

Accepted Manuscript

No evidence for intra-segment recombination of 2009 H1N1 influenza virus in swine

Maciej F. Boni, Gavin J.D. Smith, Edward C. Holmes, Dhanasekaran Vijaykrishna

PII: S0378-1119(11)00635-4
DOI: doi: [10.1016/j.gene.2011.10.041](https://doi.org/10.1016/j.gene.2011.10.041)
Reference: GENE 37081

To appear in: *Gene*

Accepted date: 20 October 2011



Please cite this article as: Boni, Maciej F., Smith, Gavin J.D., Holmes, Edward C., Vijaykrishna, Dhanasekaran, No evidence for intra-segment recombination of 2009 H1N1 influenza virus in swine, *Gene* (2011), doi: [10.1016/j.gene.2011.10.041](https://doi.org/10.1016/j.gene.2011.10.041)

This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

No evidence for intra-segment recombination of 2009 H1N1 influenza virus in swine

Maciej F. Boni^a, Gavin J.D. Smith^b, Edward C. Holmes^c, Dhanasekaran Vijaykrishna^{b,*}

^aOxford University Clinical Research Unit, Wellcome Trust Major Overseas Programme, Ho Chi Minh City, Vietnam

^bLaboratory of Virus Evolution, Duke-NUS Graduate Medical School, 8 College Rd, 169857, Singapore

^cCenter for Infectious Disease Dynamics, Department of Biology, The Pennsylvania State University, Mueller Laboratory, University Park, Pennsylvania 16802

*Corresponding author: vijay.dhanasekaran@duke-nus.edu.sg

Abstract

Hao (2011) reported that the PB2 genes of three swine influenza A viruses were likely generated through homologous recombination between two closely related parental strains. However, we show that Hao's observation is an artifact of incorrect taxon sampling arising through the lack of an appropriate evolutionary context. Through rigorous phylogenetic analyses we explain the evolutionary origins of these stains and confirm the lack of any statistical support for intra-segmental recombination.

Abbreviations

- HK – Hong Kong
- PCR – polymerase chain reaction
- PB2 – polymerase basic protein 2
- Sw – swine

Main Text

A recent paper by Hao (2011) is one of several recent studies in which the application of computational methods has provided supposed evidence for homologous intra-segment recombination in influenza virus (Gibbs et al., 2001; He et al., 2008, 2009). However, all such studies have been hampered by contamination with a laboratory strain (He et al., 2008), more likely explanations for the observed phylogenetic patterns via lineage-specific rate variation (Worobey et al., 2002), or an implausible scenario of recombination occurring among viruses isolated decades apart (Boni et al., 2008). In addition, many laboratories sequence the longer polymerase and surface protein genes from two separate PCR amplifications. In a small number of cases, each PCR reaction can amplify genes from two different viruses that have co-infected a host. In these cases,

purported recombinant breakpoints are easily identifiable as the overlap between the two PCR segments. More generally, systematic recombinant searches controlling for sequence quality have to date found little evidence for homologous recombination among human and avian influenza viruses (Boni et al., 2008; Boni et al., 2010).

The 10 influenza A viruses analyzed by Hao (2011) were isolated from a swine abattoir in Hong Kong (Vijaykrishna et al., 2010). Phylogenetic analyses of all gene segments, including independent analyses on the two putatively recombinant regions of the PB2 segment 1–1197 and 1198–2277 [Fig. 1B,C], show that these viruses belong to the 2009 H1N1 pandemic virus lineage and were most likely introduced into swine independently on at least three different occasions. Importantly, these trees show no close phylogenetic relationship among the putative parental sequences and recombinant sequences as you would expect under a plausible scenario of recombination (e.g. see Fig. 3 in Boni et al. (2010)).

We compared the variable sites in the PB2 gene of these viruses with the consensus of the 2009 pandemic H1N1 viruses ($n = 2700$) isolated from humans (Fig. 1A). Contrary to Hao's Fig. 2, our comparison illustrates that the 10 swine viruses isolated from four different time points show four distinct sets of mutations that were likely accrued independently through random mutation. In particular, the 19 mutations of the three putative recombinants (Sw/HK/NS1810/2009, Sw/HK/NS1809/2009 and Sw/HK/189/2011) that are highlighted in blue and orange in Hao's Fig. 2 are ancestral and shared by almost all 2009 H1N1 pandemic viruses isolated from humans. These 3 viruses differ from the consensus in only four or five mutations (Fig. 1A). In Hao's analysis, the absence of an appropriate out-group artificially makes the plesiomorphic states appear derived and hence wrongly suggestive of recombination.

