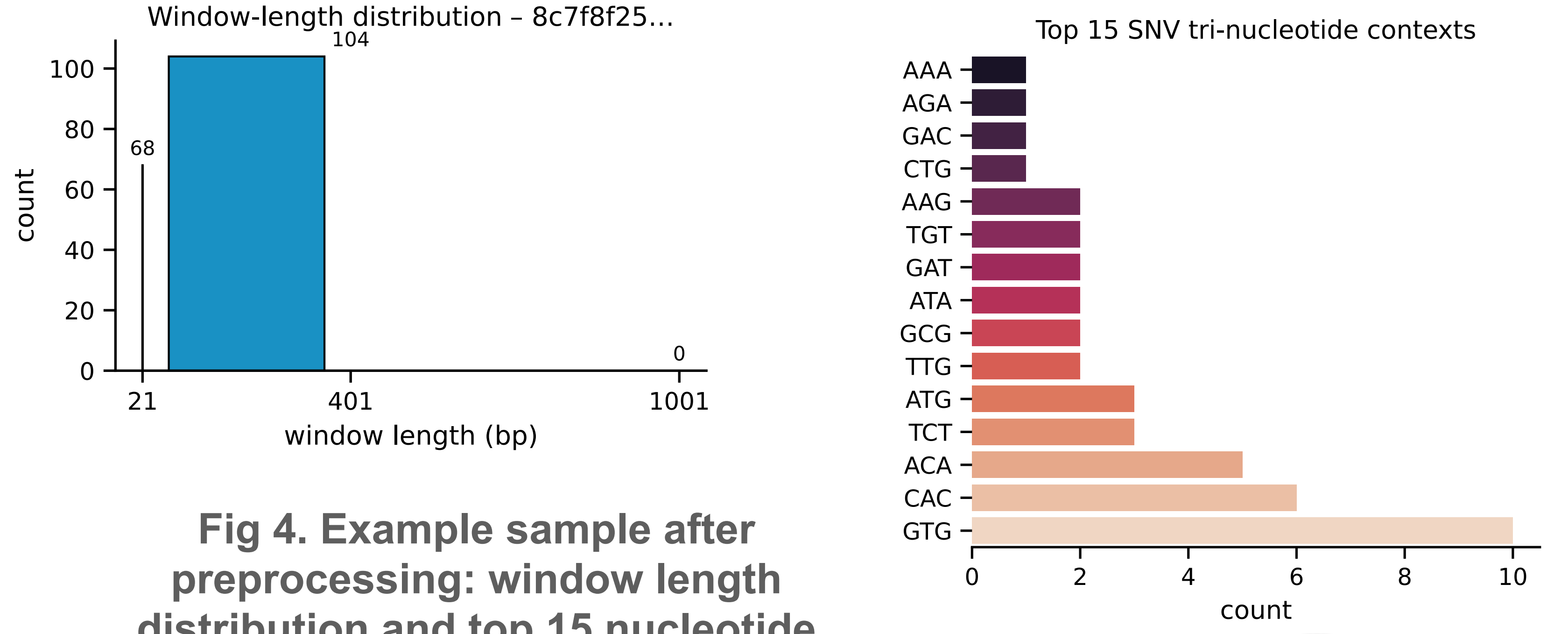


Fig 4. Example sample after preprocessing: window length distribution and top 15 nucleotide contents