Data science for genomic surveillance: predicting and evaluating pathogen risk from genomic data

PI: Eric J. Ma, Runstadler Lab, MIT

Summary

With pathogen genomic surveillance, we want to be able to predict pathogen risk (phenotype) from its sequence (genotype). However, “risk” is a nebulous term. We think that “risk” should be decomposed into biochemically measureable phenotypes, each of which inform different aspects of a pathogen’s risk. Following this line of thinking will allow us to model and predict future pathogen risk in a mechanistic and interpretable fashion. To catalyze the achievement of this goal, we will start by generate a freely-available, gold-standard and standardized dataset that matches influenza neuraminidase genotype to measured neuraminidase drug resistance (DR) phenotype. This will be accomplished by generating libraries of neuraminidase mutant proteins, and measuring their resistance to neuraminidase inhibitors in a standardized enzyme assay. We will then model the folding of the neuraminidase protein variants, and feed them into a deep learning model to learn the “biochemical fingerprints” that quantitatively predict its resistance. Bn10 funding for generating this dataset and carrying it through to an interpretable model will allow us to catalytically close the data gap between pathogen sequence and the assessment of its “risk”. Ultimately, we hope to expand this dataset beyond neuraminidase inhibitor resistance to other biochemically measurable aspects of influenza risk, and beyond influenza to other pathogens. We anticipate that beyond its public health applications, biochemists, drug development groups, and evolutionary biologists will be able to leverage this high quality and rich data resource for further research.

Rationale for Bn10 funding

The idea of generating a phenotype catalogue is not new, and has been done before with the HIV Drug Resistance Database. Yet, funding remains limited for such efforts, mainly because of the cost and scale of experimentation involved. By changing the format of an existing assay into a scalable format, our proposal aims to deliver insight into one protein's phenotype within the $40K scale of the Bn10. This builds towards the vision of an epidemiology dashboard with insights gained from safe and scalable biochemical assays. The catalytic step we take here is in "going deep" with one phenotype – where we generate a library of influenza neuraminidase variants, measure their drug resistance phenotype in a low-cost and high-throughput fashion, and develop interpretable deep learning models to predict novel neuraminidase sequence phenotypes.

Members of the Broad Institute have spearheaded initiatives to sequence emerging and current viral pathogens (e.g. the Viral Genomics group, the Sabeti lab). Our proposal is in line with the Broad *Next10* goals by developing more scalable methods to recognize virulent states, integrated with the use of genomic sequencing data. This will further cement the Broad's leadership in data generation, management, and analysis. Additionally, the techniques described in this proposal utilize concepts from the Massively Parallel Reporter Assay (MPRA), which was invented at the Broad institute, and we can foresee more technology development partnerships between the Runstadler lab and the BTL.

Project Plan

Currently, pathogen risk determination is done by looking for genetic markers experimentally associated with a particular phenotype[1]. These genetic markers are determined in an ad-hoc fashion, and reported mutations do not include their genetic context (i.e. epistatic interactions) that give rise to their phenotypes. Additionally, there are reports of conflicting effects of mutations[2]. A catalogue of experimentally determined protein phenotypes from a library of genotypes, has the potential to:

1. improve the predictive capacity of public health agencies, informing the deployment of precision strategies for containing the spread of infectious disease agents,
2. reduce the need for public health agencies to conduct expensive screens of circulating strains,
3. accelerate drug development by providing genotype and phenotype data for chemical modelling, and
4. further deepen our basic knowledge of infectious disease biology.

Apart from the HIV Drug Resistance Database[3], which has a series of viral protein genotypes matched with quantitative drug resistance phenotype measurements, this resource does not exist for any other pathogens, preventing rational risk assessment for a variety of viral pathogens. This project will kick-start the development of a comprehensive database of viral phenotypes, first starting with influenza neuraminidase resistance to three neuraminidase inhibitors (oseltamivir, zanamivir and peramivir), then expanded to other influenza virus proteins' phenotypes, and finally to other viral pathogens. This data will be made freely available, through the Broad Institute, and will be useful to a wide range of individuals, from epidemiologists to molecular biologists to data scientists.

With this proposal, rather than going "wide" (i.e. developing the technologies for measuring multiple phenotypes using the funding received), we aim to go "deep", demonstrating how we can generate the necessary data for a single protein and one of its phenotypes, and take it through to an interpretable and predictive machine learning model. Beyond this initial effort, we envision that it can then be integrated with other models into an epidemiology dashboard.

