Modeling Immune Tolerance with Antibody Language Models: Simulating B Cell Selection and Somatic Hypermutation

Benjamin Nemoz, Karenna Ng and Bryan Briney

Scripps Research Immunology & Microbiology

Introduction

New B cells, each expressing a unique B cell receptor (**BCR**), are generated by the human immune system at an astonishing rate of **10–20 million per day** through somatic recombination. During affinity maturation in germinal centers, BCRs are further refined by somatic hypermutation (**SHM**), often at high rates. A selection process, peripheral tolerance, silences B cells with reduced affinity for their cognate antigen or newly acquired autoreactivity. We sough of applying **machine learning** (**ML**) approaches to understanding this process of SHM followed by immune selection. To address this, we generated extensive extremely high-accuracy antibody sequence datasets and aimed at developing antibody-specific large language models (**AbLLMs**) to mimic the SHM process governing B cell maturation in GCs.

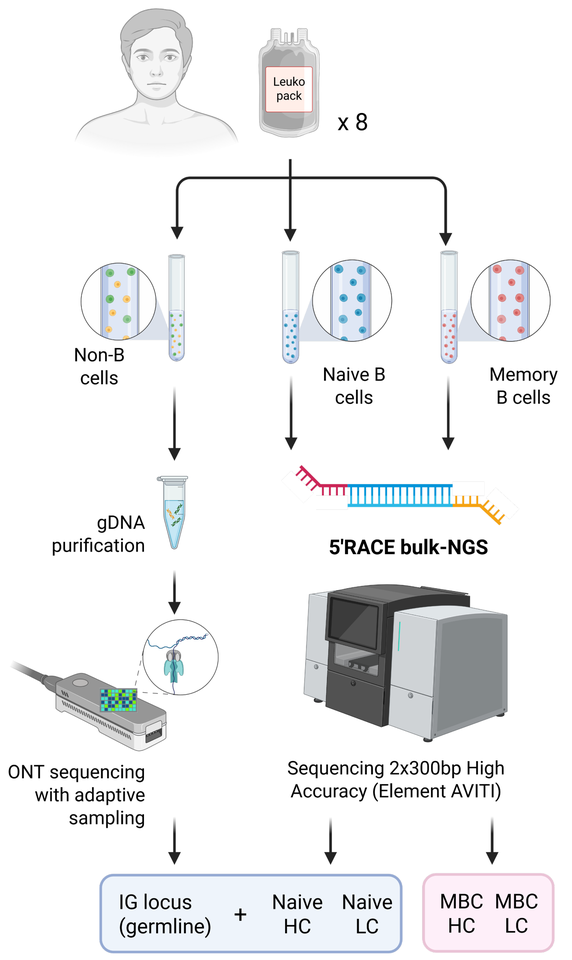

Figure 1: PBMCs from 8 healthy donors have been used to generate highly accurate database of antibody sequences.

Sequence data generation

PBMCs (500M per donor; n=8) were magnetically separated into non-B cells, naïve B cells, and memory B cells (**MBCs**) via **MACS**. Naïve and memory subsets underwent **5′RACE**-based bulk NGS of unpaired IGH/IGL chains, ensuring unbiased repertoire capture. Per donor, two biological replicates (6–8 technical replicates each) were pooled and sequenced (CloudBreak FS, Element AVITI, 600 cycles, 2×300 bp, 10 bp indexes; ~100M reads/donor).

Genomic DNA (**gDNA**) from non-B cells was **ONT**-sequenced (1 donor/flow cell, 72h, adaptive sampling targeting IG loci). Reads were basecalled (Dorado, super-accuracy mode) and assembled into contigs (Flye or Clair/EPI2ME). Germline IG inference was performed from gDNA using **Digger** and from recombined naïve B cell RNA using **IgDiscover**. Intersecting inferences yielded donor-specific IG databases.

Finally, MBC reads were deduplicated, UMI-corrected, and annotated using **AbStar** against donor-matched custom references, yielding high-accuracy unpaired IGH/IGL datasets along with inferred germline sequences.

AbLLMs training

We explored different model architectures using a sample dataset of **270k** sequences. Among explored architectures, **autoregressive architectures** seemed to yield the best results. We have adapted the **tokenization** of input sequences to use germline sequences and encode both position (IMGT numbering) and amino-acid mutation. Models are trained only on the prediction of mutation tokens (Figure 2). Three approaches were tested: ① conventional tokenization with joint prediction of position and mutation in one pass (**Multi-Task Learning, MTL**), ② newly proposed tokenization with joint prediction in one pass (standard autoregressive approach), and ③ conventional tokenization with separate prediction of position and mutations in two passes (**sequential approach**). In both MTL and sequential approaches, conventional tokenization separately encoded position and AA mutation.

