Automated Generation of HML Files for HLA Typing Results

Overview

Help the Cornejo Lab with an urgent research task! We are participating in the 19th International HLA & Immunogenetics Workshop (IHIWS 2026) in Numazu, Japan, from May 19 to 24, 2026. This conference is the leading meeting for HLA sequencing and immunogenetics research, bringing together academic researchers and commercial vendors to collaboratively establish new standards and benchmarks for future HLA research and product development.

The primary initiative of the IHIWS conference involves the independent sequencing of diverse human cell lines from the Fred Hutchinson Cancer Research Center's BPB biorepository. The thorough characterization of these samples by independent researchers will establish new open-source ground truth consensus datasets to help evaluate the accuracy of emerging HLA typing technologies.

In preparation for this conference, the Cornejo lab ordered de-identified DNA from the Hutch and generated high-quality HLA DNA sequence data. To participate in the conference, we need to share our data in an XML-based file format called "Histoimmunogenetics Markup Language (HML)." This standardized format enables the sharing of HLA genetic information, which the conference organizers can easily parse to evaluate the performance of different methods and technologies. Our lab is new to the HLA typing space and has not previously worked with this file format. We need help automating the generation of these files!

Goal

Develop a Python program that parses HLA typing results and produces a validated .hml (Histoimmunogenetics Markup Language) file.

Deliverables

1. One validated .hml file conforming to the official HML 1.0.1 schema
2. A flexible, command-line-executable Python program for writing .hml files, with inline comments explaining the code logic
3. A README.md file describing usage instructions and example commands

Inputs

1. HLA star allele calls for 32 samples (data/allele_output.csv)
2. HLA coding DNA sequences for 32 samples (data/all_CDS.fasta)
3. HLA full gene DNA sequences for 32 samples (data/all_gene.fasta)
4. Additional metadata. As you become familiar with the .hml format, please create a list of the additional information you need from the lab to populate the file.

Details

- Ideally, the Python program will be flexible enough to be applied to additional datasets.
    - It would be great if the program used argparse for input specifications and customization
- It would be great if the program were hosted and distributed on GitHub

Helpful Resources

- nmdp Public XML Schemas: https://schemas.nmdp.org/
- HML publication: https://doi.org/10.1016/j.humimm.2015.08.001
- Formatting specifications: https://www.sciencedirect.com/science/article/pii/S0198885915004346
- Schema: https://schemas.nmdp.org/spec/hml/1.0.1/hml_1_0_1-example7-ngsFull.xml
- Data: https://drive.google.com/drive/folders/1U5IN4tNtFJPvrJzzm4NKyizTbJlcmCzc

- Think of this as the file format for sharing genetic test results, what does a genetic test report look like? This is what the format looks like and how this info would be retained and stored clinically → healthcare marketability, learning how genetic test reports are created and distributed and stored.