The analysis presented by Hao also contains a pitfall common to recombination analyses: that the data-driven process of searching for a recombinant is rarely matched with a statistical test designed for that particular search process. This type of error is easily committed when looking both for mosaicism and phylogenetic incongruence with strong bootstrap support. First, the OnePop program (Hao, 2010) uses a classic sliding window approach to detect regions where ancestry-informative sites cluster, but it selects the window with the lowest p -value without correcting for this in the search process. In other words, the p -value is reported as if a fixed window size were being used, but the tested statistic (clustering of informative sites) is computed using the best window size – i.e. the one most likely to yield a recombination signal. The OnePop p -value reported in Table 1 of Hao (2011) is thus misleading; it would be much higher if we computed the probability that the best window yielded a particular clustering of informative sites. The simplest nonparametric single-breakpoint recombination statistic ($\Delta_{m,n,1}$ from Boni et al (2007)) for the pattern observed in Hao's Fig. 2 gives $p=0.0035$, while a simple Mann-Whitney test on these informative sites gives $p=0.0133$. These are no doubt significant as single tests, but they are always performed in a series of hundreds of tests (or more), and their true significance is difficult to ascertain as no appropriate multiple comparisons correction exists for exhaustive triplet testing in sequence data. This is one reason why recombination events in large data sets are best demonstrated by phylogenetically incongruent trees.

The trees in Hao's Fig. 3 do demonstrate a phylogenetic incongruence compatible with recombination, however the selection of particular closely related sequences for this rather low-diversity-tree has made closely-related clades appear strongly supported (i.e. distinct from each other) in a bootstrap analysis. We suspect that the phylogenetic incongruence is caused by the differential accumulation of mutations on either sides of the putative breakpoint (Fig. 1A); putative Parent 1 has accumulated mutations solely on the 5' side (nt 1-1197) of the putative breakpoint,

while putative Parent 2 has accumulated additional mutations and on the 3' side (nt 1198-2277) of putative breakpoint. Indeed, if all 32 sequences from Vijaykrishna et al. (2010) are included in the analysis, the phylogenetic incongruence persists but is no longer supported by high bootstrap values (Fig. 2), and the recombination signal is further weakened if the tree is rooted. Removing sequences and recalculating bootstrap values is not a valid solution to this problem, and it should not be used to demonstrate that a specific clade has strong support for monophyly. A similar effect can be seen in Hao's Fig. 6, where statistical power of recombination detection methods is somewhat reduced when distant outgroups are included in phylogenies. Removing the outgroups increases statistical power, but this figure does not show if the false positive rate also increases.

Hao notes that detecting recombination among closely related sequences is very difficult. We agree. In fact, if closely-related influenza viruses had a high propensity to recombine, we may never be able to statistically demonstrate that this process was occurring, because all generated recombinants would be equally well explained by a random mutation model. In this case, we would have no statistical basis for believing that the observed mutational patterns were a product of recombination.

In summary, the patterns observed in Hao's Fig. 2 could have been produced by mutation alone. Neither the p -values for mosaic signals nor the observed phylogenetic patterns support a hypothesis of homologous recombination among these sequences. Simply put, it is difficult to find recombination among similar sequences because there is insufficient phylogenetic information in the data.

Acknowledgements

We acknowledge support from Wellcome Trust grant 089276/B/09/Z (M.F.B), NIAID contract HHSN266200700005C (G.J.D.S.), the Agency for Science, Technology and Research and the Ministry of Health, Singapore (G.J.D.S. and D.V.).

References

- Boni M.F., Posada D., Feldman M.W., 2007. An exact nonparametric method for inferring mosaic structure in sequence triplets. *Genetics* 176, 1035-47.
- Boni, M.F., Zhou, Y., Taubenberger, J.K., Holmes, E.C. 2008. Homologous recombination is very rare or absent in human influenza A virus. *J. Virol.* 82, 4807–4811.
- Boni, M.F., de Jong, M.D., van Doorn, H.R., Holmes, E.C., 2010. Guidelines for identifying homologous recombination events in influenza A virus. *PLoS One* 5, e10434.
- Gibbs, M.J., Armstrong, J.S., Gibbs, A.J. 2001. Recombination in the hemagglutinin gene of the 1918 “Spanish Flu”. *Science* 293, 1842–1845.
- Hao, W., Richardson, A.O., Zheng, Y., Palmer, J.D. 2010. Gorgeous mosaic of mitochondrial genes created by horizontal transfer and gene conversion. *Proc. Natl. Acad. Sci. USA* 107, 21576–21581.
- Hao, W. 2011. Evidence of intra-segmental homologous recombination in influenza A virus. *Gene* 481, 57-64.
- He, C.Q. et al., 2008. Homologous recombination evidence in human and swine influenza A virus. *Virology* 380, 12–20.
- He, C. Q. et al., 2009. Homologous recombination as an evolutionary force in the avian influenza A virus. *Mol. Biol. Evol.* 26, 177–187.
- Stamatakis, A. 2006. RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics* 22, 2688–269.