Aim 1: Adapt a known neuraminidase activity assay to develop a bacterial-based biochemical screen for neuraminidase inhibitor resistance.

In order to screen for neuraminidase inhibitor resistance, we require an experimental assay amenable to rapid screening at scale without needing specialized equipment. The NA-Fluor assay (Thermo Scientific), which is used for testing influenza neuraminidase inhibitor resistance, is currently designed to be a 96-well microplate assay. While amenable for high-throughput processing, this is not currently in a format amenable for quick screening of a large library of neuraminidase mutants for their susceptibility to inhibitors.

To achieve our goal of rapid screening, we will modify the NA-Fluor assay into a LacZ-like blue/white screen, where we the expressed neurminidase will generate a fluorescence signal in the presence of the NA-Fluor assay substrate when viewed under UV light. We will first test whether a single neuraminidase, expressed intracellularly, can be detected with *E. coli* when plated on nutrient agar with the NA-Fluor assay substrate spread on its surface. We will also explore whether exporting the neuraminidase extracellularly with a bacterial protein export tag can help with the dynamic range of neuraminidase activity detection.

In addition to qualitative inspection, we will also use image analysis methods to quantify neuraminidase activity. Briefly, we will design and 3-D print a custom mount for a Raspberry Pi camera module with a UV light filter, which can be mounted on top of a UV transilluminator. We will also use scikit-image, an image analysis library for the Python programming language, to perform image segmentation and brightness quantification.

To establish assay accuracy, we will test mutants of the active site of the neuraminidase that are known to disable neuraminidase activity, and verify that their activity is quantitatively diminished.

To assay neuraminidase inhibition, replica plates of transformed *E. coli* cells carrying the neuraminidase variant will be treated ± inhibitor. In order to assay resistance, we will be looking for mutants that exhibit a large quantitative difference in fluorescence level ± inhibitor, as well as known mutants which have decreased fluorescence activity (without inhibitor) relative to a wild-type control. This will be compared against the standard micro-well plate assay format to establish a standard curve for comparison and determination of assay dynamic range.

Aim 2: Screen a large (~10^3-10^4) library of neuraminidase mutants for inhibitor resistance.

There are two strategies that we will use to generate the neuraminidase mutants. Our primary strategy is to use tiled DNA synthesis. Here, the neuraminidase gene is split into 150-mer overlapping tiles. Using microarray synthesis, individual point mutations are encoded into oligomer tiles, and are assembled using overlap extension PCR. The fallback strategy is to use error-prone PCR to generate a library of neuraminidase mutants, and clone them in a pooled fashion with a short 20-mer sequence barcode. The purpose of the short barcode is to act as a sequencing surrogate to avoid expensive Sanger sequencing across the 1.4 kb neuraminidase gene[a]. We will use PacBio sequencing with deep coverage over each mutant-barcode set to establish the mutant-barcode correspondence. The mutant plasmids will be transformed into *E. coli* cells. Replica plating followed by image-based measurement of fluorescence will be used to determine inhibitor resistance. Select mutant colonies will be identified for (1) neuraminidase activity level in the absence of inhibitors (i.e. ranging from weak to strong) and (2) a range of drug resistance phenotypes (i.e. ranging non-resistant to strongly resistant). We will sequence their barcodes to identify the variant associated with that quantified phenotype. For further validation, select mutants will be chosen, and their neuraminidase resistance will be quantified using the microplate format inhibition assay.

Aim 3: Deep Learning to Model and Identify Biochemical Fingerprints that Predict Drug Resistance

The Harvard Intelligent and Probabilistic Systems group has developed a deep learning method for learning the fingerprints of a chemical structure that are most predictive of some phenotype[4]. By using structural models of molecules, the models are highly interpretable (unlike ensemble machine learning methods). We will extend this method, in collaboration with their group, to identify the biochemical fingerprints of the neuraminidase protein structure that are most predictive of drug resistance phenotypes.

In order to do this, we will create homology models of the sequenced variants, and convert their modeled structure into a protein interaction graph. We will adapt their neural fingerprinting Python package to accept a protein interaction graph (instead of a chemical structure graph). Following this, we will train a convolutional neural network to model and identify the biochemical regions on the protein structure that are most indicative of high or low drug resistance. For model evaluation, not only will we do standard cross-validation with the data on hand, we will generate a number of synthetic mutants *in silico*, and predict their drug resistance phenotype using our trained model. We will then use site-directed mutagenesis to create these individual mutants, and measure their resistance values using both the plating screen and microplate formats.

[a] The nature of the assay does not permit pooled sequencing, which will dramatically reduce sequencing costs; we expect that the development of a pooled assay will help solve this problem. The number of isolates actually screened will depend on the budgetary room available after initial experiments are done, though we have budgeted for a minimum of 2,000 isolates at ~$5/sequence. We anticipate bringing the cost down by bulk purchasing sequencing orders ahead of time.

We will also build a web-based interface that presents the predictions along with matched sequence-resistance data. In the interest of open science, all code involved will be released publicly to Github, and versions will be archived using Zenodo.

Future Directions: Beyond Neuraminidase Inhibitor Resistance

Whether an influenza virus is "risky" or not is defined by a combination of underlying biochemical phenotypes. Beyond drug resistance, and beyond the scope of the Bn10 proposal, we will test other aspects of influenza, such as replication rate, dampening of host immune factors, and recognition by current antibody repertoires. Our proposed efforts to expand the toolkit of biochemical assays are addressed in a separate Bn10 proposal.

Deliverables

By the end of the 1-year period, we expect to deliver:

1. A barcoded plasmid library of influenza neuraminidases, which may be of interest and leverage-able by other groups.
2. Genotype-phenotype data and associated machine learning models, delivered on a web interface.
3. A manuscript describing the method, insights, and public interface generated from this large-scale phenotyping effort.

