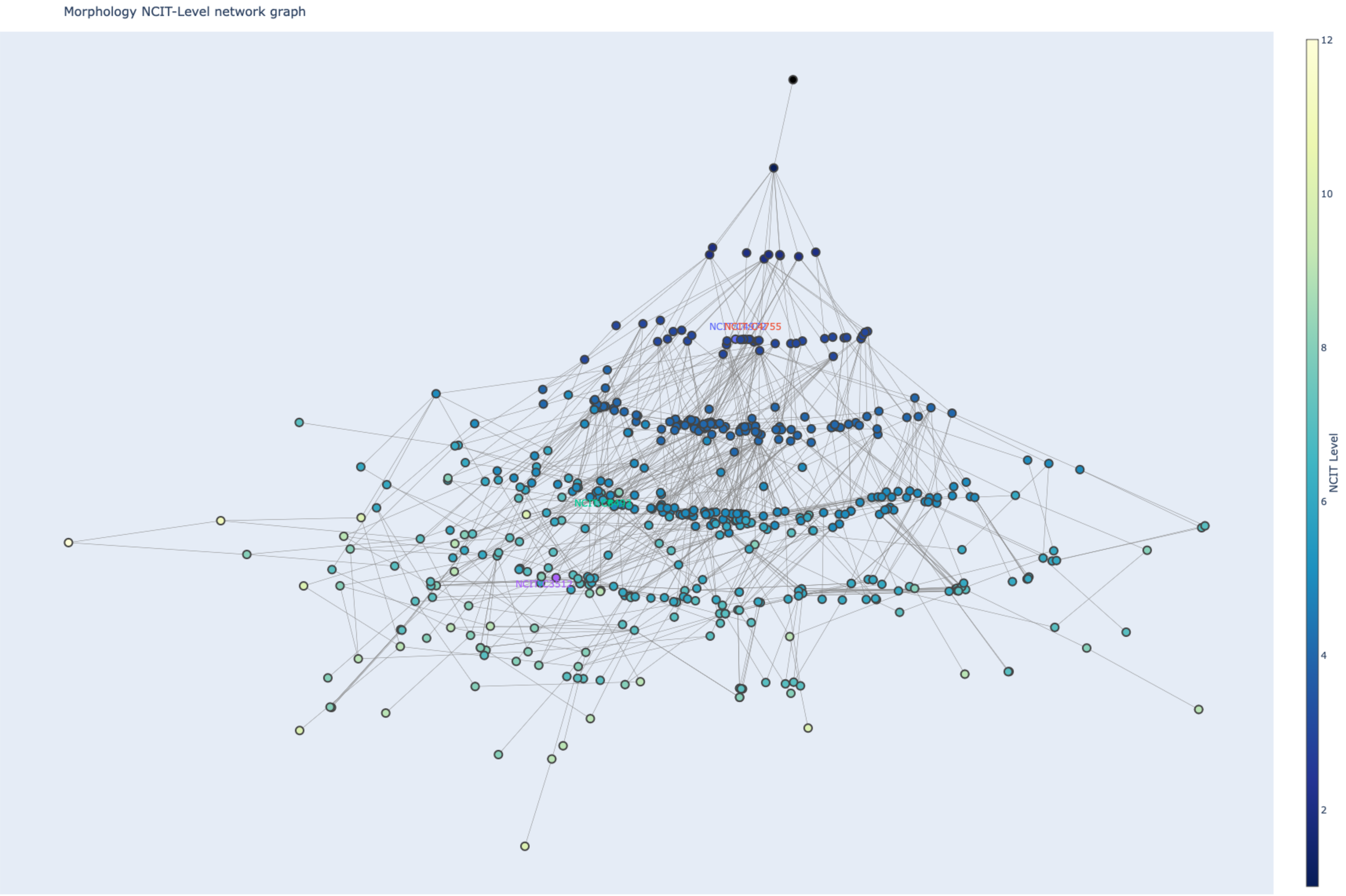
Transfer Learning for Large-Scale Genomic AI in Cancer Genomics

