Reconciling disparate estimates of viral genetic diversity during human influenza infections

To the Editor — A key question in the study of influenza virus evolution is how rapidly viral genetic variation arises within infected humans[1,2]. Recently, several studies have measured influenza's within-host genetic diversity in large cohorts of infected humans through high-throughput deep sequencing[3–6] (Supplementary Table 1). These studies disagree 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] have estimated that within-host genetic diversity is high, and 200–250 viral genomes are transmitted between individuals. However, several recent studies conducted in Wisconsin[3], Michigan[6], and Washington[7] that used similar methodologies have estimated lower levels of viral genetic diversity. In particular, the Michigan study has estimated a narrow transmission bottleneck of just one or two viral genomes[6]. We sought to examine whether technical differences in the underlying deep-sequencing datasets or the methods used to analyze them might explain the disparate estimates of within-host viral genetic diversity. We identified 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, thus indicating that some reads are incorrectly assigned.

To systematically compare the results across studies, we used the same computational framework to reanalyze raw sequencing data for four large-scale studies of influenza's within-host genetic diversity, together encompassing more than 500 acute human infections[3–6]. For each study, we applied the same variant-calling thresholds as those in the Hong Kong study[4], identifying sites with a minimum coverage of 200, at which a nonconsensus base exceeds a frequency of 3% in the sequenced reads at that site (Supplementary Note). We averaged the variant frequencies between sequencing replicates when available but otherwise used an analysis pipeline that was as similar as possible across studies to ensure comparable estimates of within-host genetic diversity[8,9].

Our analysis recapitulated the major results reported in the Hong Kong study (Supplementary Fig. 1). In both the original study and our reanalysis, 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 Fig. 1a). Across the hemagglutinin gene, the original Hong Kong study and our reanalysis of that study's data identified the same patterns of within-host variation (Supplementary Fig. 1b).

Our analysis also identified major differences between the Hong Kong dataset and the other studies. We found little within-host viral variation in the other three datasets (Supplementary Fig. 2a), in line with the studies' stated conclusions[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 are shared among more than half of the patients at 42 sites in the H3N2 genome and nine sites in the pdmH1N1 genome (Fig. 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 may be a result of genuine similarity in the mix of viruses that infect epidemiologically unrelated humans in Hong Kong. However, they might 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 therefore should always be assigned to the same infected human (Fig. 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 (Fig. 2b). When we analyzed the 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 (Fig. 2c). (Fig. 1 and Supplementary Fig. 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, because 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 (Fig. 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 demultiplexing or downstream computational analysis, and it 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 on the basis of 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 for all samples in the study (Fig. 2d) and identified four disjoint sets of samples for which read pairs are split extensively within sets but never between sets. Further analysis of the FASTQ headers showed that all of the sequencing reads from each cluster are derived from the same flow-cell lane. Poon et al.[4] reported that the samples were amplified in duplicate and that replicates were sequenced on distinct flow-cell lanes. Indeed, we found that each set of samples

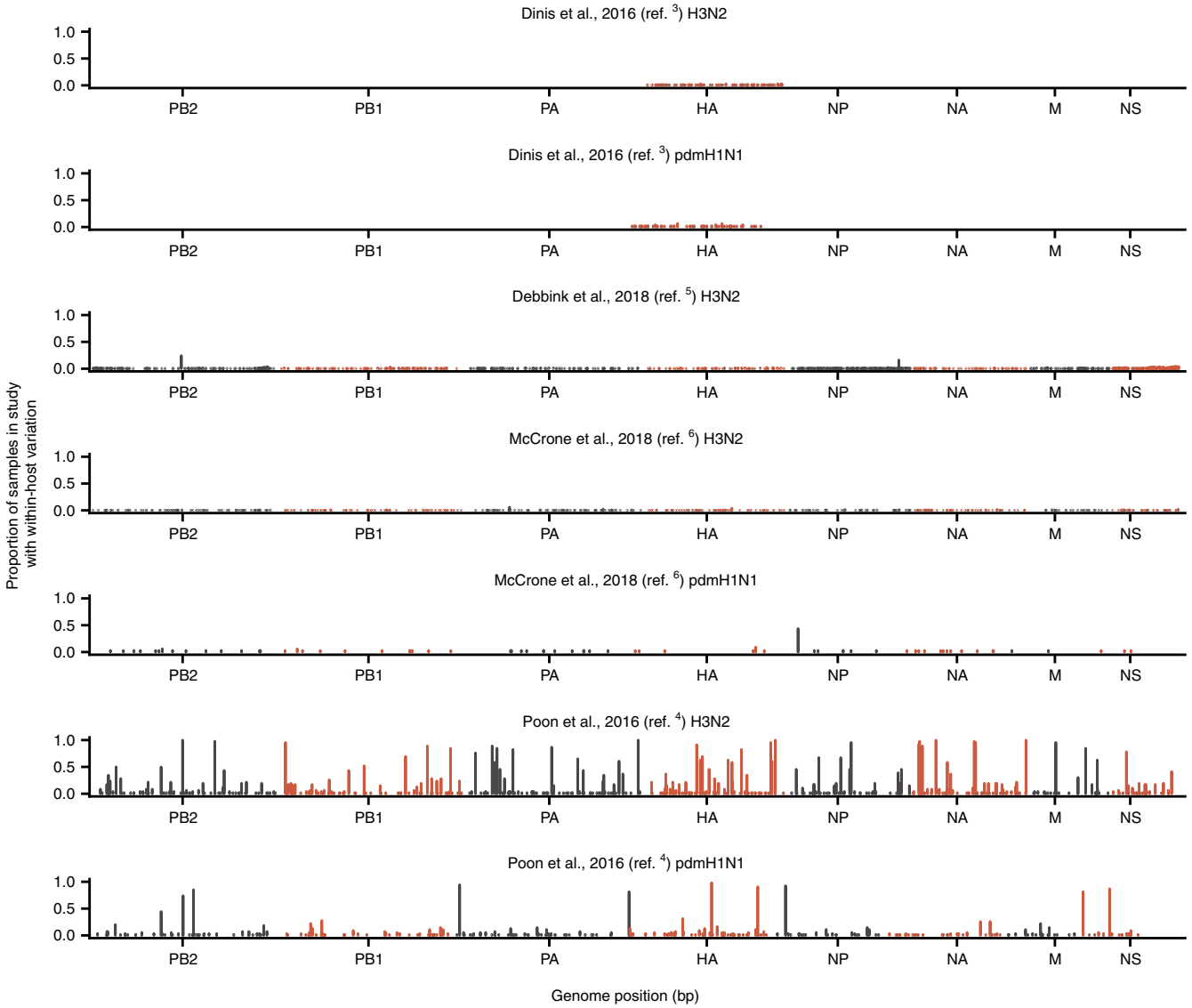

Fig. 1 | Comparison of shared within-host viral genetic diversity in four large-scale deep-sequencing studies of human influenza virus. Proportion of samples in each study in which we identified within-host variation at each genome site. For each sample, we identified within-host variants present at a frequency of at least 3% at sites with a minimum sequencing coverage of 200 reads. Our reanalysis is consistent with the previously reported results of each study: we found little shared genetic diversity in the data from the studies by Dinis et al.[3], Debbink et al.[5], and McCrone et al.[6], but we observed high shared genetic diversity in the data from the study by Poon et al.[4]. PB2, polymerase basic 2; PB1, polymerase basic 1; PA, polymerase acidic; HA, hemagglutinin; NP, nucleoprotein; NA, neuraminidase; M, matrix; NS, nonstructural. We note that Dinis et al.[3] sequenced only HA.

corresponds almost exactly to one set of replicate samples for one of the two influenza subtypes sequenced in this study (Fig. 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 reanalysis of the data and reached similar conclusions (Supplementary Note). Altogether, these analyses suggest that the 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 directly determine whether the first read, the 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 findings from other studies (Fig. 2e and Supplementary Fig. 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 an appearance of high levels of shared genetic diversity between unrelated individuals. For instance, at a

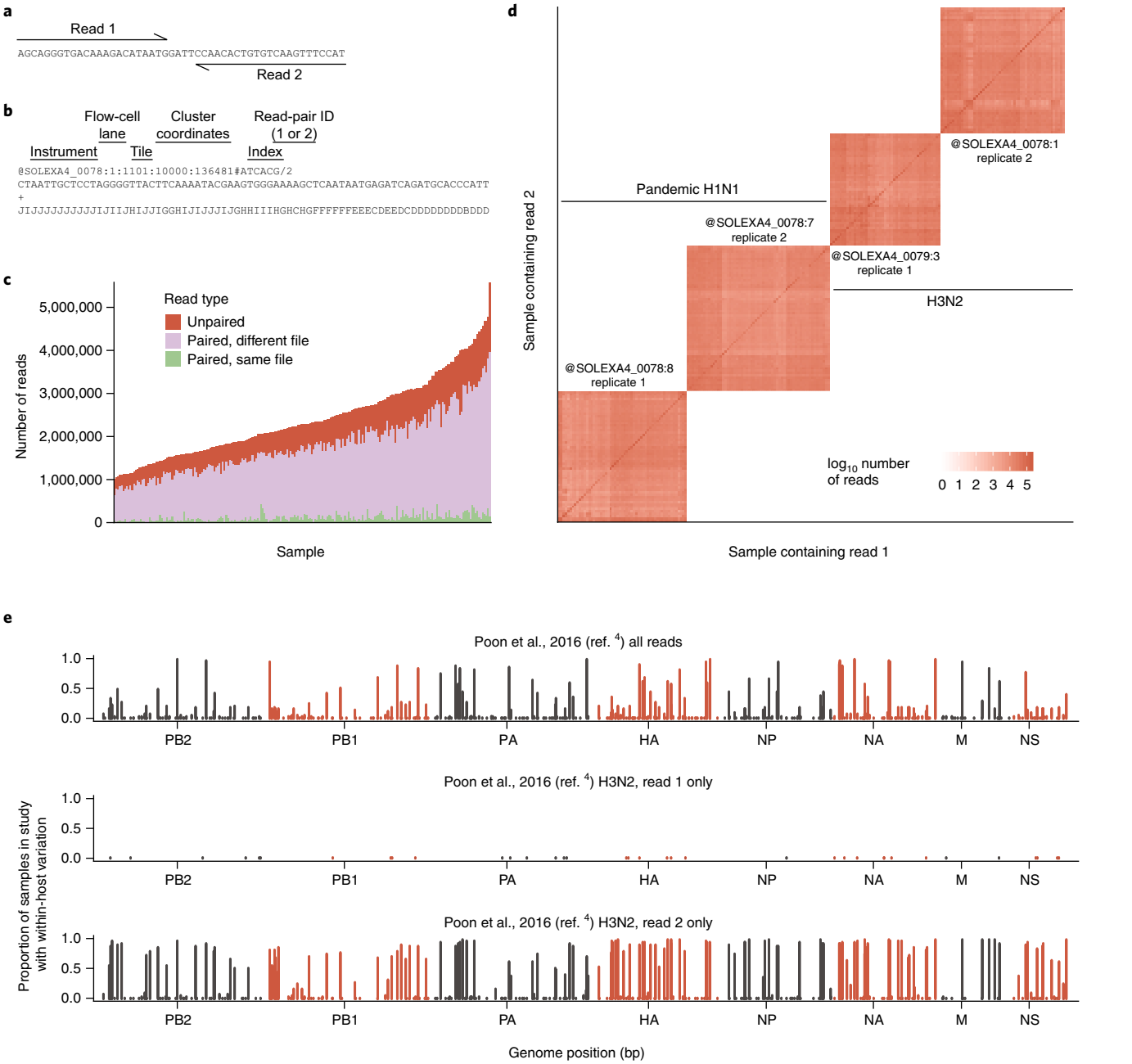

Fig. 2 | Paired-end-sequencing reads are frequently split between samples that were run on the same sequencing lane. a, Paired-end-sequencing reads are derived from the same physical DNA molecule. **b**, The FASTQ header for each sequencing read provides information about the sequencing instrument, flow-cell lane, tile, cluster coordinates, and sequencing index for each read, as well as whether the read is the first or second member of a read pair. **c**, Sequencing reads from the Hong Kong dataset are frequently split between distinct biological samples. **d**, Hierarchical clustering of the number of read pairs split between each pair of samples in the Hong Kong study. Sequencing reads from the Hong Kong dataset are split among four distinct clusters of samples. All sequencing reads in each cluster are derived from the same flow-cell lane and correspond to one set of replicate samples for one of the two influenza subtypes sequenced in the study. **e**, Proportion of samples for which we identified within-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.

site in the influenza genome where some individuals exclusively have nucleotide A and others exclusively have nucleotide T, read-pair splitting would suggest that all individuals with the 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 Fig. 1a), thus 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[4,10], compared with a Michigan household cohort study estimating a bottleneck size of one or two viral genomes[6]. Splitting of read pairs between samples would create the appearance of shared within-host variation in donor and recipient individuals in a transmission chain, thereby resulting in estimates of a looser transmission bottleneck.

