Scalable mapping of viral polymerase genotype to phenotype

PIs:

- Eric J. Ma, Runstadler Lab
- Anthony Kulesa, Blainey Lab
- Jared Kehe, Blainey and Sabeti Labs

Summary

Countermeasures to viral pandemics such as therapy development and containment are enacted only once an outbreak emerges. The future of epidemiology is to address new threats preemptively. The deployment of high throughput sequencing methods (e.g. Illumina and nanopore sequencing, automated field sample processing) allow us to obtain circulating viral genotypes at large scales but predicting outbreak risk from genetic information alone is largely intractable. Measuring biochemical properties like replication rate and antibody binding closely associated with pathogenicity (*1*), could enable preemptive risk assessment but are limited in throughput to a 1-by-1 basis. With over 100,000 pathogen genotypes recorded in public sequence databases, and thousands being added every year, developing assays that yield biochemical pathogenicity information at this scale is a critical challenge.

We aim to develop a toolbox of massively parallel biochemical measurements (e.g. antibody binding, genome replication, drug resistance) to lower costs and increase throughput by >2 orders of magnitude. **As a catalytic first step, we will develop and conduct a pooled measurement of viral replication rate from >10,000 Influenza polymerase variants in a single assay, matching the scale of available genotypes (~30,000 Influenza genotypes)**. Our assay and reagents will be readily adaptable to emerging RNA viruses (e.g. Zika), establishing a Broad Institute resource that can be rapidly deployed as new viral genotypes are sampled. With this proof of concept we aim to acquire funding to develop analogous assays for other biochemical properties, bringing rapid risk prediction and preemptive countermeasure development closer to reality.

Rationale for Bn10 Funding

We envision a future where epidemiologists forecast viral outbreaks in real-time to counter their evolution and spread. Currently, epidemiology is mostly reactive: only after an outbreak arises like the pandemic Influenza H1N1 (2009), Ebola (2014), or Zika (2016) do we collect data on a large scale to analyze virulence and drug susceptibility.

The Broad is pioneering genomic surveillance efforts to sample and sequence outbreaks as they occur. During the 2014 Ebola outbreak, Institute core member Pardis Sabeti and colleagues sequenced ~100 Ebola virus genomes, identifying key epidemiological parameters such as evolution rate (*2*).

With the falling cost of sequencing and innovations in sample collection, barriers have shifted to gaining functional insight into the many mutations observed. Of ~30,000 Influenza genomes in public databases, only a handful have any link to quantitatively measured biochemical (*3*-*9*). Additionally, the Los Alamos HIV Sequence Database has over 300,000 publicly deposited sequences, but the largest database of measured biochemical phenotypes contains only ~1000 HIV variants (*10*). The gap between genotype and phenotype is clear; without the capability to generate such data at scale, it will be impossible to generate predictive models of pathogen risk, and surveillance will remain reactive in posture.

To surmount these barriers, we propose to develop a toolbox of pooled biochemical assays of viral fitness. Biochemical readouts (e.g. antibody binding, genome replication, drug resistance) obtained in a massively parallel and inexpensive way can provide a direct indication of pathogenicity. Bn10 funding will catalyze the development of this toolbox by funding our first tool: a low-cost, high-throughput assay of replication rate for >10,000 rationally designed Influenza A polymerase variants**.** This critical first step will position us to apply for more funding to continue to develop similar assay strategies for immunogenicity (antibody binding), and drug resistance (enzyme activity).

We foresee the growth of this toolbox launching new collaborations with researchers throughout the Broad, and directly engaging the Broad Scientific Vision to understand and predict infection outcome, evolution, and spread. Bn10 funding also catalyzes our involvement as bioengineers in the Infectious Disease Program, from which we expect many new future collaborations.

Project Plan

Being a zoonotic pathogen, Influenza's reservoir hosts are wild birds, and pathogenicity is limited until the virus adapts to human cells. The virus polymerase replication rate is a key biochemical function closely associated with pathogenicity (*11*). A popular gold-standard assay used to measure this property co-expresses luciferase with viral UTRs in human cells (*12*, *13*). The viral polymerase replicates the luciferase mRNA, and fitter variants create a brighter signal. This assay has been used to sensitively identify novel mutations that vastly improve the fitness of the polymerase during host switching from birds to human (Figure 1).