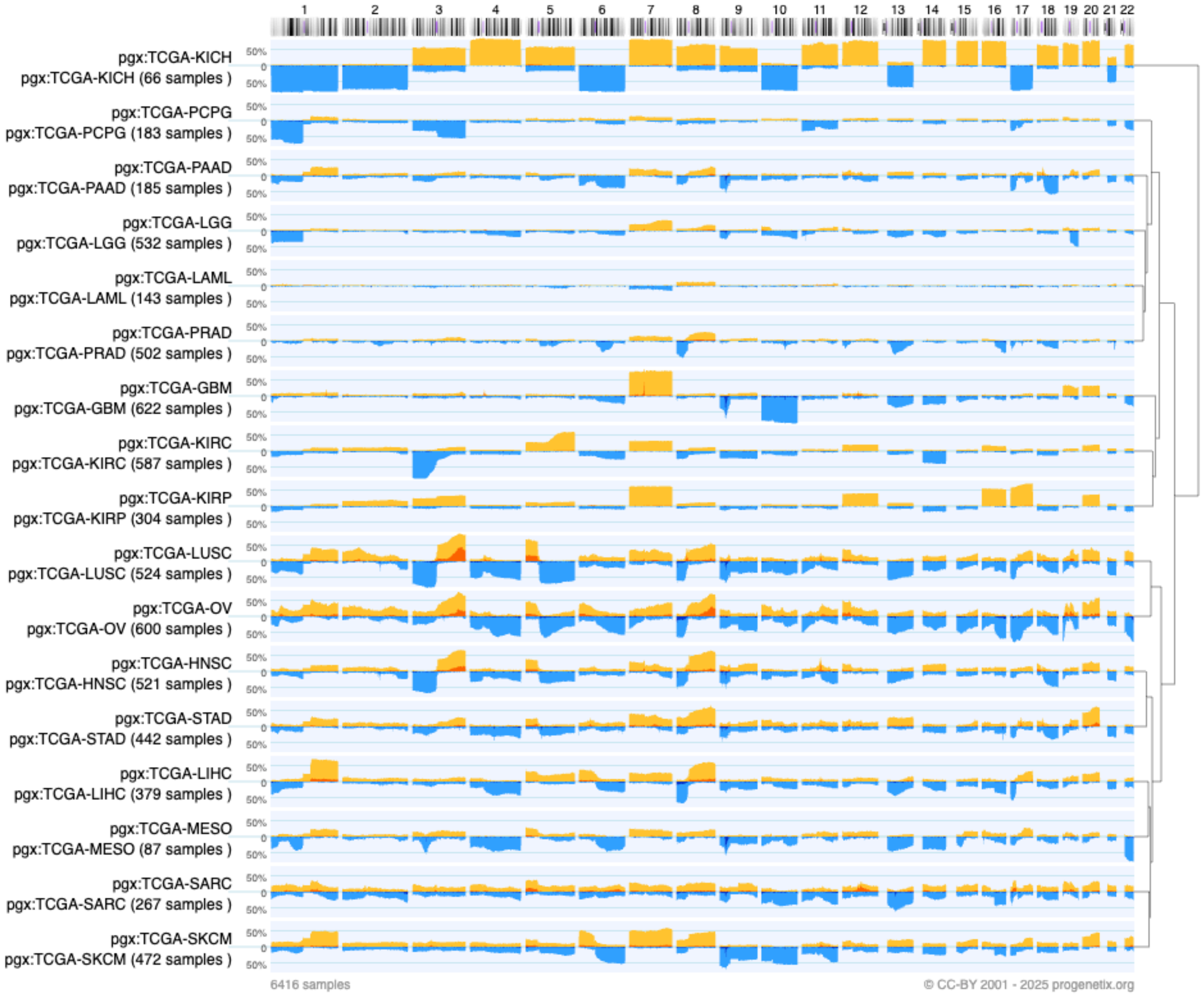
Jiahui Yu / Jun 14

Theoretical Cytogenetics and Oncogenomics Research | Methods | Standards

Genomic Imbalances in Cancer - Copy Number Variations (CNV)