Stamatakis, A., Hoover P., Rougemont, J. 2008. A rapid bootstrap algorithm for the RAxML Web servers. *Syst. Biol.* 57, 758–771.

Vijaykrishna, D. et al., 2010. Reassortment of pandemic H1N1/2009 influenza A virus in swine. *Science* 328, 1529.

Worobey, M., Rambaut, A., Pybus, O. G., Robertson, D.L. 2002. Questioning the evidence for genetic recombination in the 1918 “Spanish Flu” virus. *Science* 296, 211.

Figure Legends

Fig. 1. Lack of evidence for intra-segment recombination in the PB2 segment of 2009 pandemic influenza A viruses isolated from swine in Hong Kong. The alignment shows the variable nucleotides in the PB2 gene of swine influenza A viruses compared to a 99% consensus of 2009 human pandemic viruses (n>2000) (A). Maximum likelihood phylogenies of nucleotide regions 1–1197 (B) and 1198–2277 (C) of the coding region of the PB2 gene were generated using the best-fit nucleotide substitution model in RAxML (Stamatakis, 2006; Stamatakis et al, 2008). Putative parental strains are colored in red and blue, while the putative recombinant viruses are shaded in grey. Scale bars represent nucleotide substitutions per site. Bootstrap values generated from 500 maximum likelihood bootstrap replicates are shown at branch nodes.

Fig. 2. Maximum likelihood phylogenies of the two putative regions of the PB2 gene of the 32 pandemic influenza A viruses isolated from swine in Hong Kong. Both trees inferred with RAxML using 1000 bootstrap replicates (Stamatakis, 2006; Stamatakis et al, 2008). Red and blue sequences are the parental clades P1 and P2, respectively, as described by Hao (2011).