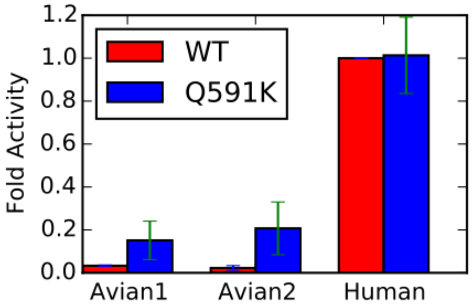

Figure 1. Activity (in human cells) of two avian-origin (Avian1, Avian2) and a human-origin (Human) Influenza polymerase, with (red) or without (blue) a single point mutation. Activity values are normalized to the wild-type human polymerase. The Q591K mutation increases avian-origin polymerase activity, but has no effect on human polymerases, clearly showing that it is sufficient but not necessary for increasing polymerase activity in human cells.

While it is the gold-standard assay, the luciferase reporter technique is low throughput and currently performed on a 1-by-1 basis in well-plates. Even with automation and robotics, the 1-by-1 basis limits the scope of investigation into Influenza fitness, costing >$10/variant assayed[1]. In contrast, a pooled assay where all variants are measured in parallel could drastically improve throughput and cost.

Here we propose to adapt the Massively Parallel Reporter Assay (MPRA) (*14*) to create a pooled assay for replicative activity of ~10,000 Influenza A polymerase variants, reducing costs and improving throughput by >2 orders of magnitude.

Deliverables

We foresee the following deliverables:

1. Protocols for our pooled assay, generalizable to all viruses where the gold-standard luciferase assay is used (e.g. Hepatitis C (*15*))
2. A plasmid library with balanced representation of optimally designed barcodes, and Influenza A UTR regions for rapid deployment to new sequences
3. A publicly available dataset of genotype-phenotype relationships for ~10,000 Influenza A polymerase variants, representative of all known genotypes from the Influenza Research Database (*16*)

[1] This cost is estimated from an R21 proposal (previously submitted by Eric Ma & Jon Runstadler), computed by averaging the cost of materials over the total number of variants that was expected to be synthesized.

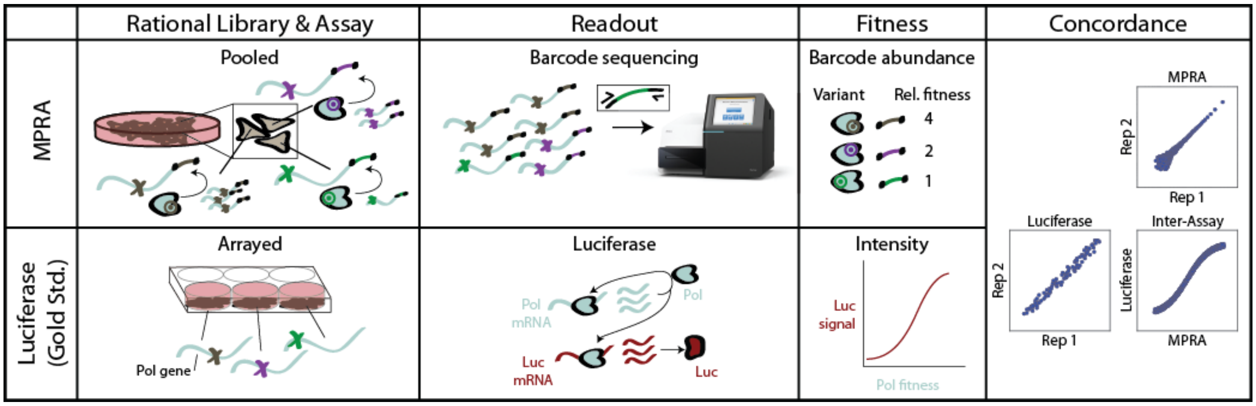

Figure 2. Experimental schematic for our proposed experiments (MPRA) and its comparison to the gold-standard luciferase assay. Assay design, readout, and fitness interpretation are shown. Concordance within and between experimental assays will be measured.

Our pooled assay strategy parallelizes the gold-standard luciferase assay to >10,000 variants

In the current gold-standard assay, a viral polymerase exponentially replicates both its own mRNA and a coexpressed luciferase gene fused to viral 5' and 3' UTRs. Fitter polymerase variants replicate more luciferase mRNA, resulting in more luciferase protein and a brighter optical signal. We propose to substitute this optical readout with a sequenceable reporter. Rather than using luciferase as a proxy, we will directly assess the fitness of a polymerase from the abundance of its own RNA. The length of the polymerase gene itself (~2kb) is too large for a single Illumina read, so each variant will receive a short 3' barcode (~20bp). By measuring the abundance of these short barcodes in parallel using standard RNA-seq protocols, all variants can be assayed in a single experiment (Figure 2).

Random sequence barcodes are assigned via random ligation and PacBio sequencing

In order to measure fitness of each polymerase based on the abundance of its barcode sequence, the correspondence between polymerase variants and barcodes must be known ahead of time. In a pooled reaction, we will randomly ligate each polymerase variant to a 3' barcode. We will sequence using PacBio to cover the full gene and barcode (full length ~2kb) in a single read. Given the size of the variant library (~10^4), read length (~2kb), and desirable error rates (<1 error/variant), we expect that the full library can be sequenced in one SMRT cell.

The number of variants (~10,000) is far smaller than the available barcoding sequence space (>10^{12} for a 20bp barcode), so we expect all barcodes to be distinguishable with up to 3 substitution errors in barcode detection (viral polymerase replication error, sequencing error, etc). The barcodes will be ligated just after the stop codon, and just before the 3' UTR of the polymerase gene, such that it does not affect expression or translation of the polymerase.

Comparisons to gold-standard will determine assay noise and systematic errors

To rigorously evaluate our assay against the gold-standard luciferase assay, we will compare dynamic range, technical noise, and systematic sources of error. To cover the full dynamic range of both assays, we will take two orthogonal strategies. First, we will establish a small library of polymerase variants (~24) that span the fitness spectrum through site-directed mutagenesis. Second, we will titrate the polymerase expression using an inducible promoter to cover 2 decades of expression.

To measure correspondence between our assay and the luciferase assay, cells cultured in well-plates will be transfected with a luciferase construct (with viral UTRs) and a polymerase variant in our reduced library (~24), or exposed to a different polymerase induction level. The luciferase signal will be measured from each well, and then each will be processed for RNA-sequencing of the sequence reporter. To determine the relationship between technical noise and number of cells and number of barcode transcripts sequenced, we will bootstrap sequencing reads from each well to levels ranging from 10-10,000 cells. By running independent sequencing runs, we will explore the effect of different normalization strategies on reducing batch effects.

To measure barcode abundance independent of differences in amplification efficiency, barcode cDNA will be ligated to short, random Unique Molecular Identifiers (UMIs) to allow digital counting of the unique molecules.[2] We expect that this strategy will minimize technical noise arising from amplification differences or noise in sequencing. This is a common strategy to achieve maximum accuracy in single-cell whole transcriptome sequencing (*17*).

Sequencer	**# reads/run**	**$/run**	**# variants/ run**	**$/variant**
MiSeq	~$1x10^7$	$600-2000	5,000-10,000	$0.06-0.4
HiSeq	~$1x10^8$	$2000-4000	50,000-100,000	$0.02-0.08

Microarray synthesis allows parallel design of 10,000 variants
The scale of our assay admits both rationally designed and randomly mutated polymerase variants. We plan to synthesize a library of ~10,000 polymerase variants representative of current and past circulating strains (in the Influenza Research Database).

Our synthesis strategy involves tiled assembly of 150bp oligos synthesized via microarray. For each 150-bp tile of the polymerase ORF, we will synthesize all known non-synonymous variants available in the database. Variant tiles will be mixed with wild type sequence tiles, and full polymerase ORFs will be randomly assembled by overlap extension PCR such that each full ORF has 1-3 variant tiles. We will achieve this synthesis scale by applying the Broad Technology Labs' 12K microarray oligo synthesis service.

