Published in final edited form as:

Nat Genet. 2019 September ; 51(9): 1298–1301. doi:10.1038/s41588-019-0349-3.

Reconciling disparate estimates of viral genetic diversity during human influenza infections

Katherine S. Xue^{1,2,3}, Jesse D. Bloom^{1,2,3,4}

¹Division of Basic Sciences, Fred Hutchinson Cancer Research Center, Seattle, Washington, USA.

²Computational Biology Program, Fred Hutchinson Cancer Research Center, Seattle, Washington, USA.

³Department of Genome Sciences, University of Washington, Seattle, Washington, USA,

⁴Howard Hughes Medical Institute, Seattle, Washington, USA.

A key question in the study of influenza-virus evolution is how rapidly viral genetic variation arises within infected humans, and how much of this genetic diversity is maintained during transmission^{1,2}. Recently, several studies have measured influenza's within-host genetic diversity in large cohorts of infected humans using high-throughput deep sequencing (Supplementary Table 1)^{3–6}. These studies have disagreed in their estimates of influenza's within-host genetic diversity. In a *Nature Genetics* letter titled “Quantifying influenza virus diversity and transmission in humans” analyzing a household cohort in Hong Kong, Poon et al.⁴ estimated that within-host genetic diversity is high and 200–250 viral genomes are transmitted between individuals. However, several recent studies conducted in Wisconsin³, Michigan⁶, and Washington⁷ that used similar methodologies have estimated lower levels of viral genetic diversity. In particular, the Michigan study estimates a narrow transmission bottleneck of just 1–2 viral genomes⁶. We sought to examine whether technical differences in the underlying deep-sequencing datasets or the methods used to analyze them explain the disparate estimates of within-host viral genetic diversity. We identify an anomaly in the Hong Kong data that provides a technical explanation for these discrepancies: read pairs from this study are often split between different biological samples, indicating that some reads are incorrectly assigned.

Correspondence should be addressed to J.D.B (jbloom@fredhutch.org).

Author Contributions

K.S.X. and J.D.B. conceptualized the study and wrote the report. K.S.X. performed the analyses, some of which were independently re-implemented by J.D.B.

Competing Interests Statement

The authors declare no competing financial interests.

Data Availability

We downloaded sequencing data generated by the Hong Kong study⁴ from <https://www.synapse.org/#!Synapse:syn8033988>, following the methods of a study that re-analyzed data from the Hong Kong study to estimate transmission bottleneck sizes using a new analytical method⁸. We obtained sequencing data for the Wisconsin study³ by personal communication. We downloaded sequencing data for the other studies from SRA BioProject PRJNA344659⁵ and PRJNA412631⁶. See Life Sciences Reporting Summary for more details.

To systematically compare the results across studies, we used the same computational framework to re-analyze raw sequencing data for four large-scale studies of influenza's within-host genetic diversity, together encompassing more than 500 acute human infections³⁻⁶. For each study, we applied the same variant-calling thresholds as the Hong Kong study⁴, identifying sites with a minimum coverage of 200 at which a non-consensus base exceeds a frequency of 3% in the sequenced reads at that site (see Supplementary Note). We averaged variant frequencies between sequencing replicates where available but otherwise used an analysis pipeline that was as similar as possible across studies to ensure comparable estimates of within-host genetic diversity.

Our analysis recapitulates the major results reported in the Hong Kong study. Supplementary Figure 1 shows within-host variation in the hemagglutinin gene in H3N2 patients in our re-analysis of the study's data, in the same format as the second figure of the original publication⁴. In both the original study and our re-analysis, the same within-host variant is often present at similar frequencies in multiple, epidemiologically unrelated individuals. Moreover, the minority variant in one group of samples is typically the majority or consensus variant in the remaining samples (Supplementary Figure 1A). Across the hemagglutinin gene, the original Hong Kong study and our re-analysis of that study's data identify the same patterns of within-host variation (Supplementary Figure 1B).