CNV profiles heterogeneity vs cancer classification





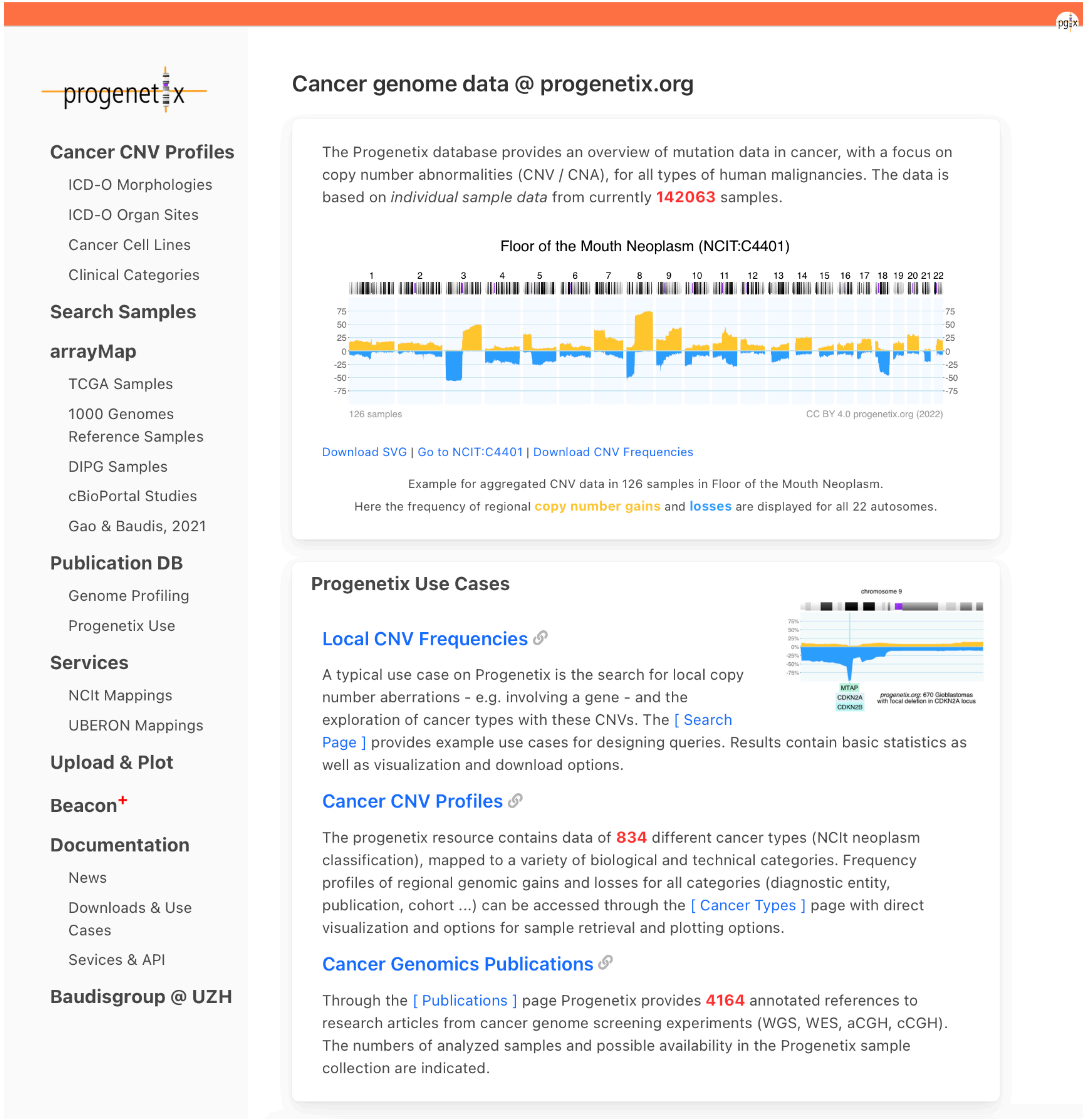
progenetix.org

Cancer Genomics Reference Resource

- ***open*** resource for oncogenomic profiles
- over **150'000 cancer CNV profiles**
- SNV data for some series (e.g. TCGA)
- more than **900 diagnostic types**
- inclusion of reference datasets (e.g. TCGA)
- standardized encodings (e.g. NCIt, ICD-O 3)
- identifier mapping for PMID, GEO, Cellosaurus, TCGA, cBioPortal where appropriate
- core clinical data (TNM, sex, survival ...)
- data mapping services

Universität Zürich^{UZH}

SIB Swiss Institute of Bioinformatics



Cancer genome data @ progenetix.org

The Progenetix database provides an overview of mutation data in cancer, with a focus on copy number abnormalities (CNV / CNA), for all types of human malignancies. The data is based on *individual sample data* from currently **142063** samples.

Floor of the Mouth Neoplasm (NCIT:C4401)

Download SVG | Go to NCIT:C4401 | Download CNV Frequencies

Example for aggregated CNV data in 126 samples in Floor of the Mouth Neoplasm.
Here the frequency of regional **copy number gains** and **losses** are displayed for all 22 autosomes.