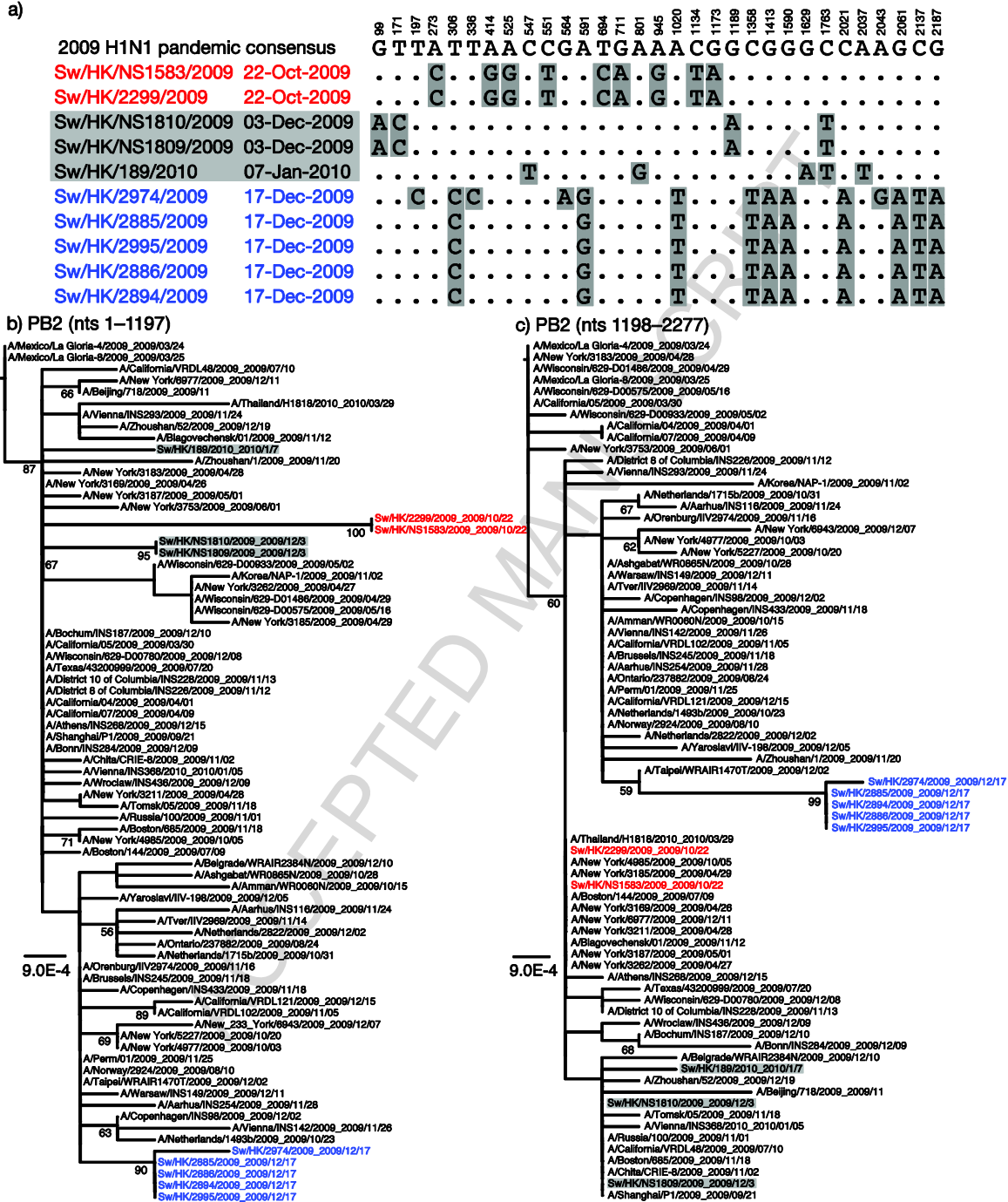


Fig. 1

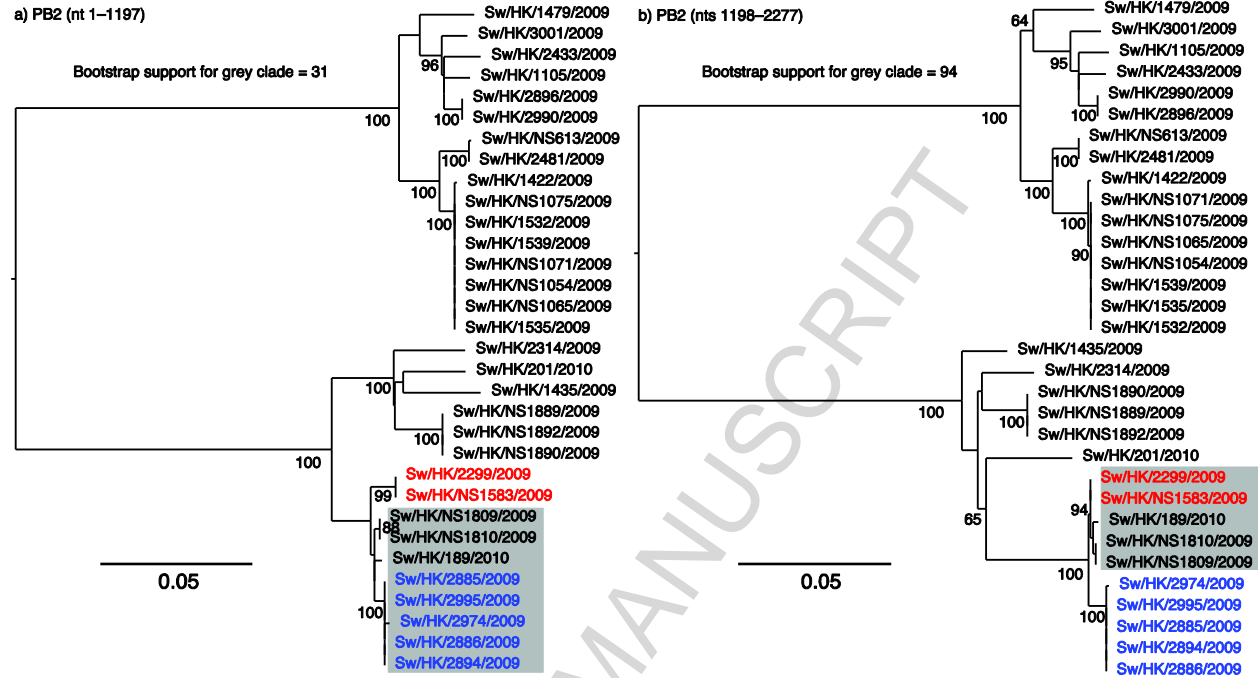


Fig. 2