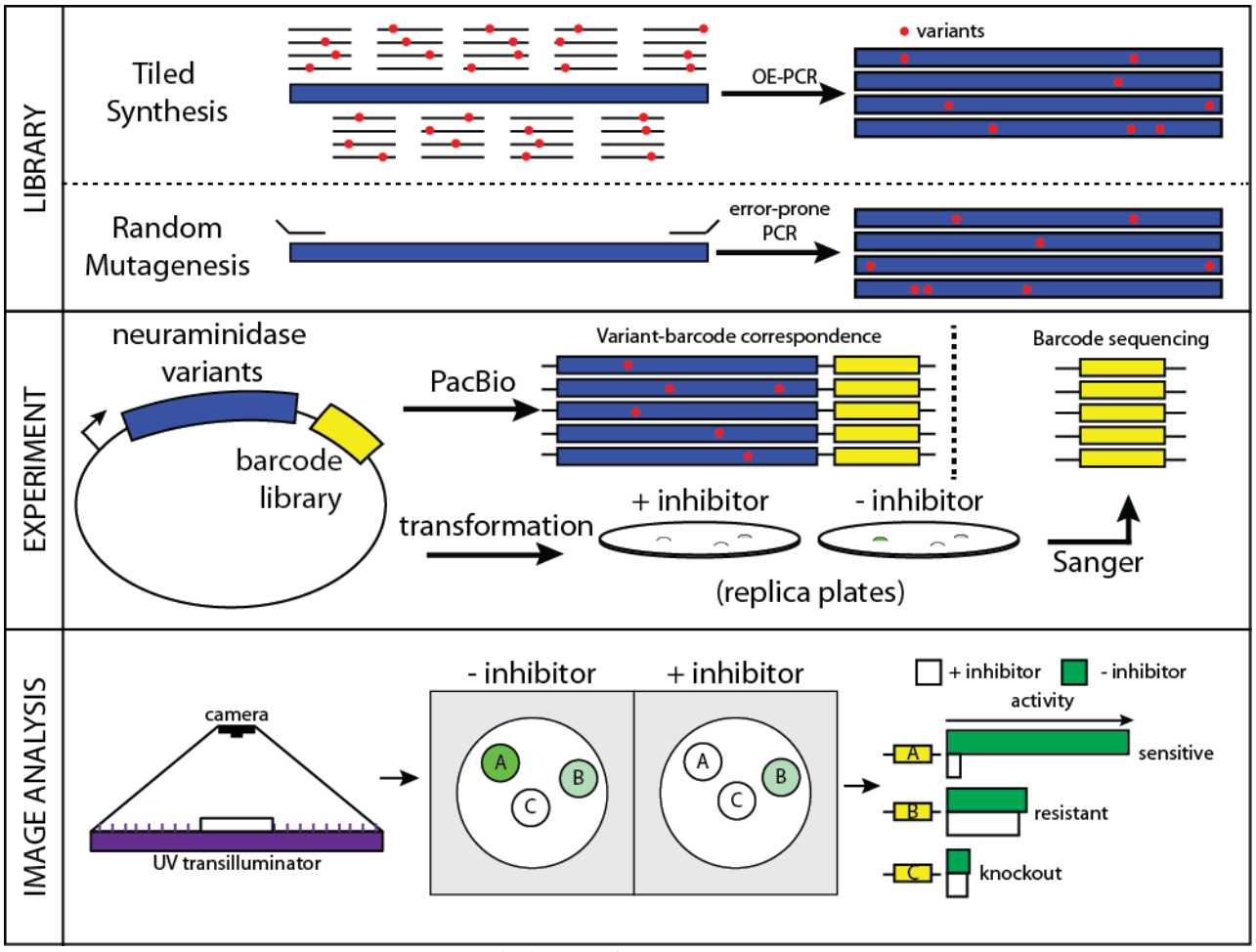

Figure 1. Experimental schematic from variant generation to screening.

Broad Platform Usage

We expect to partner with the Broad Technology Labs to use their PacBio sequencer to establish variant-barcode correspondence, as well as use the microarrays for gene synthesis. However, in the event that the BTL is at capacity, we will also be able to access PacBio sequencing via the BioMicroCenter at MIT.

We expect that the data will be of interest to the emerging Data Science platform at the Broad, and will spark discussions and collaborations on new quantitative biology problems that emerge, such as those surrounding the problems of inferring epistatic interactions from sparsely populated data.

Space Requirements

Laboratory space at the Broad is not required, as the requisite experiments can be carried out in the Runstadler Laboratory.

Regulatory Compliance

No regulatory compliance issues will need to be solved, as we are dealing with non-replicative agents under a purely biochemical system. No patient data will be used either.

Statement of Engagement at the Broad

Though my main research is conducted on the MIT campus, my current engagement at the Broad Institute has grown considerably over the past year. I have attended the Infectious Disease Program and Stats/Math Reading Club (SMRC) and Models, Inferences and Algorithms (MIA) meetings, and have given talks at individual research group meetings (Sabeti group) and at Broad seminars (Infectious Disease Program). Additionally, I have delivered a hands-on workshop on how to do machine learning in Python, co-organized with the Broad NextGen. If the proposal was awarded, engagements of this type will not change. Through this Bn10, I anticipate bringing a closer collaboration between the Runstadler lab and other groups at the Broad, such as the Sabeti lab, the viral genomics group, and the Data Science Program.

Budget Justification

Category	Items	Upper-bound cost
Materials	NA-fluor kits (x3)	$3,000
Materials	DNA synthesis (x1)	$6,000
Materials	PCR kits (x2)	$2,000
Materials	Chemically competent cells	$1,000
Materials	Standard Bacteriology Materials (LB, agar, plates)	$2,000
Methods	PacBio Sequencing (multiple runs)	$4,000
Methods	Sanger sequencing (2000 runs for 2000 variants)	$10,000
Computation	Amazon EC2 compute time	$2,000
Manpower	UROP Stipends (summer + 2 semesters)	$10,000
	Total	$40,000.00

References

1. Fan, S. *et al.* Novel residues in avian influenza virus PB2 protein affect virulence in mammalian hosts. *Nat Commun* **5,** 5021 (2014).
2. Hurt, A., Leang, S., Speers, D., Barr, I. & Maurer-Stroh, S. Mutations I117V and I117M and Oseltamivir Sensitivity of Pandemic (H1N1) 2009 Viruses. *Emerging Infect. Dis.* **18,** (2012).
3. Rhee, S.-Y. *et al.* Human immunodeficiency virus reverse transcriptase and protease sequence database. *Nucleic Acids Res.* **31,** 298–303 (2003).
4. Duvenaud, D. *et al.* Convolutional Networks on Graphs for Learning Molecular Fingerprints. *arXiv.org* **cs.LG,** (2015).