Potential Fallbacks/Alternatives
Alternatives to pooled sequencing strategy
If we encounter difficulties in our assay strategy (small number of variants dominating over all others), we propose an alternative strategy using flow cytometry that maintains the pooled nature and increased throughput. Briefly, the luciferase reporter will be substituted with a GFP reporter such that fitter polymerases correspond to a brighter GFP signal. Cells will be sorted into GFP expression bins, and then the barcodes in each bin will be sequenced to identify the polymerase variants.

[2] To maximize our dynamic range, we expect to sequence our library to saturation such that the expected number of unique molecules (UMI-tagged and amplified) not observed is less than 1. To achieve these statistics, we expect to read each UMI-tagged molecule ~20 times. Assuming a fitness distribution where the mean barcode abundance is ~10 times the lower bound, we expect the average number of counts per barcode to be ~2000 = (20 UMIs x ~10 molecules at lower bound x 10 mean/lower bound).

Alternatives to tiled library construction

In the event of difficulties in our tiled synthesis strategy, we aim to use error-prone PCR to generate a library of random polymerase variants. This has the advantage of being conceptually and logistically simpler to execute. To circumvent the problem of generating large numbers of non-functional variants, we will first apply a selection step. We will allow the polymerases to self replicate competitively in cell culture, and then create a viral cDNA library that only contains the ~10,000 most fit variants.

Future Work

Beyond the scope of this proposal, we aim to both generalize this assay to other viruses (e.g. hepatitis C (*15*)) and begin developing other new tools for pooled biochemical assays of viral fitness. For example, recent innovations in yeast display (*18*) and pooled protein binding assays (*19*) could measure the affinity of anti-Influenza antibodies to novel strains at scale.

Broad Platform Usage

We expect to enlist the assistance of the Broad Technology Lab for both the microarray synthesis of our polymerase variants and PacBio sequencing. The assay we have proposed also builds on MPRA, a previous Broad Technology Labs innovation, and we intend to rely on their expertise during development. We have consulted Scott Steelman, and can provide a letter of support if necessary.

As our assay relies on the Illumina MiSeq and HiSeq, we expect to make use of walk-up sequencing offered by the Broad Genomics Platform.

We also expect to engage the upcoming Data Science Platform (DSP) at the Broad for help building predictive models of pathogenicity from data generated by our assay. Unraveling epistatic genetic effects revealed in our data will likely develop new computational problems.

Space Requirements

No additional laboratory space is required. All necessary work will be carried out in the Blainey and Runstadler lab space.

Regulatory Compliance

All polymerase cloning and replication experiments will be conducted in BL2 labs, namely the Blainey Lab and the Runstadler Lab. The Runstadler lab has approved BL2+ facilities for all viral work included in this proposal.

No patient samples will be used.

Statement of Engagement at the Broad

Eric Ma's main research is conducted on the MIT campus under affiliate member Jonathan Runstadler, but his current engagement at the Broad Institute has grown considerably over the past year. He attends the Infectious Disease Program and Stats/Math Reading Club (SMRC) and Models, Inferences and Algorithms (MIA) meetings, and has given talks at individual research group meetings (Sabeti group) and at Broad seminars (Infectious Disease Program). Additionally, he has delivered a hands-on workshop on how to do machine learning in Python, co-organized with the Broad NextGen.

Tony Kulesa is a full time PhD student in the laboratory of core faculty member Paul Blainey. He regularly attends and program meetings in Cell Circuits, Science of Therapeutics, and Infectious Disease, where he presented twice in 2015. With Georgia Lagoudas, he led the writing workshop to support trainees through the Bn10 round 3 process and has supported several Bn10 applicants as a Communication Fellow.

Jared Kehe recently joined the Broad community to pursue his PhD under the co-advisement of core faculty members Paul Blainey and Pardis Sabeti. In addition to attending Infectious Disease Program meetings at the Broad, he is actively involved in the Center for Microbiome Informatics and Therapeutics directed by Broad Institute core faculty Ramnik Xavier and MIT professor Eric Alm. As a board member of the Center’s student-run Microbiome Club, he fosters community-wide seminars and discussions on topics ranging from microbial ecology to infectious disease.