Progenetix Use Cases

chromosome 9
75%
50%
25%
0%
-25%
-50%
-75%
MTAP
CDKN2A
CDKN2B
progenetix.org: 670 Gioblastomas with focal deletion in CDKN2A locus

Local CNV Frequencies

A typical use case on Progenetix is the search for local copy number aberrations - e.g. involving a gene - and the exploration of cancer types with these CNVs. The [Search Page] provides example use cases for designing queries. Results contain basic statistics as well as visualization and download options.

Cancer CNV Profiles

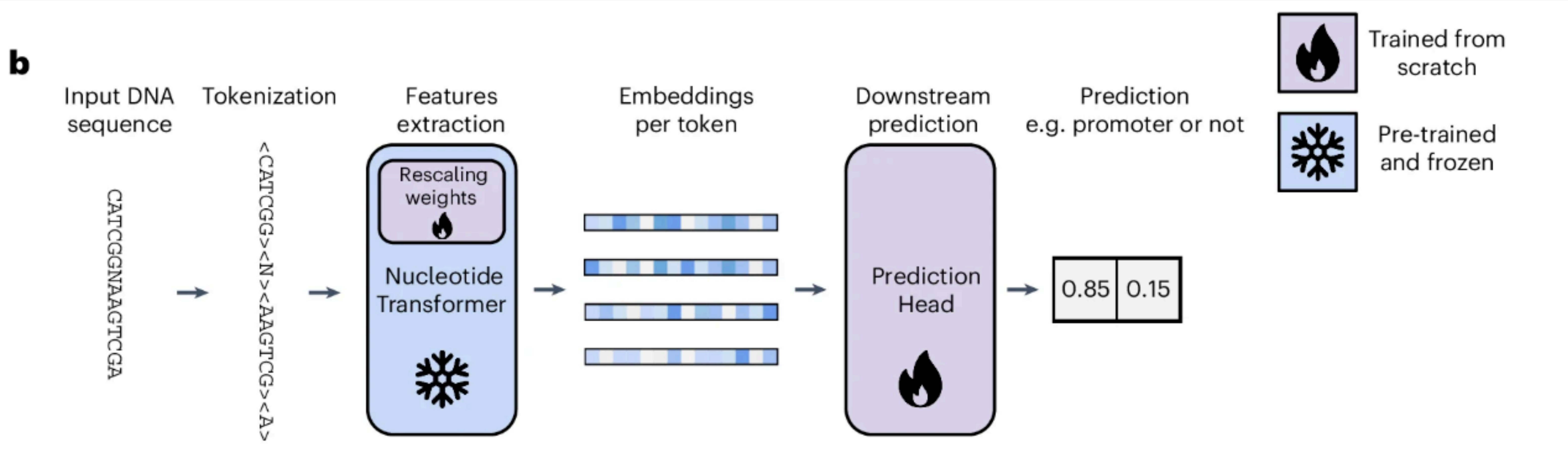
The progenetix resource contains data of **834** different cancer types (NCIt neoplasm classification), mapped to a variety of biological and technical categories. Frequency profiles of regional genomic gains and losses for all categories (diagnostic entity, publication, cohort ...) can be accessed through the [Cancer Types] page with direct visualization and options for sample retrieval and plotting options.

Cancer Genomics Publications

Through the [Publications] page Progenetix provides **4164** annotated references to research articles from cancer genome screening experiments (WGS, WES, aCGH, cCGH). The numbers of analyzed samples and possible availability in the Progenetix sample collection are indicated.

What is a Gemomic Foundation Model?

- What are genomic foundation models?
  - Self-supervised on terabases of DNA
  - Predicts masked K-mers or next token
  - Produces dense embeddings transferable to variant effect, TF binding, etc.
- Can we adapt such a model to real cancer WGS at scale?



Why Genomic Foundation Models?

Challenge	**How a foundation model *possibly* helps**
Sparse functional lables	Self-supervised pre-training (MLM / next-token) harvests unlabeled genomes; downstream fine-tuning needs fewer labels.
Cross-study bias (caller / pipeline differences)	Learns from the sequence context itself; a future dataset processed with a different caller lands in the same embedding space.
Finds hidden information between genome context and variants (captures long-range context information and handles variants uniformly)	Transformer context ≥ 6 kb (NT) or 100 kb (Evo2); a single embedding pipeline covers SNVs, indels, and SV; learn subtle context–variant relationships
Imbalanced Sample size across tumour types	Transfer learning lets us: pre-train on pan-cancer WGS → fine-tune on rare cancers with as few samples