Our analysis also identifies major differences between the Hong Kong dataset and the other studies. We find little within-host viral variation in the other three datasets, in line with these studies' stated conclusions (Supplementary Figure 2A)^{3,5,6}. Furthermore, only the Hong Kong dataset contains high-frequency within-host variants that are shared between epidemiologically unrelated individuals. In data from the Hong Kong study, the same within-host variants were shared among more than half of the patients at 42 sites in the H3N2 genome, and 9 sites in the pdmH1N1 genome (Figure 1). In contrast, we identified no such sites of extensively shared genetic variation among patients in the other three studies. These results show that the large discrepancies between the Hong Kong study and other published work cannot be accounted for solely by methodological differences in variant calling pipelines.

The extensive shared genetic diversity in the Hong Kong study could result from genuine similarity in the mix of viruses that infect epidemiologically unrelated humans in Hong Kong. But they could also reflect cross-contamination or other abnormalities in the underlying sequencing data. In the course of our analysis, we identified abnormalities in the raw sequencing data from the Hong Kong study that can explain the apparently high levels of shared viral genetic diversity across different infected individuals. The deep sequencing for this study used paired-end Illumina reads. Both reads in a pair come from the same molecule of PCR-amplified viral genetic material, and so should always be assigned to the same infected human (Figure 2A). Illumina software assigns standard headers to each FASTQ-format sequencing read. These header lines contain information about each read, including the sequencing lane, a unique read-pair identifier, and whether a read is the first or second member of a pair (Figure 2B). When we analyzed FASTQ headers in the raw sequencing data for the Hong Kong study, we found that paired-end sequencing reads were frequently split between samples assigned to different individuals (Figure 2C). (Figure 1 and

Supplementary Figure 1 were generated by analyzing the sequencing data from the Hong Kong study as single-end data.) For instance, the read @SOLEXA4_0078:1:1101:10000:101622#ATCACG/1 was associated with study subject 737-V1(0), whereas its pair @SOLEXA4_0078:1:1101:10000:101622#ATCACG/2 was associated with study subject 741-V1(0), an epidemiologically unrelated individual.

It is biologically impossible for reads in a pair to be associated with distinct individuals, since both reads originate from the same DNA molecule. Across all samples, 70% of reads had corresponding pairs in a FASTQ file assigned to a different individual, and 25% of reads were not part of an identifiable pair (Figure 2C). Only 5% of the 500 million sequencing reads in this study were associated with the same sample as their corresponding pairs. This splitting of read pairs between samples indicates a problem in the sample index de-multiplexing or downstream computational analysis, and can be considered a form of technical cross-contamination.

Importantly, the problem appears to be with how read pairs were assigned to samples rather than with the FASTQ headers. We found that 93% of the read pairs reconstructed based on FASTQ header information mapped concordantly to the H3N2 or pandemic H1N1 influenza genome—that is, both reads in a pair mapped to the same gene segment in the expected relative orientation.

We analyzed patterns of read-pair splitting between all samples in the study (Figure 2D). We identified four disjoint sets of samples for which read pairs are split extensively within sets, but never between sets. Further analysis of FASTQ headers showed that all of the sequencing reads from each cluster were derived from the same flowcell lane. Poon et al.⁴ report that samples were amplified in duplicate and that replicates were sequenced on distinct flowcell lanes. Indeed, we find that each set of samples corresponds almost exactly to one set of replicate samples for one of the two influenza subtypes sequenced in this study (Figure 2D). This finding was robust to the computational analysis pipeline: the first author generated all of the figures in this paper, but the last author conducted an independent re-analysis of the data to reach similar conclusions (see Supplementary Note). Altogether, these analyses suggest that read pairs are split extensively between samples of a given influenza subtype in the Hong Kong study.

Without access to the full computational pipeline for the Hong Kong study, we cannot determine directly whether the first read, second read, or both members of split read pairs were assigned to samples incorrectly. However, when we analyzed only the first read of each pair, we found low within-host diversity, in line with other studies (Figure 2E, Supplementary Figure 2B). In contrast, the second read of each pair was responsible for the high viral diversity reported in the Hong Kong study. These results suggest that the second member of each read pair may have been incorrectly assigned, and the first member may more accurately represent the low levels of within-host viral diversity.