Budget Justification

Type	Item	Capacity	Cost
Methods	PacBio Sequencing and Library Construction	2 runs	$8,000
Materials	Microarray synthesis	2 runs	$6,650
Methods	Illumina Sequencing (capacity for ~10 runs)	~10 runs	$20,000
Materials	Cell Culture Reagents	N/A	$5,000
Total			$39,650

References

1. M.-S. Song *et al.*, The polymerase acidic protein gene of influenza a virus contributes to pathogenicity in a mouse model. *J. Virol.* **83**, 12325–12335 (2009).

2. D. J. Park *et al.*, Ebola Virus Epidemiology, Transmission, and Evolution during Seven Months in Sierra Leone. *Cell*. **161**, 1516–1526 (2015).

3. B. W. Jagger *et al.*, The PB2-E627K mutation attenuates viruses containing the 2009 H1N1 influenza pandemic polymerase. *MBio*. **1**, e00067–10–e00067–10 (2010).

4. W. Song *et al.*, The K526R substitution in viral protein PB2 enhances the effects of E627K on influenza virus replication. *Nat Commun*. **5**, 5509 (2014).

5. G. Gabriel, V. Czudai-Matwich, H.-D. Klenk, Adaptive mutations in the H5N1 polymerase complex. *Virus Res.* **178**, 53–62 (2013).

6. S. Fan *et al.*, Novel residues in avian influenza virus PB2 protein affect virulence in mammalian hosts. *Nat Commun*. **5**, 5021 (2014).

7. L. V. Gubareva, D. V. Novikov, F. G. Hayden, Assessment of hemagglutinin sequence heterogeneity during influenza virus transmission in families. *J. Infect. Dis.* **186**, 1575–1581 (2002).

8. S. E. Hensley *et al.*, Hemagglutinin receptor binding avidity drives influenza A virus antigenic drift. *Science*. **326**, 734–736 (2009).

9. J. C. Pedersen, Hemagglutination-inhibition test for avian influenza virus subtype identification and the detection and quantitation of serum antibodies to the avian influenza virus. *Methods*

Mol. Biol. **436**, 53–66 (2008).

10. R. W. Shafer, Rationale and uses of a public HIV drug-resistance database. *J. Infect. Dis.* **194 Suppl 1**, S51–8 (2006).

11. B. W. Leung, H. Chen, G. G. Brownlee, Correlation between polymerase activity and pathogenicity in two duck H5N1 influenza viruses suggests that the polymerase contributes to pathogenicity. *Virology*. **401**, 96–106 (2010).

12. A. Lutz, J. Dyall, P. D. Olivo, A. Pekosz, Virus-inducible reporter genes as a tool for detecting and quantifying influenza A virus replication. *J. Virol. Methods*. **126**, 13–20 (2005).

13. W. Zhu *et al.*, A reporter system for assaying influenza virus RNP functionality based on secreted Gaussia luciferase activity. *Virol. J.* **8**, 29 (2011).

14. A. Melnikov, X. Zhang, P. Rogov, L. Wang, T. S. Mikkelsen, Massively Parallel Reporter Assays in Cultured Mammalian Cells. *J Vis Exp*, e51719–e51719 (2014).

15. J.-C. Lee *et al.*, A cell-based reporter assay for inhibitor screening of hepatitis C virus RNA-dependent RNA polymerase. *Anal. Biochem.* **403**, 52–62 (2010).

16. R. B. Squires *et al.*, Influenza Research Database: an integrated bioinformatics resource for influenza research and surveillance. *Influenza and Other Respiratory Viruses*. **6**, 404–416 (2012).

17. S. Islam *et al.*, Quantitative single-cell RNA-seq with unique molecular identifiers. *Nature Methods*. **11**, 163–166 (2014).

18. T. A. Whitehead *et al.*, Optimization of affinity, specificity and function of designed influenza inhibitors using deep sequencing. *Nat. Biotechnol.* **30**, 543–548 (2012).

19. L. Gu *et al.*, Multiplex single-molecule interaction profiling of DNA-barcoded proteins. *Nature*. **515**, 554–557 (2014).