Landscape of current genomic AI

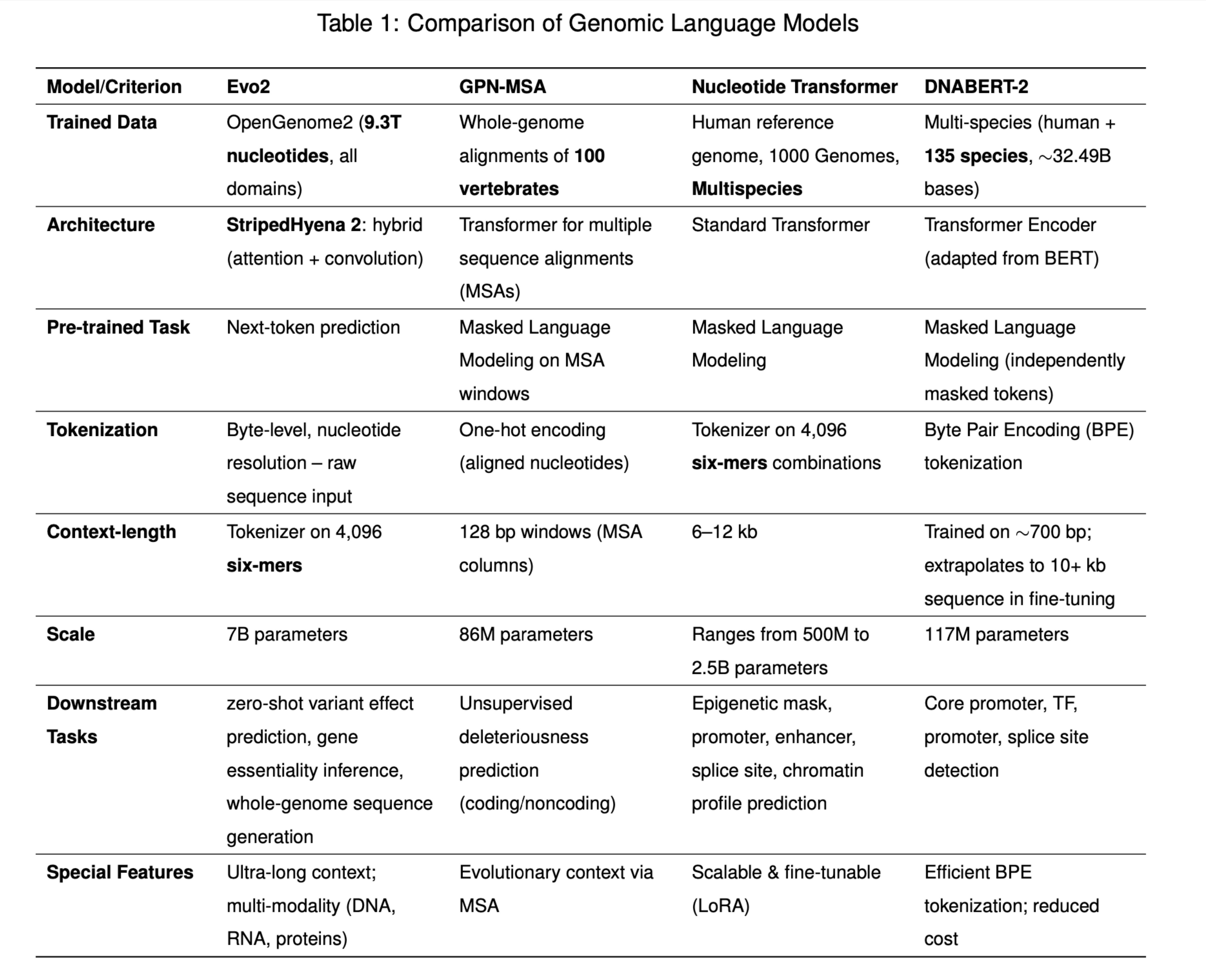


Table 1: Comparison of Genomic Language Models

Model/Criterion	Evo2	GPN-MSA	Nucleotide Transformer	DNABERT-2
Trained Data	OpenGenome2 (**9.3T nucleotides**, all domains)	Whole-genome alignments of **100 vertebrates**	Human reference genome, 1000 Genomes, **Multispecies**	Multi-species (human + **135 species**, ∼32.49B bases)
Architecture	**StripedHyena 2**: hybrid (attention + convolution)	Transformer for multiple sequence alignments (MSAs)	Standard Transformer	Transformer Encoder (adapted from BERT)
Pre-trained Task	Next-token prediction	Masked Language Modeling on MSA windows	Masked Language Modeling	Masked Language Modeling (independently masked tokens)
Tokenization	Byte-level, nucleotide resolution – raw sequence input	One-hot encoding (aligned nucleotides)	Tokenizer on 4,096 **six-mers** combinations	Byte Pair Encoding (BPE) tokenization
Context-length	Tokenizer on 4,096 **six-mers**	128 bp windows (MSA columns)	6–12 kb	Trained on ∼700 bp; extrapolates to 10+ kb sequence in fine-tuning
Scale	7B parameters	86M parameters	Ranges from 500M to 2.5B parameters	117M parameters
Downstream Tasks	zero-shot variant effect prediction, gene essentiality inference, whole-genome sequence generation	Unsupervised deleteriousness prediction (coding/noncoding)	Epigenetic mask, promoter, enhancer, splice site, chromatin profile prediction	Core promoter, TF, promoter, splice site detection
Special Features	Ultra-long context; multi-modality (DNA, RNA, proteins)	Evolutionary context via MSA	Scalable & fine-tunable (LoRA)	Efficient BPE tokenization; reduced cost

- Data: range from human to multi-species alignments.
- Scale: parameter sizes vary widely, from 86M to 7B.
- Context length: handles sequences from 128 bp to 10+ kb.
- Pre-training tasks: primarily use masked language modeling (MLM) or next-token prediction to learn sequence patterns.
- Downstream Applications: predicts variant effects, regulatory elements (promoters, enhancers), splice sites, and chromatin profiles.

Pipeline Overview

Data Collection and Preprocessing

Clinical Annotation

Cancer WGS sequence data

Annotated

Variants profiles

CATCGGNAAGTCGA
ATTGCGANAGTAGAT
CTAGCTANAATCGAT
ATTGCGANAGTAGAT

Tokenization

<CATCGG><NAAGTC><GA><CLS>

Fine-tuning

MLM

<CATCGG><MASK><GA><CLS>

Contrastive Learning

<CXTCGG><NXXGTC><GX><CLS>

Pretrained Nucleotide Transformer

Embeddings per token

Feature Extraction

Downstream Tasks

Supervised Cancer Classification

SV prediction

Mutation Impact Scoring

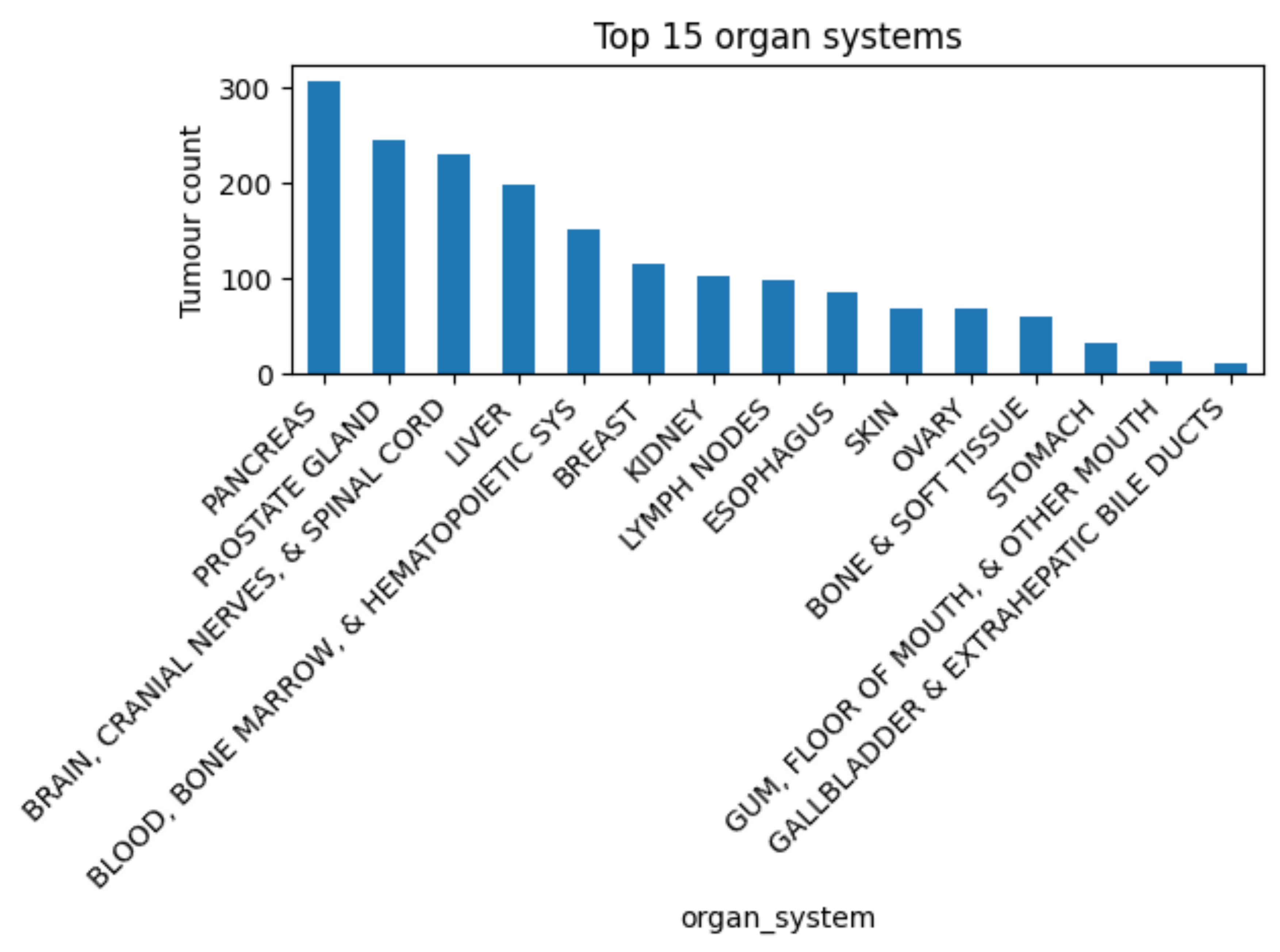
Dataset Overview

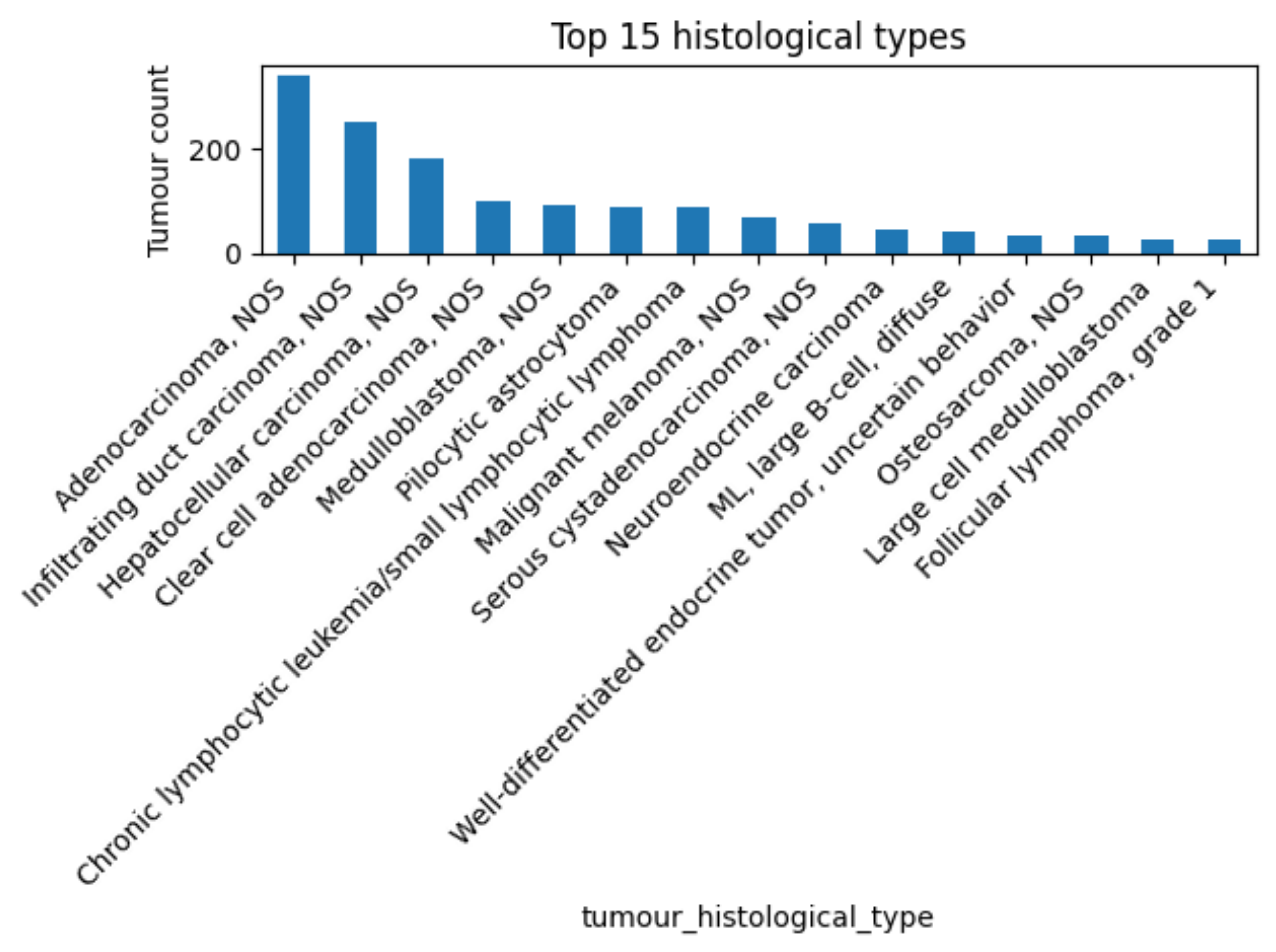
- *Pan-Cancer Analysis of Whole Genomes (PCAWG)* – an international WGS compendium of primary tumors and matched normals.

- The PCAWG miniBAM collection comprises 1788 matched tumor-normal whole-genome pairs, each reduced to only the reads supporting called variants (SNVs ±10 bp, indels ±200 bp, SV breakpoints ±500 bp).

  - 25 organ systems: pancreas, prostate gland, brain/cranial nerves & spinal cord, liver, and hematopoietic & lymphoid (top 5)

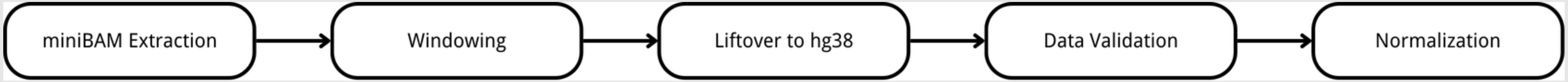
  - 47 histological subtypes (ICD-O-3): Adenocarcinoma, Infiltrating duct carcinoma, Hepatocellular carcinoma, Clear cell adenocarcinoma, Medulloblastoma (top 5)





Preprocessing

- Input: a set of 6 kb DNA windows (one per variant) for both tumor and matched normal.
  - For each somatic variant
    - Tumor window: 3000 bp upstream + somatic-alt allele + 3000 bp downstream
    - Matched-normal window: 3 000 bp upstream + normal/germline allele + 3 000 bp downstream.



Fine-tuning

- **Dual-Task Training**:
    - **Masked Language Modeling (MLM):** randomly mask 15% of tokens in each sequence and train the model to predict them, as in standard self-supervision. This helps the model refine its understanding of DNA context and recover mutations.
    - **Contrastive Pairing Task:** use a contrastive loss to bring together the representations of tumor vs normal sequence from the *same variant* and push apart those from *different variants*. The model learns to recognize that tumors/normal from the same locus are inherently related while any two sequences from different loci should be distinct.

Potential Downstream Applications

- **Cancer Type Classification:** Using the fine-tuned model's embeddings, predict a tumor's origin or subtype from its somatic mutation pattern.

- **Mutation Impact Scoring:** Beyond classification, the fine-tuned model could serve as a general predictor of variant effect – e.g., outputting an embedding that correlates with pathogenicity impact.

- **Structural Variant Breakpoint Prediction:** The model can be applied to detect or classify structural variants. For example, given a genomic region, the model might predict the likelihood of an SV breakpoint or distinguish true oncogenic rearrangements from artifacts. By training on known SV breakpoints (versus random genomic loci), the model's attention to sequence context may help identify hotspots prone to breakage. It could also predict the partner sequence of a breakpoint by embedding similarity (e.g., pairing two break-ends that should join).

- **Multi-modal Integration (Future):** While not covered in detail, these sequence embeddings could be combined with other data (gene expression, clinical features) for integrated models.

Future Steps & Challenges

- Will a model pre-trained on 1000G human genome data (germline genomes) transfer to somatic WGS?
    - Tumor mutations are out-of-distribution compared with healthy germline variation.
- Is 6kb context enough?
    - SV break-ends ±5–10 kb; enhancer–promoter loops span >10 kb.
- Class imbalance: 5 tumor types supply >50 % of samples → model could be biased toward these signatures.
- Evaluation
    - Downstream tasks benchmarks.
    - Cross-reference validation and domain shift test.