This splitting of read pairs between unrelated samples has important consequences for estimates of viral genetic diversity within human infections. Even if each individual were infected with a clonal population of influenza virus, read-pair splitting would create the

appearance of high levels of shared genetic diversity between unrelated individuals. For instance, at a site in the influenza genome where some individuals exclusively have nucleotide A and others exclusively have nucleotide T, read-pair splitting would make it seem as though all individuals with majority identity A have minority variant T and vice versa, even in the absence of genuine within-host variation. The high-frequency shared viral diversity within human hosts in the Hong Kong study corresponds closely to what would be expected from read-pair splitting (Supplementary Figure 1A), suggesting that this abnormality may be responsible for the published results.

Read-pair splitting may also explain why the Hong Kong household cohort study estimates a loose transmission bottleneck for human influenza virus of 200–250 viral genomes^{4,8}, compared to a Michigan household cohort study that estimates a bottleneck size of 1–2 viral genomes⁶. Splitting of read pairs between samples would create the appearance of shared within-host variation in donor and recipient individuals in a transmission chain, resulting in estimates of a looser transmission bottleneck.

Our finding of read-pair splitting in the Hong Kong dataset provides a technical explanation for major discrepancies in recent studies of the genetic diversity of human influenza viruses. If we exclude the Hong Kong study, then all other studies report low levels of within-host genetic diversity for human influenza virus^{3,5,6}.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

We thank P. Green for helpful comments on the manuscript. K.S.X is supported by the Hertz Foundation Myhrvold Family Fellowship. The work of J.D.B was supported by grant R01AI127893 from the NIAID of the NIH. J.D.B. is an Investigator of the Howard Hughes Medical Institute.

References

1. Xue KS, Moncla LH, Bedford T & Bloom JD Within-Host Evolution of Human Influenza Virus. *Trends Microbiol.* (2018). doi:10.1016/J.TIM.2018.02.007
2. McCrone JT & Lauring AS Genetic bottlenecks in intraspecies virus transmission. *Curr. Opin. Virol* 28, 20–25 (2018). [PubMed: 29107838]
3. Dinis JM et al. Deep Sequencing Reveals Potential Antigenic Variants at Low Frequencies in Influenza A Virus-Infected Humans. *J. Virol* 90, 3355–65 (2016). [PubMed: 26739054]
4. Poon LLM et al. Quantifying influenza virus diversity and transmission in humans. *Nat. Genet* 48, 195–200 (2016). [PubMed: 26727660]
5. Debbink K et al. Vaccination has minimal impact on the intrahost diversity of H3N2 influenza viruses. *PLOS Pathog.* 13, e1006194 (2017). [PubMed: 28141862]
6. McCrone JT et al. Stochastic processes constrain the within and between host evolution of influenza virus. *Elife* 7, e35962 (2018). [PubMed: 29683424]
7. Xue KS, Greninger AL, Pérez-Osorio A & Bloom JD Cooperating H3N2 Influenza Virus Variants Are Not Detectable in Primary Clinical Samples. *mSphere* 3, e00552–17 (2018). [PubMed: 29299533]

8. Sobel Leonard A, Weissman DB, Greenbaum B, Ghedin E & Koelle K Transmission Bottleneck Size Estimation from Pathogen Deep-Sequencing Data, with an Application to Human Influenza A Virus. *J. Virol* 91, JVI.00171–17 (2017).
9. McCrone JT & Luring AS Measurements of Intrahost Viral Diversity Are Extremely Sensitive to Systematic Errors in Variant Calling. *J. Virol* 90, 6884–6895 (2016). [PubMed: 27194763]
10. Illingworth CJR et al. On the effective depth of viral sequence data. *Virus Evol.* 3, vex030 (2017). [PubMed: 29250429]

Figure 1. Comparison of shared within-host viral genetic diversity in four large-scale deep-sequencing studies of human influenza virus.

Proportion of samples in each study in which we identified within-host variation at each genome site. For each sample, we identified within-host variants that were present at a frequency of at least 3% at sites with minimum sequencing coverage of 200 reads. Our re-analysis is consistent with the previously reported results of each study: we find little shared genetic diversity in the data from the Dinis et al. (2016), Debbink et al. (2017), and McCrone et al. (2018) studies, but we observe high shared genetic diversity in the data from the Poon et al study.

host variation at each genome site when analyzing both reads for a pair, just read 1, or just read 2. For each sample, we identified within-host variants that were present at a frequency of at least 3% at sites with minimum sequencing coverage of 200 reads.

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript