nature portfolio

Corresponding author(s): Guido van Marle and Stephen Barr

Last updated by author(s): 07/02/2025

Reporting Summary

Nature Portfolio 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 Portfolio policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the 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
☐	☒	A statement on 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 statistical parameters 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

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	Bedtools (v2.25.0), bioawk (awk version 20110810), bowtie2 (version 2.3.4.1), and restrSiteUtils (v1.2.9) were used to collect and analyse genomic DNA sequences and features, MUSCLE (MEGAX version 10.1.7), TrimAl (version 1.2), WebLogo (version 3.6), FACSDiva (software v8.0.1), Quantstudio (version 1.6.1).
Data analysis	Integration sites in non-B DNA motifs were identified using gquad R package (version 2.1-1; https://cran.r-project.org/web/packages/gquad/index.html); FlowJo (v10.4.2)

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 and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Portfolio 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 description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our policy

Data Availability
Integration site locations in the human genome were obtained from the GRCh37/hg19 database (https://hgdownload.soe.ucsc.edu/downloads.html). The source data generated in this study are provided in the Supplementary Information/Source Data file.

Code Availability
Bedtools (v2.25.0) (https://github.com/arq5x/bedtools2/releases), bioawk (awk version 20110810) (https://github.com/lh3/bioawk), bowtie2 (version 2.3.4.1) (https://github.com/BenLangmead/bowtie2), and restrSiteUtils (v1.2.9) (https://github.com/chasberry/integration-site-MRCs) were used to collect and analyse genomic DNA sequences and features. Integration sites in non-B DNA motifs were identified using the gquad R package (version 2.1-1) which is freely available as a standalone software package from the Comprehensive R Archive Network (https://cran.r-project.org/web/packages/gquad/index.html). This package provides functions for predicting non-B DNA. The key characteristics of the code and details pertaining to the test dataset can be found in the gquad documentation.

Research involving human participants, their data, or biological material

Policy information about studies with human participants or human data. See also policy information about sex, gender (identity/presentation), and sexual orientation and race, ethnicity and racism.

Reporting on sex and gender

People living with HIV (PLWH) were enrolled from a cohort of 5 HIV-1 seropositive individuals who self-reported their sex as male and men who have sex with men (MSM) followed at the Southern Alberta Clinic (SAC), Calgary, Alberta (Canada) between the years 1993 to 2010. Frozen brain samples originated from a separate unmatched cohort of 8 PLWH who were HIV-1 seropositive, not on antiretroviral therapy, three individuals self-reported as male, the sex of the other 5 individuals was not reported, and all of who succumbed to AIDS-defining illnesses. Sex and gender were not considered in the study design or analyses.

Reporting on race, ethnicity, or other socially relevant groupings

The 5 PLWH self-reported as men who have sex with men. Race, ethnicity or other socially relevant grouping information was not collected or used in the analyses. No race, ethnicity or other socially relevant grouping information was collected from the 8 deceased PLWH from which the brain tissue samples were collected.

Population characteristics

Participants were prospectively followed and assessed for plasma viral load and CD4+ T counts were performed for each individual during each visit. Upper and lower gastrointestinal endoscopies were performed in order to collect biopsies of tissues from the esophagus, stomach, duodenum, and colon. Samples were cryopreserved during shipment and stored at -70°C within 1 hour of collection. PBL/PBMC were isolated from blood and stored in liquid nitrogen. The month and year of collection and type of antiretroviral therapy received was documented. Participants received non-suppressive monotherapy or dual therapy with the nucleoside reverse transcriptase (NRTIs) such as azidothymidine (AZT/zidovudine), dideoxyinosine (ddI) prior to the study and during the study as previously described. Four participants from the SAC cohort also received HAART/ART at a later timepoint in treatment. Deceased PLWH from who the brain tissue was retrieved were not receiving antiretroviral therapy.

Recruitment

HIV-1 seropositive participants voluntarily enrolled at the Southern Alberta Clinic (SAC), Calgary, Alberta (Canada) and consented to be prospectively followed and assessed for plasma viral load and CD4+ T counts and undergo upper and lower gastrointestinal endoscopies to collect biopsies of tissues from the esophagus, stomach, duodenum, and colon over multiple visits between 1993 and 2010. Brain tissue samples were obtained from participants who donated their brains for research to Dr. Power.

Ethics oversight

Ethical approval for use of human esophagus, PBL/PBMC, stomach, duodenum and colon tissue were obtained from the Conjoint Health Ethics Research Board (CHREB, protocol approval #: REB15-1941) at the University of Calgary (Calgary, Alberta, Canada). Ethical approval for use of human brain tissue was obtained from the University of Alberta Human Ethics Committee, protocol approval #: Pro00002291. Participants signed an informed consent upon enrollment.

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Field-specific reporting

Please select the one below that is 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/documents/nr-reporting-summary-flat.pdf

Life sciences study design

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

Sample size

Our sample size was determined based on availability, feasibility, and biological relevance while ensuring sufficient statistical power to detect meaningful trends in HIV-1 integration site preferences across tissues. While small sample sizes are a limitation, the study design emphasizes in-depth genomic analyses rather than broad epidemiological trends.Observed trends across multiple individuals and tissues strengthen the reliability of our conclusions despite limited sample availability.

Data exclusions

No data was excluded

Replication

In this study, we analyzed HIV-1 integration sites across multiple tissues from a limited number of individuals, making direct replication of our findings impossible. The nature of our experimental design, which relies on patient-derived samples, inherently limits our ability to replicate results in an independent cohort. Each individual in our study harbors a unique set of HIV-1 integration sites, influenced by factors such as infection history, immune response, and viral reservoir composition.

Another key limitation to replication is the rarity of comparable sample sets. The availability of historical HIV-1 cases with untreated or non-suppressive therapy infection, particularly those with brain tissue samples and matched anatomical tissues, is exceptionally limited (to our knowledge). Ethical considerations and the widespread implementation of antiretroviral therapy have further restricted access to similar patient populations. Given these constraints, obtaining an independent set of samples that mirrors our cohort is not feasible, unless a cohort with similar samples already exists.

Despite the inability to replicate our findings in a separate cohort, we have taken multiple measures to ensure the reproducibility and reliability of our conclusions. First, by analyzing multiple tissues from each individual, we assessed tissue-specific integration trends rather than relying on a single dataset per person. This approach strengthens our ability to detect consistent integration site preferences across anatomical locations. Moreover, the identification of our top integration site hotspot in multiple other independently published HIV integration site datasets, suggests biological relevance of some findings across cohorts. Additionally, while direct replication is not possible, our results align with previously reported HIV-1 integration site preferences in blood and other tissues. The enrichment of HIV-1 integration sites near gene-dense regions and regulatory elements has been observed in prior ex vivo and in vitro studies, supporting the validity of our conclusions.

Randomization	All participants were randomly selected.
Blinding	Blinded tissue samples were processed and sequenced to determine integration site profiles.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology and archaeology
☒	☐ Animals and other organisms
☒	☐ Clinical data
☒	☐ Dual use research of concern
☒	☐ Plants

Methods

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

Plants

Seed stocks	*Report on the source of all seed stocks or other plant material used. If applicable, state the seed stock centre and catalogue number. If plant specimens were collected from the field, describe the collection location, date and sampling procedures.*
Novel plant genotypes	*Describe the methods by which all novel plant genotypes were produced. This includes those generated by transgenic approaches, gene editing, chemical/radiation-based mutagenesis and hybridization. For transgenic lines, describe the transformation method, the number of independent lines analyzed and the generation upon which experiments were performed. For gene-edited lines, describe the editor used, the endogenous sequence targeted for editing, the targeting guide RNA sequence (if applicable) and how the editor was applied.*
Authentication	*Describe any authentication procedures for each seed stock used or novel genotype generated. Describe any experiments used to assess the effect of a mutation and, where applicable, how potential secondary effects (e.g. second site T-DNA insertions, mosiacism, off-target gene editing) were examined.*