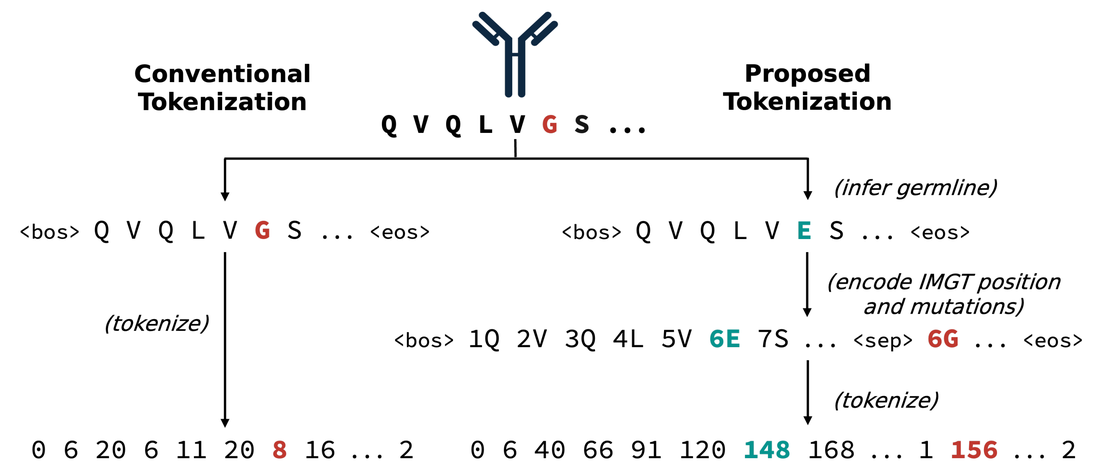

Figure 2: Proposed SHM tokenization. Conventional tokenization methods (left) use one token per amino acid and do not include positional encodings. Our proposed tokenization (right) explicitly encodes IMGT position with each amino acid and represents mutations (red) separately from the inferred germline (teal) to allow for mutation prediction with full sequence context in an autoregressive model architecture. Models are trained only on the predictions of mutation tokens (red).

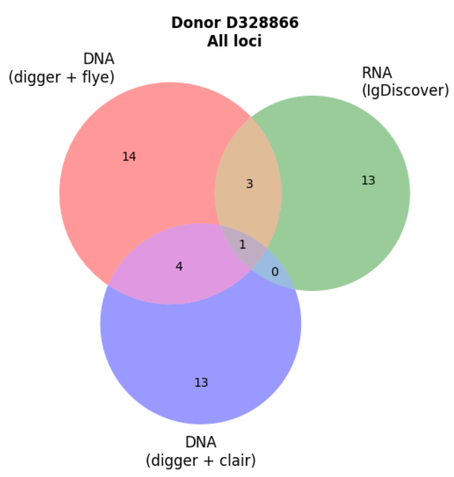

Figure 3: Intersection of germline gene results obtained using different methods. Red: inference from gDNA using Digger from contigs generated using Flye. Blue: inference from gDNA using Digger from contigs generated using clair/EPI2ME pipeline. Green: results obtained when inferring germline genes from RNA sequences using IgDiscover

Results

Sequencing experiment yielded a total of 637M reads, corresponding to **136M** unique high-quality antibody sequences. Inference of germline sequences from both gDNA and recombined RNA yielded diverging results with major discrepancies (Figure 3). We are currently investigating this to determine the ground truth for each donor. Preliminary training of AbLLMs models suggested the most suitable approach is to use separate encoding for position and mutation. Reduced scale models show that **initial mutations can be correctly predicted from germline sequences** (Figure 4 left). However, predicting higher-order mutation form low-level mutated sequences proves harder than expected. The **localization of predicted mutations** clustering around highly mutated regions (i.e., CDRs, Figure 4 right) suggest that SHM can be correctly learnt and predicted by AbLLMs upon adequate training. Further refinement of model architectures and training of full-scale models will ultimately enable antigen-agnostic prediction of somatic-hypermutations, allowing the prediction of antibody maturation if combined with antigen-driven selection predictive models (*in silico GC*).

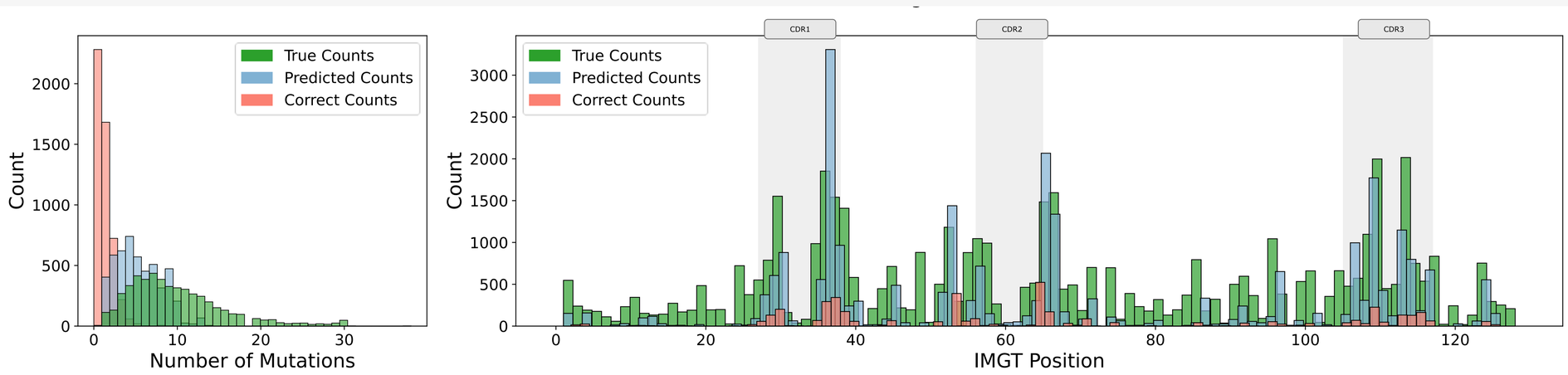

Figure 4: Left: Total number of unique true and predicted mutations per sequence. Right: Number of predicted mutations by IMGT position. Insertions (denoted in IMGT by a number followed by a letter) were collapsed into the bin of their corresponding IMGT number only for clarity.

Scripps Research Science Changing Life

NIH

CV SB Center for Viral Systems Biology

multi-omics MOVE vaccine evaluation