Clark Center Room E300
318, Campus Drive
Stanford University
Stanford, CA, 94305
USA
Email: fabio.zanini@stanford.edu

September 24, 2017

To the attention of the Early Career Clinical Research Symbiont Award Committee

Dear Sir or Madam:

As a graduate student in the group of Richard Neher at the Max Planck Institute for Developmental Biology in Tübingen, Germany, I have collected a comprehensive dataset on genome evolution of the Human Immunodeficiency Virus (HIV). In collaboration with Prof. Jan Albert at the Karolinska Institute in Stockholm, Sweden, we have processed more than 70 plasma samples from 10 untreated HIV human subjects: each infection is followed longitudinally for up to 15 years. We then deep sequenced the virus population from each sample, describing to an unprecedented level of detail the massive genetic changes that occur during natural HIV infections. The dataset is exceptional because nowadays antiviral treatment is administered immediately upon HIV diagnosis, lowering viral titers to undetectable levels; this makes it impossible to study the genetic mutations accumulated by the virus, the very same mutations that make HIV so difficult to eradicate. The dataset is also remarkably complete, since it covers the whole viral genome with great sequencing depth; this allows for many kinds of secondary analyses beyond the scope of the initial publication (DOI: 10.7554/eLife.11282).

The publication sparkled interest in a mixed audience spanning from medical doctors to computational data scientists. Traditionally, the 100,000,000 sequencing reads would be deposited into a sequence read archive (SRA). However, the complex structure of the dataset – different patients, time points, more than 10,000 genetic sites of interest, a number of clinical metadata – is largely obscured in this process, creating a high barrier for downstream research. To facilitate sharing, in addition to uploading the reads to an SRA, I also developed a web application, hosted at the URL `https://hiv.biozentrum.unibas.ch/`, that greatly simplifies data exploration. The website allows both human interaction, for instance via animated plots, and automated mining of data and metadata via a coherent programming interface. It is anecdotal but revealing that I have been able to leverage the same website to discuss with both wet lab virologists and machine learning experts. The application is designed such that everything is an endpoint; this makes it trivial to source the data from any connected device, from a personal laptop to a virtual machine in the cloud. Moreover, the backend powering the website has a modular architecture that allows easy expansion for derived projects – for instance, it has since be used to estimate time of infection of HIV patients. The source code of the application is also available at `https://github.com/iosonofabio/hivwholeweb`.

Several secondary analyses accessing the web application have been reported. I attach the bioRxiv preprint of Hartl. *et al.* (2016): the authors scan the dataset extensively to understand the molecular mechanisms driving natural selection in HIV evolution. In absence of the web application, they would have had to access the sequence read archive, download all the reads, reconstruct the structure of data and metadata, and recall all the single nucleotide polymorphisms. Data cleaning would have been especially tedious but critical, because the naturally high variability of the HIV genome is entangled with sequencing errors. Thanks to the web application, however, Hartl. *et al.* could download all mutation abundance tables simply by directing their web browser to the appropriate URLs.

As more and more complex analyses are developed, I hope that both dataset and web application will continue to be used to deepen our understanding of HIV and as a springboard to study and treat other human viruses.

URL: `https://hiv.biozentrum.unibas.ch/`

Attachment: Hartl_et_al_2016.pdf

Yours Faithfully,

Fabio Zanini
Postdoc, Quake lab
Stanford University