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

Reporting Summary

Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability

We downloaded sequencing data generated by the Hong Kong study[4] from https://www.synapse.org/#!Synapse:syn8033988/, following the methods of a study that reanalyzed data from the Hong Kong study to estimate transmission-bottleneck sizes by using a new analytical method[10]. We obtained sequencing data for the Wisconsin study[3] by contacting the corresponding authors of that study. We downloaded sequencing data for the other studies from SRA BioProject PRJNA344659 (ref. [5]) and PRJNA412631 (ref. [6]). More details are provided in the Nature Research Reporting Summary. ❐

Katherine S. Xue[1,2,3] and Jesse D. Bloom[1,2,3,4]*

[1]*Division of Basic Sciences, Fred Hutchinson Cancer Research Center, Seattle, WA, USA.* [2]*Computational Biology Program, Fred Hutchinson Cancer Research Center, Seattle, WA, USA.* [3]*Department of Genome Sciences, University of Washington, Seattle, WA, USA.* [4]*Howard Hughes Medical Institute, Seattle, WA, USA.*

**e-mail: jbloom@fredhutch.org*

Published online: 25 February 2019

https://doi.org/10.1038/s41588-019-0349-3

References

1. Xue, K. S., Moncla, L. H., Bedford, T. & Bloom, J. D. *Trends Microbiol.* **26**, 781–793 (2018).
2. McCrone, J. T. & Lauring, A. S. *Curr. Opin. Virol.* **28**, 20–25 (2018).
3. Dinis, J. M. et al. *J. Virol.* **90**, 3355–3365 (2016).
4. Poon, L. L. M. et al. *Nat. Genet.* **48**, 195–200 (2016).
5. Debbink, K. et al. *PLoS Pathog.* **13**, e1006194 (2017).
6. McCrone, J. T. et al. *eLife* **7**, e35962 (2018).
7. Xue, K. S., Greninger, A. L., Pérez-Osorio, A. & Bloom, J. D. *MSphere* **3**, e00552–17 (2018).
8. McCrone, J. T. & Lauring, A. S. *J. Virol.* **90**, 6884–6895 (2016).
9. Illingworth, C. J. R. et al. *Virus Evol.* **3**, vex030 (2017).
10. Sobel Leonard, A., Weissman, D. B., Greenbaum, B., Ghedin, E. & Koelle, K. *J. Virol.* **91**, e00171–17 (2017).

Acknowledgements

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 supported as an Investigator of the Howard Hughes Medical Institute.

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 reimplemented by J.D.B.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s41588-019-0349-3.

nature research

Corresponding author(s): Jesse Bloom

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see Authors & Referees and the Editorial Policy Checklist.

Statistical parameters

When statistical analyses are reported, confirm that the following items are present in the relevant location (e.g. figure legend, table legend, main text, or Methods section).

n/a	Confirmed	
☒	☐	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☒	☐	An indication of whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☒	☐	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☒	☐	A description of all covariates tested
☒	☐	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☒	☐	A full description of the statistics including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☒	☐	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☒	☐	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated
☒	☐	Clearly defined error bars *State explicitly what error bars represent (e.g. SD, SE, CI)*

Our web collection on statistics for biologists may be useful.

Software and code

Policy information about availability of computer code

Data collection	Not applicable.
Data analysis	A mixture of common, open-source bioinformatics software and custom code was used to perform the analyses in this study. All custom code, as well as the scripts used to run the standard bioinformatics software, is available in Github repositories described in the manuscript.

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors/reviewers upon request. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

All sequencing data analyzed in this study is publicly available or was obtained through personal communication with the authors of the publication describing the dataset.

Field-specific reporting

Please select the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/authors/policies/ReportingSummary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	Not applicable. All analyses were conducted on previously published data and did not involve statistical calculations.
Data exclusions	No sequencing data from the four previously published studies were excluded from the analysis.
Replication	The two authors independently coded the analyses underlying Figure 2 of the manuscript, and both versions of the analysis are available in public Github repositories.
Randomization	Not applicable. No study organisms were used.
Blinding	Not applicable. No study organisms were used.

Reporting for specific materials, systems and methods

Materials & experimental systems

n/a	Involved in the study
☒	☐ Unique biological materials
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology
☒	☐ Animals and other organisms
☒	☐ Human research participants

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging